UFRRJ

INSTITUTO DE AGRONOMIA

CURSO DE PÓS-GRADUAÇÃO EM AGRONOMIA – CIÊNCIA DO SOLO

TESE

Análise de Fontes de Incerteza na Modelagem Espacial do Solo

Alessandro Samuel-Rosa

2016

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE AGRONOMIA
CURSO DE PÓS-GRADUAÇÃO EM AGRONOMIA – CIÊNCIA DO SOLO

ANÁLISE DE FONTES DE INCERTEZA NA MODELAGEM ESPACIAL DO SOLO

ALESSANDRO SAMUEL-ROSA

Sob orientação de
Lúcia Helena Cunha dos Anjos

e co-orientação de
Gustavo de Mattos Vasques
e
Gerardus Bernardus Maria Heuvelink

Tese submetida como requisito parcial para obtenção do grau de **Doutor** no Curso de Pós-Graduação em Agronomia – Ciência do Solo, Área de Concentração em Ciência do Solo.

Seropédica, RJ, Brasil
Fevereiro de 2016

UFRRJ / Biblioteca Central / Divisão de Processamentos Técnicos

631.4
S193a
T

Samuel-Rosa, Alessandro, 1985-
Análise de fontes de incerteza na modelagem espacial do solo / Alessandro Samuel-Rosa – 2016.
278 f.: il.

Orientador: Lúcia Helena Cunha dos Anjos.
Tese (doutorado) – Universidade Federal Rural do Rio de Janeiro, Curso de Pós-Graduação em Agronomia – Ciência do Solo.
Bibliografia: f. 155-182.

1. Ciência do solo – Teses. 2. Mapeamento do solo – Teses. 3. Solos – Classificação – Teses. 4. Solos – Análise – Teses. I. Anjos, Lúcia Helena Cunha dos, 1957-. II. Universidade Federal Rural do Rio de Janeiro. Curso de Pós-Graduação em Agronomia – Ciência do Solo. III. Título.

Este documento foi criado usando o sistema LaTeX de preparação de documentos para composição de alta qualidade originalmente desenvolvido por Leslie Lamport a partir do sistema de formatação TeX criado por Donald Knuth.
O formato final deste documento foi obtido usando a classe `UFRuralRJ`, uma adaptação livre das classes `mdtufsm` e `iiufrgs` para a formatação de documentos acadêmicos produzidos na Universidade Federal Rural do Rio de Janeiro (UFRRJ) de acordo com as recomendações contidas na terceira edição do *Manual de instruções para organização e apresentação de dissertações e teses na UFRRJ*, publicado no ano de 2006.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE AGRONOMIA
CURSO DE PÓS-GRADUAÇÃO EM AGRONOMIA – CIÊNCIA DO SOLO

ALESSANDRO SAMUEL-ROSA

Tese submetida como requisito parcial para obtenção do grau de **Doutor** no Curso de Pós-Graduação em Agronomia – Ciência do Solo, Área de Concentração em Ciência do Solo.

TESE APROVADA EM 24/02/2016.

____

Gustavo de Mattos Vasques. Ph.D. EMBRAPA
(Presidente)

____

Marcos Bacis Ceddia. Dr. UFRRJ

____

Wenceslau Geraldes Teixeira. Ph.D. EMBRAPA

____

Ronaldo Pereira de Oliveira. Ph.D. EMBRAPA

____

Maria Leonor Ribeiro Casimiro Lopes Assad. Ph.D. UFSCar

PREFACE

I was never sure about what a thesis should consist of: I worked on so many things during the four years of my doctorate that I found myself somewhat lost when I had to decide what to write in the thesis. There are official documents suggesting *how* the thesis should be written, but not exactly *what* should be written – I find the definitions somewhat vague. For example, the manual of our University states that a “thesis consists of the result of a research which is presented as the final requirement for the completion of a doctorate”[1], which is quite the same thing said by the International Organization for Standardization (ISO): a “document which presents the author’s research and findings and submitted by him in support of his candidature for a degree or professional qualification”[2]. I tried reading other theses to see if I could get an inspiration. I also discussed this matter with my patient supervisors Lúcia Anjos (Universidade Federal Rural do Rio de Janeiro, Brazil), Gustavo Vasques (Embrapa Solos, Brazil), and Gerard Heuvelink (ISRIC – World Soil Information, the Netherlands). Unfortunately, for one reason or another, I was never satisfied with the outcome.

At first, I was a bit desperate. Have I failed? Has everyone failed? I hoped not! Perhaps the lack of an objective, ultimate, universal definition of what a thesis should consist of meant that, as a doctorate student, it was my responsibility to construct such a definition. This idea gave me back the long-lost excitement to write my thesis. I did not want to follow a boring ritual. I wanted to have fun and be completely honest with the reader, as Richard Webster[3] had once suggested. As such, I started thinking about all steps given since the start of the doctorate, something like

Once upon a time in Seropédica...

As the title says, this thesis is about a research on the factors determining a soil map to be more or less accurate, what I call *sources of uncertainty*. Many of these sources are known, others are still unknown, and some are disregarded due to our ignorance – or by convenience. When I wrote my doctorate research project, it seemed appropriate to aim at evaluating what I understood as being the main sources of uncertainty in the process of building a model to produce soil maps, a process that I call *soil spatial modelling*. The reason was simple: soil spatial modelling using modern techniques was a growing activity in Brazilian universities and research centres, and I felt that many *soil spatial modellers* were inclined towards using the most expensive data sources as the only way of producing higher accuracy soil maps of the Brazilian territory. I was preoccupied about these ideas – which appeared to be sort of an euphoria about new remote sensors – because I believed that high quality soil maps could be produced if we simply started using the data at hand.

Defining the main sources of uncertainty in soil spatial modelling required an operational definition, which was given based on the observation that, in general, the main decisions made by soil spatial modellers concern the a) calibration observations, b) covariates, and c) model structure. The general objective of evaluating these three mains sources of uncertainty was then divided into five specific objectives:

[1] UFRRJ (2006)

[2] ISO (1986)

[3] Webster (2003)

(I) Identify appropriate calibration sample sizes and designs for soil spatial modelling;

(II) Determine the accuracy of freely available covariates and their suitability to calibrate soil spatial models;

(III) Identify appropriate covariate selection methods to build linear soil spatial models;

(IV) Assess the effect of multicollinearity among covariates on the performance of linear soil spatial models;

(V) Identify database scenarios in which non-linear soil spatial models are more efficient than linear soil spatial models.

The idea was to deal with each of the objectives separately and present the results in individual chapters of the thesis which would be submitted for publication in peer reviewed journals. The main expected result was the definition of a sound *working protocol* that would allow the construction of efficient soil spatial models. My goal was to contribute to national (Brazilian Research Network on Digital Soil Mapping – RedeMDS) and international (GlobalSoilMap and Global Soil Information Facilities – GSIF) initiatives, while generating a significant amount of bibliographic material to support the teaching of modern soil spatial modelling techniques in soil classes at Brazilian universities.

With time it became clear that the five objectives and the expected results were too ambitious. I certainly was overwhelmed by the knowledge of the multiple sources of uncertainty, and felt compelled to develop a very thorough study. But I forgot that a doctorate includes more activities than those planned in the research project: you take classes, prepare grant proposals, write reports, help colleagues, get involved in other projects – such as adapting the LaTeX[4] class used to compile this thesis –, publish the papers of your master thesis, train undergraduate students, create and maintain the newsletter of a scientific group, read many articles and books, learn a couple of computer languages, start a relationship, get sick, and so on. Then, one day you realize that two years are already gone by and you still are preparing the database with which you will develop your case studies.

I know that I was particularly lucky for most of the soil and covariate data already being available for my use. This is because I have decided very early to continue using the data that I collected during my master so that I could go deeper into the details of modern soil spatial modelling techniques. Looking back, I think that this was the right decision. However, the resources needed to properly organize the data before I could actually use it were considerable. This effort was in line with my original intent of defining a working protocol for constructing soil spatial models, which I guess to have achieved, at least partially. Then I realized that I also needed to make my research the most reproducible as possible. The way to go was to make a thorough description of all data processing steps, including making available all computer scripts so that they could be reused by other people. The result was thousands of lines of computer code, mostly on R[5], which I used to indirectly access most of other computer programs. These computer scripts have shown to be invaluable for my own applications, and I have always hoped that other people would find them useful as well. But I then learned that many well known methods of data analysis/processing are not used simply because they are not implemented in a (single) software package. As such, making only the computer scripts available did seem to be a poor solution. Developing and maintaining a software package in

[4] See more about LaTeX in Wikipedia. The LaTeX class that I have adapted to compile this thesis is available in GitHub.

[5] See more about R in Wikipedia.

the most popular environment for data processing and analysis, i.e. R, was a natural decision. Although being fun, programming took a lot of resources!

A significant amount of resources was also spent preparing a description of the soil-forming factors and processes that determine the soil spatio-temporal distribution in the study area where the case studies were to be developed. Such a description is what I call *conceptual model of pedogenesis*. This was another effort in line with the definition of a working protocol because I believe that soil spatial modelling is not only about making maps, but also constructing soil knowledge. Within the scope of the thesis, this knowledge was expected to serve the development of an experiment devoted to meeting the third objective of the research project. My intent was to compare automated covariate selection methods with the use of expert knowledge. Preliminary tests were conducted with a few experts to help planning the experiment, which was believed to be a complex one. Preliminary results were encouraging, but since I needed to give more attention to the first and second objectives, I had to temporarily stop working on the third objective.

There also was my poor knowledge on some known topics, which sometimes took me to the wrong direction. For example, I wanted to evaluate how much more accurate a soil map is when more accurate covariates are used (second objective), the reason being that I was concerned with the fact that the covariates too are in error. As such, I collected field data to validate the covariates and correct them for any systematic errors. Only later, discussing with Gerard, I understood that 1) the validation data was poor, and 2) in soil spatial modelling the covariates are generally taken as they are. The latter is like assuming that the covariates were measured without error – otherwise a technique called error propagation analysis (or uncertainty analysis) can be employed to take that error into account. As such, the second objective of the research project needed to be reformulated in terms of how to define the different covariates that we had at hand. After many discussions we still were unable to reach a satisfactory solution, which did not prevent the study from being developed. Quite interesting results were produced, but presenting them also was a challenge: many models and covariates had been compared, and we wanted to have a summary way of presenting them, preferentially a figure. We came up with a figure that we later called a *model series plot*, i.e. a figure that depicts a series of models ordered according to some chosen summary performance statistic. For the purpose of our study, that was a useful figure, and I hope that the readers will understand how to interpret it. The reviewers of our paper were fundamental for improving the description of the model series plot. Fortunately, they were also able to help deciding upon a proper definition for the differences observed between the covariates that were being compared. The study was not exactly about their accuracy, but about how they were produced and their level of spatial detail.

Devising an experiment to evaluate the influence of sample design on the accuracy of soil maps also was a challenge. Most soil data used for soil spatial modelling were produced in the past century (legacy data) with observation locations purposively selected by soil spatial modellers using tacit rules. As such, I wanted to build an algorithm composed of a set of objective decision rules that would produce spatial samples similar to those produced by a soil spatial modeller. I would then simulate budget scenarios for sampling and see how different spatial samples would perform regarding soil map accuracy. But how to devise such an algorithm? I interviewed the soil spatial modellers that produced the soil data to be used to conduct the case studies, carried out a point pattern analysis of the resulting spatial sample configuration, and explored psychological concepts to understand the whys of the locations of the sampling points. A lengthy study was carried out, which provided evidence that many poorly understood factors influence the decision of soil spatial modellers on where to make soil observations. From one perspective, this enables one to plan more efficient soil observation campaigns. However, it did

not help finding a practical solution for the problem that we had at hand. Perhaps it was more appropriate to explore the existing, less complex algorithms that produce spatial samples using more objective decision rules formulated with basis on conceptual and operational factors.

We then invited Dick Brus (Alterra, the Netherlands) to participate devising the experiment to evaluate the influence of sample design on the accuracy of soil maps. After some talks and a bibliographic review, we decided for using sampling algorithms that are based on the so-called spatial simulated annealing which are commonly used to produce spatial samples for soil spatial modelling. The problem was that we did not know of spatial simulated annealing being implemented in any free and open source software package in a way that could meet the requirements of our study. Again, the solution was to work on our own implementation of spatial simulated annealing, which resulted in a second package for R. Having decided the sampling algorithm to work with, we needed to choose a sound method to take sampling costs into account. Because the access time to sampling points usually is the major cost component in soil sampling, Gerard and I thought of coupling with spatial simulated annealing an algorithm to solve the problem of travelling from one sampling point to the next with the least cost[6]. This would be a good piece of work, but we soon realized that solving the travelling problem was impossible given the available resources. The goal of taking sampling costs into account ended up being dropped out.

After some time working on the sampling experiment, which at the time seemed simpler than ever, we came to learn that the sampling algorithms that we had chosen had weaknesses – apparently like any algorithm. So, we thought that, perhaps, we could improve on those algorithms! A literature review suggested that we were correct and there was room for algorithmic improvements. Working on these improvements took a lot of resources, and I guess we came up with interesting, sound solutions. We only needed to know if the algorithmic improvements had any practical added value before evaluating the influence of sample design of prediction accuracy. It seemed appropriate to carry out two experiments, the first to evaluate the algorithmic improvements, the second to compare the algorithms. Again, the results were promising, specially from the algorithmic point of view. With regard to prediction accuracy, we cannot make high claims because the algorithms were tested using a single case study. But this gap should be easily filled since we have made our software package freely available for anyone to use. The negative side of these important developments is that carrying out the experiment to evaluate covariate selection methods became impossible due to the remaining resources available. This was a pity because I visited Murray Lark at the the British Geological Survey (BGS) headquarters in Nottingham, UK, to discuss about that experiment.

Like it happened with the third objective, there was not enough resources to conduct experiments to meet the fourth and fifth objectives. I think the topic of the fourth objective is a very important one, directly related to the problem of selecting covariates, which I really wanted to deal with. I would have been very happy if I could discuss the topic at least partially, but perhaps partial solutions may not be very useful. As for the fifth objective, I believe that developing a sound globally relevant work on the topic requires using several datasets, not a single dataset as I explored in my research project.

As a consequence of all these events, at the end of the doctorate, I was able to meet only the first and second *scientific* objectives – *scientific* in the sense of answering research questions – of my doctorate research project. This may seem little but only because the research project was too ambitious. Aside from meeting the original scientific objectives, I guess I made other important contributions that one may call *technical* contributions – *technical* in the sense of practical application – that were part of the original goal of defining a working protocol for

[6] See about the *travelling salesman problem* at Wikipedia.

soil spatial modelling. This includes, for example, documenting the soil and covariate data, as well as their processing steps, and describing the soil-forming factors that determine the soil spatio-temporal distribution in the study area. I know that these technical contributions do not help meeting any of the original scientific objectives. The same applies to the two `R`-packages that I developed, and the experiment conducted to understand how soil spatial modellers decide upon where to observe the soil, which were not originally planned in the research project. But I guess this is still valid, perhaps very important, as it seems to be common in any scientific research: you end up doing many things that are quite different from those that you were originally planning to do.

So... this is, more or less, the *story* of the research that I have carried out in collaboration with my supervisors and co-authors during my doctorate. Perhaps one will find this story biased towards the negative aspects of my doctorate. This is not entirely false, specially because I think that I generally tend to be happier with stories that have more errors than hits – because I learn more with the former than with the latter. As such, I guess there is nothing to write in this thesis other than what I have done during the doctorate that is directly and/or indirectly related to the original research project, have it been planned or not, be it a technical or scientific contribution, completed or not. This is what I present as the final requirement for the completion of the doctorate.

I do hope that my supervisors and co-authors like the work that I have done in the past four years. Lúcia, Gustavo, Gerard, and Dick have been so patient, so understanding, so respectful, that I have no words to prepare the deserved thanks. Each one of them with a different background, a different life story, a different perspective, working in a different part of the world... I learned a lot with them: soil science, mathematics, statistics, informatics, English, politics and science, human relations, and much more. What a pleasure experience working with the four of you!

I also hope that the outcome of my doctorate is of interest for the supporting institutions, because I would never write this thesis without their support. These are:

- Universidade Federal Rural do Rio de Janeiro, through the Post-Graduate Course in Agronomy – Soil Science and Department of Soil Science, for providing a solid soil science education, unconditionally supporting my research, and finding the means to guarantee my participation in several international events;

- Ministry of Science and Technology of Brazil, through the CNPq Foundation (Process 140720/2012-0), that provided a three-year grant without which I would not be able to develop any research at all;

- Ministry of Education of Brazil, through the CAPES Foundation (Process ID BEX 11677/13-9), that funded my one-year stay in the Netherlands, where most of the research was actually developed;

- ISRIC – World Soil Information, for unconditionally supporting my research.

- Embrapa Solos, for supporting my research.

- Universidade Federal de Santa Maria, through the Department of Soil Science, for supporting my research.

- Ministry of the Environment of Brazil, for providing some of the data that we used.

I need to note that seven individuals gave important contributions during the preparation of the data that were used to develop the case studies. Four of them are from the Universidade Federal Rural do Rio de Janeiro: Fabio Paes Leme Ferreira, Anastácia Perci Campos de Almeida, Pedro de Souza Calegaro and Mauro Antônio Homem Antunes. The other three are from the Universidade Federal de Santa Maria: Ricardo Simão Diniz Dalmolin, Jean Michel Moura Bueno, and Luis Fernando Chimelo Ruiz. Along with them, I must thank the development teams and module/package authors of the many free and open source software and operating system that were used to develop the case studies.

Many other individuals have also given some form of scientific and/or technical contribution, be it through the exchange of email messages, chatting during a coffee break, a short visit to their home institutions, or any other informal occasion. Their contributions were invaluable for usually raising unforeseen questions, pointing to extremely helpful references, and helping me seeing the research problems from another perspective. Special thanks are due to Ad van Oostrum, Andreas Papritz, Bas Kempen, Bradley Miller, Chantal Hendriks, Dominique Arrouays, Edgardo Ramos Medeiros, Eloi Carvalho Ribeiro, Jorge Mendes de Jesus, Madlene Nussbaum, Marcos Angelini, Murray Lark, Nicolas Saby, Pablo Miguel, Richard Webster, Thomas Caspari, Tom Hengl, Titia Mulder, and many others that I might have forgotten.

Having the story of a doctorate to tell depends not only on scientific and/or technical contributions, but also on the unconditional support of family, friends, and colleagues. My family, which gained several new members in the last four years, was strong enough to understand my rare and short visits, and the importance of a long stay on the other side of the Atlantic. Although my mother (Elaine), father (Adir), and brother (Eduardo) have only a vague idea of what I have been working on, their support was incommensurable. South-American and European friends were invaluable: Marcos, Indira, Eloi, Thomas, Nina, Jorge, André, Thiago, and Andriéli. There was the amazing, coolest of them all, ISRIC family, which I miss very much, lead by our dearest Yolanda. There also were my postgraduate colleagues and teachers. But one friend deserves a special thanks: Manoeli! Thank you Manoeli – by destiny or a strange coincidence –, born and raised in the same small town, after completing the bachelor and master together in Brazil, we ended up sharing the same house in The Netherlands during the doctorate. There, we continued our philosophical chats about all aspects of life, the universe, and everything which we started years ago. It was a pleasure to share a full year with such a wonderful friend!

Finally, three little Bacchanalian creatures, one having only a vague idea of my work, the other two not speaking any human language, deserve my most sincere thanks: Monique, Tupã, and Dindi! Your love was fire through my veins!

Alessandro Samuel Rosa
Seropédica, February 2016.

P.S. The reader should be aware that, despite the book style and formatting can cause estrangement, it follows the standards of the Universidade Federal Rural do Rio de Janeiro – we even created a LaTeX class for that end! I agree that the chapter numbering style is awkward. So is the use of two languages (English and Portuguese). But, please, do not crucify me for the awkwardness. Leave it aside and enjoy the content!

RESUMO GERAL

SAMUEL-ROSA, Alessandro. **Análise de fontes de incerteza na modelagem espacial do solo**. 2016. 278p. Tese (Doutorado em Agronomia, Ciência do Solo). Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2016.

A modelagem espacial do solo moderna usa modelos estatísticos para explorar a relação empírica entre as condições ambientais e as propriedades do solo. Esses modelos são uma simplificação da realidade, e seu resultado (mapa do solo) estará sempre *errado*. O que um mapa do solo transmite é o que esperamos que o solo seja, reconhecendo que somos *incertos* sobre ele. O objetivo dessa tese é avaliar importantes fontes de incerteza na modelagem espacial do solo, com ênfase nos dados do solo e covariáveis. Estudos de caso foram desenvolvidos usando dados de uma bacia hidrográfica do sul do Brasil. A distribuição espacial do solo na área de estudo é variável, sendo determinada pela geologia e geomorfologia (escalas espaciais maiores) e práticas agrícolas (escalas espaciais menores). Quatro propriedades do solo foram exploradas: teor de argila, teor de carbono orgânico, capacidade de troca catiônica efetiva e densidade. Cinco covariáveis, cada um com dois níveis de detalhe espacial, foram utilizadas: mapas areais de classes de solo, modelos digitais de elevação, mapas geológicos, mapas de uso da terra, e imagens de satélite. Esses dados constituem o *conjunto de dados de Santa Maria*. Dois pacotes para `R` foram criados, o primeiro (`pedometrics`) contendo várias funções para a análise exploratória espacial de dados e calibração de modelos, o segundo (`spsann`) projetado para a optimização de amostras espaciais usando recozimento simulado. Os estudos de caso ilustraram que as covariáveis existentes são apropriadas para calibrar modelos espaciais do solo, e que o uso de covariáveis mais detalhadas resulta em modesto aumento na acurácia de predição que pode não compensar os custos adicionais. Meios mais eficientes de aumentar a acurácia de predição devem ser explorados, como obter mais observações do solo. Para esse fim, deve-se usar meios objetivos para a seleção dos locais de observação a fim de minimizar os efeitos das respostas psicológicas dos modeladores do solo a fatores conceituais e operacionais sobre o plano de amostragem. Isso porque as dificuldades conceituais e operacionais encontradas no campo determinam mudanças na motivação dos modeladores do solo entre aprendizagem/verificação das relações solo-paisagem e maximização do número de observações e cobertura geográfica. Para estimar a tendência espacial, deve ser suficiente otimizar as amostras espaciais visando somente reproduzir a distribuição marginal das covariáveis. Para otimizar configurações amostrais para estimar a tendência espacial e o variograma, e interpolação espacial, pode-se formular um problema de otimização multi-objetivo sólido usando versões robustas de algoritmos de amostragem existentes. No geral, aprendemos que uma receita única, universal para a redução da incerteza na modelagem espacial do solo não pode ser formulada. Decidir sobre formas eficazes de redução da incerteza requer, em primeiro lugar, que exploremos todo o potencial dos dados existentes usando técnicas de modelagem espacial sólidas.

Palavras-chave: Pedometria. Mapeamento Digital do Solo. Dados de Solo e Covariáveis.

GENERAL ABSTRACT

SAMUEL-ROSA, Alessandro. **Analysis of sources of uncertainty in soil spatial modelling**. 2016. 278p. Thesis (Doctor of Science in Agronomy, Soil Science). Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2016.

Modern soil spatial modelling is based on statistical models to explore the empirical relationship among environmental conditions and soil properties. These models are a simplification of reality, and their outcome (soil map) will always be in *error*. What a soil map conveys is what we expect the soil to be, acknowledging that we are *uncertain* about it. The objective of this thesis is to evaluate important sources of uncertainty in spatial soil modelling, with emphasis on soil and covariate data. Case studies were developed using data from a catchment located in Southern Brazil. The soil spatial distribution in the study area is highly variable, being determined by the geology and geomorphology (coarse spatial scales), and by agricultural practices (fine spatial scales). Four topsoil properties were explored: clay content, organic carbon content, effective cation exchange capacity and bulk density. Five covariates, each with two levels of spatial detail, were used: area-class soil maps, digital elevation models, geologic maps, land use maps, and satellite images. These soil and covariate data constitute the *Santa Maria dataset*. Two packages for `R` were created in support to the case studies, the first (`pedometrics`) containing various functions for spatial exploratory data analysis and model calibration, the second (`spsann`) designed for the optimization of spatial samples using simulated annealing. The case studies illustrated that existing covariates are suitable for calibrating soil spatial models, and that using more detailed covariates results in only a modest increase in the prediction accuracy that may not outweigh the extra costs. More efficient means of increasing prediction accuracy should be explored, such as obtaining more soil observations. For this end, one should use objective means for selecting observation locations to minimize the effects of psychological responses of soil modellers to conceptual and operational factors on the sampling design. This because conceptual and operational difficulties encountered in the field determine how the motivation of soil modellers shifts between learning/verifying soil-landscape relationships and maximizing the number of observations and geographic coverage. For the sole purpose of spatial trend estimation, it should suffice to optimize spatial samples aiming only at reproducing the marginal distribution of the covariates. For the joint purpose of optimizing sample configurations for spatial trend and variogram estimation, and spatial interpolation, one can formulate a sound multi-objective optimization problem using robust versions of existing sampling algorithms. Overall, we have learned that a single, universal recipe for reducing our uncertainty in soil spatial modelling cannot be formulated. Deciding upon efficient ways of reducing our uncertainty requires, first, that we explore the full potential of existing soil and covariate data using sound spatial modelling techniques.

Keywords: Pedometrics. Digital Soil Mapping. Soil and Covariate Data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE APÊNDICES

SUMÁRIO

1 GENERAL INTRODUCTION

Modern soil spatial modelling is based on using statistical models to explore the empirical relationship among environmental conditions and soil properties. These soil spatial models, like any other model, are nothing more than a simplification of reality. Unless we observe the soil everywhere – which would destroy the soil and render the observations useless –, no matter how large the volume of data is, or how comprehensive our background knowledge, it will *never* be possible to construct a model that explains the entire complexity of the soil. Thus, the outcome of a soil spatial model, i.e. a soil map, will *always* deviate from the "truth" – this deviation from the "truth" is what we call *error*. What a soil map conveys is what we expect the soil to be, acknowledging that there is *uncertainty* about it.

Because soil spatial modellers aim at using the available resources to produce the most accurate representation of the soil, a sensible research programme is to investigate the main causes for soil maps being more or less *uncertain*. There are many sources of uncertainty in soil spatial modelling, such as the errors that result from using a poor statistical model or from making interpolations and extrapolations to predict soil properties at unvisited locations. Another important source of uncertainty is the data used to assess the empirical relationship among environmental conditions and soil properties: covariate and soil data.

The general objective of this thesis is to evaluate important sources of uncertainty in soil spatial modelling with emphasis on soil and covariate data. This general objective can be divided into specific objectives and their respective research questions:

(I) Determine the suitability of freely available covariates to calibrate soil spatial models.

   (a) Does the use of more detailed covariates result in considerably more accurate soil maps?

   (b) How does incorporation of spatial dependence in a soil spatial model compares to the gain in prediction accuracy obtained from using more detailed covariates?

   (c) Are the answers to these research questions consistent across soil properties?

(II) Identify the factors that determine how field soil spatial modellers select soil observation locations.

   (a) Which factors are considered for deciding upon the location of soil observations, and do they have a pedological origin?

   (b) Do the factors play the same role along the course of the soil observation process?

   (c) Can point pattern analysis help understanding the purposive sampling strategy traditionally employed by field soil spatial modellers?

(III) Identify appropriate calibration sample sizes and designs for soil spatial modelling.

   (a) Can the conditioned Latin hypercube sampling algorithm be improved, and does this improvement deliver more accurate soil spatial predictions?

   (b) Which are the most theoretically sound sampling algorithms for spatial trend estimation, variogram estimation, and spatial prediction when we know very little about the soil spatial variation?

* The Portuguese version of this General Introduction is included in Appendix C.

(c) Can these sampling algorithms be used to construct a general purpose sampling algorithm?

The thesis is composed of eight chapters where each of the above mentioned objectives are met. Although there is a logical sequence in their presentation, all chapters were planned so that they could be read separately. This means that there is some overlap between them, i.e. repeated information. References to specific sections of other chapters using coloured (blue) hyperlinks are common.

Chapter I is a commented review of the literature on soil spatial modelling and its main sources of uncertainty. The review starts with a discussion about the efforts made by soil spatial modellers to raise awareness about the importance of soil spatial information. These efforts seem to have fuelled a global scientific demand for up-to-date, high resolution soil spatial information. The chapter continues with a description of soil spatial modelling along human history, suggesting that the goal of producing soil maps remains more or less the same since the Neolithic Revolution (ca. 10 000 years). The chapter closes with the main sources of uncertainty.

Chapter II describes the soil data included in the *Santa Maria dataset*, which was used to develop the case studies presented in this thesis. The Santa Maria dataset is composed of $n = 410$ soil observations compiled from studies carried out between 2004 and 2013. These studies aimed at producing semi-detailed soil and land use maps, and modelling topsoil carbon stock and vulnerability to erosion. A comprehensive description of the covariate data included in the Santa Maria dataset, and their processing, is given in Chapter III. Chapter IV presents the conceptual model of pedogenesis (in Portuguese), which consists of a description of the study area that includes an explicit description of soil-forming factors (climate, geology, geomorphology, hydrology, land use, and vegetation) and processes that determine the soil spatio-temporal distribution. Beyond describing the data used in the thesis, the goal of these chapters, along with the conceptual model of pedogenesis, is to provide the basis for future soil spatial modelling exercises in the study area, and to serve as examples for new soil spatial modelling studies developed elsewhere.

Based on an article published in the peer reviewed journal Geoderma, Chapter V serves the purpose of meeting the first objective of the thesis and answering its respective research questions. The prediction performance of linear soil spatial models calibrated using covariates (area-class soil maps, land use maps, geological maps, digital elevation models, and satellite images) available in two levels of detail is evaluated. The influence of taking the spatial dependence of the residuals into account is assessed as well.

Chapter VI presents an approach that aims at helping to understand the purposive sampling strategy traditionally employed by field soil modellers, i.e. free survey. This is important because many soil spatial modelling projects rely on legacy data, i.e. soil data produced previously and made available (publicly or not), whose observation locations were purposively selected by soil spatial modellers using poorly documented tacit rules. The chapter is designed to answer the research questions of the second objective of the thesis. Point pattern analysis is used to characterize the spatial sample configuration, whereas theories borrowed from Psychology are used to elaborate on the subjective factors involved in selecting soil observation locations.

Objective 3 and its research questions are addressed in Chapter VII and Chapter VIII. In Chapter VII, three improved sampling algorithms are compared to the original conditioned Latin hypercube sampling algorithm on how they affect geographic coverage, estimated model parameters and prediction accuracy. The influence of sample size is also discussed. Chapter VIII presents the most efficient sampling strategies for spatial trend estimation, variogram estimation, and spatial prediction when we know very little about the soil spatial variation. The chapter closes with a new general purpose sampling algorithm that aims at the three objectives jointly.

The sequence of eight chapters is closed with General Conclusions where I highlight the main research results and contributions of the study. Next, there are two appendices, both devoted to the description of the two packages for `R` developed to support the thesis: `spsann` (Appendix A) and `pedometrics` (Appendix B). The first was designed for the optimization of sample configurations using spatial simulated annealing. The second includes miscellaneous functions that were put together for ease of use. All literature references are presented under a unique list of Bibliographic References at the end of the thesis.

2 CHAPTER I

MODERN SOIL SPATIAL MODELLING AND ITS SOURCES OF UNCERTAINTY

2.1 RESUMO

Os esforços da comunidade da ciência do solo têm motivado a comunidade científica a reconhecer a importância do solo para a humanidade e para o ambiente a nível local, regional e global. Os *modeladores espaciais do solo* parecem estar sendo capazes de convencer os tomadores de decisão e formuladores de políticas públicas sobre a importância de produzir e atualizar as informações do solo. Para essa finalidade, os modeladores espacias do solo tem usado o *modelo misto de variação espacial*. O modelo misto de variação espacial integra aspectos dos métodos "tradicionais" de modelagem espacial do solo, baseados no *modelo discreto de variação espacial*, além de técnicas geoestatísticas, mais formalmente o *modelo contínuo de variação espacial*. Como tal, o modelo misto de variação espacial explora o conhecimento dos fatores de formação do solo, bem como o fato de que o solo é um meio contínuo. Ele também reconhece que mapas do solo *sempre* desviam da "verdade", o que significa que um mapa do solo transmite o que esperamos que o solo seja, não a nossa certeza sobre ele. As fontes de nosso incerteza sobre o solo são muitas. Por exemplo, os dados do solo e covariáveis usados para calibrar modelos espaciais do solo são uma importante fonte de incerteza. Dados do solo pode conter erros e representar pobremente a população da qual foram amostrados. A influência dos dados das covariáveis na nossa incerteza sobre o solo se expressa através da sua fraca correlação com as propriedades do solo. Finalmente, a forma do modelo utilizado para explicar empiricamente as estruturas de variação espacial de pequena a grande escala pode determinar consideravelmente a precisão das nossas predições espaciais. Como não podemos eliminar a incerteza de um mapa do solo, o nosso conhecimento sobre o solo será *sempre* limitado. Apesar disso, modelos espaciais do solo ainda são necessários para guiar nossas ações do dia a dia.

Palavras-chave: Demanda por Informação do Solo. Modelo Misto de Variação Espacial. Dados do Solo e Covariáveis. Estrutura do Modelo.

2.2 ABSTRACT

The efforts of the soil science community have motivated the scientific community to recognize the importance of soils for humanity and the environment at the local, regional, and global levels. *Soil spatial modellers* seem to be able to convince policy and decision makers about the importance of producing and updating soil information. For that end, soil spatial modellers have been using the *mixed model of spatial variation*. The mixed model of spatial variation integrates aspects of "traditional" methods of soil spatial modelling, based on the *discrete model of spatial variation*, as well of geostatistical techniques, more formally the *continuous model of spatial variation*. As such, the mixed model of spatial variation explores the knowledge of soil-forming factors as well as the fact that the soil is a continuous media. It also acknowledges that soil maps *always* deviate from the "truth", which means that a soil map conveys what we expect the soil to be, not our certainty about it. There are many sources for our uncertainty about the soil. For instance, the soil and covariate data used to calibrate soil spatial models is an important source of uncertainty. Soil data can have errors and poorly represent the population from which it has been sampled. The influence of the covariate data on our uncertainty about the soil expresses itself through the poor correlation with soil properties. Finally, the form of the model used to empirically account for the structures of spatial variation from the small to the large scales can greatly determined the accuracy of our spatial predictions. Because we cannot eliminate the uncertainty of a soil map, our knowledge about the soil will *always* be limited. Despite of this, soil spatial models are still needed to guide our every-day actions.

Keywords: Demand for Soil Information. Mixed Model of Spatial Variation. Soil and Covariate Data. Model Structure.

2.3 DEMAND FOR SOIL SPATIAL INFORMATION

Many soil spatial modellers[7] have complained for many years about the decreasing interest in producing and updating soil information, not only in Brazil (DALMOLIN, 1999; KER, 1999; KER; NOVAIS, 2003; MENDONÇA-SANTOS; SANTOS, 2003; RAMOS, 2003; ESPINDOLA, 2008; SAMUEL-ROSA, 2012), but in many countries around the world (BASHER, 1997; HARTEMINK; MCBRATNEY, 2008; GRUNWALD, 2009; SANCHEZ et al., 2009; FINKE, 2012). Several reasons were presented to explain the general lack of interest in producing and updating soil information after the 1980s: the use of specialized taxonomic terminology by soil spatial modellers was abusive; information conveyed by soil maps was too limited due to its qualitative nature; policy and decision makers were unaware of the usefulness of soil information and dynamicity of soil; applied scientific research came to be preferred over basic scientific research; soil spatial modelling largely ignored environmental applications other than agriculture; lack of communication between soil spatial modellers and the general public; among others. But everyone seem to agree on one point: governments understood that producing and updating soil information was too costly. Cutting down the budget for soil spatial modelling fundamentally was an economic decision.

Since the last decade, soil scientists in general have launched many initiatives to make soil become a hot topic (HARTEMINK; MCBRATNEY, 2008). For example, the United Nations (UN) declared 5 December the World Soil Day and 2015 the International Year of Soils "in an effort to raise awareness and promote more sustainable use of this critical resource". Soil spatial modellers created a global consortium, the GlobalSoilMap, with the goal of producing "a new digital soil map of the world using state-of-the-art and emerging technologies". The Food and Agriculture Organization (FAO) launched a Global Soil Partnership (GSP) for "leading to the adoption of sustainable development goals for soils". The International Union of Soil Sciences (IUSS) created a working group, funded by the United States Department of Agriculture (USDA) to develop a Universal Soil Classification System, "a common language to describe soils that can be used internationally". An Intergovernmental Technical Panel on Soils (ITPS) was formed with soil experts from all regions of the world "to provide scientific and technical advice and guidance on global soil issues to the Global Soil Partnership". The Bill & Melinda Gates foundation handed out an $18 million grant "to map most parts in Sub-Saharan Africa, and make all Sub-Saharan Africa soil data available". Soil spatial modellers at the International Soil Reference and Information Centre (ISRIC) launched its Global Soil Information Facilities (GSIF), a "framework for production of open soil data", which has already output 250 m-resolution soil maps with global coverage. In Brazil, soil spatial modellers created the Brazilian Network for Research in Digital Soil Mapping (RedeMDS) with the objective of "generating synergy among Brazilian soil scientists to advance research in digital soil mapping".

The efforts of the soil science community have motivated the scientific community to recognize the importance of soils for humanity and the environment at the local, regional, and

[7] The use that I give to the expression *soil spatial modeller* throughout this thesis is approximately equivalent to expressions traditionally used in the academic world such as soil scientist, soil surveyor, soil taxonomist, geostatistician, pedometrician, soil investigator, soil mapper, and so on. In this thesis, a soil spatial modeller is any person that *constructs* an explanation – a model – of the observed spatial soil variation using the tools and techniques available at his/her time and place. The goal of a soil spatial modeller is to construct a model that is simple yet able to produce an accurate representation of the spatial soil variation given the available resources and its intended application. I call this activity *soil spatial modelling*. Accordingly, I understand that those that are excluded from the academic world such as peasants, farmers, indigenous populations, and so on, are soil spatial modellers as well, although their modelling of the soil is not the focus of this thesis.

global levels (SANCHEZ et al., 2009; KEMPEN, 2011; OMUTO et al., 2013). Soil scientists, whose presence in public administration through scientific and technical advisory boards appears to grow, seem to have been able to convince policy and decision makers about the importance of producing and updating soil information. For example, in Brazil, the Federal Court of Accounts (TCU), in collaboration with Embrapa Soils and other soil science related institutions, held a Soil Governance Conference, where a new National Program for Soil Survey and Interpretation of Brazil (PRONASOLOS) was announced, with an expected budget of R$8 billion[8]. In the sequence, Embrapa Soils created a working group, with soil spatial modellers from other government institutions and universities, and the first PRONASOLOS report with proposal and goals was completed on December 2015. Unfortunately, it is not clear whether the national soil survey program will truly be restarted because, like happened in the end of the 1980s, it essentially is an economic decision. Apparently, other areas of soil science have not been receiving much more attention and/or funding than soil spatial modelling. It is also not clear whether the soil science community efforts have brought about a renewed recognition of the importance of soils for humanity and the environment among the general public.

2.4 MODERN SOIL SPATIAL MODELLING

Technology plays a determinant role on how we perceive the world around us – see, for example, Hartemink (2009). When early farmers, during the Neolithic Revolution, ca. 10 000 years ago, first observed that soil properties varied in space, they probably soon figured out that such variation was related to other environmental features and influenced crop yields. That early, rough, approximate understanding – a *model* – of soil spatial variation certainly was fundamental for choosing – *predicting* – the most appropriate locations to start and maintain human settlements, some of which became great, long-standing empires (MAZOYER; ROUDART, 2008; BREVIK; HARTEMINK, 2010; CHURCHMAN, 2010). Archaeological research provides evidence that several of these empires had more formal *soil classification systems* and *soil survey programs*, in most cases for taxation purposes (BARRERA-BASSOLS; ZINCK, 2003) – a practice that lasts till today.

A lot happened since the Neolithic Revolution (BREVIK; HARTEMINK, 2010) – from bone to spacecraft, as in Kubrick's *2001: A Space Odyssey* –, and the knowledge constructed with multiple soil spatial studies was fundamental for the development of agriculture and increase of food production – although many farmers still live in Neolithic conditions (MAZOYER; ROUDART, 2008). If we adopt an integrative view, soil maps produced during this long period of human history seem to fit into what we call today the *discrete model of spa-*

[8] The initiative to restart the national soil survey program coincides with the currently increased, government funded, economically driven, historical pressure to occupy the Cerrado and Amazon biomes, considered "the last agricultural frontier" (CORREIA, 2005; MACARINI, 2005; SILVA, 2005; CARVALHO et al., 2009; BATLLE-BAYER et al., 2010; MARTINELLI et al., 2010; SCHNEIDER; PERES, 2015). It is traditionally argumented that transforming parts of the Cerrado and Amazon biomes into agricultural land is needed to eradicate poverty in Brazil and to feed a growing world population. This is one of the argument used by the Brazilian politicians who are in favour of changing the Brazilian legislation to easy the acquisition of up to 100 000 ha of agricultural land in these regions by multinational corporations to produce cellulose and paper (SECOM, 2015). Unfortunately, the parts of the Brazilian territory that compose "the last agricultural frontier" suffer from severe social inequalities intensified by a long history of land conflicts fuelled by the conservative development model adopted in Brazil (COMISSÃO PASTORAL DA TERRA, 2015; SCHNEIDER; PERES, 2015; FERNANDES, 2016), where the benefits of economic growth are not shared by all people. In regions such as this, the problem of food insecurity (and other societal problems) is likely more due to the lack of political will than to the lack of food (FAO, 2005; FAO, 2009; FAO, 2015). As such, one might wonder whether restarting the national soil survey program decoupled from a deep-cutting agrarian reform is beneficial for the general Brazilian population.

tial variation. The discrete model of spatial variation explains the variation of soil properties in space using mutually exclusive mapping units that are separated by sharply defined, crisp boundaries (i.e. polygons) (HEUVELINK, 1996; LEGROS, 2006). The soil[9] within each mapping unit is more or less homogeneous with regard to its properties at the time of mapping. These properties, which are generally used to name the mapping unit along with other environmental features, can be characterised using one or more direct observations made within the domain of the mapping unit (WEBSTER; OLIVER, 1990; ROSSITER, 2000; LEGROS, 2006).

2.4.1 Discrete Model of Spatial Variation

A key step was given in 1886 in Russia with the formalization of the approximate understanding of the soil spatial variation using scientific parlance, i.e. the postulation of the "the basic law of soil science" by Vasily Dokuchaev (FLORINSKY, 2012): "Any... soil is always and everywhere a mere function of the following factors of soil formation: 1) the nature (content and structure) of the parent rock, 2) the climate of the given terrain, 3) the mass and character of vegetation, 4) the age of the terrain, and finally, 5) the terrain topography.". The basic law of soil science was presented 40-years later by Sergey Zakharov in the form of a general soil formation equation, which is known in the western soil science literature as (JENNY, 1941; FLORINSKY, 2012)

$$soil = f(cl, o, r, p, t, \ldots), \tag{2.1}$$

where $soil$ is the soil and its properties, cl if the climate, o are the organisms, including humans, r stands for relief or topography, p is the parent material, t is time or age of the terrain, and $\ldots$ stand for other unknown players. Dokuchaev was aware that producing empirical evidence to corroborate his basic law of soil science was difficult because data on soil formation factors was scarce. Besides, it was difficult to numerically express the relation between soil and formation factors. Despite of these difficulties, the basic law of soil science was readily adopted by soil spatial modellers because it provided a solid basis for explaining the soil spatial variation (SMITH, 1986).

An important enthusiast and supporter of the use of Equation 2.1 was Hans Jenny (1941). He believed that the large volume of already existing soil data/knowledge, which had been constructed mostly based on the basic law of soil science, needed to be organized by means of numerical laws and quantitative theories – instead of soil maps, taxonomic classifications, and soil-forming processes – to enable treating it mathematically (i.e. using empirical correlation).

[9] It is common sense among many soil scientists, not only in Brazil, that the activity of *soil mapping* concerns the production of *area-class soil maps*, also called *polygon soil maps*, *choropleth soil maps*, and more generally *soil maps*. My view is that such an understanding, which is rooted in the fact that most early soil mapping projects aimed at producing area-class soil maps – see Grunwald (2009) for a discussion –, is erroneous. Multiple times it results in the confusion between the concept of soil itself and that of soil class, soil taxon and/or soil series. In this thesis a soil map is nothing more than a graphical representation of the soil. For that end, one has to choose a finite number of properties, characteristics, attributes, features of the soil. The taxonomic classification is one such *feature*. Thus, one must bear in mind that, throughout this thesis, the term *soil* is used to mean *soil* – the uppermost layer of unconsolidated material of the Earth's surface... –, not soil class or soil taxon or soil series. One could argue that a soil class and a soil property are completely different things. For Rossiter (2000) a major difference is that soil properties are measurable characteristics of the soil, while soil classes are categories of a predefined classification. This view is not entirely correct because even measurable characteristics of the soil such as the particle size distribution are categories of a predefined classification. It follows that, in this thesis, any property, class, characteristic, attribute, feature of the soil whose spatial variations is being modelled is treated as any other *reponse variable* or *dependent variable* that enters a model.

For that end, solving Equation 2.1 depended on the soil scientist' skills to select suitable study areas and locations for making observations. But the problem continued to be that obtaining data on soil-forming factors was still difficult compared to obtaining soil data. It follows that direct application of Equation 2.1 for producing soil maps was impossible because it required soil-forming factors to be exhaustively known everywhere (JENNY, 1941). Despite the operational difficulties encountered since the postulation of the basic law of soil science and definition of Equation 2.1, the concept of soil-forming factors were employed in most of the subsequent soil spatial studies around the world, resulting in the enhancement of taxonomic classifications, theories about soil-forming processes, and production of soil maps using the discrete model of spatial variation (SCHELLING, 1970; HUDSON, 1992; BOCKHEIM; GENNADIYEV, 2000; LEGROS, 2006; KRASILNIKOV et al., 2009; HARTEMINK; BOCKHEIM, 2013).

Soil spatial modelling using the discrete model of spatial variation and the idea that soil properties were determined by soil-forming factors had its weaknesses as any other model of spatial variation. Three main weaknesses can be pointed out, all of which only were recognized and understood using post-war scientific/technological developments (HEUVELINK; WEBSTER, 2001; MCBRATNEY et al., 2003; SCULL et al., 2003). First, "soil bodies" were described as discrete, homogeneous entities – although it was recognized that soil is a continuous media whose properties vary from place to place in such a way that nearby locations have more similar soil property values than distant locations (spatial autocorrelation) –, implying that the fluxes of energy and matter across the landscape had to be understood as being partially homogeneous (within a mapping unit), partially discontinuous (between mapping units) processes. Second, the uncertainty (the acknowledgement of errors) about mapped soil properties was disregarded, meaning that a single, absolute value for each soil property would be attributed to each mapping unit ignoring that soil properties vary from place to place and that estimates can be affected by all sorts of errors. Last, but not least, some important soil spatial modelling decisions could not be efficiently shared with others by means of formal, explicit knowledge because they were largely based on the intuitive, tacit knowledge of soil modellers, i.e. the knowledge that a soil modeller has about the soil-landscape relationships and the soil modelling process but that cannot be adequately communicated, articulated by verbal (written or spoken) means. This was evidenced, for example, by the fact that different soil modellers would produce considerably different soil maps without being able to explain why (LEGROS, 2006; BAZAGLIA FILHO et al., 2013).

2.4.2 Continuous Model of Spatial Variation

Different solutions were explored during the post-war to overcome one or another weaknesses of the discrete model of spatial variation. Most of these solutions came from the new developments in the fields of mathematics, statistics and informatics. For example, those important soil spatial modelling decisions, generally taken with basis on the tacit knowledge of the soil modeller, could now be more efficiently communicated with others through the use of computers – provided they had a computer. This is because using a computer to produce soil maps requires modelling rules to be formalized in the form of a computer script, which is the mean used to establish the communication between the soil modeller (a human being) and the data processing environment (a computer).

Some soil modellers also came to understand that the error about the mapped soil property could be acknowledged using classical statistical theory. First, the definition of mapping units, which was based on the knowledge of the soil-forming factors, needed to be viewed as a modelling exercise that aims at minimizing the within-unit variance (and maximizing the between-unit variance) of a soil property (VOLTZ; WEBSTER, 1990). This was equivalent to designing controlled agronomic experiments as devised by Ronald Fisher in the United Kingdom during the 1920s, by which large plots and blocking are used to deal with the effects of short and long-range variation, respectively (WEBSTER; OLIVER, 2007). In both cases the spatial autocorrelation was regarded as being of little importance – it still is in most agricultural experiments. Then, under the assumption of spatial independence within mapping units, the most likely value of a soil property at any one point in a given mapping unit would be its mean over the multiple observations made in that mapping unit (VOLTZ; WEBSTER, 1990; CRESSIE, 1993). It follows that the uncertainty about the value of the mapped soil property is the same everywhere within a mapping unit, irrespective of the location of existing soil observations (HEUVELINK, 1996). This is because due to statistical independence and constant variance within mapping units, the expected error of predicting the value of a soil property at any location with its mean is also the same everywhere. The only requirement so that the estimated mean and variance would be fair (unbiased) was that the location of soil observations be selected using some form of randomization (probability sampling) such as it is done in controlled agronomic experiments (DE GRUIJTER; TER BRAAK, 1990).

Those solutions still required describing “soil bodies” as discrete, homogeneous entities, and continued neglecting the spatial variation and autocorrelation of soil properties. Besides, despite the statistical soundness, most soil modellers fiercely rejected to employ random sampling and *design-based* estimation. They have always preferred to select observation locations purposively according to their mental model of soil-landscape relationships because, from a pedological perspective, this seemed more reasonable. However, they were generally unaware of the biasedness of their *model-based* estimates. For other soil spatial modellers a solution was to partially ignore the knowledge on soil-forming factors and use a different model of spatial variation: the *continuous model of spatial variation*. Accordingly, the main requirement was that a soil property be treated as a regionalised variable, the outcome of a random (stochastic) spatial process (CRESSIE, 1993; WEBSTER, 2000). A comprehensive theory of *regionalised variables* was first constructed during the 1960s by the French mathematician Georges Matheron, heavily based on earlier studies of engineers, mathematician, meteorology and statistician such as the South African Daniel Krige, the Swedish Bertil Matérn, the Russian Andrey Kolmogorov, among others (KRIGE, 1951; MATÉRN, 1960; MATHERON, 1965; MATHERON; KLEINGELD, 1987; CRESSIE, 1990; WEBSTER; OLIVER, 2007).

At first, the proposition of *imagining* the soil as being the result of randomness is awkward (see, for example, Equation 2.1). However, pondering on the multitude of soil-forming factors and processes, and on the complexity of their interactions (BOCKHEIM; GENNADIYEV, 2010; GRUNWALD et al., 2011), as well as on the limitations of the existing knowledge on soil spatial variation, one can easily conclude that soil is a chaotic media (WEBSTER, 2000) – so why not treat it as such? One only had to *imagine* that the value of a soil property observed at a given location simply was one that happened to be recorded by chance among an infinitely large number of values that could have been recorded instead (WEBSTER, 2000). In other words, one had to *imagine* that, if a soil property is recorded multiple times at the same location, the recorded values would not necessarily be the same. As such, a soil property at a given location would not be described using a single absolute value, but a probability distribution function (PDF), generally the normal (Gaussian), characterised by the mean and variance of those “recorded” values. Accordingly, to describe the fact that soil property values at nearby locations are more similar than at locations further apart – they covary, vary together or jointly – would require a joint probability distribution function, characterised by the covariance (WEBSTER; OLIVER, 1990; CRESSIE, 1993). The obvious difficulty in this approach is that the PDF cannot be defined because we usually have one single value recorded at each observation location.

Different from its discrete counterpart, the continuous model of spatial variation takes into account the relative location of existing observations when predicting a soil property at a given unobserved location. In the simplest case, the most likely value of a soil property in a given location is defined by a constant mean computed over all observations made in the mapping region, plus a random variable with mean zero and (spatial) covariance that depends only on the separation distance between locations (WEBSTER; OLIVER, 1990; CRESSIE, 1993). This prediction method is known as the *best linear unbiased predictor* (BLUP), usually called *ordinary kriging*. The main idea underlying the continuous model of spatial variation is that soil property values at nearby locations are more similar than at locations further apart (WEBSTER; OLIVER, 1990). Thus, we err less if we predict the value of a soil property at a given location with a value observed at a nearby location than with a value observed at a distant location. Optimally, because the soil is a continuous media and its properties vary in space in an autocorrelated fashion, we make the most accurate prediction taking a weighted average of the values recorded at all observation locations, nearby locations receiving larger weights than distant locations. In both cases, it follows that the uncertainty about the mapped soil property is larger the farther from existing soil observations, i.e. it is spatially varying (CRESSIE, 1993).

2.4.3 Mixed Model of Spatial Variation

Availability of general-purpose computers fuelled the use and development of the continuous model of spatial variation, especially in European, North American, and Oceanian countries (HEUVELINK; WEBSTER, 2001; MCBRATNEY et al., 2003; SCULL et al., 2003). But limitations in its prediction performance and developments in remote sensing and nonlinear models helped many soil scientists to understand that the continuous model of spatial variation had limitations too. For instance, it is unable to capture abrupt changes in the values of soil properties that occur, for example, between agricultural fields, parent materials, land uses, and so on (STEIN et al., 1988; VOLTZ; WEBSTER, 1990). Because, compared to the discrete model of spatial variation, in its simplest form, the continuous model of spatial variation largely ignores the existing pedological knowledge (GRUNWALD, 2009; LARK, 2012). Soil scientists also understood that the discrete model of spatial variation was more efficient than previously thought (BREGT et al., 1987; HEUVELINK; WEBSTER, 2001). First, because it was now possible to employ the equation of soil-forming factors (Equation 2.1) for soil mapping using remote sensing products as surrogates of the factors of soil formation (MOORE et al., 1993). Second, nonlinear models enabled identifying complex spatial patterns that before could only be identified by an experienced soil scientist (MCKENZIE; RYAN, 1999). The most logical step was to combine the strengths of both discrete and continuous models of spatial variation into a single model – the *mixed model of spatial variation* –, that is, inclusion of the existing pedological knowledge and consideration of the spatial continuity of soil property values.

The mixed model of spatial variation[10] can be viewed as a generalization of previously

[10] The use that I give to the expression *mixed model of spatial variation* in this thesis is approximately equivalent to expressions such as regression-kriging, kriging with external drift, universal kriging, hybrid approach for soil mapping, pedometric mapping, digital soil mapping, predictive soil mapping, environmental correlation, geostatistical mapping, soil-landscape modelling, and so on. As far as I know, soil spatial modellers have not yet reached an agreement on how these "traditional" expressions should be used, nor what their "correct" meaning is. See, for example, Hengl (2003), McBratney et al. (2003), Scull et al. (2003). Because the expression *mixed model of spatial variation* has a stronger theoretical basis (other expressions are defined based on operational aspects), I

existing models of spatial variation, by which a soil property, here denoted $Y(\boldsymbol{s})$, at a given spatial location symbolised as $\boldsymbol{s}$ is modelled as the outcome of a spatial stochastic process (CRESSIE, 1993; HEUVELINK; WEBSTER, 2001; LARK et al., 2006). Accordingly, the model is composed of *fixed* and *random effects*. The fixed effects, a deterministic large-scale spatial trend, denoted $m(\boldsymbol{s})$, describes the portion of the spatial variation of the soil property that is explained with the factors of soil formation as suggested by the empirical correlation calculated using point soil observations and spatially exhaustive covariates[11]. The random effects, also known as stochastic residuals or latent variables, denoted $e(\boldsymbol{s})$, describe the portion of the spatial variation of the soil property that cannot be explained with the covariates but is potentially spatially correlated, the form and degree of this spatial correlation possibly being interpreted pedologically (LARK, 2012). Thus

$$Y(\boldsymbol{s}) = m(\boldsymbol{s}) + e(\boldsymbol{s}). \tag{2.2}$$

Equation 2.2 possesses a great flexibility that facilitates to explore newly developed statistical and data-mining methods, generally resulting in better performance than the constituent models alone, as well as integrating the existing pedological knowledge provided it is translated into a mathematical form (ODEH et al., 1994; ODEH et al., 1995; HEUVELINK, 1996; MCBRATNEY et al., 2000; HENGL et al., 2004; LÓPEZ-GRANADOS et al., 2005; WEBSTER; OLIVER, 2007; GRUNWALD, 2009; LARK, 2012). These features promoted the rapid popularization of the mixed model of spatial variation, and many recent large scale soil-mapping projects already successfully employed the mixed model of spatial variation (POGGIO; GIMONA, 2014; NUSSBAUM et al., 2014; HENGL et al., 2015).

2.4.4 Spatial Soil Modelling Steps

Despite the rapid technological developments observed recently, the activity of modelling the soil remained essentially the same throughout human history. Soil maps still serve the same old purpose of representing our limited understanding about the spatial organization of the soil in the natural environment in a simplified manner, as well as giving insights about how the soil came to be and how they should be managed (JENNY, 1941; HUDSON, 1992; LEGROS, 2006; BLANCO-CANQUI; LAL, 2008; GRUNWALD, 2010). For this reason, irrespective of the method/model used to produce soil maps, perhaps we should use an (integrative) expression such as *soil spatial modelling* instead of picking one out of the many currently used.

It follows that, based on the existing body of knowledge on soil spatial modelling and the currently available technologies, we should try to define what it constitutes, in practice, the soil spatial modelling activity. A suggested general (didactic) sequence of eight steps that we believe represent what is or should be done in soil spatial modelling is presented below. The first step is numbered zero (0) to emphasize its importance: we believe it to be a crucial step, where political, economical and paradigmatic decisions have to be made. Many soil spatial

understand that its adoption and use is a more appropriate choice.

[11] In the statistical literature, the term *covariate* is synonymous to *explanatory variable*, *predictor variable*, and *independent variable*, and refers to the variable that, although not being of primary interest, is used in a model because it determines the behaviour of the *response variable* or *dependent variable* (EVERITT, 2006). In soil science, a covariate has the same statistical meaning, the difference being that they assume a pedological meaning as well, i.e. they are viewed as proxies, indicators, substitutes, surrogates, approximations of the soil-forming factors due to the simple fact that the soil-forming factors – the true environmental conditions that helped shape the soil – are unknown and cannot be known. Covariates are defined *spatially exhaustive* when they cover the entire area being modelled, i.e. they are exhaustively known in the entire geographic space.

modelling projects are not implemented because this first step is not completed.

Step 0 Identify a reality or problem entity, the geographic region for which there is a demand of spatial soil information. Target soil properties are appointed as well as the required accompanying output information (e.g. metadata). A minimum required accuracy level of the produced soil information can also be defined. Key modelling decisions are taken in this step such as the support (punctual or areal), spatial resolution (and possibly the cartographic scale), coordinate reference system, etc. Depending on how well defined the demand is, the model of spatial variation can also be specified, i.e. discrete, continuous, or mixed. Data policy is discussed (What data should be public? How to make data public? How to implement the data policy?) and agreed upon. Finally, the available infrastructure, budget, time, and workforce are specified so that next steps can be appropriately planned as to fulfil the demand.

Step 1 Develop a conceptual model of pedogenesis, a verbal representation of the reality or problem entity including the explicit description of soil-forming factors and processes that determine the spatio-temporal distribution of soil properties. This requires gathering the most of the existing environmental information contained in scientific articles, technical reports, books, websites, local knowledge, as well as existing maps of the soil, land use, geology, digital elevation models, satellite images, aerial photographs, among others. Environmental information is used to articulate pedogenetic concepts. Provided that any of the existing soil data are available, an exploratory data analysis can help unravelling soil-landscape relationships. The poorer the volume of existing environmental information, and the less experienced the soil modeller is, the more important an exploratory field campaign is to help understanding the existing soil-landscape relationships.

Step 2 Define the model of spatial variation, a translation of the conceptual model of pedogenesis into a set of possible mathematical representations. Depending on how well defined the demand was, the model of spatial variation was already specified in **Step 0**. Provided the volume of existing environmental information and legacy soil data is moderate to large and/or the soil modeller is very experienced and/or the available budget allows carrying out exploratory field campaigns, a single model of spatial variation is defined, i.e. discrete, continuous, or mixed. Assuming that the mixed model of spatial variation is chosen, the statistical and/or data-mining models that will be used to represent the discrete and continuous components are specified, taking into account the feasibility of meeting their requirements given the available soil data, infrastructure, budget, time, and workforce. If multiple models or statistical and/or data-mining models are chosen, the pedological and statistical criteria for identifying the best performing model are defined.

Step 3 Prepare the modelling database, a collection of soil and covariate data needed to estimate the parameters and test the chosen statistical and/or data-mining models. In some rare cases the modelling database is already available and does not require any further improvement. If this is the case, then this step would have already been covered in **Step 1**. However, in most cases this step requires preparing a sampling plan with formally defined selection rules, making properly documented field soil observations, and running replicated laboratory analyses. Soil data from different sources are harmonized. Covariates are selected using the conceptual model of pedogenesis and empirical evidence. Both soil and covariate data are assessed regarding the need for nonlinear transformations to meet the requirements of the chosen statistical and/or data-mining models, and to improve their empirical correlation. Several of these tasks can be (and usually are) carried out with the aid of a data processing environment (a computer).

Step 4 Estimate the parameters of the statistical and/or data-mining models, a task that essentially depends on translating the set of possible mathematical representations of the conceptual model of pedogenesis into a computer representation, that is, a computer code or script. Developing a well documented computer code that describes all processing steps ease re-design, future consistency checks, correction of mistakes, and dissemination/reproducibility. Calibrated models are evaluated using statistical criteria defined in **Step 2** such as goodness-of-fit measures, as well as regarding their tenability (pedological evaluation). The latter includes visually assessing draft soil maps, and how well they represent the range of possible mathematical models. Failure in any of these evaluations suggests that the model requires adjustments, possibly more calibration data, or that it can be discarded.

Step 5 Validate the statistical and/or data-mining models, preferentially predicting the values of the modelled soil properties at a set of independent, probabilistically selected observation locations for which the true values are known. If an independent set of observation locations is unavailable, validation is performed using leave-one-out cross-validation. The best performing model is selected using the set of criteria defined in **Step 2**: this is the best mathematical representation of the reality under study given the available data and knowledge. If two or more models present similar performance and have a considerably different structure, then (1) an aggregated version of these models is constructed or (2) parsimony is considered to elect the simpler model with fewer variables, steps, rules, etc. If previous steps have already allowed defining/selecting a single model, statistical validation is used only to assess model accuracy.

Step 6 Make spatial predictions, the application of the best performing model(s) to predict soil properties values at unvisited locations in the entire modelling region. If demanded, the uncertainty about the predicted soil properties values (i.e. the prediction interval) is estimated too. Maps of the target soil properties as well as the required accompanying output information (e.g. point soil observations, covariate maps, uncertainty maps, metadata, computer scripts) are delivered to the users of the soil information, and possibly used to populate a spatial soil information system, where they are made available for inspection using different visualization techniques. Provided there is infrastructure, budget, time, and workforce available, modelling steps can be re-designed and the outputs updated at the user request.

Step 7 Improve on the conceptual model of pedogenesis, the use of the knowledge gained during the previous steps until the production of the soil property maps, which give insights about the reality or problem entity under study, to correct and/or improve the description of soil-forming factors and processes that determine the soil spatio-temporal distribution. If demanded, the reformulated conceptual model of pedogenesis is delivered to the users of the soil information as well to help in scenario analysis and decision making.

2.5 SOURCES OF UNCERTAINTY IN SOIL SPATIAL MODELLING

Soil spatial models, like any other model (BOX, 1976), are nothing more than a simplification of reality. The reason for this is the fact that the existing knowledge and data are *always* (very) limited – we have to try our best with the available resources. This means that soil spatial models are unable to explain the spatial soil variation in its entirety, but only a small part of it (HEUVELINK, 1998; LEGROS, 2006). When we use a spatial soil model to produce continuous representations of soil properties across space, i.e. soil maps, these continuous representations will inexorably deviate from the "truth". What the soil map presents is our most likely expectation about the soil properties – not our *certainty* about them. The deviation from the "truth" is what we call *error*.

Many examples from the soil spatial modelling literature show that, irrespective of the soil property, spatial soil models have a quite variable predictive performance. In general, these models explain between 15 % (or less) and (rarely more than) 75 % of the soil spatial variation (MOORE et al., 1993; ODEH et al., 1994; GESSLER et al., 1995; MCKENZIE; RYAN, 1999; GOBIN et al., 2001; SUMFLETH; DUTTMANN, 2008; SUN et al., 2012; ROSSEL et al., 2013; NUSSBAUM et al., 2014; HENGL et al., 2015; GASCH et al., 2015; HEUNG et al., 2016). Thus, because we cannot eliminate the uncertainty of a soil map, they can always be considered as wrong, the difference being that some might be useful.

2.5.1 Soil Data

Soil modellers aim at producing the most accurate representations of the soil given the available resources. Thus, a reasonable research program is the identification of the causes for soil maps being more or less accurate. For instance, the error that results from making extrapolations and interpolations to predict soil properties at unvisited locations is an important source of uncertainty (HEUVELINK; PEBESMA, 1999; REFSGAARD et al., 2006). In the case of data-centred soil models, such as the mixed model of spatial variation, these errors are larger the farther we are from the existing observations. Thus, the most efficient way of reducing these errors generally is to increase the number of observations and improve the spatial coverage of the mapping region (BRUS; HEUVELINK, 2007). However, most soil spatial modelling projects must rely on using only *legacy* soil data, i.e. soil data produced many years or decades ago (KEMPEN et al., 2009; HENGL et al., 2014; POGGIO; GIMONA, 2014; NUSSBAUM et al., 2014; MULDER et al., 2016).

Legacy soil data have weaknesses that can affect prediction accuracy. For example, the size of the impact of using outdated soil data for predicting the current status of soil properties is unknown. However, we can expect that this impact will be larger on the accuracy of predictions of soil properties that are relatively more sensitive to land-use and climatic changes such as the organic carbon content. This is one of the reasons for the high uncertainty of global estimates of organic carbon storage in soils (HENGL et al., 2014). A second weakness is the generally poor "quality" of soil legacy data for current uses. A poor quality arises from documentation inconsistencies, diversity of field and laboratory methods, use of different measurement units and nomenclature, absence of spatial positioning data, varying levels of precision and accuracy, among others (RIBEIRO et al., 2015). As for the use of outdated data, how much the use of non-standardised and non-harmonised soil data affects soil spatial predictions remains unknown. Finally, legacy soil data can poorly represent the mapping region due to bias in the selection of observations locations. This location bias comes from soil modellers sampling preferentially in complex and less known areas (to gain knowledge of soil-landscape relationships) or by convenience in most easy access areas (to optimize the use of the available resources) (ROSSITER, 2000; DE GRUIJTER et al., 2006).

Some times the budget of the soil spatial modelling project (fortunately) includes (additional) sampling. Then soil modellers have to decide upon the number and spatial configuration of the sample (DE GRUIJTER et al., 2006; WEBSTER; LARK, 2013). But this is not an easy task, except if the goal is on modelling very few soil properties that can be rapidly measured using field sensors, and for which a model of spatial (co)variation can be assumed known (MARCHANT; LARK, 2006). In most cases, several difficult-to-measure soil properties have to be modelled, many of which have a poorly known structure of spatial (co)variation. If using the mixed model of spatial variation, the chosen spatial sample configuration has to be appropriate for estimating the spatial trend (HENGL et al., 2003; MINASNY; MCBRATNEY, 2006) and the variogram model (WARRICK; MYERS, 1987; WEBSTER; OLIVER, 1992; LARK, 2002), making spatial predictions (YFANTIS et al., 1987; WALVOORT et al., 2010) and, (cross)validating spatial predictions (BRUS et al., 2011), four often conflicting objectives. Besides, one must be careful not to decide upon collecting an insufficient or an exaggerated number of soil samples. Under-sampling can result in soil models with large uncertainty, while over-sampling can produce modelling benefits that do not outweigh the sampling costs (VAN GROENIGEN et al., 1999).

Soil data can uniformly cover the geographic and/or feature spaces[12], but at an observation density that still is unable to capture the main structures of correlated spatial variation of a soil property. This is likely to happen when the range of correlated spatial variation is relatively short (small scale variation) compared to the size of the mapping region (large scale variation) (BURROUGH, 1983). Soil data can also have significant laboratory and positional errors (NELSON et al., 2011). Besides, data on some soil properties may naturally contain more errors due to conceptual definitions and analytical procedures employed. Take particle-size distribution as an example. First, the errors are propagated to the fraction obtained by difference (silt). Second, pre-treatments, such as organic matter oxidation, can change mineral structure (MIKUTTA et al., 2005). And third, ignoring that the particle-size distribution is a compositional datum can introduce bias in the predictions (LARK; BISHOP, 2007).

2.5.2 Covariate Data

Another important source of uncertainty is the covariate data used to calibrate the soil models. Their effect appears as a poor correlation with soil properties being modelled, resulting in poor predictions. Several are the possible reasons for the covariate data to be poorly correlated with the soil property. One of the them is the fact that the covariate data also contain varying levels of error (HEUVELINK et al., 1989) which could make them poor proxies of the soil-forming factors. These errors derive from the various methods of data generation, analytical procedures and, inherent characteristics of each site. For example, digital elevation models usually present larger errors in areas with steep slopes, rough terrains, and high altitude, and with dense forest cover or urbanized (FLORINSKY, 1998; TOUTIN, 2000; FISHER; TATE, 2006). Interpolation of elevation data using kriging can produce spurious artefacts (HENGL; EVANS, 2009) compared to hydrologically consistent algorithms (HUTCHINSON, 1989). And stereoscopic correlation techniques generally produce digital elevation models with poorer quality

[12] The *feature space*, also known as *attribute space*, is the multi-dimensional mathematical space defined by the set of covariates used in a soil spatial modelling exercise. Specifically, if the number of covariates is equal to p, then the feature space is a p-dimensional mathematical space. In contrast, the *geographic space* is the bi-dimensional mathematical space defined by the geographic coordinates of the mapping region. Both spaces are *finite* in the sense that they are bounded by the set of geographic coordinates that define the geographic limits of the study area (or mapping region) and any existing non-mapping area. However, while the feature space is a *discrete space* – the set of possible values attributed to the covariates is limited to the set of known existing covariate values –, the geographic space is a *continuous space* – any pair of values can be attributed to the geographic coordinates provided it is bounded by the geographic limits of the study area and any existing non-mapping area.

than interferometric synthetic aperture radar (HIRT et al., 2010).

The method used to select which covariates to include in the soil spatial model can also be a reason for soil maps to be in more or less error. This is because every method can select considerably different sets of covariates, perhaps including a few that are poorly correlated with soil properties. The selection of covariates is especially important because the number of available (uncertain) covariates is large, many of which possibly being statistically redundant.

A pedologically sound approach for the selection of covariates is the elicitation of the knowledge of a few experts (LARK et al., 2007; MEYER; BOOKER, 2001). An expert is every soil modeller with long-term practical experience in soil spatial modelling and deep knowledge of the mapping region, as expressed in the conceptual model of pedogenesis, as well as of the soil and covariate data. However, it may be that the understanding about the mapping region may be limited to a point that enables building only a very poor conceptual model of pedogenesis. Take for example a geomorphologically complex, unstable landscape that consistently suffers from natural and/or anthropogenic alterations. Establishing soil-landscape relationships is very difficult in such circumstances, especially if the geomorphological complexity is increased as the landscape is rejuvenated (STRECK et al., 2008). Similar uncertainty regarding the soil-landscape relationships can arise in very old, stable surfaces of the tropics that have gone through many environmental modifications, but were not affected by Pleistocene glaciations (MCKENZIE; GALLANT, 2006). These landscapes commonly have polygenetic soils that present properties reflecting today and ancient vegetation and climate (PAIN; OILIER, 1995; KER, 1998). One could wonder whether expert soil modellers would be more efficient at identifying the covariates with the highest predictive power than sound statistical inference in such circumstances.

A commonly used approach to select the covariates to enter a soil spatial model is automated selection. Automated covariate selection algorithms have always been relatively easy to use (DRAPER et al., 1971) and are available in most software packages (HARRELL, 2001). They generally deliver satisfactory results and are needed to automate soil spatial modelling routines (HENGL et al., 2014). Each of them employs different search strategies. For example, some methods analyse all possible combinations of covariates. Others simulate the process of natural selection (ANDERSEN; BRO, 2010). Cross-validation selects covariates that produce the best predictions on test sets (GUYON; ELISSEEFF, 2003). Other methods take into account the order in which the covariates are added to (forward selection) or removed from (backward elimination) the model (LARK et al., 2007). The stepwise method adds and removes covariates until no further addition or removal results in significant changes in the model (DRAPER; SMITH, 1998). Some prefer to use dimensionality reduction techniques to reduce the number of covariates (MASSY, 1965) before running the covariate selection algorithm (TEN CATEN et al., 2011; HENGL et al., 2014). Several other methods exist and have been used in soil spatial modelling (POGGIO et al., 2013; NUSSBAUM et al., 2014). Because each algorithm employs a different search strategy, they generally select different sets of covariates possibly with varying degrees of correlation with the soil property being modelled. The best covariate selection algorithm can be identified, for example, using sound statistical inference made on external validation data. However, this is rarely done and the most popular algorithm, or that implemented in the available software package is used. Besides, there are evidences that the number of problems associated with using automated covariate selection methods, and dimensionality reduction techniques, can be greater than the number of advantages (FARRAR; GLAUBER, 1967; JACKSON, 1993; CHATFIELD, 1995; EDIRISOORIYA, 1995; HARRELL, 2001; JOLLIFFE, 2002; PERES-NETO et al., 2005; LARK et al., 2007; RATNER, 2010).

Despite the difficulties associated with using uncertain covariates and automated selection algorithms, their contribution to the error in soil spatial modelling is poorly understood. For example, more accurate covariate data may not necessarily be a better proxy of the soil-forming factors. In very old and stable surfaces of the tropics such as described above, current soil properties are more due to past than to present environmental conditions. Present day covariate data are not indicators of such environmental conditions. Indeed, “degrading” the accuracy of the covariates to produce multiple representations of the landscape from moderate to large spatial scales can yield more accurate predictions (BEHRENS et al., 2010). However, this multiplies the number of covariates, increasing the need to better understand the effects of using automated selection algorithms in soil spatial modelling.

2.5.3 Model Structure

The last source of uncertainty discussed here is the structure of the model of spatial variation. Such a source of uncertainty arises from the fact that soil spatial variation can be modelled using very different spatial models (Section 2.4). Based on the existing knowledge and available data and resources, the soil modeller must decide between modelling the deterministic (discrete model) or the stochastic (continuous model) or both components (mixed model) of soil spatial variation. Optimally, the best model structure is selected based on statistical inferences made using external validation data. However, like covariate selection algorithms, soil modellers usually select a single model, generally the most popular or that implemented in the available software package.

The deterministic component of spatial variation, which uses information on the soil-forming factors (i.e. covariates), usually has a larger ability to explain the structure of the large scale soil spatial variation. For that end, besides the covariates (Section 2.5.2), soil modellers have to choose the form of the model that will be used to account for the empirical relation between covariates and soil properties. Early soil spatial modelling was based solely on using *mental models* (or tacit models) that could account for many complex effects of the covariates on soil properties (HUDSON, 1992). Later, due to developments in statistics, these mental models were replaced with *linear models* (MOORE et al., 1993; ODEH et al., 1994). These linear models account for the effects of the covariates on the soil property being modelled only through a weighted sum of the individual effects of all covariates (HARRELL, 2001).

Developments in informatics introduced new forms of modelling soil-landscape relationships. Many soil spatial modellers started to employ nonlinear models such as classification and regression trees, artificial neural networks, random forests, support vector machines, among many others (MCBRATNEY et al., 2000; HEUNG et al., 2016). The main reason for the ready adoption of these nonlinear models is that they usually have a greater ability to capture more complex site-specific soil-landscape relations (GRUNWALD, 2009), possibly yielding more accurate predictions.

Modelling the structure of the small scale soil spatial variation usually requires employing stochastic models. This can be done using both the continuous and mixed models of spatial variation. Here the soil modeller has to decide upon the form of the spatial covariance function (exponential, spherical, circular, Gaussian, linear, etc) (WEBSTER; OLIVER, 2007). Again, the decision can be based on the existing pedological knowledge and on using statistical inferences made with validation data. However, different spatial covariance functions yield spatial predictions with similar accuracies, although the visual aspect of the resulting maps is considerably different (LARK, 2000a). This also happens when modelling the deterministic component of spatial variation (HEUNG et al., 2016). A solution is to aggregate the predictions of the many different models, but this will not necessarily reduce our uncertainty.

2.6 FINAL CONSIDERATIONS

The sources of uncertainty in soil spatial modelling are multiple and the list that has been presented here is far from being comprehensive – others will do a better job. The main message is that we *cannot* eliminate the uncertainty of a soil map and, as such, our knowledge about the soil will *always* be limited. Despite of this, soil spatial models – and the entire body of human knowledge – are still needed to guide our every-day actions. So, instead of eliminating uncertainty, the real quest is for understanding it and its sources. For instance, if it happens to us to gain knowledge that a source of uncertainty systematically affects our soil spatial models in a very specific manner, then a corrective measure can be taken right away.

The most comprehensive way of dealing with the multiple sources of uncertainty is *error propagation analysis*, also called *uncertainty analysis* (HEUVELINK et al., 1989; TAYLOR, 1997). This is done taking our uncertainty into account through all modelling steps, seeing how it propagates, and evaluating its impact on the uncertainty about the output soil map. This would allow us identifying the main source of uncertainty, so that we could try to take corrective measures to improve its quality. But such an exercise is cumbersome, and the efforts required may not outweigh the benefits, this being one of the main reasons why it is rarely carried out (HEUVELINK et al., 2006; NELSON et al., 2011) – including this thesis. Another important reason for error propagation analysis being unpopular is the common ignorance and lack of understanding – perhaps prejudice – about error and uncertainty (WECHSLER, 2003; HEUVELINK, 2005).

Considering the difficulties with implementing a sound error propagation analysis – given the available resources –, it seems reasonable to devise studies limited to only a few sources of uncertainty. The knowledge on the sources of uncertainty gathered here suggests that the calibration and covariate data, and the model structure should be the focus of such studies. In general, the main decisions made by soil spatial modellers concern these three topics. However, fine tuning the model structure seems to strongly depend on the existing soil and covariate data. For example, capturing the moderate to large scale structures of soil spatial variation largely depends on using covariates that depict environmental conditions at multiple scales. This also requires the soil data to cover the feature and geographic spaces. On the other hand, capturing the structure of the short range spatial variation depends on having soil observations at short distances. Thus, it may be that evaluating the impact of soil and covariate data on prediction accuracy should precede that of the model structure. Besides, gathering soil and covariate data have a larger weight on the costs of soil spatial modelling that choosing the model structure.

3 CHAPTER II

THE SANTA MARIA DATASET. PART I – SOIL DATA

[*] Collaborated in the preparation of this chapter: Pablo Miguel (UFPel), Jean Michel Moura Bueno (UFSM), Ricardo Simão Diniz Dalmolin (UFSM), Andrisa Balbinot (UFSM), Lúcia Helena Cunha dos Anjos (UFRRJ), Gustavo de Mattos Vasques (Embrapa Solos), Gerard B. M. Heuvelink (ISRIC – World Soil Information), and Ad van Oostrum (ISRIC – World Soil Information).

3.1 RESUMO

O *conjunto de dados de Santa Maria* compreende dados do solo de $n = 410$ observações feitas entre 2008 e 2013 na bacia do reservatório do Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento (DNOS-CORSAN), localizada no estado brasileiro do Rio Grande do Sul. Esses dados do solo foram produzidos no âmbito de projetos de pesquisa que visam a produção de mapas semi-detalhados do solo e do uso da terra, e prever o estoque superficial de carbono no solo e sua vulnerabilidade à erosão. Todos os locais de observação foram selecionados intencionalmente ou por conveniência. Várias características ambientais foram descritas nos locais de observação, tais como o uso da terra, geologia, classificação do solo, declividade, condições de drenagem, presença de fragmentos grosseiros e afloramentos rochosos, cobertura do solo com vegetação, entre outras peculiaridades de cada local de observação que não foram registradas de uma forma sistemática. As amostras do solo foram submetidos à análise de laboratório para determinar o conteúdo de carbono orgânico no solo, a distribuição do tamanho de partículas, a densidade e o conteúdo de bases (cálcio, magnésio, potássio e sódio) e acidez trocáveis. A capacidade de troca de cátions efetiva foi calculada como a soma das bases e acidez trocáveis. Os dados do solo estão disponíveis gratuitamente em um repositório hospedado no GitHub. Eles incluem a identificação de todos os locais de observação, as suas coordenadas geográficas, e dados de campo e de laboratório. O número de repetições de laboratório e o desvio padrão amostral também são fornecidos.

Palavras-chave: Modelagem espacial do solo. Amostragem intencional. Dados legados. Descrição do solo no campo. Análise laboratorial do solo.

3.2 ABSTRACT

The *Santa Maria dataset* comprises soil data from $n = 410$ soil observations made between 2008 and 2013 in the catchment of the reservoir of the *Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento* (DNOS-CORSAN), located in the southern Brazilian state of Rio Grande do Sul. These soil data were produced within the scope of research projects that aimed at producing semi-detailed soil and land use maps, and predicting topsoil carbon stock and vulnerability to erosion. All observation locations were selected purposively or by convenience. Several environmental features were described at the observation locations, such as land use, geology, soil classification, slope, drainage condition, presence of coarse fragments and rock outcrops, soil coverage with vegetation, among other peculiarities of each observation location that were not recorded in a systematic way. Soil samples were submitted to laboratory analysis to determine the soil organic carbon content, particle size distribution, bulk density, and the content of exchangeable bases (calcium, magnesium, potassium, and sodium) and acidity. The effective cation exchange capacity was calculated as the sum of exchangeable bases and acidity. The soil data is freely available in a repository hosted in GitHub. These include the identification of all observation locations, their geographic coordinates, and field and laboratory data. The number of laboratory replicates and the sample standard deviation is provided as well.

Keywords: Spatial soil modelling. Purposive sampling. Legacy data. Soil field description. Soil laboratory analysis.

3.3 INTRODUCTION

The *Santa Maria dataset* comprises soil data from $n = 410$ soil observations made between 2004 and 2013 in the catchment of the reservoir of the *Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento* (DNOS-CORSAN), henceforth called *DNOS catchment*, located in the southern border of the Plateau of the Paraná Sedimentary Basin, in the city of Santa Maria, state of Rio Grande do Sul, Brazil. Soil observations cover the northern sector of the DNOS catchment – an area of ±2000 ha, which corresponds to ±60 % of the entire catchment. These soil data were produced during the development of research projects that aimed at producing semi-detailed soil and land use maps (cartographic scale of 1:25 000) (PEDRON, 2005; MIGUEL, 2010; SAMUEL-ROSA et al., 2011; MIGUEL et al., 2011), and predicting topsoil carbon stock and vulnerability to erosion (SAMUEL-ROSA, 2009; MOURA-BUENO, 2012; MIGUEL, 2013).

This chapter presents a thorough description of the soil data contained in the Santa Maria dataset. A thorough description of the procedures for soil sampling and description, as well as the analytical methods employed[13] is given. Soil data is also described using summary plots with descriptive statistics.

The chapter is divided in four sections. Section 3.4 presents general information about the data, as well as the structure of the database where they have been stored and managed. Next, Section 3.5 describes how field soil observation was carried out in each of the different project that contributed with soil data to the Santa Maria dataset. Methods used to describe the soil in the field are described in Section 3.6. Section 3.7 closes the chapter with a description of the laboratory methods used to produce data on physical and chemical soil properties.

3.4 DATABASE STRUCTURE

The soil data contained in the Santa Maria dataset, as well as the code used in its processing, is freely available in the web-based Git repository <https://github.com/samuel-rosa/dnos-sm-rs-general/>. This is the same repository containing the ground control data (Section 4.4) The repository has the following folder structure:

```
dnos-sm-rs-general
|- code/                  # source code folder
|  - R/                   # R source code folder
|    - general.R          # R source code file
|
|- data/                  # soil data folder
|  - fieldData.csv        # field data file
|  - fieldMetadata.csv    # field metadata file
|  - labData.csv          # laboratory data file
|  - labMetadata.csv      # laboratory metadata file
|- README.md              # description of the repository
```

Soil data files are available as comma-separated values (CSV) files. The identification

[13] The reader should be aware that soil science evolved in Brazil following a somewhat different pathway than in the countries of the northern hemisphere due to the specific soil features of tropical and subtropical regions. Methods have been adapted along the years, possibly leading to nomenclature mismatches. The reader is invited to contribute to solve any problems in this document.

of all observation locations, their geographic coordinates, and field and laboratory data are contained in files `fieldData.csv` and `labData.csv`, respectively. Files `fieldMetadata.csv` and `labMetadata.csv` contain the metadata. The coordinate reference system (CRS) is `WGS 84`, coded `EPSG:4326` by the European Petroleum Survey Group (EPSG).

Every soil property is identified with a code composed of three or four capital letters. For example, soil organic carbon is identified with `ORCA`. A column containing the number of laboratory replicates is identified with the code of the soil property followed by the letter "N". The column containing the sample standard deviation is identified in the same manner, but using "SD". For example, `ORCA_N` and `ORCA_SD`.

3.5 FIELD SAMPLING

The Santa Maria dataset is composed of three subsets which are described in the next three sections. Together, these subsets yield a sampling density of about ± 0.18 observations per hectare, with an average separation distance between two neighbouring points of 180 m, minimum and maximum separation distances of 18 and 328 m, 95 % of neighbouring observations being separated by more than 49 m.

3.5.1 Subset I

The first subset ($n = 340$, Figure 3.1) was produced between 2008 and 20011 as part of projects that aimed at producing semi-detailed soil and land use maps, and predicting topsoil carbon stock and vulnerability to erosion (SAMUEL-ROSA, 2009; SAMUEL-ROSA et al., 2011; MIGUEL et al., 2011; MOURA-BUENO et al., 2012; SAMUEL-ROSA et al., 2013). The researchers faced several difficulties with a budget cut and shortage of workforce. They also had restricted access to several areas due to geographic barriers and prohibition of access by some landowners. These difficulties forced the researchers to reduce the originally aimed number of observations ($n = 500$) during the development of the project.

All observation locations were selected purposively or by convenience. Tacit knowledge (Section 2.4.1 and Section 7.5.1) was the main tool to choose the observation locations, a process that was carried out in the office using 1 m spatial resolution Google Earth® imagery of the years of 2008 and 2009. The main goal of the researchers was to obtain a sample that they understood as being representative of the different landforms, land uses, and soil taxa present in the DNOS catchment. They also wanted the observations to be spread throughout the entire DNOS catchment.

At the observation locations, the researchers defined an area of $\pm 100\,\mathrm{m}^2$ within which they opened three soil pits up to a depth of 20 cm. Soil samples were collected up to a depth of 20 cm, the depth being measured with a ruler. The resulting sampling depth of Subset I varies from 2 to 20 cm, with an average of 17 cm. This variation of the vertical sampling support[14] was

[14] *Sample support* refers to the shape, size and orientation of sampling units, the latter being the smallest single entity that we are able or choose to observe in the universe of interest, i.e. the sampling region. A discrete universe such as a forest is defined by the collection of these single entities. However, by definition, such single entities have no real, physical existence in continuous universes such as the soil – their "existence" require our prior, more or less arbitrary, definition of their shape, size and orientation. This definition is usually based on theoretical and practical considerations. For example, the sampling unit can be defined as a roughly polygonal block that is large enough to encompass the pattern of small-scale local vertical (≤ 2 m) and horizontal (1–10 m^2) variability of soil properties – a pedon. Depending on the size of the sampling unit relative to the universe of interest, the sample support is referred to as *areal* or *point* sample support. A pedon of 10 m^2 area observed in an agricultural field of

not a problem for the researchers because their goal was to sample the *topsoil*. The topsoil was defined as the topmost soil layer, with a thickness equal or inferior to $20\,\mathrm{cm}$, being the soil layer most susceptible to degradation induced by poor agricultural practices and land use changes.

Soil samples from the three pits opened in each sampling area were used to produce a composite sample which was used for laboratory analyses. Subsurface soil features were observed with an auger in each pit, and the average (continuous variables) or most common (categorical variables) value recorded. Note that soil sampling was done using an areal horizontal support – an area of $\pm 100\,\mathrm{m}^2$. However, the shape and exact area of the sampling units are unknown, and georeferencing took place at point support.

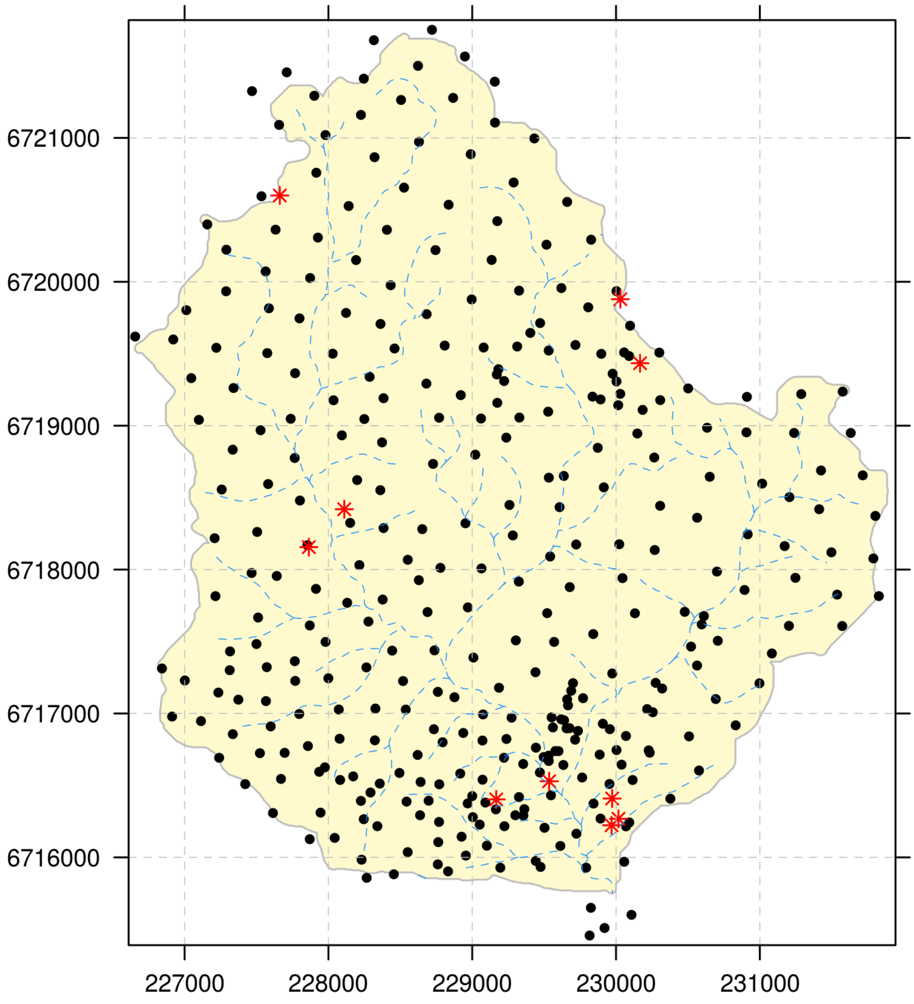


Figure 3.1: Spatial distribution of the soil observations contained in *Subset I* ($n = 340$, black solid circles) and *Subset III* ($n = 10$, red stars) of the Santa Maria dataset. The drainage network is shown in the background (blue dashed line) to give an idea of how the locations of soil observations is related to terrain features.

Georeferencing was done in the field using a Global Navigation Satellite System (GNSS) receiver with a horizontal positional error of less than $8\,\mathrm{m}$ positioned approximately at the centre of the sampling area. Sometimes, the horizontal positional error was larger than $8\,\mathrm{m}$ due

1 ha corresponds to the *areal sample support*. However, the same pedon observed in a catchment of $200\,\mathrm{ha}$ would correspond to the *point sample support*.

the effects of vegetation, terrain, and satellite configuration. In these cases, observation locations were georeferenced in the office using 1 m spatial resolution Google Earth® imagery with positional horizontal error of 6 m (Table 4.5).

Every observation was identified with a number in increasing order, following the order in which the observations were made (1–340). A total of 17 field campaigns were carried out, yielding an observation density of about 18 observations per km^2 (Chapter VI).

3.5.2 Subset II

The second subset ($n = 60$, Figure 3.2) was produced in the years 2012 and 2013, and was intended to constitute an independent dataset for validation purposes. Because of the many access limitations (geographic barriers and prohibition by landowners) and shortage of workforce, budget, infrastructure and time faced in previous field campaigns, researchers chose to employ transect (cluster) sampling (MIGUEL et al., 2011; MOURA-BUENO et al., 2012; SAMUEL-ROSA et al., 2013). They started defining the population of transects using their knowledge of the study area, taking into account the factors that they thought determined the spatial distribution of soil properties. Each researcher (three) delineated $m = 60$ easily accessible, straight transects of 400 m following the spatial gradients of selected environmental features (topography, geology, vegetation, land use, and soils), totalling 180 transects. Accordingly, knowledge of existing roads, human settlements, water bodies, and other access limitations was used as well. The activity was carried out using 1 m spatial resolution Google Earth® imagery of the years of 2008 and 2009.

Twelve out of the $m = 180$ transects were randomly selected using as many iterations as necessary until there were no intersecting transects, and there was at least one transect in each of the three major morphostructural units of the DNOS catchment (*Planalto*, *Rebordo do Planalto*, and *Depressão Periférica*) (Chapter IV). Finally, $n = 5$ observation locations, separated by equidistant intervals of 100 m, were selected in each transect. Observation locations were named with a number in increasing order, following the order in which the observations were made, starting from 341 (341–400).

The location of the observations was identified in the field using a GNSS receiver with a horizontal positional error of less than 8 m. Soil sampling and description was carried out using the same procedure used with *Subset I*, except for the fact that a single soil pit was opened within a radius of 2 m from the predefined observation location. More accurate geographic coordinates were collected in the field using a Differential Global Positioning System (DGPS) with a horizontal positional error of less than 1 cm.

3.5.3 Subset III

The third subset ($n = 10$) contains data compiled from the studies of Pedron (2005) and Miguel (2010), specifically from the uppermost A horizon of modal soil profiles (point support) whose locations were purposively selected using tacit knowledge after a preliminary area-class soil map had been produced and/or the observations included in *Subset I* had been made.

Pedron (2005) and Miguel (2010) aimed at observation locations that they understood as being most representative of the soil mapping units depicted in their respective area-class soil maps. A single soil sample was taken from each of the described soil horizons and used for laboratory analysis. The resulting thickness of the uppermost A horizons varies from 12 to 30 cm, with a mean of 22.6 cm. Georeferencing was carried using a GNSS receiver with

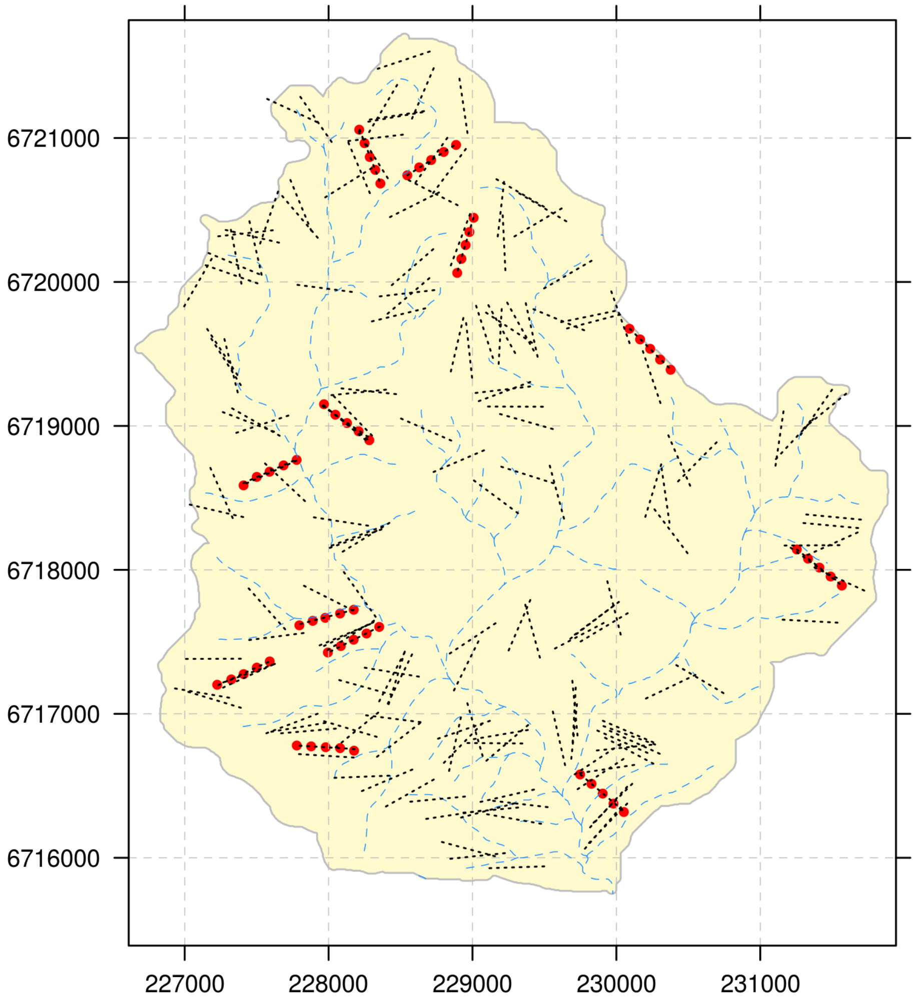


Figure 3.2: Three soil spatial modellers manually drew 180 straight transects (black dotted lines) aligned in the direction of maximum expected spatial variation of environmental conditions. They avoided locations where it was known that geographic barriers or landowners would impede the access to make soil observations. Twelve transects were probabilistically selected using simple random sampling to yield $n = 60$ validation observations (red solid circles) separated by equidistant intervals of 100 m. The drainage network is shown in the background (blue dashed line) to give an idea of how the location and direction of transects is related to terrain features.

a horizontal positional error of less than 8 m positioned at the observation location. Data are identified in the Santa Maria dataset using the same identification that was used in the studies from which they were compiled.

3.6 FIELD DESCRIPTION

Several environmental features were described at the observation locations. This sections present a summary description of how this description was done, specially for subsets I and II, which have not been documented before. For subset III, a thorough explanation of how field description was done is given by Pedron (2005) and Miguel (2010).

Despite the different origins of the datasets, soil sampling and description guidelines are very similar. As such, merging field descriptions from subset III with those of subset I and II was easy, rarely requiring conceptual translations and adaptations – this practice is reported when used.

Finally, the code used in the database to identify each of the variables described in the field is presented between parenthesis using fixed-width or monospace font.

3.6.1 Land Use and Vegetation

Land use (`LAND`) was assessed at the time of sampling using data collected in the field. Five land uses were identified using nomenclature of FAO (2006) (Figure 3.3):

`animal husbandry` Native grasslands used for animal husbandry.

`crop agriculture` Annual and biannual crop agriculture.

`forestry` Plantations of *Eucalyptus spp.* and *Pinus spp.*.

`native forest` Primary or secondary native forests.

`shrubland` Abandoned areas with predominance of shrub-sized vegetation, known in Brazil as *capoeira*.

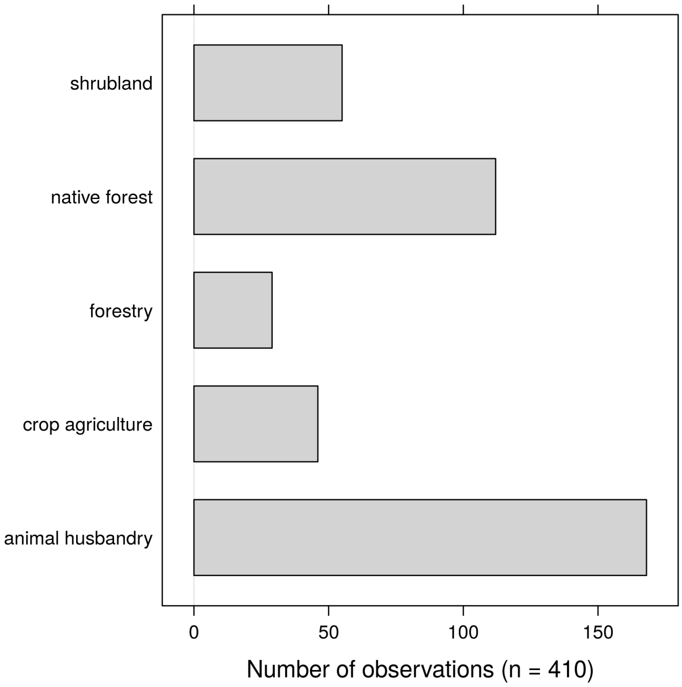


Figure 3.3: Distribution of land use types in the Santa Maria dataset. Most soil observations were made in areas used for animal husbandry although more that half of the area is occupied with native forest (Section 4.9).

Other land uses are observed in the study area such as human settlements and water bodies (Section 4.9). However, due to access constraints, soil observations were not made in areas under these land uses.

3.6.2 Geology

Soil parent material (`PARENT`) was inferred in the field from direct observation of soil properties and local environmental features. Two classes were identified using nomenclature of (FAO, 2006):

`igneous` Soil derived from the *in sutu* weathering of igneous rocks.

`sedimentary` Soil derived from the *in sutu* weathering of sedimentary rocks, or from sediments of igneous and/or sedimentary rocks.

Underlying geologic formation (`GEO`) and lithology (`LITHO`) were inferred based on soil properties and environmental features observed in the field, and on existing area-class soil maps (Section 4.6) geologic maps (Section 4.8).

3.6.3 Soil Classification

The most likely taxon (`TAXON`) of the Brazilian System of Soil Classification (SiBCS) (SANTOS et al., 2013) was inferred in the field using data obtained from direct observation of soil properties (20 cm-deep soil pits and auger holes down to the diagnostic subsurface horizon or bedrock) and local environmental features. These data were then interpreted using the bases and concepts of the SiBCS to identify the most likely taxon up to the second taxonomic level of the SiBCS. Further levels of the SiBCS were not considered because making any sort of inference would require data that were not observable in the field. Eleven taxa were identified (Figure 3.4):

`CX` Cambissolo Háplico. Moderately developed soil.

`GX` Gleissolo Háplico. Poorly drained greyish soil with a somewhat constant clay content throughout the profile.

`PA` Argissolo Amarelo. Soil with significant increase of the clay content with depth, and with a yellowish B horizon.

`PBAC` Argissolo Bruno-Acinzentado. Soil with significant increase of the clay content with depth, and with an upper B horizon slightly darker than the lower B horizons.

`PV` Argissolo Vermelho. Soil with significant increase of the clay content with depth, and with a reddish B horizon.

`PVA` Argissolo Vermelho-Amarelo. Soil with significant increase of the clay content with depth, and with a reddish-yellowish B horizon.

`RL` Neossolo Litólico. Poorly developed soil.

`RQ` Neossolo Quartzarênico. Deep sandy soil derived from sediments.

`RR` Neossolo Regolítico. Poorly to moderately developed soil.

`RY` Neossolo Flúvico. Poorly developed soil derived from alluvial sediments.

`SX` Planossolo Háplico. Poorly drained greyish soil with significant increase of the clay content with depth.

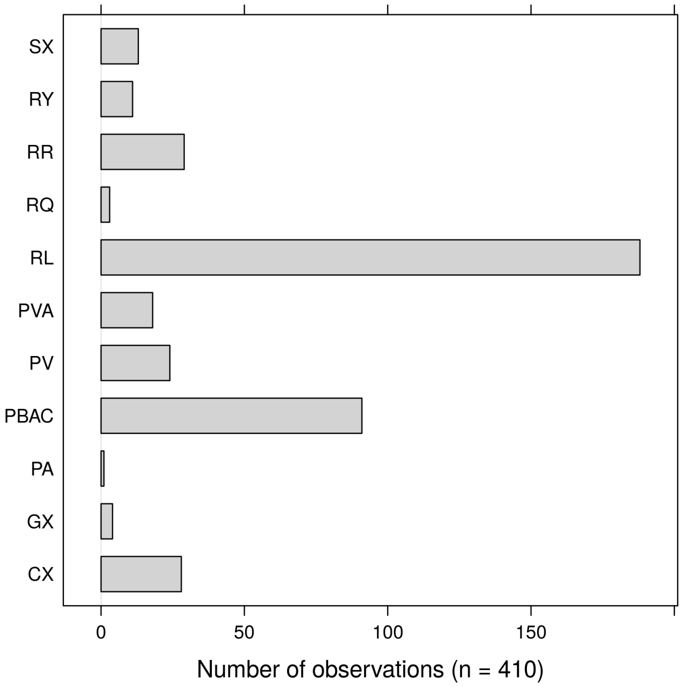


Figure 3.4: Distribution of soil taxa in the Santa Maria dataset. Most soil observations were classified as Neossolo Litólico and Argissolo Bruno-Acinzentado. The proportions approximately agree with the information conveyed by existing area-class soil maps (Figure 4.2).

3.6.4 Slope

The slope gradient (`SLOPE`, %) was measured using a clinometer, the observer and target being at a constant height above the ground (Figure 3.5). The distance between observer and target was between 30 m (dense forests) and 50 m (open fields).

3.6.5 Drainage

Soil drainage status (`DRAIN`) was inferred visually from soil features observed with an auger using the classification scheme proposed by Santos et al. (2013). Four drainage classes were identified (Figure 3.6):

`poorly` Water is removed from the soil so slowly that the profile remains wet for much of the time.

`somewhat poorly` Water is removed slowly from the soil, so that it remains wet for a significant period, but not during most of the year.

`moderately well` Water is removed from the soil somewhat slowly, so that the profile remains wet for small but significant period of time.

`well` Water is removed from the soil with ease but not rapidly.

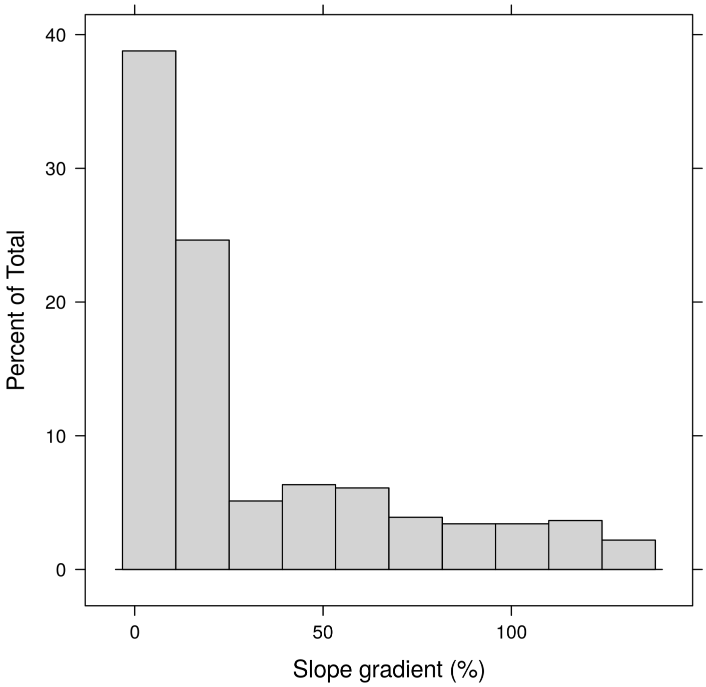


Figure 3.5: Distribution of slope gradient in the Santa Maria dataset. Most soil observations were made in areas with a slope gradient $< 25\,\%$.

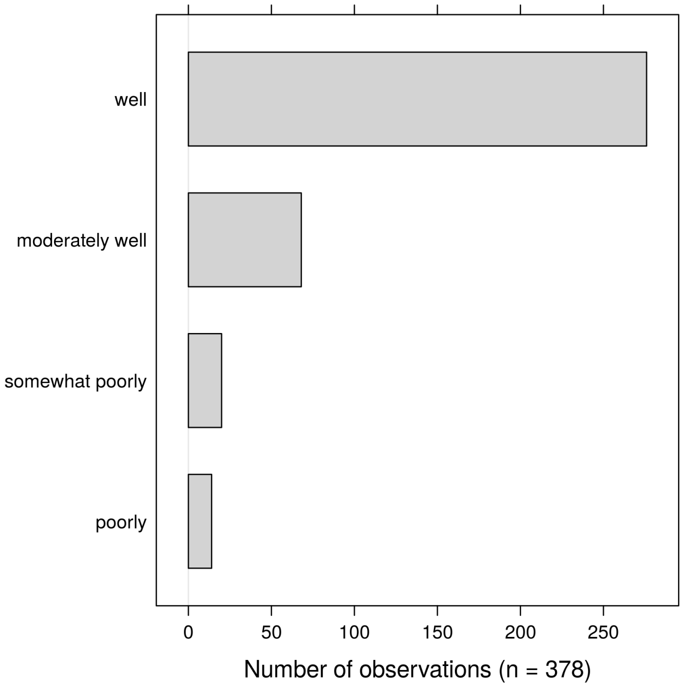


Figure 3.6: Distribution of drainage classes in the Santa Maria dataset. Most soil observations were made in areas with well drained soil.

3.6.6 Coarse Fragments and Rock Outcrops

Presence of coarse fragments (FRAG) – soil material of diameter $> 2\,\text{mm}$ – was described as a binary variable, that is, a value of 1 (one) was annotated when coarse fragments were present, and 0 (zero) otherwise. The same approach was adopted to describe the presence of rock outcrops (ROCK). The quantity of coarse fragments (GRAVEL, %) was estimated visually in some observation points.

It is worth noting that the approach employed to describe the presence of coarse fragments and rock outcrops is not in line with the standard soil description guidelines currently used in Brazil. The reason for recording only their presence/absence is that the actual content was not of primary interest at the time of sampling.

3.6.7 Canopy

Soil coverage with vegetation, an estimate of the density of stand or plant cover, (CANOPY) was inferred visually in the field using three classes:

low $< 25\,\%$

medium $25\text{–}75\,\%$

high $> 75\,\%$

3.6.8 Additional Information

Additional information was recorded at each observation location during the field campaigns. They refer to peculiarities of each observation location and were not recorded in a systematic way.

3.7 LABORATORY ANALYSIS

Several laboratory analysis were performed with the soil samples collected in the DNOS catchment. This sections present a summary description of how this analyses were done, specially for subsets I and II, which have not been documented before. For subset III, a thorough explanation of how laboratory analyses were done is given by Pedron (2005) and Miguel (2010).

Despite the different origins of the datasets, laboratory analyses protocols are very similar. As such, merging the results of laboratory analyses from subset III with those of subset I and II was easy, rarely requiring the use of conversion factors – this practice is reported when used. In all three datasets, soil samples were air dried, crushed and passed through a 2 mm-sieve prior to laboratory analyses. For datasets I and II, one or more laboratory replicates were used to enable calculating analytical errors.

The same coding standard used with field description variables is used here, i.e. the code used in the database is presented between parenthesis using fixed-width or monospace font.

3.7.1 Soil Organic Fraction

The soil organic carbon content (ORCA, $\text{g}\,\text{kg}^{-1}$) was determined using wet combustion (YEOMANS; BREMNER, 1988; MEBIUS, 1960; TEDESCO et al., 1995; CLAESSEN et al.,

1997) (Figure 3.7).

Sample aliquots of 0.050–0.500 g were placed in glass digestion tubes (80 ml). The amount of sample used varied according to the ORCA estimated by visual interpretation of soil colour. Every digestion tube received an aliquot of 10 ml of sulfochromic solution[15] [0.067 mol L^{-1} potassium bichromate solution ($K_2Cr_2O_7$) in the presence of concentrated sulphuric acid (H_2SO_4)] and a small reflux funnel to avoid loss of reagent during digestion. A digestion block with capacity for 40 samples was used: 36 tubes with soil sample plus 3 tubes with blank plus 1 tube with H_2SO_4 and a thermometer for temperature check. Digestion at 150 °C last 30 min. Three blanks were prepared and set aside at room temperature to estimate the loss of reagent due to heat in the digestion block.

After digestion the tubes were set aside at room temperature to cool down. Next, the solution was transferred to Erlenmeyer flasks (250 ml) with 60 ml of distilled water and 2 ml of concentrated orthophosphoric acid [H_3PO_4] and 3 drops of 1 % diphenylamine. The solution was titrated using 0.1 mol L^{-1} ammonium ferrous sulphate ($FeSO_4(NH_4)_2 \cdot 6\,H_2O$) until persistent green colour. The results were multiplied by 1.11 to correct the estimated soil organic carbon content to the standard analytical method (dry combustion).

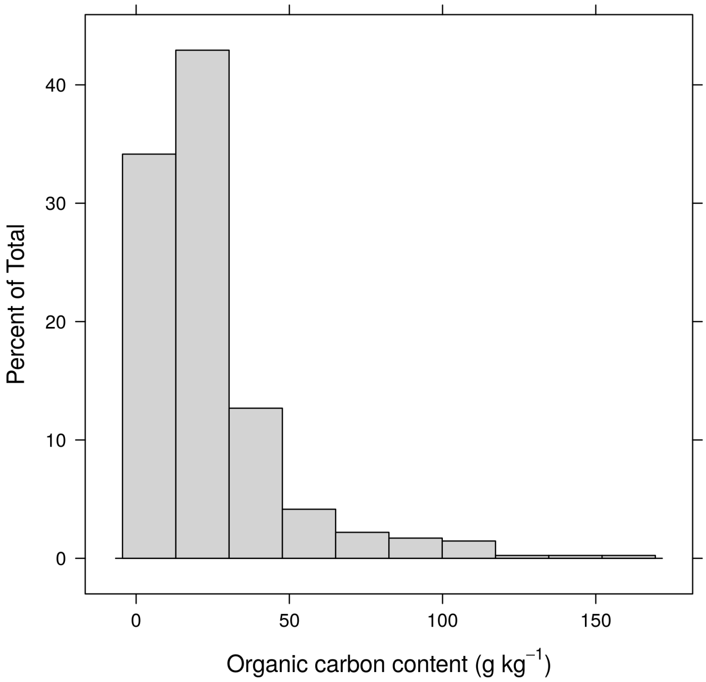


Figure 3.7: Distribution of organic carbon content (g kg^{-1}) in the Santa Maria dataset.

Observations compiled from Pedron (2005) had their soil organic matter content determined instead of the soil organic carbon content. Sample aliquots of 2.5 ml were placed in Erlenmeyer flasks (50 ml). Every Erlenmeyer flask received an aliquot of 15 ml of 0.067 mol L^{-1} sulfochromic solution ($Na_2Cr_2O_7 + H_2SO_4$). The flasks were heated in a water bath at 75–80 °C during 30 min and shaken for 5 min. A water aliquot of 15 ml was added to the flask and let overnight (15–18 h).

In the next day, an aliquot of 3.0 ml was sampled to a small plastic cup with 3.0 ml of distilled water. The absorbency of the supernatant was measured at 645 nm. The estimated soil organic matter content was transformed to soil organic carbon content assuming that 58 % of the

[15] See a detailed description of the sulfochromic solution, or chromic acid, at Wikipedia.

soil organic matter is composed of organic carbon. The resulting transformed ORCA value was assumed to be equivalent to soil organic carbon content measured using the standard analytical method (dry combustion). The results are expressed using a volume-basis and were converted to a mass-basis using a 1:1 relation because the mass of the sample aliquot used in the analyses is unknown.

3.7.2 Particle Size Analysis

Particle size analysis was performed using the pipette method (CLAY, < 0.002 mm, $\mathrm{g\,kg^{-1}}$), with the sand fraction (SAND, 0.053–2 mm, $\mathrm{g\,kg^{-1}}$) determined by wet sieving, and the silt fraction (SILT, 0.002–0.053 mm, $\mathrm{g\,kg^{-1}}$) calculated by difference (CLAESSEN et al., 1997). The analytical procedure includes adaptations[16] of the method of the Soil Conservation Service of the United States Department of Agriculture (SOIL CONSERVATION SERVICE, 1972) made by the Soil Physics Laboratory of the *Universidade Federal de Santa Maria* (SUZUKI et al., 2004b; SUZUKI et al., 2004a).

Before particle size analysis, all soil samples with organic matter content $> 5\,\%$ – estimated assuming that $58\,\%$ of the soil organic matter is composed of organic carbon – were submitted to oxidative treatment with hydrogen peroxide (H_2O_2). For this end, an aliquot of 20 g of soil was placed in a 1000 ml Beaker, to which 15 ml of distilled water and 5–15 ml of 30 % H_2O_2 were added – the larger the ORCA, the smaller the quantity of H_2O_2 added to avoid loss of material due to the strong frothing that usually takes place during the first addition of H_2O_2. The mixture was homogenized and covered with a watch glass and allowed to stand for 12 h, when another 5–15 ml of H_2O_2 was added. The addition of 5–15 ml of H_2O_2 continued at approximately regular intervals of 6 h until the mixture did not present any more effervescence (frothing). The Beaker was transferred to hot plate at a temperature of 50 °C to evaporate the excess water and remove any remaining H_2O_2. All H_2O_2-treated soil material was transferred quantitatively to a tared 100 ml glass container and put to dry overnight in an oven at 45 °C. When the soil material was completely dry, the glass container was transferred to a desiccator chamber where it was let to cool for 30 min. Then, the glass container was weighted with 0.001 g accuracy to estimate quantity of soil material remaining after the H_2O_2 treatment.

For particle size analysis, a sample aliquot of 20 g (or approximately 20 g when the material was submitted to oxidative treatment with H_2O_2) was placed in a 100 ml glass container (height: 10.5 cm; diameter: 2.75 cm; weight: 85 g). Two nylon spheres with a diameter of 1.71 cm and weighting 3.04 g (density: $1.11\,\mathrm{g\,cm^{-1}}$) were added to act as physical disaggregating agents. Then, an aliquot of 10 ml of $1\,\mathrm{mol\,L^{-1}}$ sodium hydroxide (NaOH) solution was added to act as chemical dispersing agent along with 40 ml of distilled water. The glass container was closed with a plastic cap, manually shaken for 10 s, and placed in a horizontal mechanical shaker with capacity for 85 samples. The suspension was left to stand overnight (10 h). In the next day the suspension was submitted to horizontal mechanical agitation during 4 h at 120 cycles per minute.

After horizontal agitation, the suspension was poured in a plastic graduated cylinder with capacity for 1000 ml using a glass funnel and a metal sieve to hold the two nylon spheres. The suspension in the graduated cylinder was completed to 1000 ml and homogenized using a hand stirrer (30 s). The suspension was allowed to stand on a vibration-free table to settle the clay

[16] As far as I know, a comprehensive description of this method has not been published so far, neither in Portuguese nor in English. You can visit the homepage of the Soil Physics Laboratory of the Universidade Federal de Santa Maria at <https://coral.ufsm.br/fisicadosolo/> to get more information about the method or contact their developers.

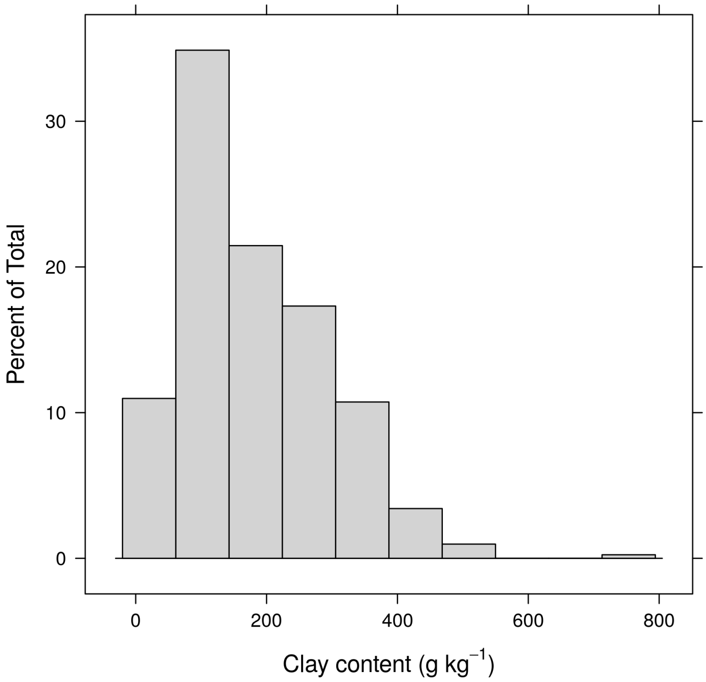


Figure 3.8: Distribution of clay content (g kg^{-1}) in the Santa Maria dataset.

fraction (< 0.002 mm) to a depth of 5 cm. The settling time needed was calculated using the Stokes' law with the temperature measured in a graduated cylinder filled with distilled water. After settling, an aliquot of 50 ml of the suspension was collected at a depth of 5 cm using a hand volumetric pipette. The aliquot was transferred to a tared moisture Beacker and put to dry overnight at 105 °C. When the material was completely dry, the Beacker was transferred to a desiccator chamber where it was let to cool for 30 min. Then, the Beacker was weighted with 0.001 g accuracy to estimate `CLAY` (Figure 3.8).

The suspension remaining in the plastic graduated cylinder after the 50 ml-aliquot had been collected was passed through a 0.053 mm sieve. The material retained in the sieve was washed thoroughly with water tap. The remaining material was quantitatively transferred to a tared moisture Beacker and put to dry overnight at 105 °C. When the material was completely dry, the Beacker was transferred to a desiccator chamber where it was let to cool for 30 min. Then, the Beacker was weighted with 0.001 g accuracy to estimate `SAND`.

3.7.3 Soil Density

The bulk soil density (`BUDE`, Mg m^{-3}) at the soil surface was determined using the core method with a metallic ring (height: 3 cm; diameter: 5 cm) (CLAESSEN et al., 1997) (Figure 3.9).

The bulk soil density was not determined in the locations where the soil was very shallow or stony.

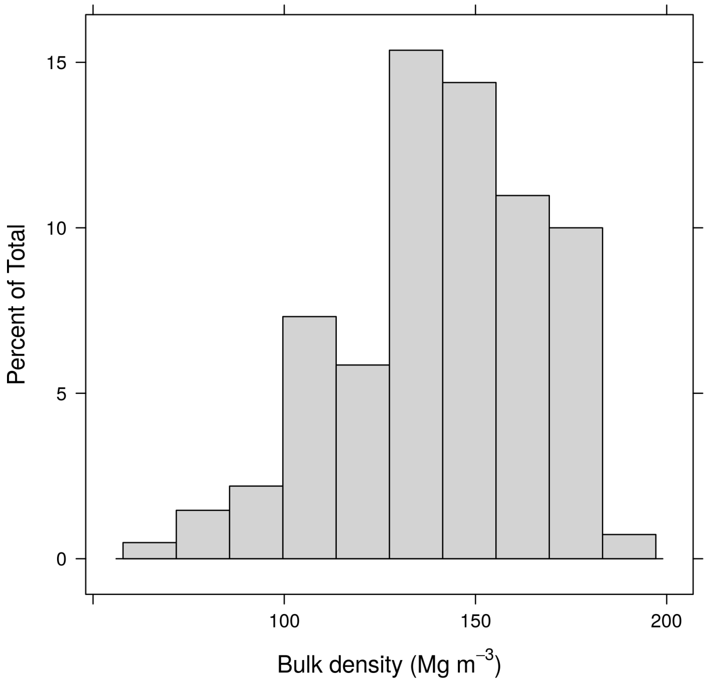


Figure 3.9: Distribution of bulk density (Mg m^{-3}) in the Santa Maria dataset.

3.7.4 Exchangeable Bases and Acidity

The exchangeable calcium (CALC, mmol kg^{-1}) and magnesium (MAGN, mmol kg^{-1}) were determined by atomic absorption spectroscopy after extraction with 1.0 mol L^{-1} KCl solution (CLAESSEN et al., 1997). The exchangeable sodium (SODI, mmol kg^{-1}) and potassium (POTA, mmol kg^{-1}) were extracted with a 0.05 mol L^{-1} HCl solution plus 0.025 mol L^{-1} H_2SO (Mehlich-1 solution). Both were quantified by means of flame atomic emission spectrometry (TEDESCO et al., 1995).

The exchangeable acidity (EXAC, mmol kg^{-1}) was extracted using the same 1.0 mol L^{-1} KCl solution used to extract the exchangeable calcium and magnesium. It was determined by titrimetry with 0.025 mol L^{-1} NaOH solution (CLAESSEN et al., 1997).

The effective cation exchange capacity (ECEC, mmol kg^{-1}) was defined as the sum of exchangeable bases and exchangeable acidity (Figure 3.10), i.e.

$$\mathtt{ECEC} = \mathtt{CALC} + \mathtt{MAGN} + \mathtt{POTA} + \mathtt{SODI} + \mathtt{EXAC}.$$

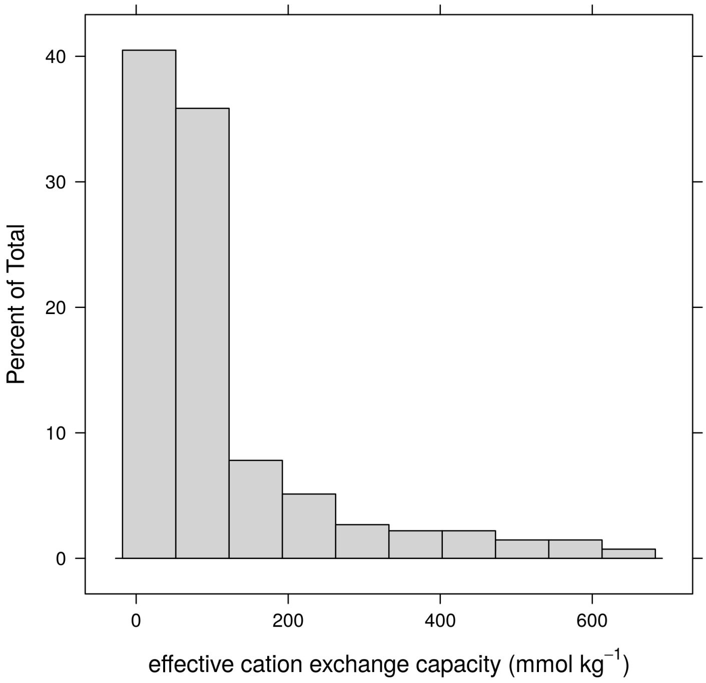


Figure 3.10: Distribution of effective cation exchange capacity (mmol kg^{-1}) in the Santa Maria dataset.

4 CHAPTER III

THE SANTA MARIA DATASET. PART II – COVARIATE DATA

* This chapter is based on the studies *Identifying and correcting oblique striping in the Topodata digital elevation model*, presented at the XXXIV Brazilian Congress of Soil Science (SAMUEL-ROSA et al., 2013), and *Evaluation of freely available ancillary data used for detailed soil mapping in Brazil*, presented at the EGU General Assembly 2014 (SAMUEL-ROSA et al., 2014a). Also collaborated in the preparation of this document: Pablo Miguel (UFPel), Jean Michel Moura Bueno (UFSM), Ricardo Simão Diniz Dalmolin (UFSM), Lúcia Helena Cunha dos Anjos (UFRRJ), Gustavo de Mattos Vasques (Embrapa Solos), and Gerard B. M. Heuvelink (ISRIC – World Soil Information).

4.1 RESUMO

O *conjunto de dados de Santa Maria* compreende uma lista de mais de $p = 100$ covariáveis espacialmente exaustivas produzidas nos 1980s, 1990s, and 2000s cobrindo a bacia do reservatório do Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento (DNOS-CORSAN), localizada no estado sulista brasileiro do Rio Grande do Sul. Essas covariáveis foram derivadas de cinco fontes disponíveis gratuitamente em dois níveis de detalhe espacial: mapas areais de classes de solo (escala cartográfica de 1:100 000 e 1:25 000), modelos digitais de elevação (dados orbitais de radar com resolução espacial de 90 m e dados de curvas de nível com escala cartográfica de 1:25 000), mapas geológicos (escala cartográfica de 1:50 000 e 1:25 000), mapas de uso da terra (escala cartográfica de 1:25 000 e 1:2000), e imagens de satélite (resolução espacial de 30 e 5 m). Esses dados de covariáveis são o resultado de projetos que visaram modelar várias características ambientais e foram realizados como parte de iniciativas locais (solo, geologia, uso da terra), regionais (terreno, uso da terra) e globais (terreno, uso da terra) de mapeamento. Os dados de covariáveis estão disponíveis gratuitamente em uma base de dados do GRASS GIS hospedada nos servidores do MEGAsync, enquanto o código-fonte utilizado na sua produção e processamento está disponível gratuitamente em um repositório hospedado no GitHub.

Palavras-chave: Mapas areais de classes de solo. Modelos digitais de elevação. Mapas geológicos. Mapas de uso da terra. Imagens de satélite.

4.2 ABSTRACT

The *Santa Maria dataset* comprises a list of more than $p = 100$ spatially exhaustive covariates produced in the 1980s, 1990s, and 2000s covering the catchment of the reservoir of the *Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento* (DNOS-CORSAN), located in the southern Brazilian state of Rio Grande do Sul. These covariates were derived from five sources that are freely available in two levels of spatial detail: area-class soil maps (cartographic scale of 1:100 000 and 1:25 000), digital elevation models (airborne radar data with 90 m spatial resolution and contour line data at a cartographic scale of 1:25 000), geological maps (cartographic scale of 1:50 000 and 1:25 000), land use maps (cartographic scale of 1:25 000 and 1:2000), and satellite images (30 and 5 m spatial resolution). These covariate data are the outcome of projects that aimed at modelling various environmental features and were carried out as part of local (soil, geology, land use), regional (terrain, land use), and global (terrain, land use) mapping initiatives. The covariate data is freely available in a GRASS GIS database hosted in MEGAsync servers, while the source code used in its production and processing is freely available in a repository hosted in GitHub.

Keywords: Area-class soil maps. Digital elevation models. Geologic maps. Land use maps. Satellite images.

4.3 INTRODUCTION

The *Santa Maria dataset* is a data set comprising spatially exhaustive covariate data produced in the 1980s, 1990s, and 2000s covering the catchment of the reservoir of the *Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento* (DNOS-CORSAN), henceforth called *DNOS catchment*, located in the southern border of the plateau of the Paraná Sedimentary Basin, in the city of Santa Maria, state of Rio Grande do Sul, Brazil. Most of the covariate data cover only part of the DNOS catchment, mainly the northern sector, which has an area of ±2000 ha, corresponding to ±60 % of the entire catchment. These covariate data are the outcome of projects that aimed at modelling various environmental features and were carried out as part of local (soil, geology, land use), regional (terrain, land use), and global (terrain, land use) mapping initiatives.

This chapter presents a thorough description of the covariate data contained in the Santa Maria dataset. These are area-class soil maps, digital elevation models, geological maps, land use maps, and satellite images. A thorough description of the procedures for their production, as well as the processing methods employed is given. The original sources of the covariate data are freely available at the producers databases or in public libraries.

The chapter is divided in seven sections. Section 4.4 presents general information about the covariates, as well as the structure of the database where they have been stored and managed. Next, Section 4.5 describes how ground data was collected to help processing and validating the existing covariate data. Area-class soil maps included in the Santa Maria dataset are described in Section 4.6, while Section 4.7 deals with the digital elevation models. Geologic maps and land use maps are described in Section 4.8 and Section 4.9, respectively. Section 4.10 closes the chapter with a description of the satellite images included in the Santa Maria dataset.

4.4 DATABASE STRUCTURE

Preparing the covariate data to enter the Santa Maria dataset required creating two databases. The first contains only vector (point) data, specifically point ground control data, as well as the code used in its processing. This database is freely available in the web-based Git repository <https://github.com/samuel-rosa/dnos-sm-rs-general/>. This is the same repository containing the soil data (Section 3.4). The repository has the following folder structure:

```
dnos-sm-rs-general
|- code/              # source code folder
|   - R/              # R source code folder
|     - general.R     # R source code file
|
|- data/              # data folder
|   - gcpData.csv     # ground control data file
|   - gcpMetadata.csv # ground control metadata file
|- README.md          # description of the repository
```

Ground control data files are available as comma-separated values (CSV) files. The identification of all observation locations, their geographic coordinates, and elevation data are contained in the file `gcpData.csv`. File `gcpMetadata.csv` contain the metadata. The coordinate reference system (CRS) is `SIRGAS 2000 / UTM zone 22S`, coded `EPSG:31982` by the European Petroleum Survey Group (EPSG).

The second database contains both vector (point, line and polygon) and raster (image) data. This database is freely available in MEGAsync servers, and is structured as a GRASS GIS database[17]. The database has the following folder structure:

```
dbGRASS                # GRASS GIS database
|- dnos-sm-rs          # Santa Maria dataset
|   - PERMANENT/          # general definitions folder
|   - predicitons/        # data folder
```

All covariate data in the GRASS GIS database are harmonized to a reference grid of 5 m grid size. The CRS is `WGS1984 / UTM zone 22S`, coded `EPSG:32722`.

4.5 GROUND CONTROL POINTS

All covariates were validated prior to their use. Horizontal (positional) validation was performed using a set of $n = 14$ validation points, here called ground control points (GCP), spread throughout and beyond the limits of the DNOS catchment (Figure 4.1). The location of the GCPs was defined based on the existence of easily identifiable geographical markers across the covariates, including road intersection, fence corners, and property entrances.

Positional validation was performed comparing the x- and y-coordinates of GCPs (observed value) with the coordinates of the respective geographical markers visually identified on the covariates (predicted value). The differences in the observed and predicted x- and y-coordinates were used to calculate the mean error (ME, m), mean absolute error (MAE, m), and mean squared error (MSE, m) to evaluate the if there were differences in the accuracy and precision between coordinates. The error vector (or module, the euclidean distance between two points) and its azimuth (the orientation of the error vector) were computed as well for every point. The mean of the error vector and its azimuth give the size and orientation of the systematic error present in the covariates, while the square root of the mean squared error vector (RMSE) is a measure of the uncertainty about the true position of the covariate in the geographic space.

The field location of the GCPs is as follows:

GCP 01 In Santa Maria, on the right side of the concrete dam of the DNOS-CORSAN reservoir, 6 m before reaching the bridge over the spillway, distant 3 m from the centre of the road that descends from the BR-158 federal highway.

GCP 02 In Itaara, at the entrance of the Estrada do Perau, Rua Gralha Azul, in the centre of the roundabout, between the outdoor and the tree (*Cedrella fissilis*), close to the BR-158 federal highway.

GCP 03 In Itaara, near the speed monitoring radar of the BR-158 federal highway, distant 520 m from the UFO Museum, opposite the Fruteira da Esquina, opposite to the cell phone tower and to the entrance of the gravel pit DallaPasqua, located 4 km away from the site.

GCP 04 In Santa Maria, at the entrance of the street that leads to the cemetery of the Campestre do Menino Deus neighbourhood, on the right side of Perau Road going towards the BR-158 federal highway, aligned with the façade of the residences, 2.2 m away from the front

[17] Detailed information about the structure of a GRASS GIS database can be found in GRASS GIS help pages.

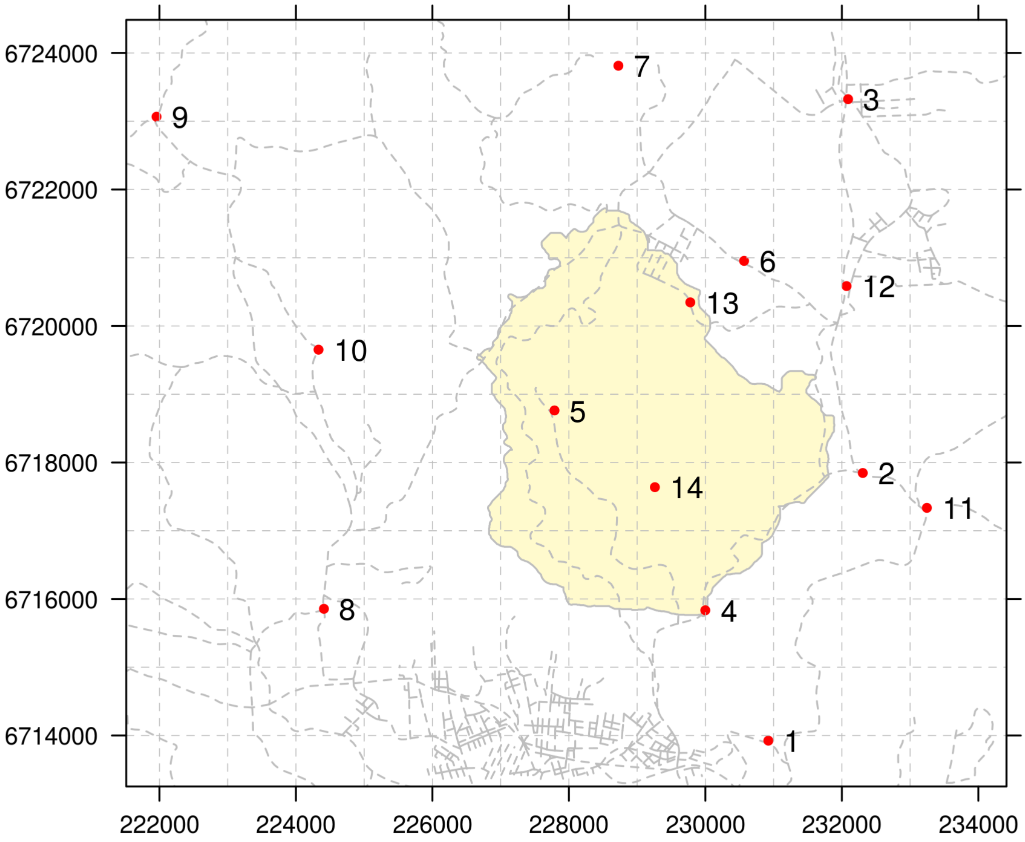


Figure 4.1: Spatial distribution of the ground control points ($n = 14$, red dots) used for the horizontal positional validation of covariates included in the Santa Maria dataset. The location of the ground control points is highly determined by the local road network (dashed grey line) surrounding the northern sector of the catchment of the reservoir of the *Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento* (lemon chiffon-filled polygon).

wall, 6.5 m away from the curb of the Perau Road, 50 m away from the bridge over the Vacacaí-Mirim River.

GCP 05 In Santa Maria, at the entrance of the Rancho do Amaral, next to the gate, on the right size, off the road, 1 m away of a palm tree and 1 m away from the stone wall.

GCP 06 In Itaara, Etelvina Avenue, on the roadside, 2.5 m away from the road centre, aligned with the fence that separates the native forest from the *Citrus* sp. orchard located across the road.

GCP 07 In Itaara, Estação Pinhal Locality, on the left side of the beginning of the road that gives access to the DallaPasqua gravel pit, under the electricity transmission network.

GCP 08 In Santa Maria, Santo Antão District, at the beginning of the road that gives access to the Caturrita Waste Treatment Centre, opposite the Municipal Elementary School Intendente Manoel Ribas.

GCP 09 In São Martinho da Serra, Água Negra Locality, at the bifurcation of the road coming from Santa Maria and giving access to the Campinas Locality, close to the bus stop, at the roundabout in the middle of bifurcation, 40 m away from the Piquete Laçador Jorge R. da Silva, in front of Ronaldo's Market.

GCP 10 In Santa Maria, on the road towards São Martinho da Serra, after Santo Antão Chapel, on the outside of a curve, near two small trees, aligned with the fence that marks the boundary between two properties occupied with native grass. In front of the property with two houses, one with two floors and four small lakes in the backyard.

GCP 11 In Itaara, at the entrance to the Rincão dos Minello Locality, at the beginning of the road that gives access to Brita Pinhal, next to the BR-158 federal highway, 5 m away and in front of the road cut exposing the sandstone rocks of the Botucatu Formation, 15 m away from the post of the electricity transmission line of the AES company.

GCP 12 In Itaara, at the federal highway BR-158, in front of the SOCEPE's Lake, in the city entrance, near the Ricardo's Bar and Warehouse, shifted 1 m into the side walk in relation to the alignment of the posts of the electrical network.

GCP 13 In Itaara, at Vila Etelvina, with a vineyard upstream, aligned with the fence that divides two properties located on the opposite side of the road, the one on the left covered with native/exotic woods and the other on the right occupied with field of annual crops.

GCP 14 In Itaara, in the road that goes towards the property of Mr. Antoninho Luccas, soon after the steep climb, where there is only pavement on the tracks, at the end of the forest and beginning of native grass, aligned with the fence on both sides of road, 1 m away from the left corner fence post.

Attribute validation of soil, geologic, and land use maps, and digital elevation models was done using a set of $n = 60$ validation points located along $m = 12$ linear transects (Section 3.5.2). The procedures for obtaining soil, geologic, land use, and elevation data at these validation points is described in Section 3.6. Such a validation exercise was carried out because these maps originally had no accompanying validation information.

4.6 AREA-CLASS SOIL MAPS

Two area-class soil maps are included in the Santa Maria dataset. The first of them (`SOIL_100`, Figure 4.2a) was published at a cartographic scale of 1:100 000 (AZOLIN; MUTTI, 1988). Existing area-class soil maps and technical reports (BRASIL, 1973; AZOLIN, 1977; MACIEL FILHO et al., 1987a; MACIEL FILHO et al., 1987b; ABRÃO et al., 1988), and sparse field observations were used to elaborate the preliminary legend of the soil map. Aerial photographs (cartographic scale of 1:60 000) were used to produce the first draft of the soil map. Field checks of soil polygons (i.e. mapping units) was done along the road network (i.e. by convenience sampling). These observations were used to estimate the composition (occurrence and spatial distribution of soil taxa) of soil mapping units. They were also used to review the first draft of the soil map. The final version of `SOIL_100` was prepared using topographic maps originally published at a cartographic scale of 1:50 000 and resampled to a cartographic scale of 1:100 000. Soil classification followed the criteria adopted by the Brazilian soil science community at that time (BRASIL, 1973; CAMARGO et al., 1982; CARVALHO, 1982; LEMOS; SANTOS, 1982; OLMOS; CAMARGO, 1982). Identification of soil taxa was performed based on morphological features, analytical data compiled from existing technical reports, and analysis of soil samples collected from soil profiles observed along the road network. Description of each soil mapping unit includes the estimated area (ha) and the approximate taxonomic composition (%).

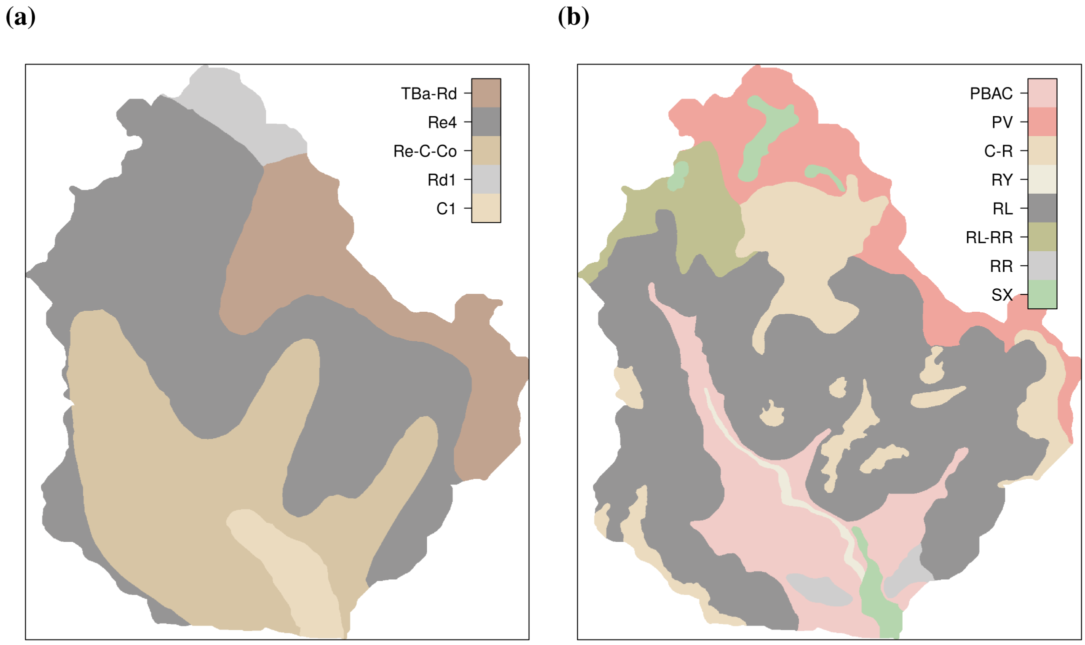


Figure 4.2: Area-class soil maps (a) `SOIL_100` and (b) `SOIL_25` used to derive indicator covariates included in the Santa Maria dataset. Legend abbreviations and derived indicator variables are described in the text.

The second area-class soil map (`SOIL_25`, Figure 4.2b) included in the Santa Maria dataset (MIGUEL, 2010) was prepared at a cartographic scale of 1:25 000. Satellite images of the year of 2009, with 2.4 m-spatial resolution, produced by the Digital Globe® QuickBird satellite[18], freely available for visualization in Google Earth®, were used to produce the first draft of the soil map. Existing area-class soil maps and technical reports (PEDRON, 2005; POELKING, 2007; STÜRMER, 2008) were used to help defining the preliminary soil map legend. Field observations (soil pits and boreholes) were made in more than 350 locations using a purposive sampling approach (see Chapter II and Chapter VI). These observations helped to identify six modal (representative) soil profiles. Soil sampling and description of modal soil profiles, and laboratory analyses of soil samples, followed the standard protocols adopted in Brazil (CLAESSEN et al., 1997; SANTOS et al., 2005). Soil classification was performed following the criteria of the Brazilian System of Soil Classification (SiBCS) (SANTOS et al., 2006). The final version of the map was prepared using the same satellite images freely available for visualization in Google Earth® and manually-digitalized topographic sheets published at a cartographic scale of 1:25 000 (DSG, 1992b; DSG, 1992a). Description of soil mapping units includes only the most common soil taxon, followed by morphological and laboratory data of modal soil profiles.

The area-class soil maps went through different preprocessing routines. The original `SOIL_100` is available only in analogical format, what required its digitalization. Georeferencing was carried out using the GDAL Georeferencer plug-in in QGIS (GDAL DEVELOPMENT TEAM, 2013; QUANTUM GIS DEVELOPMENT TEAM, 2013). Intersections between all meridians and parallels ($n = 9$) were used as control points to adjust a second order polynomial model. Resampling was performed using the cubic resampling method. Soil polygons

[18] Detailed information about the Digital Globe® QuickBird satellite can be found in Wikipedia.

and their attributes were also manually digitalized in QGIS. Because of the coarseness of the cartographic map scale (cartographic scale of 1:100 000), most geographical markers used to locate validation GCPs could not be identified and positional validation was performed using only $n = 4$ GCPs. Estimated error statistics suggest that there are large positional errors in all directions, with an RMSE of 114 m and a mean azimuth of 128° (Table 4.1).

Table 4.1: Error statistics of the horizontal positional validation of `SOIL_100` using $n = 4$ ground control points.

Statistics	x-coord	y-coord	Error vector	Azimuth
Mean, m	30	-36	105	128°
Absolute mean, m	58	64	-	-
Squared mean, m	7241	5712	12953	-

The original `SOIL_25` is available in digital format in the personal database of the author (MIGUEL, 2010). A topology check performed using the Topology Checker plug-in in QGIS identified that there were many gaps and overlaps between polygons. This required a topological edition prior to the use of `SOIL_25`. There also was a mismatch between the boundary of `SOIL_25` and the actual boundary of the study area as estimated using `ELEV_10` (Section 4.7). This occurred because the database used to produce `SOIL_25` included 2.4 m-spatial resolution Google Earth® imagery and topographic maps, which are data sources that differ considerably in their positional accuracy (Section 4.7 and Section 4.9). To avoid data losses, all boundary gaps were manually filled using the closest mapping unit. Boundaries of soil polygons were defined based on land use (`LU2009`, Section 4.9) and topographic data (contour lines, Section 4.7) as it was done for the original map (MIGUEL, 2010). New delineations were checked and approved without modifications by the author of the original map. Because `SOIL_25` includes very few geographical markers, its positional validation was not possible with the available GCPs. However, the RMSE is expected to vary between 8 and 114 m across the DNOS catchment as a result of the different errors present in the data sources used in its production.

Both `SOIL_100` and `SOIL_25` were cropped to the bounding box of the DNOS catchment, and resampled to match the reference grid using the nearest neighbour resampling method to maintain data integrity. Each category was named with the code of the respective mapping unit in the original map. Prior to validation in the attribute space, class codes of `SOIL_100` were changed to match soil taxa codes of the current Brazilian System of Soil Classification using a standard correlation table (SANTOS et al., 2006). The overall purities of both soil maps are not considerably different. A reason for this could be that validation was performed considering only the second level of the SiBCS – it is likely that `SOIL_25` would outperform `SOIL_100` if validation data included soil taxa up to the fourth level of the SiBCS. The low overall purity of `SOIL_100` and `SOIL_25` (31.67 and 30.00 %, respectively) is likely due to several sources of error. First, the small number of modal soil profiles used to produce both maps that might have resulted in an optimistic view of the homogeneity of each mapping unit. Second, soil classes of `SOIL_100` were translated to the more recent SiBCS using only a standard correlation table (SANTOS et al., 2006) and expert knowledge because the survey report does not include analytical soil data. Last, soil classes at the $n = 60$ validation sites were inferred in the field using only morphological soil properties and the general concepts of the SiBCS.

Five ($p = 5$) covariates were derived from `SOIL_100` as described below (including the soil classes according to the original and updated classification (AZOLIN; MUTTI, 1988; SANTOS et al., 2013), and the international classification (IUSS WORKING GROUP WRB, 2007)):

SOIL_100b Shallow soil (*Re4*) with low to high base saturation covering mountainous terrain (Solo Litólico Eutrófico/Distrófico relevo montanhoso; Neossolo Litólico Distrófico/Eutrófico; Distric/Eutric Leptosol).

SOIL_100c Association (*Re-C-Co*) of shallow soil with high base saturation located in steep terrain (Solo Litólico Eutrófico relevo forte ondulado; Neossolo Litólico Eutrófico; Eutric Leptosol), low weathered soil (Cambissolo Eutrófico; Cambissolo Háplico Eutrófico; Eutric Cambisol), and colluvial deposits.

SOIL_100d Association (*TBa-Rd*) of deep, well-structured, low base saturation soil (Terra Bruna Estruturada álica; Nitossolo; Nitisol), and shallow soil (Solo Litólico; Neossolo Litólico; Leptosol).

SOIL_100e Shallow soil (*Rd1* and *Re4*) with low to high base saturation (Solo Litólico Distrófico/Eutrófico; Neossolo Litólico Distrófico/Eutrófico; Distric/Eutric Leptosol) located in undulating to mountainous terrain.

SOIL_100f This covariate includes the best soil mapping units for row crop agriculture among those identified in the soil survey, that is *TBa-Rd*, described above, and *C1*, which is composed of low weathered soil developed in lower landscape positions, close to drainage channels (Cambissolo Eutrófico; Cambissolo Háplico Eutrófico; Eutric Cambisol).

Covariates derived from SOIL_25 are presented below. Mapping unit *RY*, composed mainly of soil developed from fluvial deposits (Neossolo Flúvico; Fluvisol) does not appear due to the small area that it occupies.

SOIL_25a Moderately deep soil (*PBAC*) derived from sedimentary rocks, with abrupt textural change and low base saturation (Argissolo Bruno-Acinzentado; Alisol).

SOIL_25b Deep soil (*PV*) derived from igneous rocks, with moderate textural gradient, and low base saturation (Argissolo Vermelho; Acrisol).

SOIL_25c Low weathered soil (Cambissolo Háplico; Cambisol) and shallow soil with low to high base saturation (Neossolo Litólico/Regolítico Eutrófico/Distrófico; Eutric/Distric Leptosol/Regosol) (*C-R*).

SOIL_25d Shallow soil (*RL*) with low to high base saturation (Neossolo Litólico Eutrófico/Distrófico; Eutric/Distric Leptosol).

SOIL_25h This covariate includes the mapping units with the best soils for row crop agriculture among those identified in the soil survey (*PBAC*, *PV*, and *SX*). *PBAC* and *PV* are as described above. *SX* is composed of moderately deep soil derived from sedimentary rocks, with abrupt textural change, low base saturation, and which are saturated with water for long periods of the year (Planossolo Háplico; Planosol).

SOIL_25i This covariate includes all three mapping units (*RL*, *RL-RR*, and *RR*) composed mainly of shallow soils (Neossolo Litólico and Neossolo Regolítico; Leptosol and Regosol).

SOIL_25j This covariate includes all four mapping units (*PV*, *RL*, *RL-RR*, and *C-R*) composed mainly of soil derived from igneous rocks.

4.7 DIGITAL ELEVATION MODELS

Two digital elevation models (DEMs) are included in the Santa Maria dataset as sources of terrain covariates. The first DEM (`ELEV_10`, Figure 4.3a) is the result of the interpolation of the contour lines of the most recent topographic sheets produced by the Brazilian Army (cartographic scale of 1:25 000) that cover the DNOS catchment (DSG, 1980; DSG, 1992a; DSG, 1992b). Topographic sheets were digitalized and georeferenced using the GDAL Georeferencer plugin in QGIS. Intersections between all meridians and parallels (about $n = 160$ per topographic sheet) were used as control points to adjust a third order polynomial model. Resampling was performed using the cubic resampling method. All contour lines, peaks, lakes and rivers, and their respective attributes within a distance of 1000 m from the boundary of the DNOS catchment were manually digitized and stored in vector format. After digitalization, the original coordinate reference system (EPSG:31982 – SIRGAS2000 / UTM zone 22S) of all vector files was transformed to WGS1984 / UTM zone 22S (EPSG:32722) using the `R`-package `rgdal` (BIVAND et al., 2013).

The horizontal positional validation of topographic maps was performed using the $n = 14$ GCPs. According to Brazilian legislation, high-quality map standards require that at least 90 % of the GCPs have horizontal positional errors smaller than 13 m, and an overall horizontal error (i.e. RMSE) smaller than 8 m at the cartographic scale of 1:25 000 (BRASIL, 1984). Estimated validation statistics show that the overall observed horizontal error (RMSE = 65 m) is larger than those established by current regulations (Table 4.2). The mean error vector is larger than 60 m with an azimuth of 63°. Both x- and y-coordinates are positively biased, but the largest error occurs in the x-coordinate (50 m). Similar ME and MAE values suggests that there is a strong systematic positional error. An affine transformation was employed using the `R`-package `vec2dtransf` (CARRILLO, 2012) to correct this systematic error. Model parameters were adjusted using the same set of GCPs used for the validation in the geographic space.

Table 4.2: Error statistics of the horizontal validation of topographic maps (cartographic scale of 1:25 000) using $n = 14$ ground control points.

Statistics	x-coord	y-coord	Error vector	Azimuth
Mean, m	50	27	63	63°
Absolute mean, m	50	32	-	-
Squared mean, m^2	3088	1180	4268	-

Interpolation of the raster surface with 5 m grid size was performed using the function `Topo to Raster` in ArcGIS® software by ESRI, which includes an interpolation method based on ANUDEM (HUTCHINSON, 1989). Vector files of contour lines (`multiline`), drainage network (`multiline`), lakes (`polygons`) and peaks (`points`) were used to generate a hydrologically sound DEM, that is, a DEM without spurious depressions and giving a precise representation of the hydrological data. Next, the interpolated DEM was imported into GRASS GIS, where a neighbourhood average filter was used to remove stair-like artefacts. A window of 7×7 pixels was used because it removed a significant amount of the artefacts and did not affect the derived boundary of the study area (see more bellow).

The vertical datum of the DEM was transformed from the local datum to a global datum. The geoidal models MAPGEO2010 (IBGE, 2010) and EGM1996 (LEMOINE et al., 1998) were used to calculate the geoidal undulation for the local and global datums, respectively. MAPGEO2010 is optimized to estimate geoidal undulations in the Brazilian territory, while EGM1996 is a gravitational model of the Earth and is used as the vertical datum for Shuttle Radar Topography Mission (SRTM) products. The following equation was used:

$$h = H + N, \tag{4.1}$$

where h is the ellipsoidal height (height above the reference ellipsoid that approximates the surface of the planet), H is the orthometric height (height above the imaginary surface called geoid and commonly referred to as mean sea level), and N is the geoidal undulation. Ellipsoidal heights estimated by MAPGEO2010 are referenced to the world ellipsoid of 1980, while EGM1996 estimates ellipsoidal heights referenced to the world ellipsoid of 1984. Because the difference between both ellipsoids is of the order of millimetres, it can be assumed that both models estimate the same ellipsoidal height. Therefore, if $h_{\text{EGM1996}} = h_{\text{MAPGEO2010}}$, then orthometric heights referenced to the local vertical datum can be transformed to the global vertical datum using the following equation:

$$H_{\text{EGM1996}} = H_{\text{MAPGEO2010}} + N_{\text{MAPGEO2010}} - N_{\text{EGM1996}}. \tag{4.2}$$

The difference in the geoidal undulation estimated by both models is of about 1 m in the entire DNOS catchment. Thus, transforming the vertical datum was performed adding 1 m to the raster surface interpolated from contour lines, yielding the first DEM included in the Santa Maria dataset (`ELEV_10`).

The second DEM (`ELEV_90`, Figure 4.3b) is the well known SRTM DEM ($3'' \approx$ 90 m spatial resolution) produced by the National Aeronautics and Space Administration's Jet Propulsion Laboratory in collaboration with the National Geospatial/Intelligence Agency (RODRÍGUEZ et al., 2006). The SRTM DEM version used here is the *sink-filled SRTM version* 4, prepared by the Consultative Group for International Agricultural Research (CGIAR) using the same hydrologically correct interpolation method that was used before to produce `ELEV_10` (REUTER et al., 2007; JARVIS et al., 2008). However, the only data source used was the SRTM DEM version 3 converted to point data.

The SRTM DEM was processed to match the reference grid using cubic resampling (GDAL module `gdalwarp` and GRASS module `r.resamp.interp`). This resampling method was used because it is efficient in minimizing the double-oblique stripping present in SRTM products (SAMUEL-ROSA et al., 2013). Sinks produced during datum transformation were filled using the GRASS module `r.fill.dir`. Vertical datum transformation was not necessary because elevation values of the SRTM DEM already are referenced to the global geoid model EGM1996 (orthometric heights).

A third DEM (not shown) was included in the Santa Maria dataset with the sole purpose of supporting the orthorectification – removal of the effects of the perspective of the sensor on the relative position of objects in the image – and topographic correction – removal of the effects of the topography on the reflectance values, i.e. on the illumination of the geomorphic surfaces – of satellite images (Section 4.10) (MATHER, 2004; SCHOWENGERDT, 2007). The third DEM (TOPODATA) was produced by the Brazilian National Institute for Space Research (INPE) by refining the SRTM DEM version 1 ("unfinished") to $1'' \approx$ 30 m spatial resolution

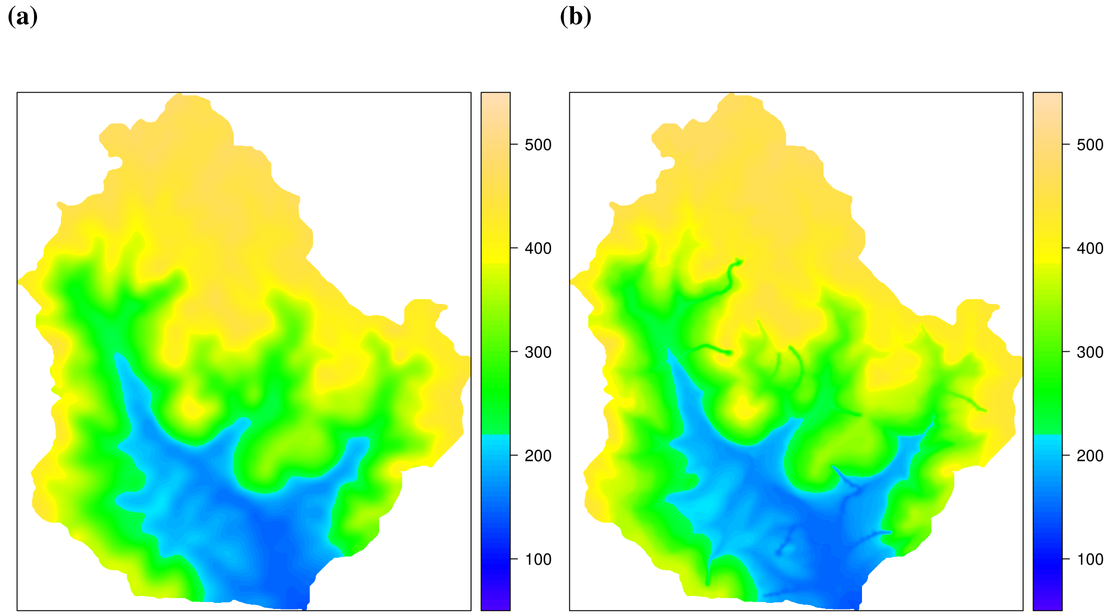


Figure 4.3: Digital elevation models (a) `ELEV_90` and (b) `ELEV_10` used to derive terrain covariates included in the Santa Maria dataset.

using ordinary kriging with a Gaussian spatial autocorrelation model (VALERIANO; ROSSETTI, 2012). Eight tiles were mosaicked (GDAL module `gdal_translate`) and the CRS transformed from WGS1984 (EPSG:4326) to WGS1984 / UTM zone 22S (EPSG:32722) using cubic resampling (GDAL module `gdalwarp`), and the sinks filled using GRASS module `r.fill.dir`. Before the atmospheric correction of satellite images, orthometric heights were converted to ellipsoidal heights using Equation 4.1, the geoidal undulation calculated with the gravitational model EGM1996. This conversion was done because orbital satellites use the WGS1984 ellipsoid as vertical datum. For the orthorectification of satellite images, the DEM was then processed using GRASS module `r.resamp.interp` with the bicubic resampling method to match the reference grid.

The three DEMs present similar vertical accuracy (Table 4.3). In the case of `ELEV_10`, which was derived from contour lines published at a cartographic scale of 1:25 000, the vertical positional accuracy does not meet the high-quality standards of current Brazilian legislation, which states that 90 % of the GCPs should have vertical errors smaller than 5 m, which is half of the distance between contour lines (BRASIL, 1984). The corresponding RMSE should be less than 3 m.

Table 4.3: Error statistics of the vertical validation of `ELEV_90`, TOPODATA, and `ELEV_10` using $n = 60$ validation points located along $m = 12$ linear transects.

Statistics	`ELEV_90`	TOPODATA	`ELEV_10`
Mean, m	-15	-17	-16
Absolute mean, m	15	17	16
Squared mean, m	350	361	374

Despite all DEMs present a similar vertical accuracy, `ELEV_10` is considered the highest quality DEM in the Santa Maria dataset. Because it was produced using information about the drainage network and location of lakes and natural depressions, it is likely to provide a better hydrological representation of the DNOS catchment. As such, `ELEV_10` was used to estimate the geographical limits (boundary) of the DNOS catchment, for which GRASS GIS modules `r.watershed` and `r.water.outlet` were employed. Because the overall deviation between the affine-corrected coordinates of topographic maps and target coordinates of GCPs is RMSE $= 29.55\,\text{m}$ – there still is uncertainty about the correct position of topographic maps – a $30\,\text{m}$ buffer was added to the estimated geographical limits of the DNOS catchment. The water outlet point used to estimate the boundary is located on the bridge that crosses the main drainage channel ($-29.658\,68°$, $-53.789\,69°$).

Eight terrain attributes were derived from each of `ELEV_90` and `ELEV_10` to produce the covariate data included in the Santa Maria dataset, the first of them being elevation (`ELEV`). The others are slope, aspect, northernness, flow accumulation, topographic wetness index, stream power index, and topographic position index.

Slope (`SLP`) and aspect (`ASP`) were calculated using GRASS module `r.param.scale`. This module calculates terrain attributes by fitting a bivariate quadratic polynomial using least squares (WOOD, 1996). It allows using different window sizes to fit the bivariate quadratic polynomial, thus including the effect of scale in the calculation of terrain attributes. Seven window sizes were used (3, 7, 15, 31, 63, 127, and 255) and the results for calculated slope can be seen in Figure 4.4. Larger window sizes result in a smoother version of the terrain attribute, while smaller windows sizes result in raster maps with more (small-scale) details. Several flat surfaces (slope of $0°$) were produced in the slope raster maps calculated using `ELEV_90` as a result of resampling the original DEM from 90 to $5\,\text{m}$. A value of $0.1°$ was added to the rasters to remove these flat surfaces.

Aspect values had to be corrected before use because GRASS module `r.param.scale` stores aspect values in the range 0–180° from West to North to East, and 0–−180° from West to South to East, when the standard procedure is to work with aspect values ranging from 0–360° clockwise. This correction was done using

$$\mathtt{ASP}_0 = \begin{cases} \mathtt{ASP}_{GRASS} + 360° & \text{if } \mathtt{ASP}_{GRASS} < 0°, \\ \mathtt{ASP}_{GRASS} & \text{else}, \end{cases} \tag{4.3}$$

and

$$\mathtt{ASP} = \begin{cases} \mathtt{ASP}_0 + 270° & \text{if } \mathtt{ASP}_0 < 90°, \\ \mathtt{ASP}_0 - 90° & \text{else}. \end{cases} \tag{4.4}$$

A second correction of aspect values involved their linearisation. This is necessary because aspect is a circular variable, that is, the beginning (0°) and end (360°) of the measurement scale have the same physical meaning. Aspect values were transformed to northernness (`NOR`), a measure of the level of exposition of a given surface to the North, a linear variable, using the equation

$$\mathtt{NOR} = abs(180° - \mathtt{ASP}). \tag{4.5}$$

Flow accumulation (ACC), also known as catchment area or contributing area, was calculated using GRASS module `r.watershed`, the resulting raster surface being multiplied by the square of the cell size (i.e. by the cell area $25\,\text{m}$) to convert to areal units. This raster surface

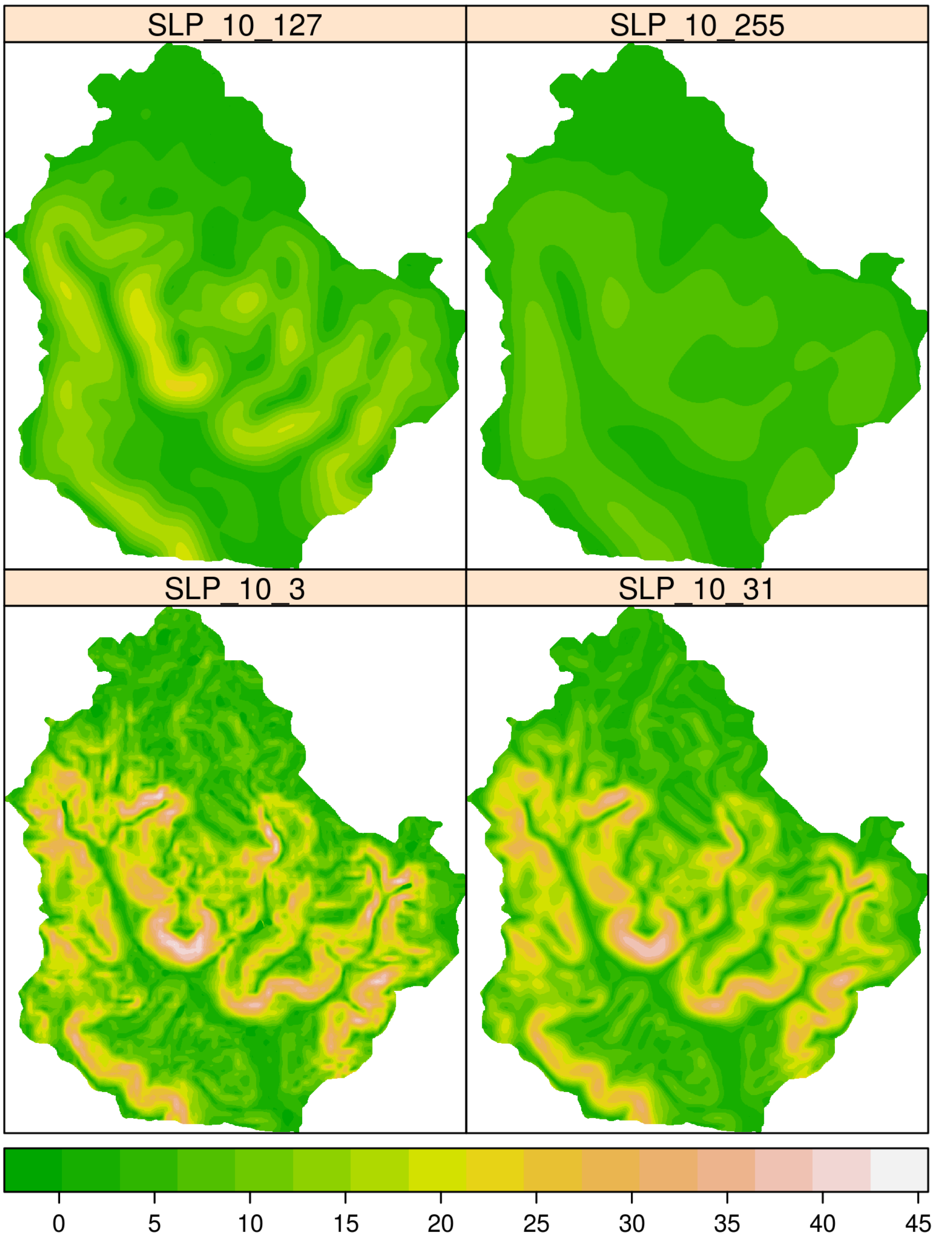


Figure 4.4: Slope `SLP` raster surfaces derived from `ELEV_10` using windows of sizes (from left to right) 3×3, 31×31, 127×127, and 255×255.

was used to calculate the topographic wetness index (`TWI`) and the stream power index (`SPI`) using

$$\text{sACC} = \frac{\text{ACC}}{5}, \tag{4.6}$$

$$\texttt{TWI} = log\frac{\text{sACC}}{tan(\texttt{SLP})}, \tag{4.7}$$

and

$$\texttt{SPI} = log(\text{sACC} \times tan(\texttt{SLP})), \tag{4.8}$$

where sACC is the specific catchment area, 5 m is the cell size, and SLP is the slope raster surface calculated using seven different window sizes.

The topographic position index TPI was calculated with SAGA library ta_morphometry. Different values of maximum radius were used to include the effect of scale, all of them related to the window sizes used to calculate previous terrain attributes. A minimum radius value of 3 m was used in all calculations.

A total of $p = 1 \times \texttt{ELEV} + 7 \times (\texttt{NOR}, \texttt{SLP}, \texttt{TWI}, \texttt{SPI}, \texttt{TPI}) = 36$ covariates were defined using the terrain attributes derived from ELEV_90 and ELEV_10.

4.8 GEOLOGIC MAPS

Geologic data comes from the two most recent geologic maps (GEO_50 – Figure 4.5a, and GEO_25 – Figure 4.5b) published at the cartographic scales of 1:50 000 and 1:25 000 (GASPARETTO et al., 1988; MACIEL FILHO, 1990). Both of them were produced based on the most recent topographic sheets produced by the Brazilian Army at the cartographic scales of 1:50 000 and 1:25 000 (DSG, 1980; DSG, 1992a; DSG, 1992b). Alike topographic maps, geologic maps were available only in analogical format, and were hand digitized and georeferenced in QGIS. Intersections between all meridians and parallels ($n = 16$) were used as control points to adjust a second order polynomial model. Resampling was performed using the cubic resampling method. The original CRS (EPSG:31982 – SIRGAS2000 / UTM zone 22S) was transformed to match the reference grid using the R-package rgdal (BIVAND et al., 2013).

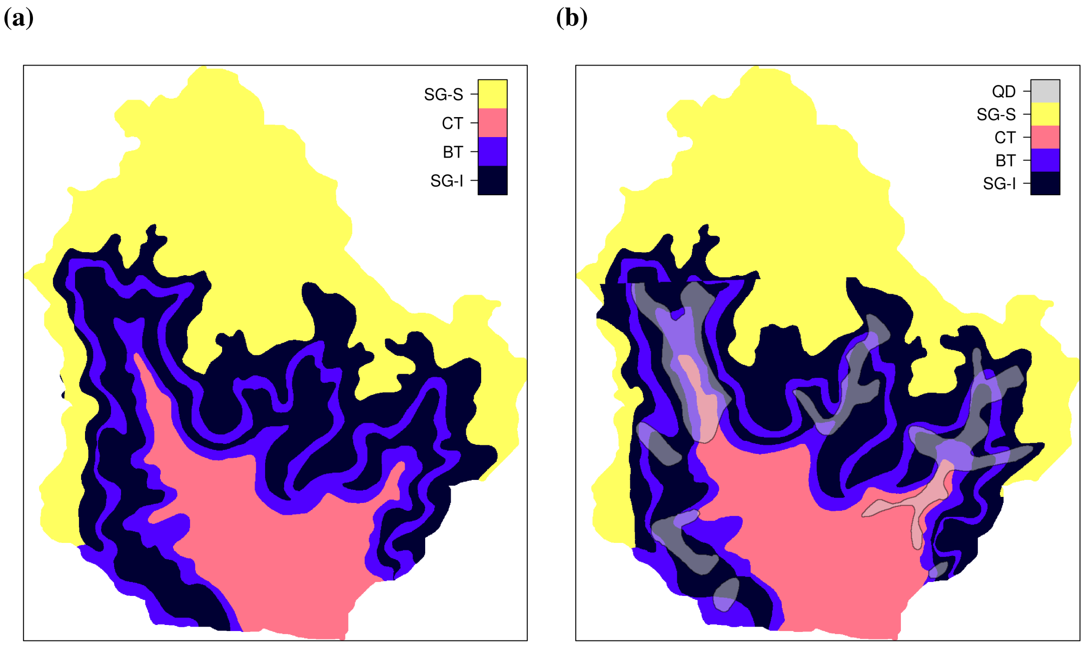


Figure 4.5: Geologic maps (a) GEO_50 and (b) GEO_25 used to derive indicator covariates included in the Santa Maria dataset. Legend abbreviations and derived indicator variables are described in the text.

The positional validation of geologic maps was performed using $n = 8$ (GEO_50) and $n = 5$ (GEO_25) GCPs, respectively. Validation statistics show that the positional accuracy of neither geologic maps meets the high-quality standards of the current Brazilian legislation (Table 4.4). Estimated RMSE are 147 and 69 m for GEO_50 and GEO_25, respectively, when the maximum RMSE recommended are 25 and $13\,\mathrm{m}$, respectively. For GEO_50, the lowest accuracy is found in the y-coordinate, while for GEO_25, the x-coordinate is the least accurate. Validation statistics suggest that there is a strong systematic error, which probably was propagated from the topographic maps used to produce the geologic maps. Therefore, the same strategy (affine transformation) used to remove the systematic positional error of the topographic maps was employed on geologic maps. Due to the lack of GCPs, model parameters were adjusted using the same set of GCPs used for the validation. The estimated uncertainty (RMSE) of the affine transformation is 86 and $22\,\mathrm{m}$ for GEO_50 and GEO_25, respectively.

Table 4.4: Error statistics of the horizontal validation of geologic maps GEO_50 and GEO_25 using $n = 8$ and $n = 5$ ground control points.

Statistics	x-coord	y-coord	Error vector	Azimuth
GEO_50 ($n = 8$)				
Mean, m	10	-102	140	169°
Absolute mean, m	43	125	-	-
Squared mean, m	3431	18067	21498	-
GEO_25 ($n = 5$)				
Mean, m	51	29	67	58°
Absolute mean, m	51	29	-	-
Squared mean, m	3457	1312	4769	-

Three ($p = 3$) indicator covariates were derived from GEO_50:

GEO_50a Inferior Sequence of the Serra Geral Formation. Composed mainly by basic igneous rocks (tholeiitic basalt and andesite). It is likely related to high soil clay content and effective cation exchange capacity.

GEO_50b Superior Sequence of the Serra Geral Formation. Composed mainly by acid igneous rocks (granophyric rhyolite and rhyodacite). It is likely related to moderate to high CLAY and ECEC.

GEO_50c Botucatu Formation. Composed mainly by aeolian sandstones. It is likely related to low CLAY and ECEC.

Four ($p = 4$) indicator covariates were derived from GEO_25, the first three of them having the same meaning of those derived from GEO_50:

GEO_25a Inferior Sequence of the Serra Geral Formation.

GEO_25b Superior Sequence of the Serra Geral Formation.

GEO_25c Botucatu Formation.

GEO_25d Quaternary deposits of fluvial, alluvial, and colluvial origin. It can help explaining the low clay content in areas where the soil is believed to have developed from igneous rocks as depicted in less detailed geologic maps.

4.9 LAND USE MAPS

The first land use map used to derive covariate data included in the Santa Maria dataset was produced by manually digitizing land use data of 1980 (`LU1980`, Figure 4.6a) published in the most recent topographic map produced by the Brazilian Army (cartographic scale of 1:25 000) (DSG, 1992b; DSG, 1992a). Most processing steps, including the correction of positional bias, are described in Section 4.7, except for the use of the nearest neighbour resampling method to match `LU1980` to the reference grid.

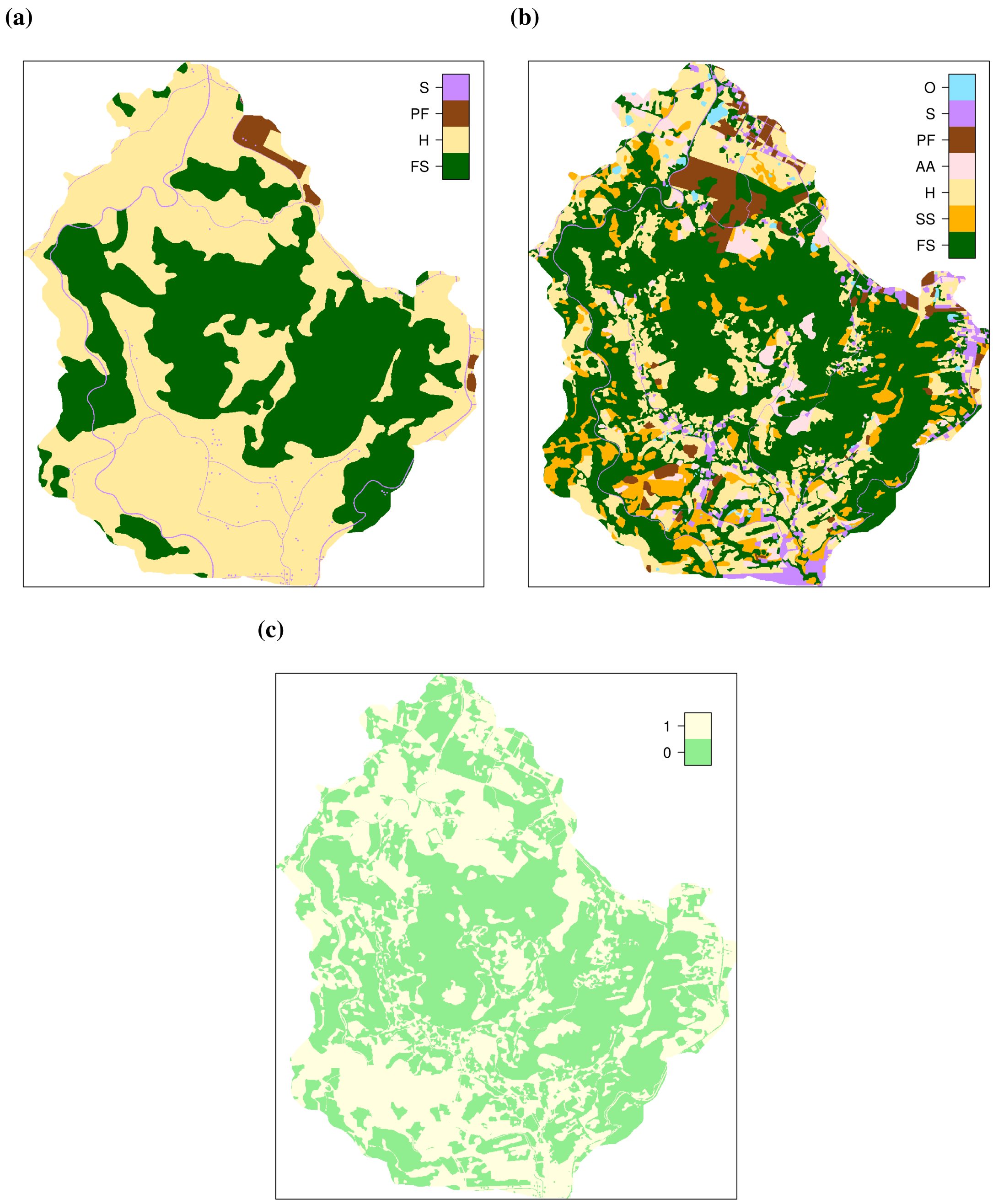


Figure 4.6: Land use maps (a) `LU1980` and (b) `LU2009` used to derive indicator covariates included in the Santa Maria dataset such as (c) `LUdiff`, the land use change between the years of 1980 and 2009. Legend abbreviations and derived indicator variables are described in the text.

The second land use map (`LU2009`, Figure 4.6b) was prepared at a cartographic scale of 1:2000 using high resolution (2.4 m) Digital Globe® QuickBird satellite images of 2008 and 2009, made publicly available in Google Earth® (SAMUEL-ROSA et al., 2011). Identification of land uses and delineation of mapping units were done manually, on the computer screen, without using any automated classification routine. Positional validation of Google Earth® imagery revealed that they have only minor systematic positional errors (Table 4.5). Despite of this, the attribute validation of both land use maps, using $n = 60$ validation points placed along $m = 12$ linear transects, showed that they have similar overall accuracy close to 70.00 %.

Table 4.5: Error statistics of the horizontal validation of Google Earth® imagery using $n = 14$ ground control points.

Statistics	x-coord	y-coord	Error vector	Azimuth
Mean, m	-1	3	6	184°
Absolute mean, m	3	5	-	-
Squared mean, m	14	57	71	-

`LU1980` was used to derive $p = 2$ covariates defined as indicator variables, with plantation forests (*PF*) and human settlements (*S*) being grouped together due to their small importance in terms of covered area (*PF*) and for not containing any soil observation (*S*). These covariates are:

`LU1980a` Native forest (*FS*), which is likely to have soils with higher fertility.

`LU1980b` Animal husbandry (*H*), the second most important land use in the study area, which is likely to have a soil fertility status lower than native forests.

Four ($p = 4$) indicator covariates were derived from `LU2009`, with plantation forests (*PF*), human settlements (*S*), and other land uses (*O*), which comprise natural and artificial water bodies, being grouped together due to their small importance in terms of covered area (*PF*) and for not containing any soil observation (*S* and *O*). These covariates are:

`LU2009a` Native forest (*FS*), as described above.

`LU2009b` Shrubland (*SS*), which is likely to have a soil fertility level above those found in areas used with annual crop agriculture and animal husbandry, but lower than in native forests.

`LU2009c` Animal husbandry (*H*), as described above.

`LU2009d` Annual crop agriculture (*AA*), which is likely to present the lowest soil fertility levels due to the usually poor management practices employed.

A seventh indicator covariate (`LUdiff`, Figure 4.6c) was derived using data from both land use maps. It consists of the land use change between 1980 and 2009, computed by checking if the land use has changed (1) or remained the same (0) in every grid cell after the 29-year period. `LUdiff` can be useful, for example, to explain the low soil organic carbon content in forest soils due to previous use with crop agriculture or animal husbandry.

4.10 SATELLITE IMAGES

Two sources of satellite images were used to produce covariate data included in the Santa Maria dataset. The first is the Thematic Mapper (TM) sensor aboard the longest-operating Earth observation satellite – Landsat 5 (Figure 4.7a). The satellite image used was acquired on 26 December 2010 and is freely available in the database of the Division of Image Generation of the Brazilian National Institute for Space Research (INPE-DGI). The image contains seven spectral bands (Table 4.6) (including a thermal band that is not included in the Santa Maria dataset), with 8 bit radiometric resolution (digital numbers from 0–255) and 30 m spatial resolution. The satellite image was orthorectified using Geomatica® OrthoEngine® with the Landsat rigorous model (Toutin's Model) (TOUTIN, 2004; PCI GEOMATICS, 2007). Due to the absence of good identifiable field GCPs, $n = 28$ GCPs were collected from Google Earth® imagery, which have a high positional accuracy in the DNOS catchment (Table 4.5). GCPs were located at easily identifiable geographical markers such as road intersections and bridges, evenly distributed throughout the image, and covering a variety of elevations, following standard recommendations (PCI GEOMATICS, 2007). The DEM used for orthorectification is TOPODATA after preprocessing as described in Section 4.7. Resampling was done using the nearest neighbour method to avoid changing the digital numbers.

After orthorectification, all bands were imported into GRASS GIS, where all other necessary corrections were performed. Radiometric correction (conversion from digital numbers to top-of-atmosphere reflectance) was done using GRASS module `i.landsat.toar`. Atmospheric correction (removal of the effects of the atmosphere on the reflectance values) was done with the 6S atmospheric model (VERMOTE et al., 1997) as implemented in GRASS module `i.atcorr` using the tropical atmospheric model, the continental aerosols model, an image-based visibility estimate of 20 km, and a constant elevation of 300 m.

The second source of satellite images is RapidEye (Figure 4.7b). Images are available through the Brazilian Ministry of the Environment (BRASIL, 2012), which has a full coverage of the Brazilian territory for 2011 and 2012. The satellite image used (tile number 2 225 403) was acquired on 16 November 2012 (second coverage). It contains five spectral bands (Table 4.6), featuring among them the so-called red-edge band, located between the red and the near-infrared bands. This spectral band is the main feature distinguishing RapidEye images from most other sources of satellite images, considered to provide additional information about the vegetation (WEICHELT et al., 2013). The satellite image has 16 bit radiometric resolution and 6.5 m spatial resolution, and was orthorectified at the source to 5 m spatial resolution using the sink-filled SRTM version 4 DEM (RAPIDEYE, 2013).

Table 4.6: Comparison between satellite images produced by Landsat 5 TM and RapidEye and the derived covariates.

Landsat 5 TM			RapidEye		
Band	Interval, nm	Covariate	Band	Interval, nm	Covariate
1 Visible	450–520	`BLUE_30`	Blue	440–510	`BLUE_5`
2 Visible	520–600	`GREEN_30`	Green	520–590	`GREEN_5`
3 Visible	630–690	`RED_30`	Red	630–685	`RED_5`
-	-	-	Red-edge	690–730	`EDGE_5`
4 Near-infrared	760–900	`NIR_30a`	Near-infrared	760–850	`NIR_5`
5 Near-infrared	1550–1750	`NIR_30b`	-	-	-
7 Mid-infrared	2080–2350	`MIR_30`	-	-	-

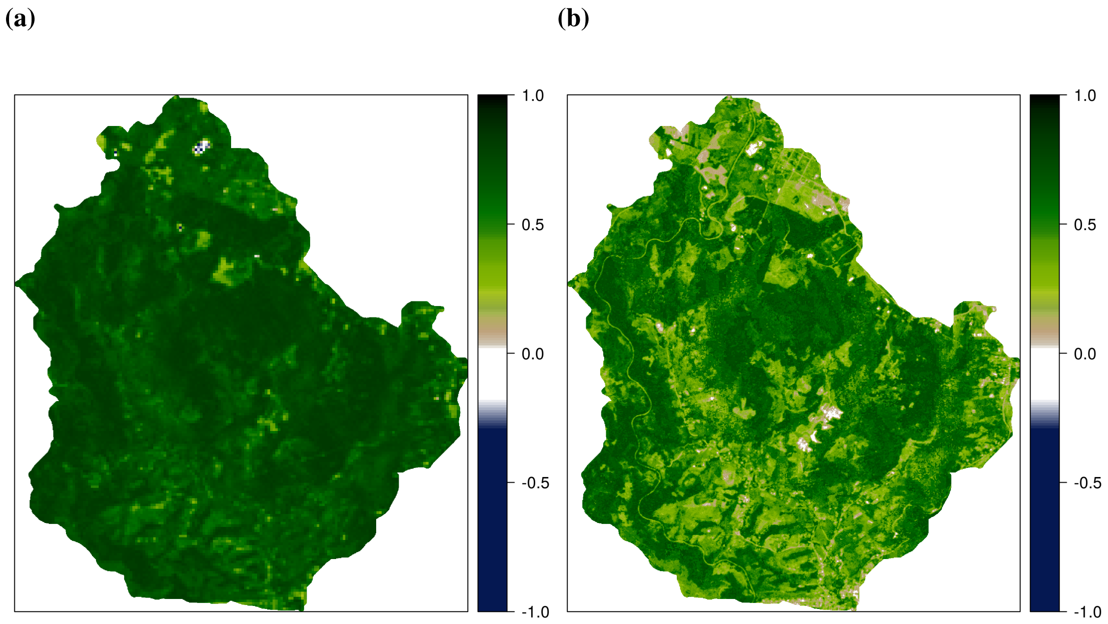


Figure 4.7: Satellite images used to derive covariates such as (a) `NDVI_30` and (b) `NDVI_5b` included in the Santa Maria dataset.

The RapidEye image was atmospherically corrected using the 6S atmospheric model (VERMOTE et al., 1997) employing the Fortran code developed by Antunes & Siqueira (2013) – GRASS module `i.atcorr` was not used because a bug was found when trying to correct the RapidEye image – assuming a tropical atmospheric model, the continental aerosols model, an image-based visibility estimate of 20 km, and a constant elevation of 300 m.

After atmospheric correction, Landsat 5 TM and RapidEye images were resampled using the nearest neighbour method to match the reference grid. Topographic correction was performed using GRASS module `i.topo.corr` with TOPODATA geometrically corrected to match the reference grid as described in Section 4.7.

Table 4.7: Error statistics of the horizontal validation of satellite images produced by Landsat 5 TM and RapidEye using $n = 14$ ground control points.

Statistics	x-coord	y-coord	Error vector	Azimuth
Landsat 5 TM				
Mean, m	31	-11	45	136°
Absolute mean, m	33	25	-	-
Squared mean, m	1494	1223	2717	-
RapidEye				
Mean, m	-25	-25	36	226°
Absolute mean, m	25	25	-	-
Squared mean, m	680	708	1388	-

Individual bands of both satellite images were defined as covariates, totalling $p = 6$ from Landsat 5 TM and $p = 5$ from RapidEye (Table 4.6). Another $p = 6$ covariates ($p = 2$

from Landsat 5 TM and $p = 4$ from RapidEye) were defined using two vegetation indices: the normalized difference vegetation index (NDVI) and the soil-adjusted vegetation index (SAVI), calculated as

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}} \tag{4.9}$$

and

$$\text{SAVI} = (1.0 + 0.5) \times \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED} + 0.5}, \tag{4.10}$$

where NIR $=$ `NIR_30a` and RED $=$ `RED_30` for the Landsat 5 TM image, and NIR $=$ `NIR_5` and RED $=$ `RED_5` or RED $=$ `EDGE_5` for RapidEye image.

5 CAPÍTULO IV

MODELO CONCEITUAL DE PEDOGÊNESE DA BACIA DO DNOS

* Colaboraram na preparação deste manuscrito: Pablo Miguel (UFPel), Jean Michel Moura Bueno (UFSM), Ricardo Simão Diniz Dalmolin (UFSM), Andrisa Balbinot (UFSM), Lúcia Helena Cunha dos Anjos (UFRRJ), Gustavo de Mattos Vasques (Embrapa Solos), e Gerard B. M. Heuvelink (ISRIC – World Soil Information).

5.1 RESUMO

O presente manuscrito apresenta o modelo conceitual de pedogênese – uma descrição explícita dos fatores e processos de formação do solo que determinam as características do solo e o seu padrão de distribuição espaço-temporal – da bacia de captação do reservatório do Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento (DNOS-CORSAN), localizada no sul do Brasil. O clima é subtropical úmido sem estação seca definida. O relevo é plano a montanhoso (variação de altitude entre 139 e 475 m acima do nível do mar), com vales encaixados que influenciam a precipitação e o fluxo radiativo nas diferentes superfícies geomórficas. A geologia é constituída pela sequência de três formações: rochas sedimentares (arenito fluvial), seguidas de rochas ígneas (basaltos-andesitos toleíticos e vitrófilos, riólitos-riodacitos granofíricos) intercaladas por rochas sedimentares (arenito eólico). Depósitos do Quaternário aparecem nas partes mais baixas. A geomorfologia atual é resultado dos processos erosivos do Terciário e Quaternário. A dissecação atual é fraca devido ao clima que favorece a instalação e permanência de vegetação exuberante. Três unidades morfoestruturais são identificadas: no topo, o Planalto, com relevo suave-ondulado a ondulado, seguido pelo Rebordo do Planalto, com ampla variação altimétrica, declividade acentuada e escarpas abruptas; na base, a Depressão Periférica, com formas agradacionais de planície fluvial. Nas partes altas, a rede de drenagem apresenta padrão bem definido, geralmente retangular, determinada pelas falhas e/ou fraturas. Já nas áreas mais baixas, devido aos processos de deposição sedimentar e erosão fluvial, sua configuração é sinuosa. Ali encontram-se um lençol freático próximo da superfície do solo e cursos de água perenes. O uso da terra para produção agrossilvopastoril foi intenso em tempos pretéritos e resultou em forte degradação do solo. O abandono de muitas áreas degradadas permitiu a regeneração da vegetação natural, resultando na atual ocupação com florestas e vegetação secundária de ±60 %. Em geral, o solo é pouco profundo devido ao predomínio de condições de forte declividade. É comum encontrar solo raso mesmo em áreas de maior estabilidade como fruto da degradação pelo uso agrícola. O solo é mais profundo no Planalto, nos terraços do Rebordo, nas coxilhas (colinas) de relevo suave-ondulado a ondulado, e nas planícies aluviais. A textura é mais fina e homogênea ao longo do perfil quando desenvolvido a partir de rochas ígneas. As características do solo nas planícies aluviais são determinadas pela presença constante de lençol freático próximo da superfície.

Palavras-chave: Província Geológica do Paraná. Bacia do DNOS. Rebordo do Planalto. Fatores de formação do solo. Pedogênese.

5.2 ABSTRACT

This document presents the conceptual model of pedogenesis – an explicit description of soil-forming factors and processes that determine the spatio-temporal distribution of soil properties – of the catchment of the reservoir of the Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento (DNOS-CORSAN), located in southern Brazil. Climate is subtropical humid without a dry season. Relief varies between plain and mountainous, with enclosed valleys (elevation ranging between 139 and $475\,\mathrm{m}$ above sea level) that determine rainfall volume and radiative flux on different surfaces. The geology is composed of a sequence of three formations: consolidated sedimentary rocks (fluvial sandstone), followed by basic and acid igneous rocks (andesite-basalt and rhyolite-rhyodacite), interlayered with consolidated sedimentary rocks (aeolian sandstone). Unconsolidated Quaternary colluvial deposits occur in the lower portions of the landscape. Current geomorphology is a result of erosive processes of the Tertiary and Quaternary. Landscape dissection is weak due to the current climate that favours the installation and maintenance of an exuberant vegetation. There are three morphostructural units: at the top, the *Planalto* (Plateau), with gently-rolling to sloping relief, followed by the *Rebordo do Planalto* (Plateau Border), with wide altimetric variation, steep slopes and abrupt cliffs; at the bottom, the *Depressão Periférica* (Peripheral Depression), composed of aggradational fluvial plains. In higher altitudes, the drainage network has a well defined pattern, generally rectangular, determined by the faults and/or fractures. In the lower areas, its configuration is sinuous due to sediment deposition and fluvial erosion, with the presence of water table close to the surface and perennial water streams. Land use for agrosilvopastoral production was intense in past times, resulting in severe soil degradation. Recent abandonment of many degraded areas allowed the regeneration of natural vegetation, resulting in $\pm 60\,\%$ of the area being now occupied with forest and secondary vegetation. The soil is predominantly shallow due to the dominance of steep slopes. Even in gently-sloping terrain it is common to find shallow soils as a result of soil degradation. Deeper soil can be found in the Planalto, in the terraces of the Rebordo, and in the small hills with gently-rolling slopes and alluvial plains. Soil texture is finer and more homogeneous throughout the soil profile in soil developed from igneous rocks. Soil features in the alluvial plains are determined by the constant presence of the water table close to the surface.

Keywords: Paraná Geological Province. DNOS Catchment. Plateau Border. Soil-forming factors. Pedogenesis.

5.3 APRESENTAÇÃO

A modelagem espacial do solo inicia com a definição de um *modelo conceitual de pedogênese*. Um modelo conceitual de pedogênese constitui uma representação verbal da realidade sob estudo que inclui a descrição explícita dos fatores e processos de formação do solo que determinam as características do solo e o seu padrão de distribuição espaço-temporal. Isso requer a reunião de toda a informação ambiental disponível e aplicação dos conceitos de relação solo-paisagem, desenvolvimento do solo em catenas, ou outro modelo teórico de explicação da variação espacial do solo.

O presente manuscrito apresenta o modelo conceitual de pedogênese da bacia de captação do reservatório do Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento (DNOS-CORSAN), localizada na divisa entre os municípios de Itaara (ao norte) e Santa Maria (ao sul), na porção sul da Bacia Sedimentar do Paraná, estado do Rio Grande do Sul, Brasil (Figura 5.1). A bacia de captação do reservatório do DNOS-CORSAN corresponde à cabeceira da bacia hidrográfica do Rio Vacacaí-Mirim, tributário do Rio Jacuí e, consequentemente, do Rio Guaíba e da Lagoa dos Patos. A bacia de captação do reservatório do DNOS-CORSAN cobre uma área de ±29 km^2 e alimenta um reservatório com volume máximo de ±3 800 000 m^3 em uma área inundada de 0.74 km^2. Este reservatório contribui com até 30 % do abastecimento de água da cidade de Santa Maria (DIAS, 2003; DILL et al., 2004; MIGUEL, 2010).

Os estudos em modelagem espacial do solo na bacia do DNOS – abreviatura de bacia de captação do reservatório do DNOS-CORSAN – iniciaram no ano de 2008 com o grupo de pesquisa *Gênese, composição e comportamento dos solos do RS*[19], sediado no Departamento de Solos da Universidade Federal de Santa Maria (UFSM). Devido à limitação de recursos, o grupo de pesquisa optou por restringir seus estudos em modelagem espacial do solo à uma parte da bacia do DNOS. A área escolhida cobre ±60 % de toda a bacia do DNOS, o que corresponde à uma área de ±18 km^2. A área de estudo foi escolhida por apresentar acesso facilitado, além de englobar as principais formações geológicas, geoformas, usos da terra, e vegetação presentes na bacia do DNOS (Figura 5.2).

5.4 CLIMA

O clima local é classificado como Cfa[20] – subtropical úmido sem estação seca definida –, com temperatura média anual de 19.1 °C. As temperaturas podem alcançar > 40 °C, no verão, e valores negativos no inverno (HELDWEIN et al., 2009). A precipitação média anual é de 1708 mm bem distribuídos ao longo do ano (MALUF, 2000). Predominam os ventos (em ordem de frequência) do quadrante leste (frio, úmido e de intensidade fraca a moderada), oeste (frio, seco e de intensidade fraca a moderada) e norte (quente, seco e de intensidade moderada a forte) (HELDWEIN et al., 2009).

O padrão predominante das chuvas é o avançado, caracterizado por ter seu pico de maior intensidade no início da precipitação (MEHL et al., 2001). As chuvas de maior intensidade ocorrem nos meses do final da primavera, verão e início do outono (MOURA-BUENO, 2012). Como resultado desse padrão, as chuvas de inverno são as menos erosivas, mesmo que o conteúdo de água do solo permaneça elevado durante todo o período. O padrão de precipitação também é condicionado pelo relevo. Observações feitas em três locais durante o ano de 2011, marcado por forte estiagem, mostram variação na lâmina total precipitada entre 1317 e 1411 mm

[19] O grupo está registado no Diretório dos Grupos de Pesquisa no Brasil (DGP) mantido pelo CNPq.

[20] Mais informações sobre o tipo climático Cfa podem ser encontrados na Wikipedia.

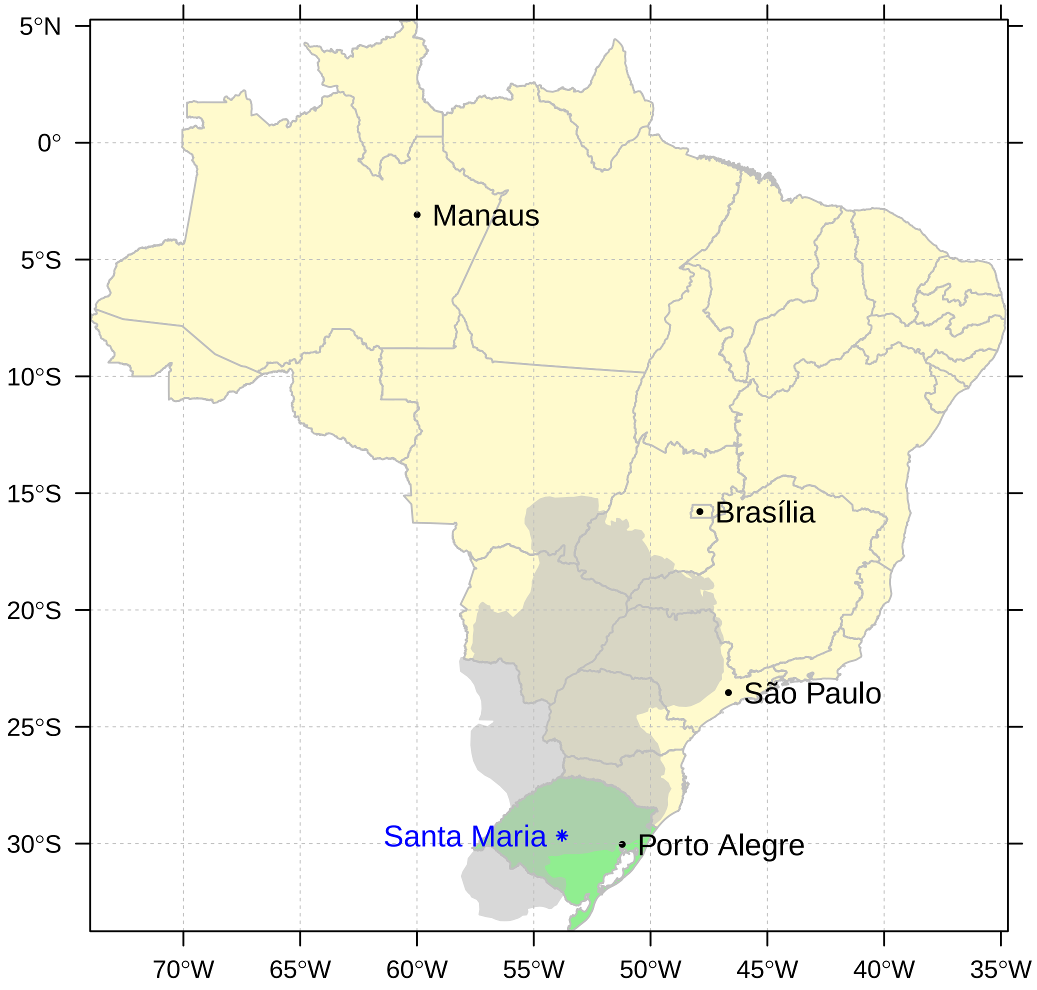


Figura 5.1: Localização da bacia de captação do reservatório do Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento no Município de Santa Maria (em azul), Estado do Rio Grande do Sul (em verde), Brasil, na porção sul da Província Geológica do Paraná (em cinza).

(MOURA-BUENO, 2012). Assim, o relevo plano a montanhoso, com vales encaixados, parece condicionar a formação de diferentes regiões microclimáticas, refletindo no volume e intensidade das chuvas (MOURA-BUENO, 2012).

O relevo também deve condicionar o fluxo radiativo que atinge as diferentes superfícies. Apesar de não haver estudos que demonstrem a efetividade desse fenômeno na área, é reconhecido que grande parte da superfície em terrenos de topografia complexa é influenciada pelo efeito de sombreamento, sobretudo nas primeiras horas da manhã e no final da tarde (OLIPHANT et al., 2003). Além disso, a declividade do terreno possui forte influência sobre o ângulo de interceptação da radiação solar pelas superfícies (BIRKELAND, 1999). Como consequência, deve ocorrer variações na temperatura e conteúdo de água no solo nas diferentes superfícies. Os meses de inverno são marcados por ainda menor disponibilidade de radiação solar devido à alta frequência de nevoeiros, sobretudo nas partes mais baixas, com valores normais de insolação de $5.1\,\mathrm{h\,d^{-1}}$ (HELDWEIN et al., 2009). Além disso, devido à variação de altitude entre 139 e $475\,\mathrm{m}$, deve ocorrer diferença na temperatura da ordem de $4\,°\mathrm{C}$ entre a parte

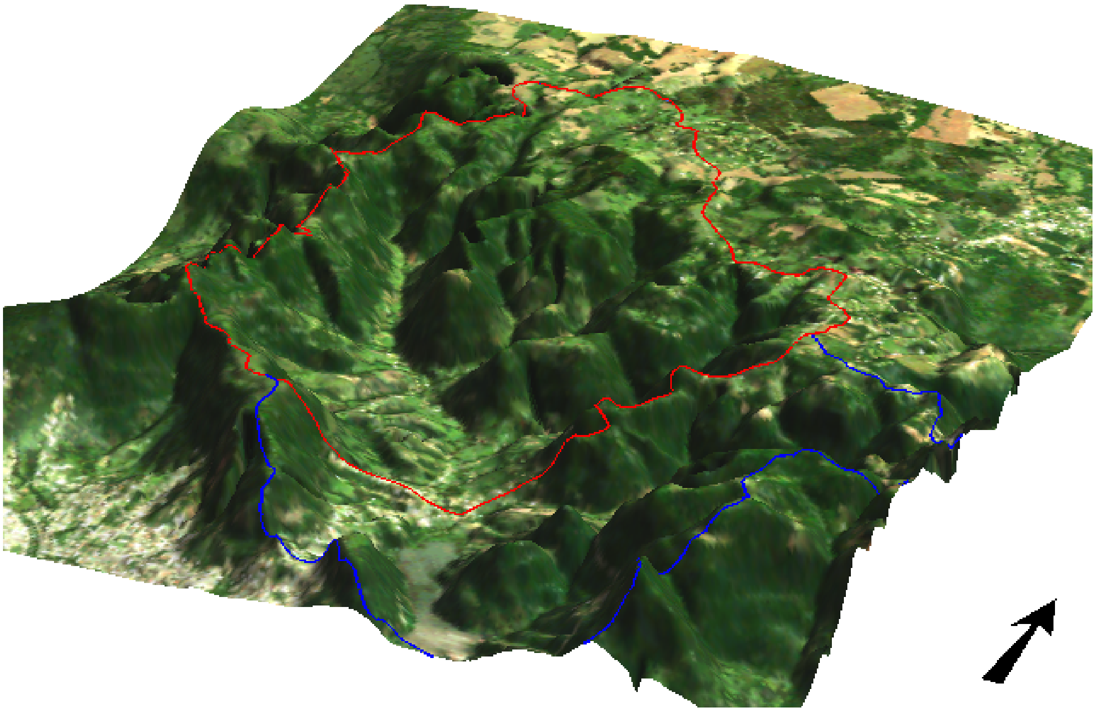


Figura 5.2: Localização da área de estudo (em vermelho) na bacia de captação do reservatório do Departamento Nacional de Obras de Saneamento-Companhia Riograndense de Saneamento (DNOS-CORSAN) (em azul). A área de estudo, cujo relevo é bastante acidentado e o uso da terra predominante é floresta natural, possui aproximadamente 4 km de largura e 5 km de comprimento, o que corresponde a ±60 % da área bacia do DNOS-CORSAN.

mais baixa e a parte mais alta (HELDWEIN et al., 2009).

5.5 GEOLOGIA

A geologia é bastante complexa, sendo constituída de três formações geológicas, além de depósitos coluviais e aluviais do Quaternário. A literatura sobre o tema é vasta (BORTOLUZZI, 1974; BRASIL, 1980; GASPARETTO et al., 1988; MACIEL FILHO, 1990; MACHADO, 1998; PIERINI et al., 2002; MARQUES; ERNESTO, 2005; MILANI, 2005; PINTO, 2005; CPRM, 2007; PEDRON, 2007; SARTORI, 2009; NASCIMENTO; PENNA E SOUZA, 2010; WERLANG et al., 2010; PEDRON et al., 2012), e uma revisão da mesma é apresentada aqui.

Na base da sequência estratigráfica, em elevações abaixo de ±200 m, está a Formação Caturrita, constituída de material sedimentar depositado em ambiente fluvial no Triássico Superior. Sua composição é diversa, apresentando seixos de siltito argiloso vermelho na base, seguido de arenito avermelhado de granulometria fina à média, composição quartzosa e matriz argilosa, podendo ainda conter considerável teor de feldspato, sobreposto por siltito e folhelho também avermelhados. Em geral, a granulometria do arenito é mais grosseira e menos argilosa na base da deposição. Devido à sua origem fluvial, a Formação Caturrita apresenta marcada estratificação cruzada acanalada e tabular. A origem fluvial também resulta em significativa variação espacial na granulometria do arenito, identificada pelo contraste entre áreas de maior cimentação e coesão, com outras de maior condutividade hidráulica. Imediatamente acima da Formação Caturrita encontra-se, ora a Formação Botucatu, ora a Sequência Inferior da Formação Serra Geral.

Em elevações entre ±200 e ±350 m está a Sequência Inferior da Formação Serra Geral (basaltos-andesitos toleíticos). As rochas básicas são de coloração cinza-escura e são constituídas de plagioclásio cálcico, clinopiroxênio, magnetita e material intersticial de quartzo e material desvitrificado. Em elevações superiores a ±350 m está a Sequência Superior da Formação Serra Geral (vitrófilos, riólitos-riodacitos granofíricos). As rochas ácidas apresentam cor cinza-clara, estrutura microcristalina e são constituídas de cristais e plagioclásio, clinopiroxênios, hornblenda uralítica e magnetita. A origem desse material remonta o Cretáceo, quando sucessivos derrames de lavas de origem vulcânica fissural ocorreram durante aproximadamente 10 milhões de anos em toda a Bacia do Paraná. Esses eventos ocorreram ao mesmo tempo em que iniciava-se a separação das plataformas continentais que hoje constituem a América do Sul e África, marcando o final da existência do supercontinente Pangeia.

O arenito eólico constituinte da Formação Botucatu é encontrado tanto assentado sobre a Formação Caturrita, como no interior da Formação Serra Geral (arenito *intertrap*). Trata-se de arenito quartzoso de granulometria fina à média, contendo feldspato alterado e cimentado por sílica ou por óxido de ferro, que lhe confere a coloração rosa-avermelhada. Sua deposição teve início no Cretáceo Inferior, período em que a Bacia do Paraná estava sob influência de clima desértico. Essa condição climática continuou durante todo o período em que ocorreram as dezenas de eventos de vulcanismo fissural, fazendo com que os mesmos fossem sucedidos por deposições eólicas de duração variável. Como a duração e a quantidade de material depositado pelos eventos de vulcanismo fissural era variável, assim como o intervalo de tempo entre cada novo evento e a intensidade das deposições de sedimentos eólicos, a espessura das camadas do arenito eólico e das rochas vulcânicas é bastante variável. Além disso, devido aos diversos eventos de subsidência que ocorreram no eixo central da Bacia do Paraná, com consequente soerguimento de suas bordas, as camadas dessas rochas possuem diferentes inclinações ao longo de sua faixa de exposição, sendo caracteristicamente ondulada e com suave tendência de inclinação para sudoeste.

As deposições do Quaternário são constituídas por depósitos coluviais e aluviais. Em elevações entre ±200 e ±300 m encontram-se depósitos coluviais de material proveniente de uma ou ambas as Formações Serra Geral (fragmentos de tamanho variado) e Botucatu. Em elevações abaixo de ±200 m são mais comuns os depósitos coluviais de uma ou ambas as Formações Botucatu e Caturrita. Esses depósitos ocorrem de maneira descontínua nas encostas. Próximo aos cursos de água na porção mais baixa da bacia e no entorno do reservatório, encontram-se depósitos fluviais recentes, geralmente constituídos de fragmentos arredondados (seixos) de tamanho variável e/ou sedimentos arenosos. Em pequenas áreas abaciadas e mal drenadas, os sedimentos apresentam granulometria mais fina.

5.6 GEOMORFOLOGIA

A área de estudo está situada na porção sul da Bacia Sedimentar do Paraná. Assim, as geoformas atuais são resultado dos processos erosivos que ocorreram durante o Terciário e o Quaternário (SARTORI, 2009), após as últimas deposições de lavas vulcânicas e de sedimentos eólicos. Durante esse período, a esculturação da paisagem foi determinada pelas alternações entre climas úmidos, semiáridos e áridos (SARTORI, 2009). Atualmente, o clima subtropical úmido favorece a instalação e permanência de vegetação mais eficiente na redução do processo de dissecação da paisagem (SARTORI, 2009; NASCIMENTO; PENNA E SOUZA, 2010). Isso permite que as superfícies geomórficas atinjam maior estabilidade e maturidade, embora o uso agrícola das terras tenha acelerado, pontualmente, os processos erosivos na área devido à limitada adoção de práticas conservacionistas (Seção 5.8).

A bacia do DNOS é abrangida pela unidade morfoestrutural do Rebordo do Planalto da Bacia do Paraná (NASCIMENTO; PENNA E SOUZA, 2010). O Rebordo do Planalto é caracterizado pela ampla variação altimétrica, declividade acentuada e escarpas abruptas, apresentando formas denudacionais com topos convexos (fluxo hídrico divergente), aguçados e em formas de escarpas. Nesses locais, as vertentes assumem forma retilínea com grande desnível (NASCIMENTO; PENNA E SOUZA, 2010), muitas vezes interrompidas por degraus ou patamares, na maioria das vezes encaixadas em falhas e/ou fraturas. Esses patamares são resultado da ação diferencial dos processos denudacionais sobre a paisagem, geralmente condicionados pela resistência do material de origem, seja ela química/mineralógica (rochas vulcânicas básicas vs. ácidas), física/granulométrica (rochas vulcânicas versus sedimentares), ou estrutural (padrão de diaclasamento vertical vs. horizontal das rochas vulcânicas) (HOLTZ, 2003; PEDRON, 2007; STRECK et al., 2008). Entretanto, em algumas situações, os patamares são formados por depósitos coluviais – mesmo nas partes altas do Rebordo do Planalto – resultantes de movimentos de massa causados por eventos pluviométricos de elevada intensidade e/ou duração (PINHEIRO; SOARES, 2004; PAISANI; GEREMIA, 2010). Nesses casos, os patamares possuem menor dimensão e maior declividade. Em outras situações, os patamares coluviais, sobretudo aqueles originados dos arenitos da Formação Botucatu, formam vertentes alongadas e com menor inclinação, que chegam até a margem dos cursos de água.

Como a bacia do DNOS se encontra em uma região de transição morfoestrutural, as características geomorfológicas das porções mais altas e mais baixas da paisagem são similares àquelas encontradas nas unidades morfoestruturais adjacentes, respectivamente, o Planalto e a Depressão Periférica. O Planalto é marcado pelo relevo suave-ondulado a ondulado, com formas denudacionais de superfícies planas com topos convexos (fluxo hídrico divergente). Nesses locais, as vertentes assumem forma convexa levemente ondulada, muitas vezes encaixadas em falhas e/ou fraturas (NASCIMENTO; PENNA E SOUZA, 2010). Já a Depressão Periférica é marcada pelo acúmulo de sedimentos provenientes do Planalto e do Rebordo do Planalto, formando planícies aluviais que se intercalam entre as coxilhas (denominação regional de colinas). Ali predominam as formas agradacionais de planície fluvial e formas denudacionais com topos convexos e superfícies planas. As últimas correspondem às coxilhas de algumas dezenas de metros de altitude, geralmente formadas sobre o substrato da Formação Caturrita e assentadas na base do Rebordo do Planalto (GASPARETTO et al., 1988). Essas coxilhas constituem divisores de água de pequena amplitude comumente usados na subdivisão da bacia do DNOS em pequenas sub-bacias (MARINS, 2004; SARTORI, 2009). Nesses locais as vertentes costumam ser alongadas e assumem a forma predominantemente côncava devido aos processos de deposição sedimentar e erosão fluvial, muito fracos sob a atual condição climática (NASCIMENTO; PENNA E SOUZA, 2010; WERLANG et al., 2010).

A grande heterogeneidade geomorfológica da bacia do DNOS se traduz em uma grande heterogeneidade textural do relevo, ou rugosidade, resultante da ação climática ao longo do tempo geológico (NASCIMENTO; PENNA E SOUZA, 2010). Entretanto, em menor escala, a ação antrópica também atuou sobre a configuração geomorfológica da área (Seção 5.8). O principal efeito se deu pela erosão da camada superficial do solo, cultivado intensivamente sem adoção de práticas conservacionistas ao longo de inúmeras décadas (MENEZES, 2008; STÜRMER, 2008; MIGUEL, 2010; SAMUEL-ROSA et al., 2011). Além disso, a abertura de caminhos para acesso às áreas de produção nos patamares do Rebordo do Planalto e nos topos de morros proporcionou a formação de canais de concentração e escoamento dos fluxos hídricos superficiais, levando à formação inicial de voçorocas. Obras de maior expressividade, como ferrovias, ruas e rodovias pavimentadas, conjuntos habitacionais e construções isoladas, as quais envolvem operações de terraplanagem e aterramento, também resultaram em modificações localizadas na geomorfologia da área. Por fim, a construção do reservatório possibilitou a formação de uma área de agradação da paisagem bastante estável em seu entorno, uma vez que o sedimento removido do Planalto e do Rebordo do Planalto já não são mais transportados a jusante.

5.7 HIDROGRAFIA

A hidrografia da área é condicionada pelas condições geomorfológicas, geológicas, pedológicas e climáticas, ao mesmo tempo em que exerce forte influência sobre a modelagem da paisagem (NASCIMENTO; PENNA E SOUZA, 2010). Como a bacia do DNOS é abrangida pela unidade morfoestrutural do Rebordo do Planalto da Bacia do Paraná (Seção 5.6), a drenagem apresenta padrão bem definido, geralmente retangular, determinado pelas falhas e/ou fraturas (BORTOLUZZI, 1974; GASPARETTO et al., 1988; NASCIMENTO; PENNA E SOUZA, 2010). A própria formação da bacia do DNOS se deve a existência de falhas e/ou fraturas (GASPARETTO et al., 1988). A principal e maior delas localiza-se no eixo central da bacia, onde atualmente está localizado o leito do Rio Vacacaí-Mirim. Quanto aos tributários do Rio Vacacaí-Mirim, a maioria possui leito assentado sobre outras falhas e/ou fraturas de menor dimensão perpendiculares àquela do eixo central da bacia.

Nas áreas mais baixas, cujas características geomorfológicas se assemelham à Depressão Periférica, a drenagem apresenta configuração sinuosa, resultado dos processos de deposição sedimentar e erosão fluvial (PAIVA et al., 2001; SUTILI et al., 2009). Como o relevo é plano a suave-ondulado, e as vertentes longas e predominantemente côncavas, o lençol freático fica próximo da superfície do solo. A variação das condições meteorológicas ao longo do ano fazem com que o lençol freático apresente flutuação significativa, mantendo o conteúdo de água do solo elevado durante os meses mais frios (menor evapotranspiração) (HELDWEIN et al., 2009). Isso também favorece a ocorrência de inundações, sobretudo nas proximidades dos cursos de água e do reservatório (GOLDANI, 2006).

Muitos cursos de água localizados na áreas cujas características geomorfológicas se assemelham ao Planalto, assim como no Rebordo do Planalto e nas coxilhas assentadas em sua base, são sazonais. Em geral, esses cursos de água estão em atividade apenas nos meses mais frios do ano, quando a disponibilidade de água no ambiente é maior, ou durante os eventos de precipitação de forte intensidade, que ocorrem nos meses de verão (HELDWEIN et al., 2009; MOURA-BUENO, 2012). Os cursos de água do Rebordo do Planalto costumam apresentar leito raso e pedregoso, muitas vezes assentado sobre rochas da Formação Serra Geral (SUTILI et al., 2009). Já os cursos de água localizados nos patamares e coxilhas costumam ser rasos, se assentados sobre rochas da Formação Caturrita, ou profundos, se assentados sobre rochas da Formação Botucatu ou depósitos coluviais, formando voçorocas. Segundo relatos de alguns dos moradores mais antigos da bacia do DNOS, muitas nascentes e pequenos cursos de água já perderam totalmente sua atividade, sobretudo quando localizadas no interior ou à jusante de áreas de uso antrópico intensivo.

Dado que a declividade e o desnível entre a parte mais baixa e a parte mais alta da área são acentuados, as cheias costumam apresentar velocidade e vazão bastante grandes (PAIVA et al., 2001; SUTILI et al., 2009). Em média, nos 7 km de extensão do Rio Vacacaí-Mirim, da nascente até o reservatório, a declividade média é de $0.03\,\mathrm{m\,m^{-1}}$. Isso representa um desnível de $\pm 210\,\mathrm{m}$, o que resulta em um tempo de concentração da bacia estimado de 3 h (PAIVA et al., 2001). Essas características causam erosão severa nas margens dos cursos de água nas áreas mais baixas (depósitos aluviais), sobretudo nos raios externos das curvas, onde a velocidade da água é maior (SUTILI et al., 2009). Em alguns trechos, os cursos de água chegaram a ter sua largura duplicada em menos de uma década, resultando no aumento da sinuosidade e do nível de fundo (PAIVA et al., 2001). Esses eventos comprometem áreas de produção agrícola, bem como a estrutura de residências e vias públicas localizadas nas margens dos cursos de água. Grande parte do material removido das margens dos cursos de água é transportado para dentro do reservatório, que já perdeu mais de um terço de sua capacidade inicial de armazenamento de água (DILL et al., 2004).

5.8 USO DA TERRA E VEGETAÇÃO

A bacia do DNOS foi intensamente ocupada em tempos pretéritos para produção agrossilvopastoril de pequeno porte (agricultura familiar) usando sistemas de cultivo convencional com aração periódica e queimada. A grande quantidade, extensão e boa distribuição da rede viária em toda a área é uma forte evidência dessa ocupação. A área também é cortada por uma estrada férrea e avizinha uma rodovia federal, ambas muito movimentadas. Entretanto, nas últimas décadas, muitos caminhos internos das propriedades rurais foram desativados, fruto do abandono de muitas áreas de produção agrossilvopastoril, a maioria delas localizada nos topos dos morros, patamares do Rebordo ou no fundo de vales, onde a manutenção dos caminhos é muito onerosa, especialmente quando distantes da sede das propriedades (Figura 4.6 e Figura 4.7). No passado, essas estradas internas possibilitavam o trânsito de pessoas e o transporte da produção agrossilvopastoril, a qual era usa para suprir as necessidades próprias, sendo o excedente comercializado às margens da estrada férrea e nas áreas urbanas do município. Atualmente, o tráfego existente está concentrado nas estradas principais e secundárias, enquanto alguns poucos caminhos internos ainda são utilizados para condução de rebanhos ou acesso a pequenas áreas de produção agrossilvopastoril. A maioria dos caminhos internos inativos está localizada no interior de áreas de floresta natural regenerada, o que dificulta a sua identificação com imagens aéreas ou orbitais. Em alguns casos, esses caminhos são utilizados em atividades de turismo, especialmente para caminhadas a pé, ou atividades esportivas com motocicletas (trilhas).

O abandono de muitas áreas de produção agrossilvopastoril possibilitou a regeneração da vegetação natural em grande parte da bacia do DNOS. Atualmente, a área ocupada por florestas e vegetação secundária (capoeira) representa ±60 % da área total (SAMUEL-ROSA et al., 2011). Isso mostra que, assim como toda a região do Rebordo do Planalto da Bacia do Paraná, o uso da terra para fins agrossilvopastoris na bacia do DNOS também foi desintensificado nas últimas décadas (SEMA/UFSM, 2001; DILL et al., 2004; POELKING, 2007; MIGUEL, 2010; SAMUEL-ROSA et al., 2011; DULLIUS, 2012; TEN CATEN et al., 2012a). As áreas de floresta são encontradas nos mais diversos estádios de desenvolvimento, com os estratos intermediário e superior concentrando a maior parte dos indivíduos. As florestas originais, e secundárias em estádio avançado de desenvolvimento, são encontradas, predominantemente, em áreas de difícil acesso. Em geral, as áreas de floresta predominam nas regiões com maior declividade e solo raso e pedregoso. Tais condições pedológicas são resultado das condições geológicas e geomorfológicas e/ou da degradação causada pelo intenso uso da terra para produção agrossilvopastoril durante inúmeras décadas usando sistemas de cultivo convencional (SAMUEL-ROSA et al., 2011).

A ocorrência de fragmentos florestais ao longo das margens dos cursos de água, estradas e próximo de edificações é comum. Nas formações mais jovens, é comum encontrar fragmentos de carvão e outros sinais do uso agrossilvopastoril num passado recente. As áreas sob vegetação secundária (capoeira) também ocorrem em toda a bacia do DNOS, predominantemente em locais de difícil acesso e acentuada declividade, geralmente interligados por caminhos internos, indicando sua utilização agrossilvopastoril no passado (SAMUEL-ROSA et al., 2011). Sua composição florística varia entre as áreas, predominando espécies de porte herbáceo e arbustivo. O solo dessas áreas costuma ser menos pedregoso que nas áreas florestadas, mas apresentam profundidade semelhante (SAMUEL-ROSA et al., 2011), possivelmente indicando que as atividades agrossilvopastoril foram encerradas antes do solo atingir seu nível máximo de degradação. Entretanto, a reserva de nutrientes do solo, assim como a sua fertilidade física, foram esgotadas a tal ponto que a vegetação secundária ainda não foi capaz de produzir melhorias significativas quando comparado com as condições originais (MENEZES, 2008; ZALAMENA, 2008).

Apesar do abandono de muitas áreas de produção agrossilvopastoril nas últimas décadas, sobretudo aquelas com maior dificuldade de acesso, os caminhos internos utilizados para o escoamento da produção, mesmo depois de inativos, continuam associados ao processo de degradação do solo. Isso ocorre porque, na grande maioria dos casos, todo o fluxo da água de escoamento superficial proveniente dos eventos de precipitação é concentrado nesses caminhos internos, onde, geralmente, quantidade significativa de resíduos vegetais está depositada. Esse resíduo vegetal, mais a fração mineral do solo carregada pela enxurrada, costuma chegar com facilidade aos cursos de água. No caso dos caminhos internos ainda utilizados na condução de rebanhos bovinos, a produção de sedimentos minerais tende a ser ainda maior, haja vista o impacto mecânico do pisoteio animal sobre a desagregação do solo. Além disso, muitas áreas florestadas e sob vegetação secundária ainda são utilizadas para o pastoreio de bovinos e equinos (SAMUEL-ROSA et al., 2011). Isso causa a degradação da camada superficial do solo, sobretudo pela sua compactação, dificultando a infiltração da água dos eventos de precipitação, o que resulta na perda da serapilheira (SCHENEIDER et al., 1978). Além disso, a própria regeneração natural é prejudicada, uma vez que plantas em estádio inicial de desenvolvimento e raízes superficiais são destruídas (SCHENEIDER et al., 1978; HACK et al., 2005). O resultado desse processo é a redução da capacidade do solo em suportar o desenvolvimento potencial da floresta (KÖNIG et al., 2002).

Dado que a bacia do DNOS também possui áreas com relevo plano a suave-ondulado e solo profundo, ±30 % de sua superfície ainda é utilizada com atividades agrossilvopastoris ao longo de todo o ano, principalmente a pecuária bovina extensiva (SAMUEL-ROSA et al., 2011). As áreas de produção animal extensiva predominam nas porções norte, sul e centro-oeste, ao longo dos cursos de água e estradas. O solo é mais profundo e menos pedregoso do que em áreas de floresta natural e vegetação secundária. É comum encontrar fragmentos de carvão e formações de microrrelevo devido ao uso de implementos agrícolas para o revolvimento do solo na maior parte desses locais, evidenciando seu uso agrícola num passado recente. Essas áreas podem ser divididas entre aquelas de campo sujo (pastagens naturais mal manejadas, com predomínio de vegetação de porte herbáceo-arbustivo) e campo limpo (pastagens naturais e perenes bem manejadas) (SAMUEL-ROSA et al., 2011). Em geral, as áreas de campo sujo estão próximas a áreas de vegetação secundária (capoeira), com relevo mais declivoso e solo mais pedregoso e raso do que aquelas de campo limpo. Assim como nas áreas de floresta e sob vegetação secundária, a dinâmica de uso dessas áreas ao longo do tempo é bastante complexa, dificultando o estabelecimento de relações diretas com a maior parte das características do solo. Entretanto, sob o ponto de vista da reserva de nutrientes e matéria orgânica, o solo pode ser considerado pobre, uma vez que a exploração pecuária é totalmente extrativista e extensiva.

As atividades agrossilvopastoris que ocupam menor extensão territorial são a agricultura e a silvicultura. As áreas de lavoura anuais e bianuais estão dispersas em toda a área, geralmente localizadas em terrenos de menor declividade e solo medianamente profundo. Entretanto, algumas áreas de produção agrícola possuem declives superiores a 50 % e solo raso e pedregoso (SAMUEL-ROSA et al., 2011). Em qualquer das situações, as condições de degradação do solo costumam ser bastante avançadas devido ao emprego de sistemas convencionais de cultivo, exceto por algumas áreas de produção olerícola. As florestas plantadas (*Eucalyptus spp.*) são implantadas em áreas com menor declividade e solo mais profundo, geralmente onde o acesso com máquinas é melhor, sobretudo pela necessidade de manejo e escoamento da produção. A área ocupada por essa atividade possui tendência de crescimento, haja vista os novos plantios existentes e o relato de alguns moradores (SAMUEL-ROSA et al., 2011). Em geral, os novos plantios são implantados em áreas de produção agropecuária, seja pelo elevado nível de degradação do solo já atingido, seja pela redução da força de trabalho das famílias devido ao êxodo rural, ou pela maior lucratividade dessa atividade.

Por fim, as obras de engenharia e assentamentos urbanos são aquelas que ocupam a menor parte da bacia do DNOS. No que diz respeito à malha viária, a maior concentração ocorre na porção sul, junto ao maior assentamento urbano, localizado no entorno do reservatório. Entretanto, diversas construções são encontradas ao longo das estradas que cortam a área, muitas das quais são pertencentes a moradores do centro da cidade de Santa Maria e são utilizadas apenas como sítios de final de semana. O número de sítios de final de semana aumentou significativamente nas últimas décadas (GOLDANI, 2006), muitos dos quais construídos em locais inapropriados, como margens dos corpos de água e áreas com forte declividade. Esse processo de urbanização desordenada, que exigiu a realização de obras de corte e aterramento dos terrenos, é um importante contribuinte da carga de sedimentos recebida pelo reservatório anualmente, aos quais somam-se os resíduos domésticos e cloacais (GOLDANI, 2006; PAIVA et al., 2001; DILL et al., 2004; MIGUEL et al., 2014). Quanto às demais obras de engenharia, destacam-se os reservatórios de água, a maioria deles de pequena extensão, utilizados para a dessedentação animal. Como os cursos de água de maior volume estão localizados na porção sul da área, a maior parte dos reservatórios de água está na porção norte (SAMUEL-ROSA et al., 2011). Além disso, a menor permeabilidade do solo e do substrato rochoso, bem como da condição topográfica, favorecem essa característica.

5.9 PEDOLOGIA

O solo da bacia do DNOS possui características com forte dependência do material de origem que, conforme descrito anteriormente, é um importante condicionante da geomorfologia e da hidrografia (NASCIMENTO; PENNA E SOUZA, 2010). Nas superfícies geomórficas mais estáveis, como no topo do Planalto (rochas vulcânicas), nos terraços do Rebordo (rochas vulcânicas e sedimentares) e nas coxilhas de relevo suave-ondulado a ondulado (rochas sedimentares), as condições ambientais são mais favoráveis ao desenvolvimento do solo em profundidade (MOSER, 1990). O contrário ocorre nas áreas de relevo mais acidentado do Rebordo, onde acredita-se que a taxa de formação do solo seja semelhante à taxa de remoção natural (MOSER, 1990; DALMOLIN et al., 2006; STÜRMER, 2008; SAMUEL-ROSA et al., 2011). Além da condição geomorfológica, a resistência da rocha ao intemperismo também condiciona o desenvolvimento do solo em profundidade nesses locais (PEDRON, 2007). Já nas planícies aluviais, as características do solo foram fortemente influenciadas pelo hidromorfismo e deposição sedimentar, geralmente resultando em solo de coloração acinzentada e maior profundidade do que nas área declivosas Rebordo (MOSER, 1990; MIGUEL, 2010). A forte influência da geologia também aparece na textura do solo (Figura 5.3). Enquanto os arenitos conferem textura arenosa ao solo, sobretudo aqueles da Formação Botucatu, as rochas vulcânicas conferem textura média ao solo, sobretudo os basaltos-andesitos toleíticos (Seção 4.8).

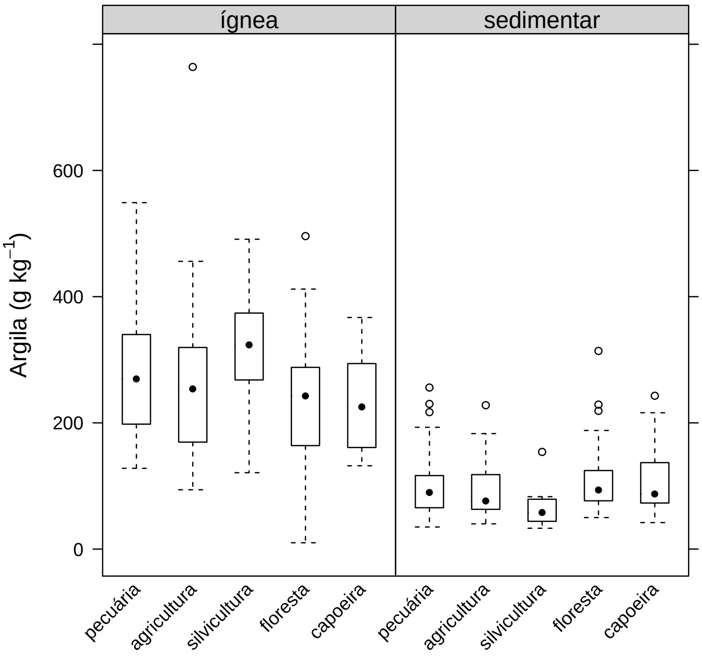


Figura 5.3: Distribuição do conteúdo de argila (g kg^{-1}) na camada superficial do solo ($\leq$ 20 cm) e sua relação com o uso da terra e vegetação (pecuária, agricultura, silvicultura, floresta e capoeira) e o tipo de material de origem (ígnea – rocha ígnea, ou sedimentar – rocha sedimentar ou sedimentos diversos). Solo desenvolvido a partir de material de origem sedimentar costuma apresentas conteúdo de argila inferior à solo desenvolvido a partir de material de origem ígnea. Em geral, essa relação é pouco influenciada pelo tipo de uso da terra. A exceção são as áreas destinadas à silvicultura localizadas no setor norte da bacia do DNOS, cujo solo desenvolveu a partir de material de origem ígnea. Ali a tendência é de o solo apresentar maior conteúdo de argila na camada superficial.

Como a bacia do DNOS possui a maior parte de sua área em condições de declividade moderada à forte, e foi intensamente ocupada em tempos pretéritos para produção agrossilvopastoril com aração e queimada periódicas (Seção 5.8), o solo é, predominantemente, pouco profundo. Nas áreas de declividade moderada à forte o solo costuma apresentar profundidade inferior à 50 cm até o contato lítico, sendo comum a ocorrência de pedregosidade e rochosidade abundantes (MIGUEL, 2010). Assim, predominam as classes taxonômicas Neossolo Litólico Distro-Úmbrico típico, Cambissolo Háplico Ta Eutrófico típico, Neossolo Litólico Eutro-Úmbrico típico e Neossolo Regolítico Distro-Úmbrico típico (Seção 4.6). O efeito do uso inapropriado do solo para produção agrossilvopastoril aparece mesmo em algumas áreas de maior estabilidade (topos de morros, patamares do Rebordo do Planalto e coxilhas) onde as condições para o desenvolvimento pedogenético são mais favoráveis (MOSER, 1990; MOURA-BUENO, 2012). Por exemplo, alguns patamares do Rebordo, inicialmente constituídos de colúvios sedimentares (arenito Botucatu) e vulcânicos (fragmentos de tamanhos variáveis), apresentam solo com superfície recoberta por fragmentos rochosos, fruto da forte erosão a que foi submetido, limitando a continuação de seu uso para atividades agrossilvopastoris (MOURA-BUENO, 2012). Essas atividades também resultaram na depleção da fertilidade do solo, marcada atualmente pelos baixos conteúdo de carbono orgânico (Figura 5.4) e capacidade de troca de cátions efetiva

(Figura 5.5).

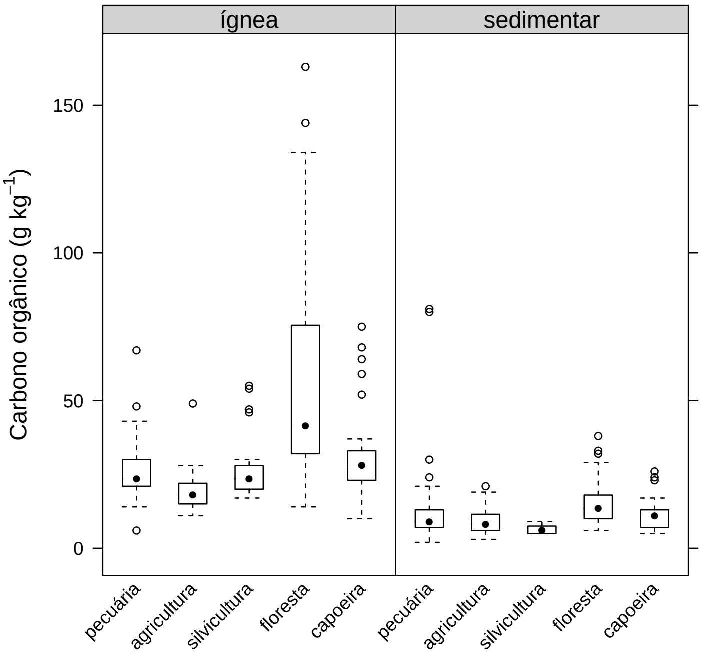


Figura 5.4: Distribuição do conteúdo de carbono orgânico (g kg^{-1}) na camada superficial do solo ($\leq$ 20 cm) e sua relação com o uso da terra e vegetação (pecuária, agricultura, silvicultura, floresta e capoeira) e o tipo de material de origem (ígnea – rocha ígnea, ou sedimentar – rocha sedimentar ou sedimentos diversos). O conteúdo de carbono é mais elevado em solo florestal, independente do tipo de rocha que deu origem ao solo. Contudo, o conteúdo de carbono é notadamente maior quando o solo tem rocha ígnea como material originário, sugerindo uma íntima relação com o conteúdo de argila do solo (Figura 5.3) e de cátions básicos liberados durante o intemperismo do material de origem, refletido na capacidade de troca de cátions efetiva do solo (Figura 5.5). A grande variação no conteúdo de carbono em solo florestal reflete a ocorrência de vários estágios de sucessão florestal.

Existe consenso de que a pequena profundidade do solo na maior parte da bacia do DNOS seja devida ao material de origem, às condições geomorfológicas e hidrológicas, bem como aos sistemas de cultivo do solo empregados ao longo de inúmeras décadas. Contudo, a diversidade de fatores e a escassez de estudos torna impossível isolar a contribuição individual de cada fator. Por exemplo, não existe estimativa quantitativa do volume total de solo perdido devido à erosão em áreas de produção agrossilvopastoril. Entretanto, acredita-se que horizontes pedogenéticos inteiros tenham sido removidos, dando origem ao processo de retrocesso pedogenético do solo (SAMUEL-ROSA et al., 2011). Esse processo reflete a involução da classificação taxonômica do solo dentro de um determinado sistema taxonômico. A comum ocorrência do táxon Neossolo Litólico com solum de pouco menos de 50 cm em inúmeras áreas de produção agrícola é usada para corroborar essa hipótese. Isso porque o valor de 50 cm para o solum é aquele usado para distinguir os taxa Neossolo Litólico e Neossolo Regolítico no Sistema Brasileiro de Classificação do Solo (SANTOS et al., 2013). A mesma hipótese foi levantada em outras regiões de topografia complexa para explicar a identificação dos taxa Cambissolo e Luvissolo onde antes observava-se o táxon Chernossolo (STRECK et al., 2008). Essa involução da classificação taxonômica seria resultado da remoção do horizonte pedogenético A chernozêmico devido à erosão acelerada pelo uso agrícola extrativista e intenso praticado por várias décadas. Em resumo, nas áreas onde a taxa de formação do solo é semelhante à taxa de erosão natural, o uso agrossilvopastoril da terra sem adoção de práticas conservacionistas invariavelmente resultaria na degradação do solo e consequente retrocesso pedogenético porque a soma das taxas de erosão natural e erosão induzida seria maior do que a taxa de formação do solo.

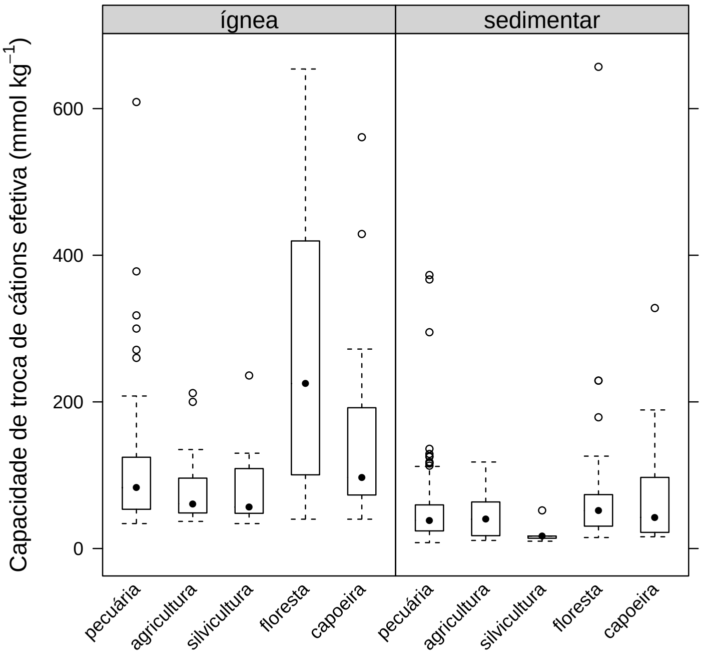


Figura 5.5: Distribuição da capacidade de troca de cátions efetiva (mmol kg^{-1}) na camada superficial do solo ($\leq$ 20 cm) e sua relação com o uso da terra e vegetação (pecuária, agricultura, silvicultura, floresta e capoeira) e o tipo de material de origem (ígnea – rocha ígnea, ou sedimentar – rocha sedimentar ou sedimentos diversos). Assim como o conteúdo de carbono orgânico do solo, a capacidade de troca de cátions é maior em solo florestal. Contudo, áreas em estádio recente de regeneração (capoeira) e pecuária também se destacam pela elevada capacidade de troca de cátions. O contrário ocorre em tipos de uso da terra onde ocorre significativa exportação dos cátions do solo, caso da agricultura e silvicultura.

As áreas com maior potencial de desenvolvimento pedogenético em profundidade possuem menor expressão territorial. Elas correspondem à paisagem menos declivosa do Planalto (rochas vulcânicas), algumas coxilhas (rochas sedimentares) e depósitos aluviais (MIGUEL, 2010). Em áreas de material de origem vulcânica, sobretudo basaltos-andesitos toleíticos, identifica-se o táxon Argissolo Vermelho Alítico típico, que caracteriza solo com horizonte superficial de textura arenosa a média, sobrejacente a um horizonte subsuperficial de textura média a argilosa (MIGUEL, 2010). Nas áreas do Planalto em que ocorrem riólitos-riodacitos granofíricos, o solo costuma apresentar menor profundidade, sendo predominantemente classificado como Neossolo Litólico e Cambissolo Háplico. Em pequenas manchas, o solo é classificado como Argissolo Vermelho-Amarelo. O menor desenvolvimento do solo em profundidade em área de riólitos-riodacitos granofíricos deve-se, em parte, à maior resistência dessas rochas ao intemperismo, uma vez que possui maior teor de sílica, sobretudo na forma de grandes de cristais de quartzo cristalizados a baixa temperatura ($<$ 600 °C) (PEDRON, 2007). Contudo, a pequena profundidade do solo também resulta da degradação causada pelo longo uso agrossilvopastoril sem adoção de práticas conservacionistas. Até o momento é impossível isolar a contribuição desses dois fatores de formação sobre as características do solo.

Em áreas de material de origem sedimentar (Formação Caturrita), o solo costuma ser classificado como Argissolo Bruno-Acinzentado Alítico abrúptico. Trata-se de solo com horizonte superficial de textura arenosa que transiciona de maneira abrupta para um horizonte subsuperficial de textura argilosa (MIGUEL, 2010). Essa descontinuidade textural geralmente é atribuída a processos pedogenéticos, sobretudo a perda (erosão lateral seletiva) e translocação (argiluviação) das partículas mais finas. Contudo, devido à complexidade geológica e geomorfológica da área, outros fatores podem ter contribuído. O primeiro deles estaria relacionado às características do próprio material de origem que, por ter sido formado em ambiente fluvial durante um período de mudança climática no Triássico, apresenta camadas deposicionais com granulometria diferenciada (PIERINI et al., 2002). Assim, a presença de camadas de siltitos e folhelhos poderia ter contribuído para a formação da descontinuidade textural. Outro fator seria uma possível contribuição dos arenitos da Formação Botucatu e *intertrap* na Formação Serra Geral. Dado que esse material está localizado em posições superiores na paisagem e são bastante suscetíveis à erosão, o mesmo poderia ter contribuído para a formação do horizonte superficial arenoso do solo. Em ambos os casos, a descontinuidade textural seria atribuída à ocorrência de descontinuidade litológica, com a diferença de que no primeiro caso as litologias pertenceriam à mesma formação geológica.

Nas áreas deprimidas da paisagem do Planalto, formando pequenas bacias de acumulação, ou nas áreas planas ao longo dos cursos de água da Depressão Periférica, o solo é classificado como Planossolo Háplico Alítico típico (MIGUEL, 2010). Trata-se de solo com horizonte superficial de textura média sobre horizonte subsuperficial de textura média a argilosa, podendo ou não apresentar horizonte intermediário eluvial. Ainda mais próximo dos cursos de água, o solo é classificado como Neossolo Flúvico Tb Eutrófico fragmentário (MIGUEL, 2010). Como o seu material de origem é diverso e possui arranjamento espacial discordante, a textura do solo é variável, mas sempre arenosa ou média, nunca argilosa, mesmo quando presente em áreas do Planalto ou do Rebordo. Por fim, com expressão territorial ainda menor, o solo é classificado como Neossolo Quartzarênico Órtico típico em alguns patamares do Rebordo, sejam eles de origem estrutural ou da deposição de colúvios do arenito da Formação Botucatu.

Atualmente, o aumento da área ocupada por florestas devido ao abandono de diversas área de produção sugere que a bacia do DNOS esteja adentrando um período de redução da degradação do solo, exceto nas áreas de expansão urbana. Os processos erosivos já apresentam caráter mais pontual, onde a geomorfologia parece possuir papel preponderante. Segundo estimativas, apenas uma pequena fração da área apresenta perdas de solo por erosão laminar acima de valores toleráveis (MIGUEL, 2010). Algumas observações até mesmo indicam que essas estimativas estão acima da perda real de solo por erosão laminar (MOURA-BUENO, 2012). Espera-se que a redução da degradação do solo contribua, num futuro próximo, para a construção de um entendimento mais acurado da relação entre o solo e os demais componentes ambientais, principalmente nas áreas florestadas recentemente abandonadas.

6 CHAPTER V

DO MORE DETAILED COVARIATES DELIVER MORE ACCURATE SOIL MAPS?

* This chapter is based on A. Samuel-Rosa, G.B.M. Heuvelink, G.M. Vasques, L.H.C. Anjos. Do more detailed environmental covariates deliver more accurate soil maps? *Geoderma*, v.243–244, p.214–227, 2015. Terms and expressions have been modified to match the standard terminology used throughout the thesis without compromising the content of the original text. Footnotes were added where a definition required correction or clarification.

6.1 RESUMO

Neste estudo nós avaliamos se investir em covariáveis espacialmente mais detalhadas aumenta a acurácia dos mapas do solo. Nós usamos um estudo de caso no sul do Brasil para mapear o conteúdo de argila (CLAY), o conteúdo de carbono orgânico (SOC), e capacidade de troca de cátions efetiva (ECEC) da camada superficial do solo de uma área de ~ 2000 ha localizada na borda do planalto da Bacia Sedimentar do Paraná. Cinco covariáveis, cada uma com dois níveis de detalhe espacial, foram usadas: mapa areal-categórico de solo, modelos digitais de elevação (DEM), mapas geológicos, mapas de uso da terra, e imagens de satélite. Trinta e dois modelos de regressão linear múltipla foram calibrados para cada propriedade do solo usando todas as combinações de detalhe espacial das covariáveis. Para cada combinação, *stepwise regression* foi usada para selecionar as variáveis preditoras incorporadas no modelo. A avaliação dos modelos foi feita usando o R-quadrado ajustado da regressão. O modelo de referência, calibrado com a versão menos detalhada de cada covariável, e o modelo com o melhor desempenho, foram usados para calibrar dois modelos lineares mistos para cada propriedade do solo. Parâmetros dos modelos foram estimados usando máxima verossimilhança restrita. Predições espaciais foram realizadas usando o melhor preditor linear não-enviesado empírico. Validação-cruzada foi usada para validar os modelos de regressão linear múltipla e dos modelos lineares mistos de referência e com melhor desempenho. Os resultados mostram que para CLAY a acurácia da predição não aumentou consideravelmente por usar covariáveis mais detalhadas. A quantidade de variância explicada aumentou apenas ~ 2 pp (pontos percentuais), menos do que obtido pela inclusão do passo de krigagem, que explicou 4 pp. Por outro lado, a predição de SOC e ECEC aumentou em ~ 13 pp quando o modelo de referência foi substituído pelo modelo com melhor desempenho. Em geral, o aumento no desempenho preditivo foi modesto e pode não sobrepor os custos adicionais do uso de covariáveis mais detalhadas. Pode ser mais eficiente investir recursos adicionais na coleta de mais observações do solo, ou no aumento do detalhe apenas da covariável que tem o efeito de aumento mais forte. Em nosso estudo, a última funcionaria apenas para SOC e ECEC pelo investimento em um mapa de uso da terra mais detalhado e, possivelmente, também em um mapa geológico e DEM mais detalhados.

Palavras-chave: Mapeamento Digital do Solo. Modelo Linear Misto. Informação Auxiliar. Seleção de Variáveis. Acurácia do Modelo. Custo do Mapeamento do Solo.

6.2 ABSTRACT

In this study we evaluated whether investing in more spatially detailed covariates improves the accuracy of soil maps. We used a case study from Southern Brazil to map clay content (CLAY), organic carbon content (SOC), and effective cation exchange capacity (ECEC) of the topsoil for a ≈ 2000 ha area located on the edge of the plateau of the Paraná Sedimentary Basin. Five covariates, each with two levels of spatial detail were used: area-class soil maps, digital elevation models (DEM), geologic maps, land use maps, and satellite images. Thirty-two multiple linear regression models were calibrated for each soil property using all spatial detail combinations of the covariates. For each combination, stepwise regression was used to select predictor variables incorporated in the model. Model evaluation was done using the adjusted R-square of the regression. The baseline model, calibrated with the less detailed version of each covariate, and the best performing model were used to calibrate two linear mixed models for each soil property. Model parameters were estimated using restricted maximum likelihood. Spatial prediction was performed using the empirical best linear unbiased predictor. Validation of baseline and best performing linear multiple regression and linear mixed models was done using cross-validation. Results show that for CLAY the prediction accuracy did not considerably improve by using more detailed covariates. The amount of variance explained increased only ~ 2 percentage points (pp), less than that obtained by including the kriging step, which explained 4 pp. On the other hand, prediction of SOC and ECEC improved by ~ 13 pp when the baseline model was replaced by the best performing model. Overall, the increase in prediction performance was modest and may not outweigh the extra costs of using more detailed covariates. It may be more efficient to spend extra resources on collecting more soil observations, or increasing the detail of only those covariates that have the strongest improvement effect. In our case study, the latter would only work for SOC and ECEC, by investing in a more detailed land use map and possibly also a more detailed geologic map and DEM.

Keywords: Digital Soil Mapping. Linear Mixed Model. Auxiliary Information. Variable Selection. Model Accuracy. Soil Mapping Cost.

6.3 INTRODUCTION

Modern soil mapping relies on the use of statistical models to produce digital representations of spatial soil distribution using point soil observations and spatially exhaustive covariates (MCBRATNEY et al., 2003; SCULL et al., 2003; FLORINSKY, 2012). Three important weaknesses in the statistical soil distribution modelling approach can be pointed out. First, it requires sufficient and appropriately distributed point soil data within the area being mapped (CARRÉ et al., 2007). Second, the model structure explores only the empirical relationship among environmental conditions and soil properties, being less comprehensive than soil-landscape process models (GRUNWALD, 2009). Last, the covariates are only approximations of the true environmental conditions that helped shape the soil. They serve only as proxies (surrogates) of the current environmental conditions, which in many cases are different from the past conditions under which pedogenesis took place (HEUVELINK; WEBSTER, 2001). In spite of these weaknesses, modern soil mapping techniques have proven very successful in the past decades in producing soil property maps that capture the main patterns of soil spatial variation (MOORE et al., 1993; MCBRATNEY et al., 2000; GRUNWALD, 2009).

More recently, there has been a growing interest in understanding how the characteristics of the covariates influence the success of soil mapping – this study contributes to this effort. It is commonly accepted that the more resources are spent on the construction of a covariate and the more spatial information it has, the more accurately it describes the environmental conditions (HUPY et al., 2004; HENGL et al., 2013). It is also generally believed that such *more detailed* covariates will be more valuable for soil mapping and lead to more accurate soil property predictions (CAVAZZI et al., 2013; MAYNARD; JOHNSON, 2014). If these more detailed covariates convey more information and represent more adequately the environmental conditions – the drivers of soil forming processes –, then it is fair to expect that they improve the accuracy of the resulting soil maps. However, some studies have shown the contrary (THOMPSON et al., 2001; ELDEIRY; GARCIA, 2008; KIM et al., 2014). For example, the window size at which DEM derivatives are calculated can be more important than the spatial resolution of the DEM (WOOD, 1996; ZHU et al., 2008; BEHRENS et al., 2010). The uncertainty about the added value of using more detailed covariates is of concern for those seeking to use resources efficiently, because using more detailed covariates generally increases soil mapping costs (SHI et al., 2012).

The objective of this study was to evaluate whether investing in more detailed covariates improves the accuracy of soil maps. The main difference of our study to previous ones is that we use a rigorous statistical approach to assess the added value of using five more detailed covariates simultaneously. We used a case study in Brazil to compare the accuracy of maps of the clay content, organic carbon content and effective cation exchange capacity of the topsoil as obtained from regression kriging on the five covariates, whereby each covariate was evaluated on two levels of spatial detail.

6.4 MATERIAL AND METHODS

6.4.1 Study Area and Soil Data

The study area constitutes a small catchment (~ 2000 ha) located on the southern edge of the plateau of the Paraná Sedimentary Basin, Rio Grande do Sul, Brazil (Figure 6.1). The climate is classified as Cfa (Köppen – subtropical humid without a dry season) with mean annual temperature of 19.3 °C, and mean annual precipitation of 1708 mm, well distributed throughout the year (MALUF, 2000). Relief varies between plain (slope between 0 and 3 %) and mountainous (slope between 45 and 100 %), and elevations range between 140 and 475 m. Geology consists of basic, intermediate and acid igneous rocks (rhyolite-rhyodacite and andesite-basalt) of the Cretaceous period, consolidated sedimentary rocks (aeolian and fluvial sandstones) of the Triassic and Jurassic periods, and non-consolidated (fluvial and colluvial deposits) of the Quaternary period (GASPARETTO et al., 1988; MACIEL FILHO, 1990; SARTORI, 2009). Native semi-deciduous forests occupy more than half of the area, followed by native grassland used for animal husbandry, semi-deciduous shrubland, annual crop agriculture, forestry (Eucalyptus), urban areas, and artificial water bodies (SAMUEL-ROSA et al., 2011).

A dataset containing $n = 350$ point soil observations collected between 2004 and 2011 (PEDRON et al., 2006; SAMUEL-ROSA et al., 2011; MIGUEL et al., 2011; SAMUEL-ROSA et al., 2013) was used in this study [21]. Sampling locations were selected purposively and by convenience (SAMUEL-ROSA et al., 2014). Three soil pits were opened within an area of ± 100 m at most sampling locations to obtain composite samples of the topsoil for laboratory analysis. Soil was collected to a depth of 20 cm or less when soil depth was smaller than 20 cm. A few observations ($n = 10$) correspond to individual samples collected up to 30 cm. Sampling depth ranges from 2 to 30 cm, with a mean of 17.3 cm. We assumed that the vertical, horizontal and temporal support differences between soil samples is negligible for the purpose of this study.

Three soil properties (fine earth fraction, < 2 mm) were explored: clay content (CLAY, g kg^{-1}), organic carbon content (SOC, g kg^{-1}), and effective cation exchange capacity (ECEC, mmol kg^{-1}). CLAY was determined by the pipette method. SOC was determined using wet digestion. ECEC was calculated as the sum of exchangeable bases plus exchangeable acidity. The soil properties selected were expected to present different patterns of spatial variation and correlation with the most dominant factors of soil formation (JENNY, 1941) in the area: organisms (*O*), relief (*R*), and parent material (*P*). CLAY was presumed to have a stronger relation with *P*, while SOC was expected to be more correlated with *O*. Because the soils of the study area were strongly eroded due to intense agriculture in the 20th century, both CLAY and SOC were also expected to be closely related with *R*. Finally, ECEC was expected to be strongly correlated with *P* and *O*, which is supported by its natural relationship with both CLAY and SOC.

Point soil data, here denoted by $Y(s)$, showed a positive skew (Figure 6.2) and was normalized, $Y'(s)$, using the Box-Cox family of power transformations, where $Y'(s) = (Y(s)^{\lambda} - 1)/\lambda$, if $\lambda > 0$, and $Y'(s) = log(Y(s))$, if $\lambda = 0$ (DIGGLE; RIBEIRO JR, 2007). Lambda (λ) values were selected empirically (FOX; WEISBERG, 2011). Because the resulting distribution of the back-transform (see Section 6.4.3.2) has no expectation when $\lambda < 0$ (RIBEIRO JR; DIGGLE, 2001), a logarithm transformation ($\lambda = 0$) was used when a negative λ was estimated (SOC and ECEC).

6.4.2 Covariates

Five freely available covariates were evaluated in this study, each with two levels of spatial detail: area-class soil maps (`soil`), geologic maps (`geo`), land use maps (`land`), digital elevation models (`dem`), and satellite images (`sat`). Each pair was composed of covariates that were produced separately from scratch using different data sources and/or production methods, thus demanding different amounts of resources (time, workforce, budget, technology, etc.).

[21] The data is available at <https://github.com/samuel-rosa/dnos-sm-rs-general>.

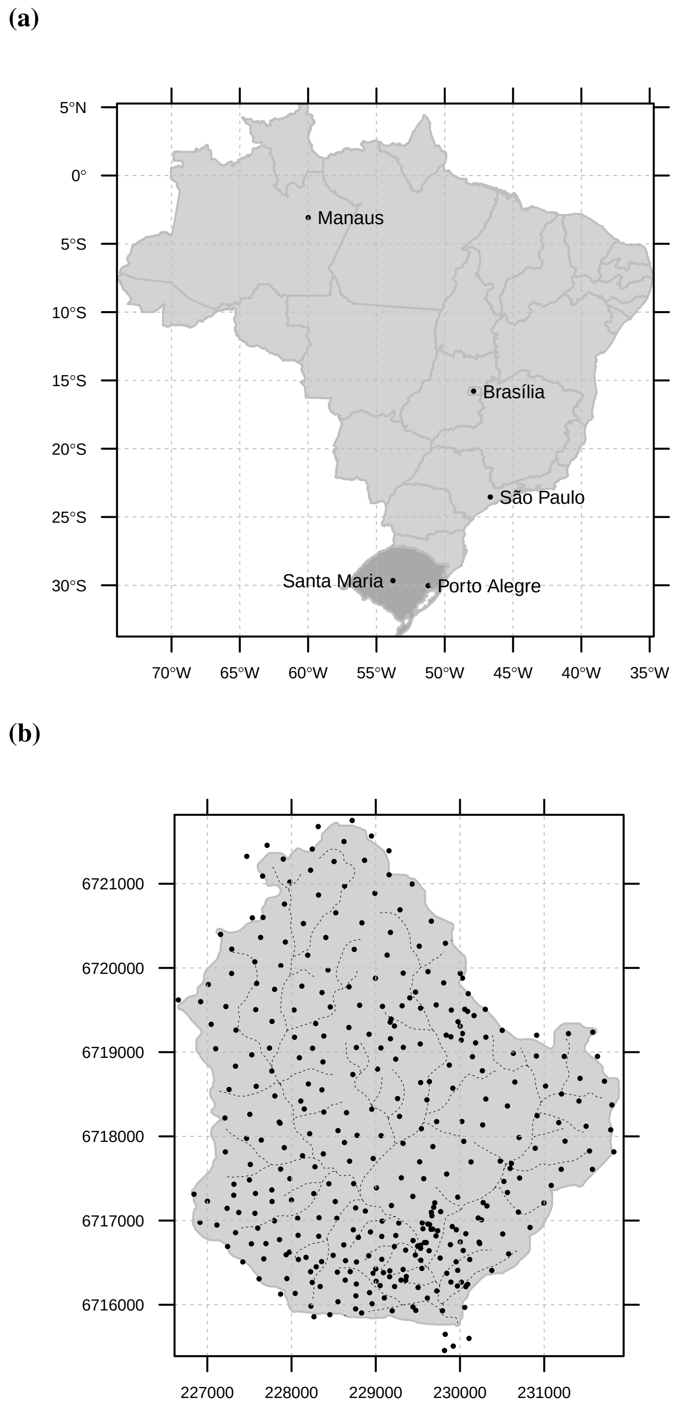


Figure 6.1: Location of the study area in Santa Maria (a) and spatial distribution of the point soil observations and drainage network (b).

In this study, the level of spatial detail of a covariate is a function of the components of its production process such as the cartographic ratio (`soil`, `geo` and `land`), spatial sampling support (`sat`), number and diversity of data sources explored (`dem`), and quantity of spatial data used (all five). Thus, the reader should bear in mind that our definition of spatial detail is broader than spatial resolution or spatial scale. It should also not be confounded with spatial support (WEBSTER; OLIVER, 2007) or thematic detail (ROSSITER, 2000).

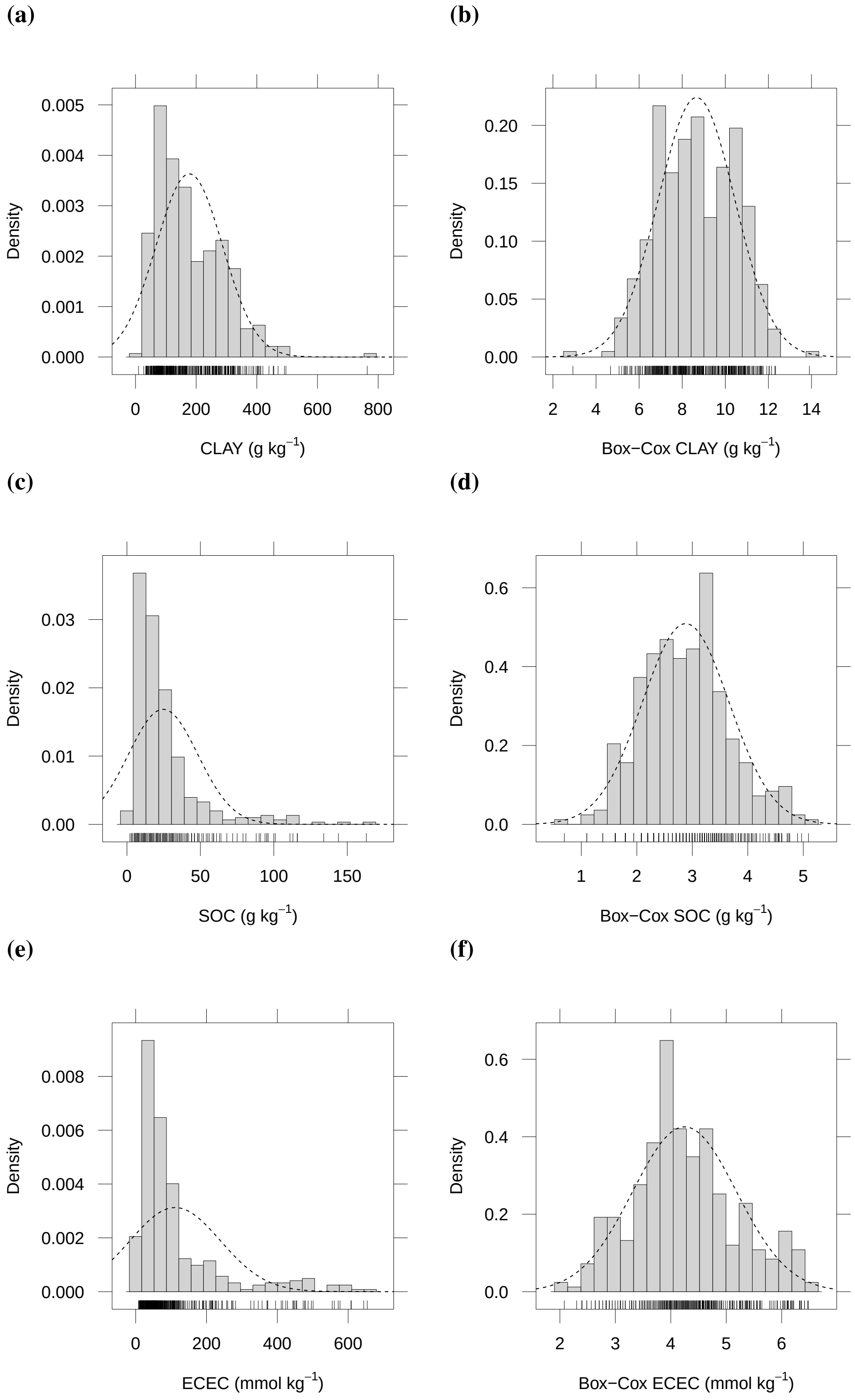


Figure 6.2: Histogram, empirical density function, and summary statistics of CLAY (a, b), SOC (c, d), and ECEC (e, f) in the original (left) and Box-Cox feature spaces (right).

The covariates were transformed to predictor variables[22] that were used in the geostatistical modelling. Since the transformation is different for categorical and continuous covariates,

[22] In statistical terms, the terms *covariate* and *predictor variable* are synonymous, and the reason for the use given in this study is purely operational.

the procedures are explained below for each type separately.

6.4.2.1 Categorical predictor variables

Area-class soil maps, geologic maps and land use maps are categorical covariates (factors). Mapping units are the k factor levels that are transformed to as many dummy (indicator, binary) variables as there are factor levels, before model calibration. Each dummy variable receives a value equal to one (1) when a given class is present, and zero (0) otherwise (EVERITT, 2006). If the number of point soil observations falling inside the spatial domain of a mapping unit is too small to accurately estimate a regression coefficient (we used a threshold of $n = 15$ observations), the mapping unit is merged with a similar mapping unit prior to calculating dummy variables. The resulting generalized categorical covariate maps are shown in Figure 6.3. The binary maps are the categorical predictor variables.

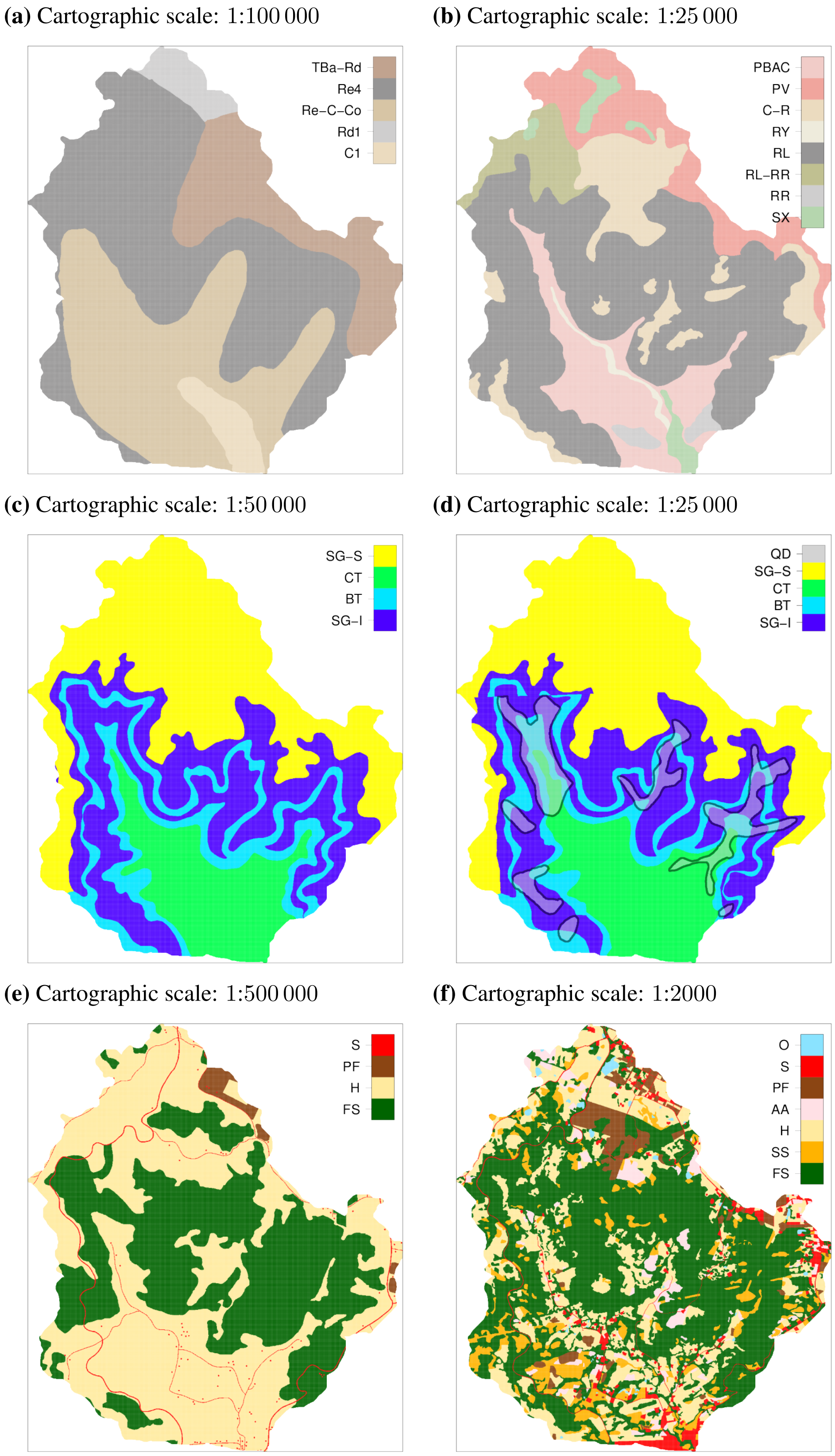
Soil maps. The less detailed soil map (`SOIL_100`) was published with a cartographic scale of 1:100 000 and has five mapping units (AZOLIN; MUTTI, 1988) (Figure 6.3a). It was produced using existing soil maps and technical reports (cartographic scale of 1:750 000) (BRASIL, 1973), aerial photographs (cartographic scale of 1:60 000), topographic maps (cartographic scale of 1:50 000), and sparse point soil observations along the road network. The more detailed soil map (`SOIL_25`) was prepared with a cartographic scale of 1:25 000 and has eight mapping units (MIGUEL et al., 2011) (Figure 6.3b). It was produced using high spatial resolution satellite images (65 cm), existing soil maps and technical reports published with a cartographic scale of 1:50 000 (POELKING, 2007) and 1:25 000 (PEDRON et al., 2006), topographic maps (cartographic scale of 1:25 000), and descriptions from ~ 350 point soil observations. Five dummy predictor variables were derived from `SOIL_100` and seven from `SOIL_25` (Table 6.1).

Geologic maps. The less detailed geologic map (`GEO_50`) was produced using topographic maps with cartographic scale of 1:50 000 (GASPARETTO et al., 1988) (Figure 6.3c). The more detailed geologic map (`GEO_25`) was produced using topographic maps with cartographic scale of 1:25 000, and includes the location of overlaying Quaternary sedimentary deposits (MACIEL FILHO, 1990) (Figure 6.3d). `GEO_25` did not cover a small part in the North of the study area, where `GEO_50` was used instead (this strategy was approved by experts on the local geology). The mapping unit of both geologic maps depicting the Caturrita Formation was used indirectly by deriving dummy predictor variables from all other individual mapping units. Three dummy predictor variables were derived from `GEO_50` and four from `GEO_25` (Table 6.2).

Land use maps. The less detailed land use map (`LU1980`) was produced by manually digitizing land use data included in topographic maps with a cartographic scale of 1:25 000 (DSG, 1980; DSG, 1992a; DSG, 1992b) (Figure 6.3e). The more detailed land use map (`LU2009`) was prepared (cartographic scale of 1:2000) by manual digitization using 65 cm spatial resolution satellite images covering the years 2008 and 2009 (SAMUEL-ROSA et al., 2011) (Figure 6.3f). Mapping units depicting human settlements and water bodies ($n = 0$) were not masked out from the prediction grid and were merged with other mapping units to derive dummy predictor variables. Five dummy predictor variables were derived from `LU2009` and two from `LU1980` (Table 6.3).

6.4.2.2 Continuous predictor variables

The less detailed DEM (`ELEV_90`) is the hole-filled SRTM DEM version 4 (JARVIS et al., 2008) (Figure 6.4a). The spatial sampling support of the SRTM DEM is $1''$ (~ 30 m),



(a) Cartographic scale: 1:100 000

(b) Cartographic scale: 1:25 000

(c) Cartographic scale: 1:50 000

(d) Cartographic scale: 1:25 000

(e) Cartographic scale: 1:500 000

(f) Cartographic scale: 1:2000

Figure 6.3: Area-class soil maps (a, b), geologic maps (c, d), and land use maps (e, f) compared in our study. The less and more detailed version are displayed at the left and right, respectively. Legend abbreviations are described in Tables 6.1–6.3.

but elevation data were aggregated to $3''$ (~ 90 m) for public release in regions outside the United States (REUTER et al., 2007). The more detailed DEM (ELEV_10) was produced by interpolating contour lines with vertical spacing of 10 m along with data about the drainage network, lakes and peaks digitized from topographic maps with cartographic scale of 1:25 000 (Figure 6.4c). Interpolation to 5 m pixel size was performed using a hydrologically correct algorithm implemented in ArcGIS software by ESRI (HUTCHINSON, 1989). Contour line artefacts were minimized using a seven by seven low-pass filter (GRASS module r.neighbors). The window size was chosen such that the smoothed DEM best matched the original contour map while also respecting the original drainage network pattern.

Table 6.1: Description of the $p = 12$ dummy predictor variables derived from the two soil maps.

Code	Mapping unit(s) included and Description[a,b]
Source: Azolin & Mutti (1988). Cartographic scale: 1:100 000. Minimum Legible Delineation: 40 ha.	
SOIL_100b	*Re4*. Shallow soil with low to high base saturation covering mountainous terrain (Solo Litólico Eutrófico/Distrófico relevo montanhoso; Neossolo Litólico Distrófico/Eutrófico; Distric/Eutric Leptosol).
SOIL_100c	*Re-C-Co*. Shallow soil with high base saturation located in strongly sloping terrain (Solo Litólico Eutrófico relevo forte ondulado; Neossolo Litólico Eutrófico; Eutric Leptosol), low weathered soil (Cambissolo Eutrófico; Cambissolo Háplico Eutrófico; Eutric Cambisol), and colluvial deposits.
SOIL_100d	*TBa-Rd*. Deep, well-structured, low base saturation soil (Terra Bruna Estruturada álica; Nitossolo; Nitisol), and shallow soil (Solo Litólico; Neossolo Litólico; Leptosol).
SOIL_100e	*Rd1* and *Re4*. *Rd1* is composed mainly of shallow soil with low to high base saturation (Solo Litólico Distrófico/Eutrófico; Neossolo Litólico Distrófico/Eutrófico; Distric/Eutric Leptosol) located in slopping terrain. This dummy predictor variable is composed of shallow soil in both sloping and mountainous terrain.
SOIL_100f	*TBa-Rd* and *C1*. *C1* is composed of low weathered soil developed in lower landscape positions, close to drainage channels (Cambissolo Eutrófico; Cambissolo Eutrófico; Eutric Cambisol). This dummy predictor variable includes the best soil mapping units for crop agriculture among those identified in the soil survey.

[a] Soil classification according to the old Brazilian classification (only for Azolin & Mutti (1988)), the current Brazilian classification, according to Santos et al. (2013), and the international classification, following IUSS Working Group WRB (2007).

[b] Minimum Legible Delineation calculated following Rossiter (2000).

Eight DEM derivatives were calculated: elevation (ELEV), slope (SLP), aspect (ASP), northernness (NOR), flow accumulation (ACC), topographic wetness index (TWI), stream power index (SPI), and topographic position index (TPI). SLP and ASP were calculated using GRASS module r.param.scale with seven window sizes (sampling support, analysis scale): 3, 7, 15, 31, 63, 127, and 255. ASP was scaled to the standard 0–360° range and orientation, and was transformed to NOR using NOR $= abs(180^\circ -$ ASP$)$. TWI and SPI were calculated using SLP calculated with different window sizes, and ACC calculated using GRASS module r.watershed. TPI was calculated using SAGA library ta_morphometry with the same seven window sizes. The combination of DEM derivatives (ELEV, SLP, NOR, TWI, SPI, and TPI) and window sizes yielded $p = 36$ continuous predictor variables from each DEM.

The less detailed satellite image was acquired by the Landsat-5 Thematic Mapper on December 26, 2010 (available at Instituto Nacional de Pesquisas Espaciais - Divisão de Geração de Imagens – INPE-DGI) (Figure 6.4b). It has 8 bit radiometric resolution and ~ 30 m spatial resolution. Spectral bands were orthorectified (Geomatica OrthoEngine) and radiomet-

Table 6.1: Description of the $p = 12$ dummy predictor variables derived from the two soil maps. (continued)

Code	Mapping unit(s) included and Description[a,b]
Source: Miguel et al. (2011). Cartographic scale: 1:25 000. Minimum Legible Delineation: 2.5 ha.	
SOIL_25a	*PBAC*. Moderately deep soil derived from sedimentary rocks, with abrupt textural change and low base saturation (Argissolo Bruno-Acinzentado; Alisol).
SOIL_25b	*PV*. Deep soil derived from igneous rocks, with moderate textural gradient, and low base saturation (Argissolo Vermelho; Acrisol).
SOIL_25c	*C-R*. Low weathered soil (Cambissolo; Cambisol) and shallow soil with low to high base saturation (Neossolo Litólico/Regolítico Eutrófico/Distrófico; Eutric/Distric Leptosol/Regosol).
SOIL_25d	*RL*. Shallow soil with low to high base saturation (Neossolo Litólico Eutrófico/Distrófico; Eutric/Distric Leptosol).
SOIL_25h	*PBAC*, *PV* and *SX*. *SX* is composed of moderately deep soil derived from sedimentary rocks, with abrupt textural change, low base saturation, and which is saturated with water for long periods of the year (Planossolo Háplico; Planosol). This dummy predictor variable includes the best soil mapping units for crop agriculture among those identified in the soil survey.
SOIL_25i	*RL*, *RL-RR* and *RR*. This dummy predictor variable includes all three mapping units composed mainly of shallow soil (Neossolo Litólico and Neossolo Regolítico; Leptosol and Regosol).
SOIL_25j	*PV*, *RL*, *RL-RR* and *C-R*. This dummy predictor variable includes all four mapping units composed mainly of soil derived from igneous rocks.

[a] Soil classification according to the old Brazilian classification (only for Azolin & Mutti (1988)), the current Brazilian classification, according to Santos et al. (2013), and the international classification, following IUSS Working Group WRB (2007).
[b] Minimum Legible Delineation calculated following Rossiter (2000).

rically corrected (GRASS module `i.landsat.toar`). The more detailed satellite image comes from the RapidEye constellation (available at Ministério do Meio Ambiente – MMA) (Figure 6.4d). It was acquired on November 16, 2012, has 16 bit radiometric resolution, 6.5 m spatial resolution, and was orthorectified to 5 m spatial resolution. Both images were atmospherically (6S atmospheric model (VERMOTE et al., 1997), GRASS module `i.atcorr`) and topographically corrected (GRASS module `i.topo.corr`). Derived predictor variables are the spectral bands (except the thermal band) and vegetation indices (normalized difference vegetation index - NDVI, and soil-adjusted vegetation index - SAVI). Eight continuous predictor variables were derived from the Landsat-5 TM image and nine from the RapidEye image.

6.4.2.3 Additional processing

Soil maps, geologic maps, land use maps, and satellite images were registered with the prediction grid (5 m pixel size) using nearest neighbour resampling. `ELEV_90` was registered using cubic resampling (SAMUEL-ROSA et al., 2013). Systematic positional errors were corrected using affine transformation (SAMUEL-ROSA et al., 2014a).

Table 6.2: Description of the $p = 7$ dummy predictor variables derived from the two geologic maps.

Code	Mapping unit(s) included and Description[a]
Source: Gasparetto et al. (1988). Cartographic scale: 1:50 000. Minimum Legible Delineation: 10 ha.	
GEO_50a	*SG-I*. Inferior Sequence of the Serra Geral Formation. Composed mainly of basic igneous rocks (tholeiitic basalt and andesite). It is likely to be related with high CLAY and ECEC.
GEO_50b	*SG-S*. Superior Sequence of the Serra Geral Formation. Composed mainly of acid igneous rocks (granophyric rhyolite and rhyodacite). It is likely to be related with moderate to high CLAY and ECEC.
GEO_50c	*BT*. Botucatu Formation. Composed mainly of aeolian sandstones. It is likely to be related with low CLAY and ECEC.
Other	*CT* depicts the Caturrita Formation, which is composed mainly of fluvial sandstones.
Source: Maciel Filho (1990). Cartographic scale: 1:25 000. Minimum Legible Delineation: 2.5 ha.	
GEO_25a	*SG-I*. Inferior Sequence of the Serra Geral Formation.
GEO_25b	*SG-S*. Superior Sequence of the Serra Geral Formation.
GEO_25c	*BT*. Botucatu Formation.
GEO_25d	*QD*. Quaternary deposits of fluvial, alluvial, and colluvial origin. It can help explaining the low CLAY of soils supposedly derived from igneous rocks.
Other	*CT* depicts the Caturrita Formation.

[a] Minimum Legible Delineation calculated following Rossiter (2000).

Table 6.3: Description of the $p = 7$ dummy predictor variables derived from the two land use maps.

Code	Mapping unit(s) included and Description[a]
Source: DSG (1980), DSG (1992a), DSG (1992b). Cartographic scale: 1:25 000.Minimum Legible Delineation: 2.5 ha.	
LU1980a	*FS*. Native forest, which is likely to have soil with higher fertility.
LU1980b	*H*. Animal husbandry, which is likely to have soil fertility status lower than native forests and is the second most important land use in the area.
Other	Plantation forests (*PF*) and human settlements (*S*).
Source: Samuel-Rosa et al. (2011). Cartographic scale: 1:2000. Minimum Legible Delineation: 100 m^2.	
LU2009a	*FS*. Native forest.
LU2009b	*SS*. Shrubland, which is likely to have SOC and ECEC level above those found in areas used with annual crop agriculture and animal husbandry, but lower than in native forests.
LU2009c	*H*. Animal husbandry.
LU2009d	*AA*. Annual crop agriculture, which is likely to have the lowest soil fertility due to the usually poor management practices employed.
LUdiff	Land use difference between 1980 and 2009. It can be useful to explain, for example, low SOC in forest soil due to previous use with crop agriculture or animal husbandry.
Other	Plantation forests (*PF*), human settlements (*S*), and other land uses (*O*), comprising natural and artificial water bodies.

[a] Minimum Legible Delineation calculated following Rossiter (2000).

6.4.3 Linear Mixed Model of Spatial Variation

We model each of the soil properties of interest as the outcome of a spatial stochastic process. The model is composed of fixed and random effects (HEUVELINK; WEBSTER,

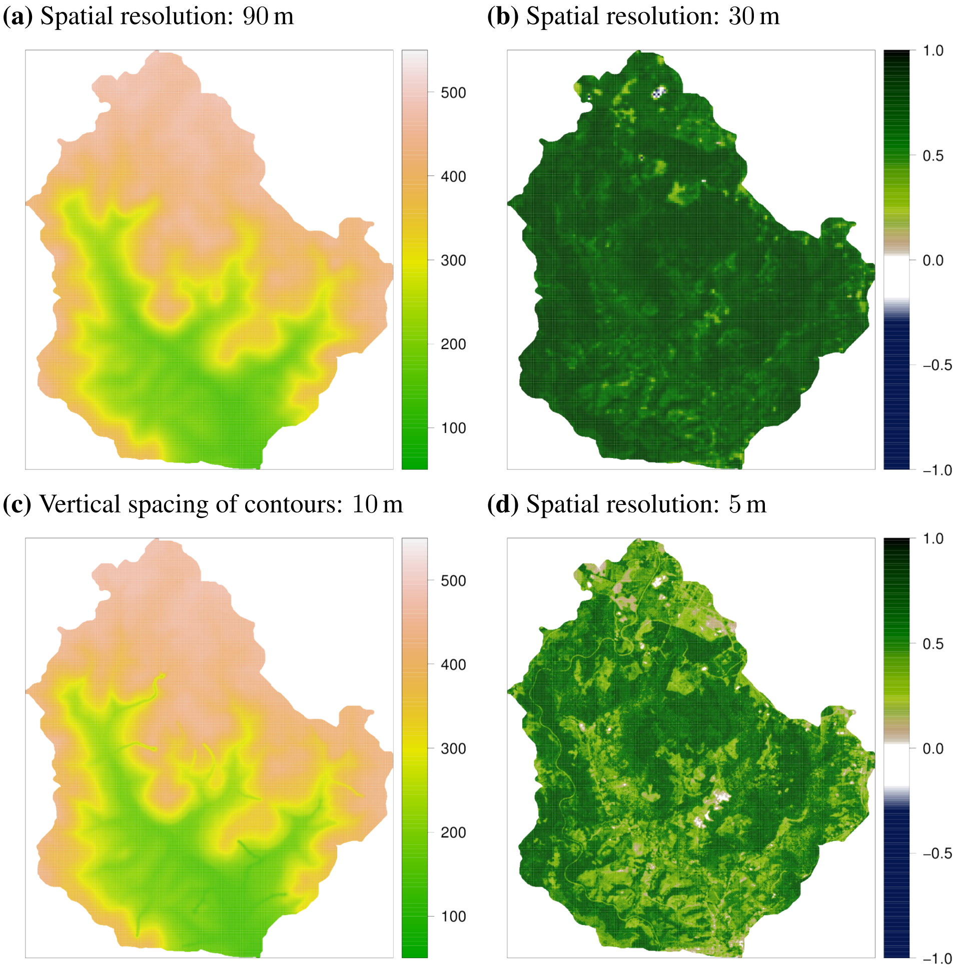

(a) Spatial resolution: 90 m
(b) Spatial resolution: 30 m
(c) Vertical spacing of contours: 10 m
(d) Spatial resolution: 5 m
500
400
300
200
100
1.0
0.5
0.0
−0.5
−1.0

Figure 6.4: Digital elevation models (a, c) and satellite images, depicted using the normalized difference vegetation index (b, d), compared in our study. The less detailed version is displayed at the top, while the more detailed version is shown on the bottom.

2001; LARK et al., 2006). We use the point soil observations and spatially exhaustive predictor variables to calibrate the model and predict the outcome of the spatial stochastic process at unobserved locations. This fixed effect (deterministic trend), $m(\mathbf{s})$, describes that part of the spatial variation of the soil property that is explained by the covariates. We assume here that is a linear function of the predictor variables. The random effect (stochastic residuals, latent variables), $e(\mathbf{s})$, describes that part of the spatial variation that cannot be explained by the covariates (CRESSIE, 1993). It is represented by a spatially correlated, Gaussian distributed random variable, that is assumed stationary in the mean and covariance. Thus, the linear mixed model of spatial variation that we employed is given by

$$Y'(\mathbf{s}) = m(\mathbf{s}) + e(\mathbf{s}) = \sum_{j=0}^{p} \beta_j \cdot X_j(\mathbf{s}) + e(\mathbf{s}), \tag{6.1}$$

where $Y'(\mathbf{s})$ is the soil property after Box-Cox transformation, $m(\mathbf{s})$ and $e(\mathbf{s})$ are defined as above, β_j are the regression model coefficients, and $X_j(\mathbf{s})$ is the regression model matrix, with $j = 0, 1, 2, \ldots, p$, p being the number of predictor variables. Variable $X_0(\mathbf{s})$ is taken as unity so that β_0 is the intercept.

6.4.3.1 Model selection

We calibrated $k = 2^5 = 32$ multiple linear regression models for each soil property (fitted using ordinary least squares, OLS) to model the deterministic trend for each combination of the five covariates (recall from Section 6.3 that each covariate is available at two levels of spatial detail, hence 2^5 combinations). The number of predictor variables used to calibrate each model varied among combinations between $p = 52$ and $p = 62$, because more detailed covariates enabled the derivation of a larger number of predictor variables (except the DEM). Backward VIF (variance inflation factor) selection followed by stepwise AIC (Akaike's Information Criterion) selection were used to select predictor variables to enter the models (SAMUEL-ROSA et al., 2014b; VENABLES; RIPLEY, 2002).

The $k = 32$ multiple linear regression models calibrated for each soil property were ranked using the ratio between the regression sum of squares and the total sum of squares. Because stepwise regression results in biased models (HARRELL, 2001), the ratio of sum of squares was adjusted (R^2_{adj}) using the number of predictor variables initially offered to enter the model instead of the reduced number of predictor variables that entered the model. Next, the five covariates were ranked based on how their level of spatial detail related with the calibration of models with improved predictive performance. The relation between the level of spatial detail of the covariates and model performance was evaluated using a graphical output called *model series plot* (R-package pedometrics, Samuel-Rosa et al. (2014b)). Pedological evaluation of predictor variables included in the models was omitted because this was beyond our objectives.

The multiple linear regression model calibrated using only the less detailed covariates, which we call the *baseline* model, and the multiple linear regression model with the highest R^2_{adj}, which we call the *best performing* model, were extended to linear mixed models of spatial variation (Equation 6.1) for each soil property. Estimation of the parameters of the linear mixed models was performed using residual (restricted, marginal) maximum likelihood (REML) (RIBEIRO JR; DIGGLE, 2001; LARK; CULLIS, 2004). The spatial correlation function adopted was the exponential function (this is equivalent to the Matérn correlation function with smoothness parameter $\nu = 0.5$ (STEIN, 1999)).

6.4.3.2 Model validation

Only the *baseline* and *best performing* multiple linear regression and linear mixed models calibrated for each soil property were validated. Model validation was performed using leave-one-out cross-validation (LOO-CV) (BRUS et al., 2011). All model parameters were re-estimated at each LOO-CV run to reduce bias (LASLETT et al., 1987). LOO-CV predicted values were back-transformed from the Box-Cox space to the original space of soil properties using stochastic simulation (CHRISTENSEN et al., 2001):

(I) each predicted value and associated prediction error variance were used to simulate $n = 20\,000$ values from a Gaussian distribution;

(II) simulated values were back-transformed using $Y(s) = (Y'(s) \times \lambda + 1)^{1/\lambda}$, if $\lambda > 0$, and $Y(s) = exp(Y'(s))$, if $\lambda = 0$;

(III) the mean and variance of back-transformed simulated values were used as the predicted value and prediction error variance in the original space of soil properties.

Five error statistics were computed from the leave-one-out cross-validation results (JANSSEN; HEUBERGER, 1995; KEMPEN et al., 2010; BRUS et al., 2011). The mean error (*ME*), which measures the prediction bias, the mean absolute error (*MAE*) and the root mean squared error (*RMSE*), which measure the prediction accuracy, the scaled root mean squared error (*SRMSE*, also known as mean squared deviation ratio), which measures how well the prediction error variance matches the squared differences between predicted and observed soil property, where $SRMSE > 1$ indicates under-estimation, while $SRMSE < 1$ indicates over-estimation, and the amount of variance explained (*AVE*, also known as coefficient of determination or ratio of scatter), which measures the fraction of the overall spread of observed values that is explained by the model. The AVE ranges from 0 to 100, where $AVE = 100$ is the optimal value.

6.4.3.3 Spatial prediction

Only the *baseline* and *best performing* linear mixed models calibrated for each soil property were used for spatial prediction. Spatial predictions at a fine grid of $\sim 800\,000$ point locations were made in the Box-Cox space using the best linear unbiased predictor (BLUP) with the empirical estimates of the random effects (EBLUP) (LARK et al., 2006). EBLUP with a fixed effect model is conceptually equivalent to kriging with external drift and regression kriging, and mathematically equivalent to kriging with external drift and universal kriging. Point predicted values and prediction error variances were back-transformed to the original soil property space using stochastic simulation as described above (Section 6.4.3.2).

6.5 RESULTS

6.5.1 Model Series Plots

The model series plot is a graphical description of the relation between the prediction accuracy of multiple linear regression models and the covariates used to calibrate them (Figure 6.5). The magnitude of improvement in prediction accuracy is depicted in the bottom panel with the R^2_{adj}. The top panel is interpreted both horizontally and vertically. In the vertical direction we identify which version of each covariate was used to calibrate a given model. The less and the more detailed versions are identified by the yellow (bright) and green (dark) colours, respectively. The *baseline* model is identified by the column containing only yellow cells, while the column with only green cells represents the model calibrated using only the more detailed version of each covariate, which we call the *most detailed* model. The first important results that we obtain from the model series plots is that a) the *baseline* model is not the model with the lowest R^2_{adj}, which we call the *poorest performing* model, and b) the *most detailed* model is not the *best performing* model.

The row-wise analysis of the model series plots shows if a model calibrated with the more detailed version of a given covariate has a higher prediction accuracy. This information is retrieved by looking at the row-wise distribution of green cells – these cells represent the $k = 16$ models calibrated using the more detailed version of a given covariate, irrespective of the version of the other covariates. The more concentrated the green cells are in the right half of the plot, the larger the relative benefit of using the more detailed version of that covariate. For example, the top row of the second model series plot shows the SOC models calibrated using the two versions of the land use map (`land`). All green cells are on the right half of the plot between rankings 1 and 16 (see the x axis). The four lower rows show that the green cells of the other four covariates are distributed along the entire ranking range (from 1 to 32). This means that the relative benefit of calibrating a SOC model with a more detailed land use map is larger

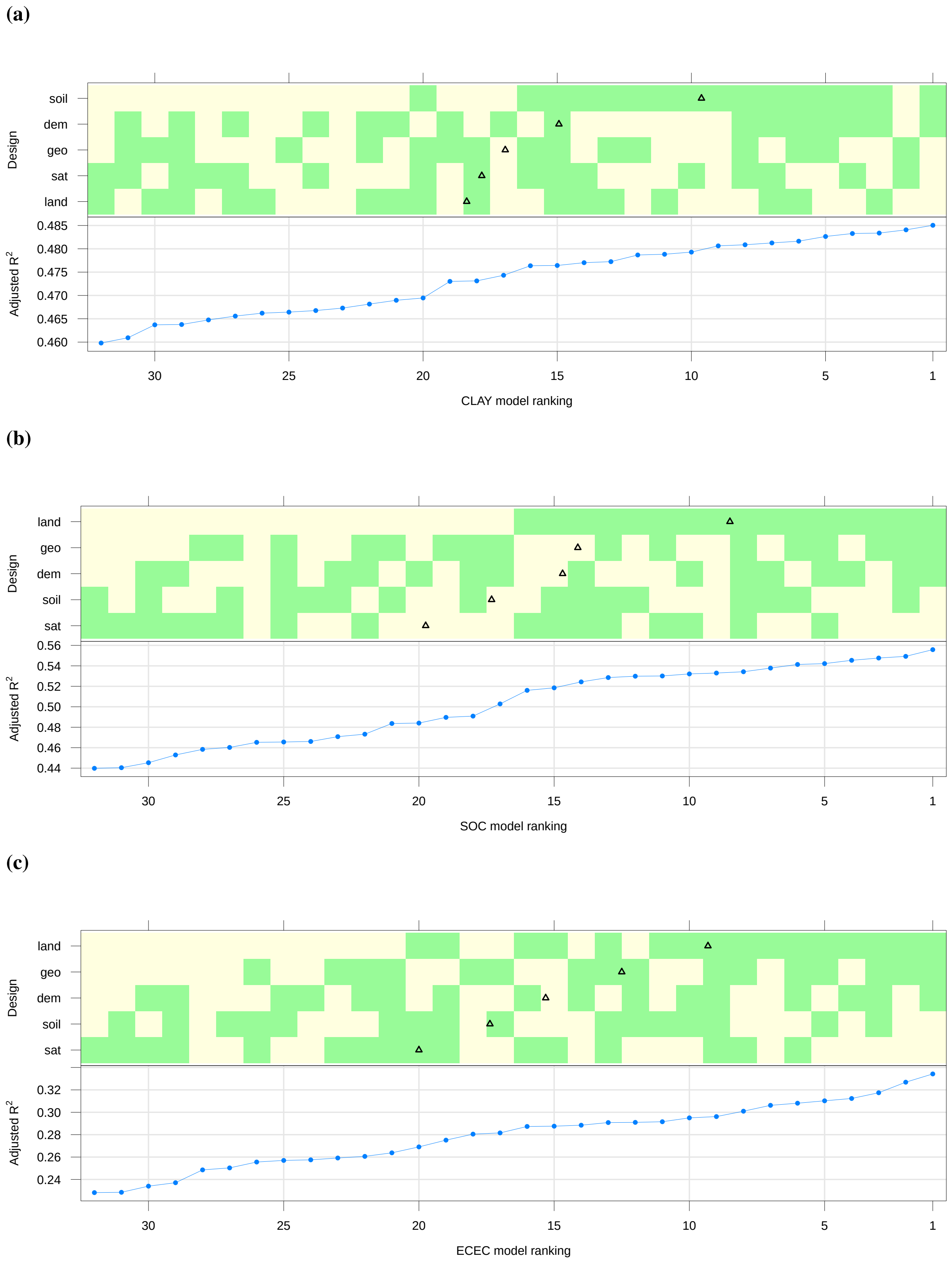


Figure 6.5: Model series plots for CLAY (a), SOC (b), and ECEC (c). The less and more detailed version of each covariate are identified with the yellow (bright) and green (dark) colours, respectively. Multiple linear regression models were ranked using their R^2_{adj}. Triangles show the mean ranking of the more detailed covariates.

compared to that of using a more detailed version of the other covariates.

The centre of the row-wise distribution of the green cells for each covariate, calculated as the mean ranking, is represented by the triangles. The mean ranking quantifies the relative benefit of using a more detailed version of each covariate. For example, the mean ranking of the SOC models calibrated using the more detailed land use map is about 8 (top row), while the mean ranking of the models calibrated using the more detailed satellite image (`sat`) is close to 20 (bottom row). Using the more detailed DEM (`dem`) is almost as beneficial as using the more detailed geologic map (`geo`) – the mean ranking of the SOC models calibrated using the more detailed version of these two covariates is about 15–16 (second and third rows). Using the more detailed version of the soil map (`soil`, fourth row) is not as beneficial as using `land`, `geo` or `dem`, but more beneficial than using `sat`. Because the covariates were ranked based on the mean rankings, the covariate displayed in the top row of each model series plot is the one which resulted in the largest improvement of the prediction accuracy when the more detailed version was used to calibrate the model – for SOC this is the land use map.

For CLAY, calibrating the models with the more detailed soil map resulted in the largest improvement of the prediction accuracy relative to the other covariates. The DEM was the second most beneficial covariate (mean ranking of 15), but the benefit of using its more detailed version was similar to that of using the more detailed version of any other covariate (mean rankings between 17 and 18). Nine models had a poorer prediction performance than the baseline model, ranked 27th, the poorest performing model being that calibrated with the more detailed land use map and satellite image. Despite these patterns, calibrating CLAY models with the more detailed version of any covariate resulted in a small improvement of the prediction accuracy, as evidenced by the small increases of the R^2_{adj}. The difference between the poorest and best performing models is less than 3 pp (percentage points). In comparison, for SOC, by simply using the more detailed land use map we already obtained a model ranked 9th, an increase of 8 pp in R^2_{adj} compared to the baseline model, ranked 24th.

The same general pattern observed for SOC models was observed for ECEC models – the more detailed land use map results in the largest improvement of the prediction accuracy. The main difference is that calibrating the models with the more detailed geologic map was slightly more beneficial for ECEC (mean ranking of 12) than for SOC (mean ranking of 14). The poorest performing ECEC model was that calibrated with the more detailed satellite image. Using only the more detailed land use map resulted in an improvement of 6 pp in R^2_{adj} (model ranked 7th), differing from the best performing model by only 2 pp. Using the more detailed version of all covariates except the soil map or satellite image resulted in increases of about 6 and 7 pp in R^2_{adj}, respectively. The baseline model was ranked as 28th, which is a higher ranking than the models calibrated with all possible combinations of the more detailed satellite image and the more detailed soil map and/or DEM.

The patterns observed in the model series plots resulted from the change (increase or decrease) of the importance of each covariate on explaining the variance when the more detailed version was used (Table 6.4). We used the *baseline* and *most detailed* models to quantify this change. Each model was refitted dropping one covariate at a time. The difference Δ between the R^2_{adj} of the model calibrated with all five q covariates ($R^2_{adj\,q=5}$) and the model calibrated without the q-th covariate ($R^2_{adj\,q=5-1}$) was calculated. The more positive ΔR^2_{adj} becomes, the more beneficial the more detailed version of the q-th covariate is for improving prediction accuracy. For CLAY, `dem` and `land` were the most important covariates in the *baseline* model, while `geo` was the least important. The importance of `soil` and `geo` increased when their more detailed version was used (change of 0.013 pp for both), while `sat`, `land` and `dem` became less important. For SOC and ECEC, `land` was not the most important covariate in the *baseline* model. But it was the covariate whose importance had the largest positive shift when the more detailed version was used (0.085 pp for SOC and 0.045 pp for ECEC). sat became less important when the more detailed version was used – see its low ranking in all model series plots. The increase of the importance of geo was larger for ECEC (0.026 pp) than for SOC (0.013 pp) – see the difference in the mean ranking of geo in the SOC (14) and ECEC (12) model series plots.

Table 6.4: The importance of each covariate[a] (ΔR^2_{adj} [b]) in the models calibrated with their less and more spatially detailed version.

Covariate	CLAY		SOC		ECEC	
	Less	More	Less	More	Less	More
soil	−0.009	0.004	−0.006	−0.008	0.011	−0.003
land	0.003	−0.002	0.003	0.088	−0.004	0.041
geo	−0.019	−0.006	−0.005	0.008	0.007	0.033
sat	−0.010	−0.016	0.018	−0.014	0.011	−0.029
dem	0.030	0.001	−0.009	0.016	−0.035	−0.041

[a] Covariate: soil - soil map, land - land use map, geo - geologic map, sat - satellite image, and dem - digital elevation model.

[b] $\Delta R^2_{adj} = R^2_{adj\,q=5} - R^2_{adj\,q=5-1}$, where q is the number of covariates included in the model. Negative values result from adjusting the R^2 using the number of predictor variables initially offered to enter the model instead of the reduced number of predictor variables that entered the model.

6.5.2 REML Fit of the Variogram Model

The small improvement in the prediction accuracy of the CLAY linear mixed model calibrated with the more detailed covariates is evidenced by Figure 6.6. The shape of the experimental variogram is very similar for both *baseline* and *best performing* linear mixed models, which is also true for SOC and ECEC. However, the sill variance had a very small reduction for CLAY compared to SOC and ECEC. The last two showed a more considerable improvement in prediction accuracy. It can also be seen that the number of point observations separated by short distances is very small, reducing the accuracy of the estimate of the nugget variance[23]. The result is that the estimated nugget variance changes rather erratically from the *baseline* to the *best performing* models, decreasing for CLAY and SOC, and increasing for ECEC.

6.5.3 Validation

The LOO-CV results indicate that the linear mixed models for CLAY are slightly positively biased, while those for SOC and ECEC are slightly negatively biased (Table 6.5). For both CLAY and ECEC, the *MAE* shows that these models are more accurate than the multiple linear regression models, suggesting that the kriging step improves the prediction accuracy.

Overall, all models had a moderate to poor prediction performance. The errors are, in absolute values, somewhat large, mainly for ECEC. The best *AVE* are about $60\,\%$ for CLAY, $50\,\%$ for SOC, and $40\,\%$ for ECEC. In general, the prediction error variance was under-estimated by

[23] To be more precise, the small number of point observations separated by short distances reduces the ability of modelling the behaviour of the variogram near the origin as a whole.

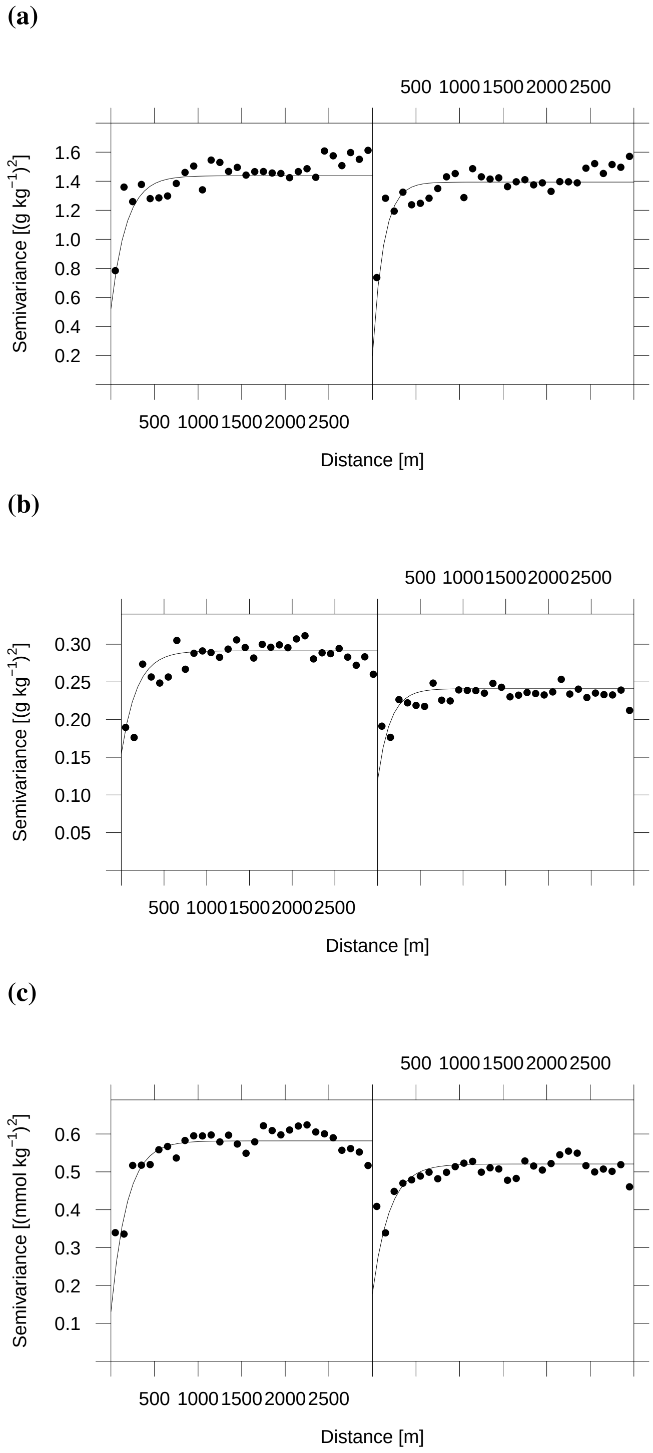


Figure 6.6: Experimental variogram (dots) and REML fit of the linear mixed models (line) for CLAY (a), SOC (b), and ECEC (c). Left – baseline model. Right – best performing model.

the linear mixed models and over-estimated by the multiple regression models. The best estimates of the prediction error variance were obtained by both CLAY linear mixed models, and the ECEC baseline linear regression model.

For CLAY, the increase in the *AVE* was larger when including a kriging step ($\Delta AVE = 3.9\,\text{pp}$) than when using more detailed covariates ($\Delta AVE = 1.6\,\text{pp}$). In the case of SOC, including a kriging step reduced the *AVE* by $3.2\,\text{pp}$, and for ECEC, both strategies increased the *AVE*

Table 6.5: Statistics[a] of the leave-one-out cross-validation of baseline and best performing multiple linear regression models (LM) and linear mixed models (LMM).

Model	Type	*ME*	*MAE*	*RMSE*	*SRMSE*	*AVE*
CLAY (g kg^{-1})						
Baseline	LM	1.31	52.1	72.1	0.89	56.8
	LMM	0.94	48.5	68.8	1.03	60.7
Best performing	LM	1.59	51.3	70.7	0.91	58.4
	LMM	1.08	47.8	68.1	1.03	61.5
SOC (g kg^{-1})						
Baseline	LM	−0.30	10.9	18.9	1.22	35.8
	LMM	−0.39	11.0	19.4	1.43	32.5
Best performing	LM	−0.20	10.1	16.9	0.91	49.0
	LMM	−0.25	10.4	17.6	1.16	44.3
ECEC (mmol kg^{-1})						
Baseline	LM	−0.88	70.6	112.4	0.97	22.3
	LMM	−0.32	63.3	101.1	1.32	37.1
Best performing	LM	−0.76	64.9	101.7	0.86	36.3
	LMM	−0.29	62.6	97.9	1.09	41.1

[a] Statistics: mean error (*ME*), mean absolute error (*MAE*), root mean squared error (*RMSE*), scaled root mean squared error (*SRMSE*, unitless), and amount of variance explained (*AVE*, percent).

(Table 6.5).

6.5.4 Spatial Prediction

Both *baseline* and *best performing* linear mixed models captured the same overall pattern of spatial variation of the soil properties (Figure 6.7). The main difference is that the spatial patterns of the different covariates used to calibrate each model produced different features in the prediction maps. For example, the CLAY map produced by the best performing model (Figure 6.7b) displays abrupt changes in the predicted values in the north-north-east due to the use of the more detailed soil map. Strongly-marked features following the stream network obtained through the use of the more detailed DEM are also observed (Figure 6.7b and Figure 6.8b).

SOC maps (Figure 6.7d and Figure 6.8d) show peculiar features in the central part of the study area, where predictions reached values as high as $507\,\mathrm{g\,kg^{-1}}$, while the maximum value in the calibration data is $163\,\mathrm{g\,kg^{-1}}$. The extremely high predicted values resulted from the inclusion of the topographic position index derived from the more detailed DEM, using a window size of 15×15 pixels (`TPI_10_15`) to model the deterministic trend. `TPI_10_15` values in the point calibration data range from -7 to $6\,\mathrm{m}$, while in the central part of the study area they range from 12 to $31\,\mathrm{m}$. Thus, feature-space extrapolation explains the extremely high predicted values for SOC. Abrupt changes in predicted SOC are also observed at locations with low to moderate SOC ($40\,\mathrm{g\,kg^{-1}}$–$80\,\mathrm{g\,kg^{-1}}$). This is caused by using the more detailed land use map.

Predicted ECEC (Figure 6.7e and Figure 6.7f) had a large dependency on land use and geologic maps. Several features observed in the prediction maps derive from these two covariates. The influence of land use is seen in the northern part, while in the western, central, and eastern parts the influence of both covariates create an irregular pattern in the spatial distribution of ECEC. It is also in these parts that the largest prediction error standard deviations occur, following the spatial pattern of the covariates.

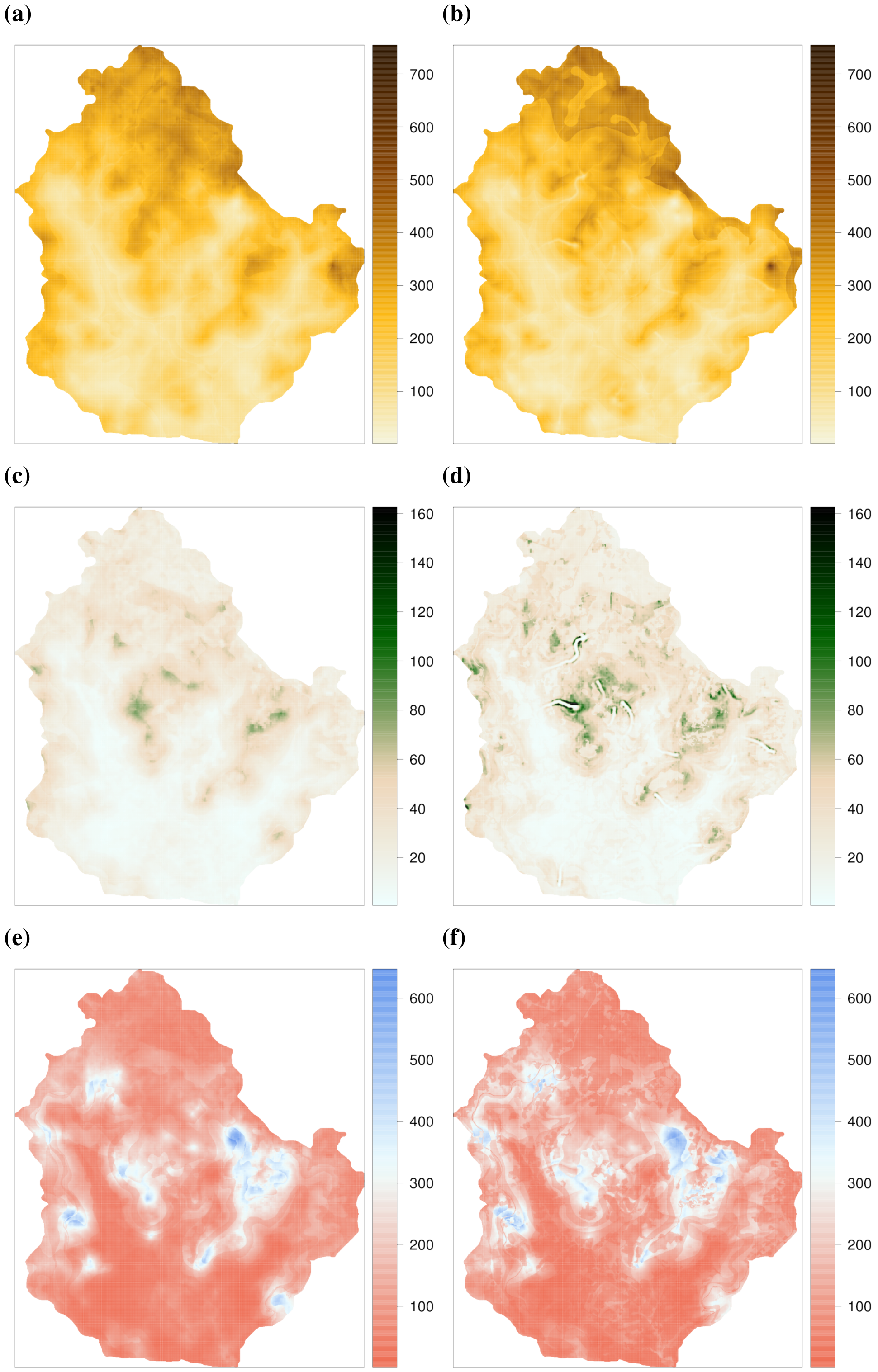


Figure 6.7: Predicted values for CLAY (g kg^{-1}) (a, b), SOC (g kg^{-1}) (c, d), and ECEC (mmol kg^{-1}) (e, f) using the *baseline* (left) and *best performing* (right) linear mixed models.

The smallest prediction error standard deviations occur at lower elevations, along the

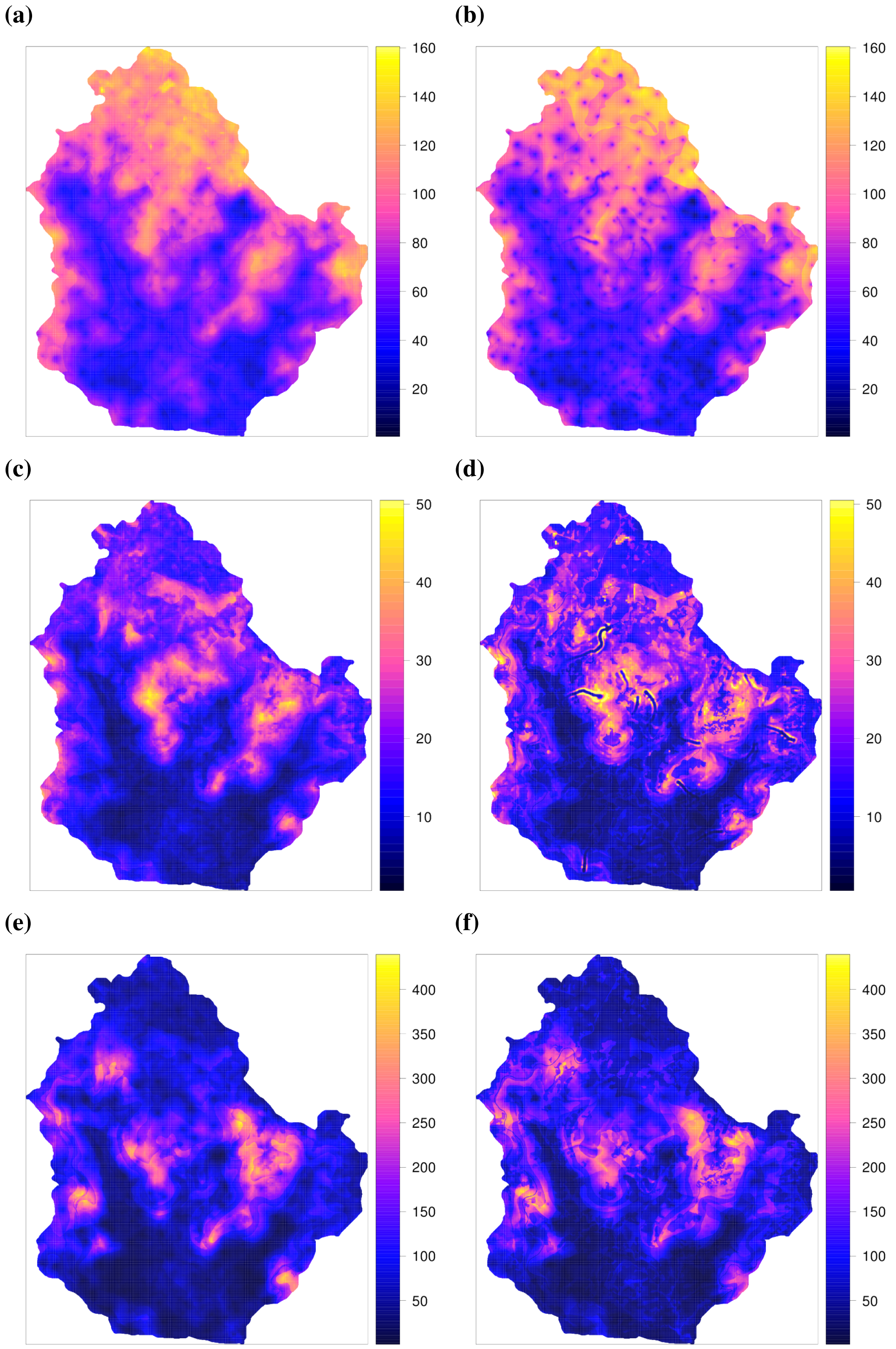


Figure 6.8: Prediction error standard deviations for CLAY ($\mathrm{g\,kg^{-1}}$) (a, b), SOC ($\mathrm{g\,kg^{-1}}$) (c, d), and ECEC ($\mathrm{mmol\,kg^{-1}}$) (e, f) using the *baseline* (left) and *best performing* (right) linear mixed models.

three main streams, and close to the water outlet in the southern part of the study area. These areas have the highest density of point soil observations used to calibrate the models, and the smallest values for all three soil properties. While the first determines the accuracy of the EBLUP, the second influences the final accuracy through the back-transformation of predicted values.

6.6 DISCUSSION

Our main goal was to evaluate whether investing in more spatially detailed covariates improves the accuracy of soil maps. We saw that calibrating the models with more detailed covariates generally has a small to moderate, but positive, impact on the predictions. The magnitude of this benefit depends on the magnitude of the increase of the spatial detail of the covariate, on the other covariates included in the model, and on the soil property. However, there seems to be a limit above which the increase of spatial detail has a negative impact on the predictions. In the next two subsections we interpret the results from a pedological perspective and assess whether the investment in more detailed covariates is worthwhile or if alternatives to improve prediction accuracy should be favoured.

6.6.1 Spatio-Temporal Controls of Soil Properties

CLAY was moderately well predicted using less detailed covariates, with small improvement when using the more detailed covariates. CLAY was expected to have a strong correlation with topography and parent material. This correlation was already considerable when the less detailed DEM and geologic map were used, and improved only marginally with the more detailed version. One sensible explanation is that the effective (actual rather than theoretical) spatial detail of the two geologic maps was similar, although they had a four-fold difference in the size of the minimum legible delineation (see Hengl & Husnjak (2006) for a discussion on effective scale). For the DEM, many studies have already suggested that its resolution may be of secondary importance when calculating DEM derivatives for soil mapping (ZHU et al., 2008; BEHRENS et al., 2010; MILLER et al., 2015). The influence of land use on CLAY is currently small due to reduction of soil erosion in the first decade of the 21st century (MIGUEL et al., 2011; TEN CATEN et al., 2012b). A moderate within-field spatial variation may exist due to past erosional processes (MOURA-BUENO, 2012), but we lack evidence of how well this source of variation was captured in the present-time point soil data.

It is worthwhile to consider the influence of the more detailed soil map on predicting CLAY. Due to its production process, the more detailed soil map derives a large amount of spatial detail from the geologic map, land use map and DEM – note that the second-best performing model for CLAY included the more detailed geologic map instead of the more detailed soil map (Figure 6.5). However, most of the additional spatial detail included in the more detailed soil map was probably based on the spatial variation of soil texture, because this is a strongly marked soil feature in the area (MIGUEL et al., 2011). Soil texture is one of the most important soil properties used by soil surveyors in the field to identify mapping units (LEGROS, 2006). These findings help explain why in the end the more detailed soil map was the most beneficial for CLAY instead of the geologic map.

SOC and ECEC were considerably better predicted when more detailed covariates were used. Our expectation that SOC and ECEC would have a strong correlation with land use was confirmed by the fact that this covariate explained a large amount of the variance and was highly beneficial for improving the predictions. Although the available point soil data are limited to the 2004–2011 period, we believe that land use changes in the last 30 years (MIGUEL et al., 2011; TEN CATEN et al., 2012b) strongly affected SOC and ECEC. Thus, the more detailed land use map is likely to have considerably improved model performance because it is up-to-date and, possibly, because it has 40 times more spatial detail than its less detailed version. Despite the fact that the two land use maps used in this study were from different time periods, which confounds the analysis, the results obtained indicate that a more detailed land use map improves the prediction of SOC. For example, the areas used for crop agriculture, which are well known for having lower SOC and ECEC (MENEZES, 2008; MOURA-BUENO, 2012), are not depicted in the less detailed land use map.

We expected SOC to have a stronger correlation with the DEM than with the geologic map due to its strong dependence on erosion, but we observed the contrary. This result may be partially explained by the fact that there is a strong relation between geology and topography in the study area (SARTORI, 2009). Due to its production process (MACIEL FILHO, 1990), the geologic maps can be interpreted as an aggregated version of a DEM. A second sensible explanation is that the effect of erosion on SOC is not that large because erosion was considerably reduced in the last decade (MIGUEL et al., 2011; TEN CATEN et al., 2012b). A last possible explanation, which integrates the previous two, is the existence of a spatial relation between SOC and CLAY, the last being strongly correlated with parent material. These relations help explain why the more detailed DEM was almost as beneficial as the more detailed geologic map for SOC predictions. In the case of ECEC, our expectation of a strong dependency on a more detailed geologic map for producing more accurate predictions was confirmed.

The observed benefit of the more detailed geologic map and DEM for making more accurate CLAY, SOC, and ECEC predictions suggests that these soil properties are spatially related in the study area. We also hypothesize that the complexity of current land use makes it difficult to achieve SOC and ECEC models with performances comparable to CLAY. One important source of variation in forested areas is its use for animal grazing (SAMUEL-ROSA et al., 2011). This influences nutrient cycling and soil nutrient availability (SCHRAMA et al., 2013). Current remote sensing technology is unable to capture the data needed to proxy the environmental conditions created by these processes.

6.6.2 Using More Detailed Covariates

More detailed covariates are usually expected to improve predictions in soil mapping (CAVAZZI et al., 2013; MAYNARD; JOHNSON, 2014). However, deciding whether to invest or not in more detailed covariates requires careful thinking and depends on case-specific elements. We generally saw improvement in the predictions in our study, but the improvement was not large and may not outweigh the costs. Also, the models calibrated with the more detailed versions of all covariates were not the best performing models. Using more detailed satellite images and land use maps degraded CLAY predictions. Although the more detailed soil map had the largest benefit for CLAY, it may be too costly and impractical since its production usually requires having available more detailed versions of all other covariates. For SOC and ECEC, simply using a more detailed land use map resulted in considerably more accurate predictions. However, the superior performance may not outweigh the extra costs because producing a more detailed land use map usually requires up-to-date field observations and satellite images. Thus, the decision to adopt a more detailed covariate for soil mapping will ultimately depend on a trade-off between the increased accuracy and the extra budget required. It may also depend on other potential applications of the covariates, but this is not our concern here.

One interesting observation is that if a less detailed covariate yields poor predictions, its more detailed version has the potential to produce larger improvement in model performance. However, this is only a potential, not a guarantee. For instance, Eldeiry & Garcia (2008) were not able to increase the $R^2 = 0.31$ of linear regression models of soil salinity by more than 0.07 points using 7.5 times more detailed satellite images. On the other hand, model performance is likely to be hardly improved using more detailed covariates if their less detailed version has already produced accurate predictions. This agrees with findings by Thompson et al. (2001) and Kim et al. (2014).

We also observed that the predictions can be degraded when using the more detailed version of covariates. In our study, this happened with the satellite image (all three soil properties), land use map (CLAY) and soil map (SOC and ECEC). A (small) benefit was observed only when these covariates were used along with the more detailed version of other covariates. As pointed out above, such a small benefit may not outweigh the increase in mapping costs. The trade-off between reducing model performance and being beneficial seems to depend on how much more spatial detail a covariate will have and on its correlation with the soil property. For example, the land use map was strongly correlated with SOC and ECEC, but not with CLAY, and its more detailed version had 40 times more spatial detail. It helped improve SOC and ECEC predictions, but degraded CLAY predictions, resulting in only a small improvement when used along with the more detailed satellite image and geologic map.

If the influence of a more detailed covariate depends on the increase of spatial detail, then the priority should be to improve the spatial detail of the most beneficial covariate. This requires solid subject area knowledge because empirical evidence from the *baseline* model may be insufficient. The most beneficial covariate is not necessarily that which explained the largest part of the variance in the *baseline* model (see Table 6.4). This occurs because increasing the spatial detail reduces the correlation between the covariate and the soil property. And also because there is little room to improve a correlation that is already high in the *baseline* model. Cavazzi et al. (2013) suggest that the more detailed covariate has an excess of detail, a "noise" that degrades the predictions. This could explain the results for `sat`: higher resolution images can resolve smaller objects (e.g. individual plants) whose spectral behaviours are highly variable, adding noise to the `sat`-soil property correlation; on the other hand, lower resolution images capture collections of objects, and thus their variation is smoothed out in the pixel, reducing noise.

According to information theory one should optimize (maximize) the correlation between the point soil data and the covariates. This was described elsewhere as matching the "phenomenon scale" (the spatial pattern of the soil property) with the "analysis scale" (the spatial pattern of the covariates) (DUNGAN et al., 2002; MILLER; SCHAETZL, 2014). Finding the "optimum" requires evaluating the strength of the correlation using covariates with different levels of spatial detail (DRĂGUŢ et al., 2009; CAVAZZI et al., 2013; MILLER et al., 2015). Our results show that this approach may be too costly and impractical. Since modern soil mapping techniques explore only the empirical relation among environmental conditions and soil properties (GRUNWALD, 2009), the "optimum" is a "conditional optimum" – conditional on the point soil data available. It does not necessarily mean that the most accurate predictions will be made, but only that there is a level of spatial detail at which the correlation between the covariate and the point soil data is at a maximum. We suggest that instead more comprehensive approaches should be used to explore the full potential of the available covariates (see Behrens et al. (2010) and Miller et al. (2015) for examples).

Finally, one must still judge whether the potential improvement in predictions is sufficient given the extra costs involved with using more detailed covariates. If the extra budget is spent on deriving more detailed covariates, we suggest that it may be better to substantially improve the detail of a less influential covariate than marginally increase the detail of the most influential covariate. However, other means to spend the extra budget should be considered. For instance, it may be more efficient to concentrate on obtaining more soil observations. These may focus on better capturing the short range spatial variation (BRUS; HEUVELINK, 2007) or improving the representation of the feature space to avoid undesirable extrapolations (MINASNY; MCBRATNEY, 2006).

6.7 CONCLUSIONS

This study has shown that:

(I) Using more detailed covariates results in only a modest increase in the prediction accuracy of linear prediction models;

(II) A more detailed covariate has a greater potential to improve prediction accuracy when the soil property is poorly predicted by its less detailed version;

(III) The impact on prediction accuracy when using the more detailed version of a less important covariate may depend on which other covariates are included in the model;

(IV) Choosing whether or not to invest in more detailed covariates depends on the strength of the relationship between the covariates and the soil property being modelled, and on the relative difference between the less detailed and the more detailed versions of the covariates.

7 CHAPTER VI

SPATIAL POINT PATTERN ANALYSIS OF SOIL SURVEY OBSERVATION LOCATIONS

* This chapter is based on the studies *An approach to help formalizing the purposive sampling strategy of classical soil surveys*, presented at the 20th World Congress of Soil Science (SAMUEL-ROSA et al., 2014), and *Spatial point pattern analysis of soil survey sampling locations*, presented at the 10th European Conference on Geostatistics for Environmental Applications (SAMUEL-ROSA et al., 2014).

7.1 RESUMO

Modeladores espaciais de campo do solo normalmente selecionam os locais de observação com base no conhecimento tácito sobre as relações solo-paisagem. O objetivo desse estudo foi avaliar o potencial da análise de padrão pontual (PPA) para ajudar a compreender a estratégia de amostragem intencional, tradicionalmente empregada por modeladores de campo do solo. Um conjunto de dados consistindo de $n = 340$ observações do solo obtidas em um exercício de modelagem espacial do solo realizado no Sul do Brasil entre 2008 e 2011 foi usado. PPA foi realizada utilizando os pacotes `spatstat` e `splancs` do `R`. A intensidade de observação do solo foi estimada utilizando um núcleo de convolução suavizador Gaussiano isotrópico. A função G não-homogênea e simulações de Monte Carlo foram usadas para avaliar o padrão espacial dos pontos de observação. Um modelo de processo de Poisson não-estacionário foi ajustado utilizando covariáveis (uso da terra e atributos do terreno) e a data de observação para avaliar como características ambientais e outros fatores intervenientes influenciam a escolha dos locais de observação. Comparações entre PPA e julgamentos elicitados dos modeladores do solo que realizaram a modelagem do solo foram usadas para validar a análise. A PPA mostrou que os locais de observação do solo são não-homogeneamente distribuídos na área de estudo. Duas áreas foram mais densamente amostradas (distância do vizinho mais próximo, NND $< 125\,\text{m}$), uma no setor Sul, a outra no setor Centro-Norte-oriental. Os pontos de observação têm uma distribuição espacial aproximadamente aleatória nessas duas áreas, enquanto que nas zonas de baixa densidade amostral eles são espaçados à distâncias aproximadamente regulares (NND $> 125\,\text{m}$). Uso da terra, estratos fisiográficos, e data de observação tem uma influência significativa sobre a distribuição espacial dos locais de observação. O uso da terra produz a maior redução na desviância ($6\,\%$), seguido pela data de observação ($4\,\%$) e estratos fisiográficos ($2\,\%$). Esses resultados estão em concordância com os julgamentos eliciados dos modeladores do solo. De acordo com eles, as áreas mais densamente amostrados são aquelas onde eles tinham um conhecimento menos consistente das relações solo-paisagem. Estas áreas foram visitadas nas primeiras e nas últimas campanhas de campo, evidenciando um efeito temporal. Por outro lado, a maioria das campanhas intermediárias foram realizadas em áreas onde eles tinham um melhor conhecimento das relações solo-paisagem. Outras campanhas intermediárias foram realizadas em áreas de difícil acesso, onde os modeladores do solo selecionaram os locais de observação por conveniência. Campanhas de campo intermediárias também coincidiram com um corte no orçamento. A interação desses fatores intervenientes reduzir a motivação dos modeladores do solo para realizar amostragem mais intensiva. Os modeladores do solo também relataram que eles realizaram uma estratificação mental da área de estudo antes de selecionar os locais de observação. Características do terreno compuseram a primeira variável de estratificação, ao passo que o uso da terra foi utilizado como variável de estratificação de segundo estágio. A estreita concordância entre a PPA e os julgamentos elicitados dos modeladores do solo sugere que PPA pode ajudar a compreender a estratégia de amostragem utilizada por modeladores do solo.

Palavras-chave: Modelagem espacial do solo. Caminhamento livre. Análise de padrão pontual. Julgamento especialista. Motivação.

7.2 ABSTRACT

Field soil spatial modellers usually select observation locations based on tacit knowledge about soil-landscape relationships. The aim of this study was to assess the potential of point pattern analysis (PPA) to help understanding the purposive sampling strategy traditionally employed by field soil modellers. A dataset consisting of $n = 340$ soil observations obtained in a soil spatial modelling exercise carried out in Southern Brazil between 2008 and 2011 was used. PPA was performed using the R-packages `spatstat` and `splancs`. Soil observation intensity was estimated using an isotropic Gaussian kernel smoother. The inhomogeneous G function and Monte Carlo simulations were used to evaluate the spatial pattern of the observation points. A non-stationary Poisson point process model was fitted using covariates (land use and terrain attributes) and the observation date to evaluate how environmental features and other intervening factors influenced the choice of observation locations. Comparisons between PPA and judgements elicited from the soil modellers who carried out the soil modelling were used to validate the analysis. PPA showed that soil observation locations are inhomogeneously distributed in the study area. Two areas were more densely sampled (nearest neighbour distance, NND $< 125\,\mathrm{m}$), one in the Southern sector, the other in the Centre-North-eastern sector. Observation points have an approximately random spatial distribution in these two areas, while in the less densely sampled areas they are spaced at approximately regular distances (NND $> 125\,\mathrm{m}$). Land use, physiographic strata, and observation date have a significant influence on the spatial distribution of observation locations. Land use yields the largest deviance reduction (6 %), followed by observation date (4 %) and physiographic strata (2 %). These results are in close agreement with judgements elicited from soil modellers. According to them, more densely sampled areas are those where they had a less consistent knowledge of soil-landscape relationships. These areas were visited in the first and last field campaigns, evidencing a temporal effect. On the other hand, most of the intermediate campaigns were carried out in areas where they had a better knowledge of soil-landscape relationships. Other intermediate campaigns were carried out in areas of poor accessibility, where soil modellers selected observation locations by convenience. Intermediate field campaigns also coincided with a budget cut. The interplay of these intervening factors reduced the motivation of the soil modellers to conduct more intensive sampling. Soil modellers also reported that they performed a mental stratification of the study area prior to selecting observation locations. Terrain features composed the first stratification variable, while land use was used as the stage-two stratification variable. The close agreement between PPA and judgements elicited from soil modellers suggests that PPA can help understand the sampling strategy used by soil modellers.

Keywords: Soil spatial modelling. Free survey. Point pattern analysis. Expert judgement. Motivation.

7.3 INTRODUCTION

Free survey has been successfully employed for many decades in soil spatial modelling. It is a model-based soil sampling method that relies on the mental model of soil-landscape relationships of field soil spatial modellers – also known as soil surveyors. This mental model – or tacit knowledge – is built with the experience gained in the field and its quality generally is directly proportional to the years of field work. The main drawback is that formalising such complex models as verbal representations – formal knowledge – is rather difficult, the reason why it is rarely or only partially done. Thus, the tacit knowledge is at the risk of being lost when the soil modeller deceases, hindering the transmission of soil knowledge to young soil modellers. The present study assesses the potential of *point pattern analysis* (PPA) as a tool to help understanding the purposive sampling strategy traditionally employed by field soil modellers.

The point pattern analysis toolbox includes multiple statistical techniques devoted to the analysis of sets of points distributed in the geographic space (DIGGLE, 2003). The main objective of these techniques is to provide empirical evidence to infer about the underlying process – a biological or other deterministic mechanism – that generated the *point pattern* (BIVAND et al., 2008). A basic assumption is that the point pattern can be treated as a realization of a spatially homogeneous (stationary) random point process in the plane (DIGGLE, 2003). Assuming that soil observations made using the free survey method come from a stochastic process is unrealistic because soil observations are made following the conceptual model of pedogenesis of soil modellers. In other words, they are the product of an unspecified deterministic process, alike many other biological mechanisms evaluated using point pattern analysis. Point pattern analysis serves as an efficient exploratory tool in these circumstances and helps choosing non-stationary models to fit spatially inhomogeneous processes (BADDELEY, 2010) alike soil observation.

Calibrating a point process model requires some knowledge of the underlying process – the biological mechanism – that generated the point pattern. For example, such knowledge can be obtained by eliciting the judgements of experts, a common practice in many fields of research (O'HAGAN et al., 2006). Usually, experts are consulted when important decisions have to be made during the development of a project or when there is great uncertainty about a phenomenon or event (MEYER; BOOKER, 2001). An expert is every individual who has a deep knowledge about a given theme. When trying to understand a soil observation process, an expert is every soil modeller that helped planning and conducting field sampling. The elicitation can be carried out in several ways (COOKE, 1991; MEYER; BOOKER, 2001; O'HAGAN et al., 2006) but due to operational issues, individual interview is the method most commonly employed. It also helps avoiding negative dominance effects due to eventual hierarchical relationships among experts (COOKE, 1991). Aggregation of expert judgements can be done using behavioural methods, where each expert evaluates the information provided by other experts until a common point is reached (ORSI et al., 2011).

Eliciting expert judgement is a key step towards understanding the underlying process that generated the point pattern of soil observation. Articulating expert information with psychological theories can provide important insights on the structure of this process. For example, it is agreed that the motivation to develop a series of tasks changes with time (BONEZZI et al., 2011; TOURÉ-TILLERY; FISHBACH, 2011a) and that the architecture of the surrounding environment influences the way space and distances are perceived (COETERIER, 1994; EPSTEIN; KANWISHER, 1998). Taking such psychological elements into account could help explaining the origins of any operational and conceptual tendencies noticed in the soil observation process as evidenced by the point pattern analysis and the interview of experts.

Environmental psychology is the field of science dedicated to the study of the relationships between human behaviour and the surrounding physical environment (BONNES; BONAIUTO, 2002). One of the recognized effects of the architecture of the surrounding physical environment on neurophysiological responses is the spatial enclosure (EPSTEIN; KANWISHER, 1998). In natural landscapes, spatial enclosure expresses itself in environments surrounded by elevated geomorphological features and rugged terrain such as in the valley bottom of mountainous regions. The effect of spatial enclosure on neurophysiological responses is complex (STAMPS; KRISHNAN, 2004), but it commonly results in the alteration of how distances are perceived (COETERIER, 1994). For instance, it can induce the perception that an enclosed space has a size larger than what it has in reality. A good example on how the perception of distances is altered due to spatial enclosure is the sensation of having walked hundreds of meters through a dense forest when only a few have been covered in reality.

Research on the psychology of goal pursuit and motivation has been carried out by many decades without coming to a common conclusion (TOURÉ-TILLERY; FISHBACH, 2011a; HULL, 1932). Modern theories claim that there are motivation shifts when a task involves the pursuit of multiple goals and describe it as the U-shaped motivation pattern of multi-goal pursuit (BONEZZI et al., 2011; TOURÉ-TILLERY; FISHBACH, 2011a). In the beginning of a multi-goal project, motivation is high, and comes from the desire to accomplish all tasks using means that follow previously established rules. This is called the *means-focused motivation*. With time the motivation to pursue the main goal decreases due to one or more factors such as diminished sense of goal achievability, decline of physiological and psychological resources (depletion) after first tasks are completed, positive goal-related emotions that reduce the efforts toward the main goal and shift the focus to another goal, and choosing to relax initial standards to save time and resources. The *outcome-focused motivation* prevails in this phase, which is the bottom of the U-shaped motivation curve. As the main goal becomes closer, the motivation increases again from the desire to accomplish all tasks using means that follow previously established rules as in the beginning of the project. Stronger adherence to previously established rules and standards at the beginning and end of a project is possibly explained by the fact that the concerns about the image that one makes of himself (self-signalling concerns) usually are greater in these two phases of goal pursuit (TOURÉ-TILLERY; FISHBACH, 2011b).

7.4 CASE STUDY

A case study was conducted using a database composed of $n = 340$ soil observations made between 2008 and 2011 in a small catchment ($20\,\mathrm{km}^2$) located in the southern border of the Plateau of the Paraná Sedimentary Basin, in the city of Santa Maria, state of Rio Grande do Sul, Brazil (Figure 7.1). Soil observations were made for the purpose of soil and land use mapping, as well as modelling soil carbon stocks and vulnerability to erosion (MIGUEL, 2010; SAMUEL-ROSA, 2009; MOURA-BUENO, 2012; MIGUEL, 2013).

Local climate is classified as Cfa (Köppen climate classification – subtropical humid without a dry season) with mean annual temperature of 19.3 °C, and mean annual precipitation of 1708 mm well distributed throughout the year (MALUF, 2000). Relief varies between plain (slope between 0 and 3 %) and mountainous (slope between 45 and 100 %), and elevations range between 139 and 475 m above sea level. There are three main geological formations which consist of (a) basic, intermediate and acid igneous rocks (rhyolite-rhyodacite and andesite-basalt) of the Cretaceous period; (b) consolidated sedimentary rocks (aeolian and fluvial sandstones) of the Triassic and Jurassic periods; and (c) non-consolidated (fluvial and colluvial deposits) of the Quaternary period (GASPARETTO et al., 1988; MACIEL FILHO, 1990; SARTORI, 2009).

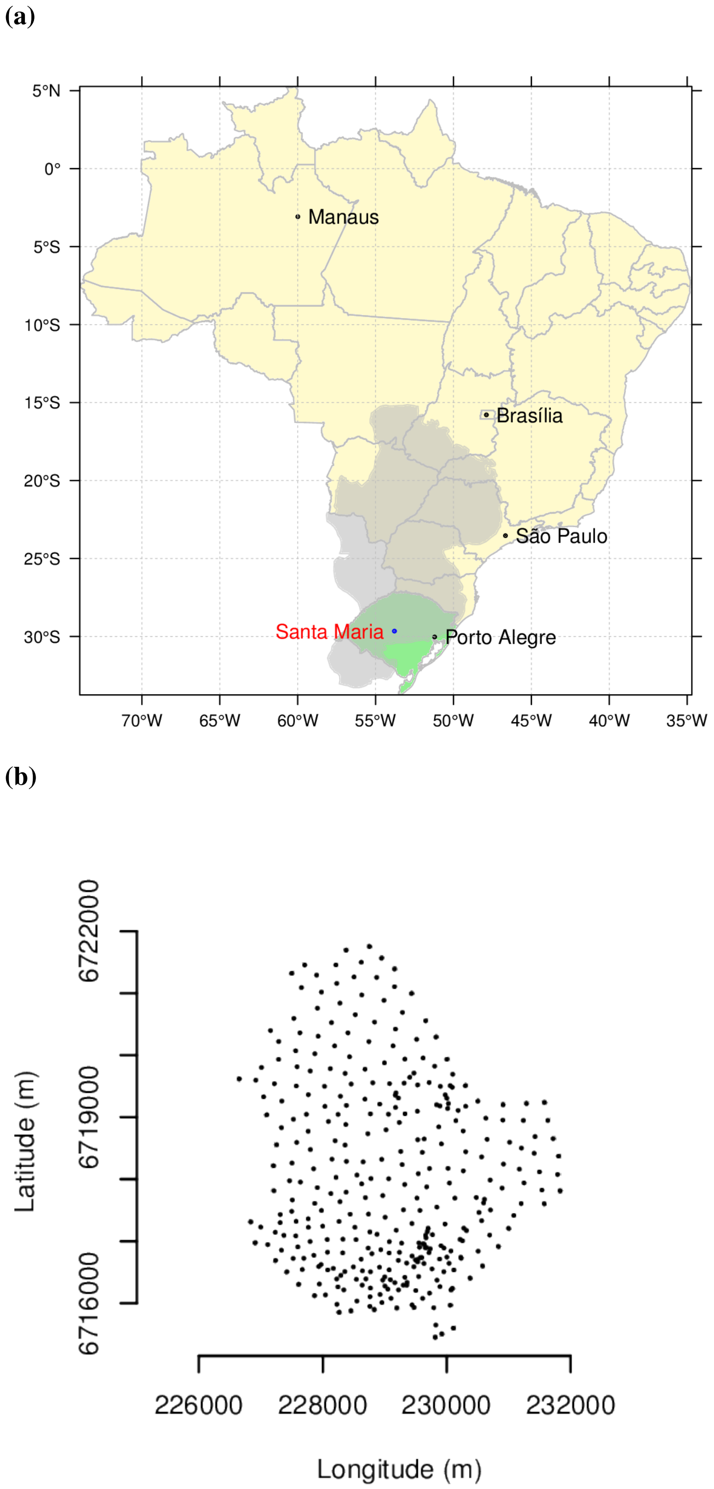


Figure 7.1: Location of the study area (a) in the central region of the Southernmost Brazilian state, Rio Grande do Sul, and the planar point pattern (b) composed of $n = 340$ soil observations made between the years of 2008 and 2011.

Forest areas occupy more than half of the study area, followed by native grassland, shrubland, farmland, forestry, urban areas and artificial water bodies (SAMUEL-ROSA et al., 2011).

7.4.1 Elicitation of Expert Judgements

Starting knowledge to understand the soil observation process was obtained eliciting the judgements of experts. An expert is every individual who has a deep, comprehensive knowledge about the object under study (MEYER; BOOKER, 2001), specifically, every soil modeller that helped planning and conducting the field campaigns that yielded the $n = 340$ soil observations in the study area. Due to operational constraints, the elicitation was carried out using email-based discussions and written narratives. First, written narratives were obtained by making a bibliographic review of published material such as articles, reports and theses. All information gathered was processed by the authors to yield a starting description of the soil observation process.

In the next step, field soil modellers were asked to prepare individual written description of the soil observation process. In their description, soil modellers were asked to point to the aspects that influenced most the selection of observation locations, as well as describe all important events and/or changes that might occurred in during the development of their project. Their description was then used to improve the existing description prepared by the authors, which was then submitted to appreciation by the field soil modellers. This step was carried out using a behavioural method (ORSI et al., 2011), specifically, email-based group discussions coordinated by the authors. These were responsible for aggregating expert judgements and continuously improving the description of the soil observation process until a common point was reached.

The resulting starting description of the soil observation process was used to define a set of covariates that could potentially be included in the point process model. These covariates are presented and described in Section 7.5.1 along with a summary description of the soil observation process, while their use to construct the point process model is explained next in Section 7.4.2.

7.4.2 Point Pattern Analysis

Point pattern analysis was carried out using the `R`-packages `spatstat` (BADDELEY, 2010) and `splancs` (ROWLINGSON; DIGGLE, 1993).

The first step consisted of estimating the observation window W, i.e. the limits of the spatial domain where soil observation were made. This was necessary because soil observation points are not constrained to the physical limits of the study area (Section 3.5). The estimate of W was entirely based on the spatial point pattern assuming that the observation points have been generated independently and uniformly distributed inside an unknown W. The maximum likelihood estimate of the unknown W is the rescaled convex hull H' of the observation points, with the scaling factor equal to $1/\sqrt{1 - \frac{m}{n}}$, where n is the number of observation points and m the number of vertices of the convex hull H (RIPLEY; RASSON, 1977).

The spatial distribution of observation points was investigated using two measures of distance between observation points. The first is the *empty space distance* or void distance, which measures the Euclidean distance from every single location in W to the nearest existing observation point. The second is the *nearest neighbour distance* (NND), i.e. the Euclidean distance from every existing observation point to its nearest existing observation point. The latter was used to construct an *Stienen diagram*, which consists of depicting every point using a circle with diameter and colour intensity proportional to its NND. The Stienen diagram was visualised in Google Earth® to check for possible correlations between environmental conditions and nearest neighbour distances.

The temporal variation of NND was evaluated computing summary statistics based on the order in which soil observations were made. The first of them was the auto-correlation coefficient of first order, $r(1)$, which measures the strength of the interaction between a soil observation made at time t and the soil observation made at time $t+1$ (PRUSCHA, 2013). Next, a moving average with a window size of $n = 34$ observation points was calculated. Finally, a third degree polynomial was adjusted to NND to describe the structure of the long-term temporal variation.

Evaluation of the spatial distribution of observation points suggested the observation intensity λ, i.e. the average density of observation points per unit area, to be spatially inhomogeneous. Contrary to the homogeneous case, where the intensity would be a function of the total number of observation points and the area of W, the inhomogeneous case requires estimating an *intensity function* that accounts for the spatial variation of the intensity of the spatial point process. Here, the intensity function was estimated nonparametrically by kernel smoothing. Specifically, an isotropic Gaussian kernel with edge effect correction was employed, the bandwidth of the Gaussian kernel adjusted to minimize the square-mean-error criterion (DIGGLE, 1985).

The hypothesis that the spatial point pattern is completely random was tested comparing the fit between observed, $\widehat{G}$, and theoretical distribution functions of the nearest neighbour distance, G. For a spatial point pattern that conforms with the hypothesis of complete spatial randomness (CSR), the relation between observed and theoretical distribution functions of the nearest neighbour distance is approximately linear. Values $\widehat{G} > G$ suggest a clustered (aggregated) spatial point pattern, while values $\widehat{G} < G$ suggest a inhibited (regular) spatial point pattern. Monte Carlo simulations were used to assess the significance of departures from the theoretical distribution function. The simulation consisted of creating $s = 99$ independent realizations of a spatial point pattern with $n = 340$ observation points independently and identically distributed with inhomogeneous intensity on the observation window. Inhomogeneity was achieved using point estimates of observation intensity calculated with a leave-one-out Gaussian kernel smoother with edge effect correction (BADDELEY et al., 2000). The empirical distribution function of the nearest neighbour distance of all $s = 99$ realization were then used to compute the global upper and lower simulation envelopes. Here, we reject the null hypothesis of CSR if the observed distribution function of the nearest neighbour distance lies outside the envelope at any moment with an exact significance level $\alpha = 1/(s+1) = 1/100 = 0.01$.

Finally, assuming that the observation points have been generated independently of each other inside W, an inhomogeneous (non-stationary) Poisson point process model was calibrated to estimate the intensity function. A series of spatially exhaustive covariates was used to define explanatory (deterministic) variables of the intensity function. These covariates served as proxies of the factors that determined the selection of observation locations by field soil spatial modellers. For that end, a generalized linear model with a log link was fitted by maximizing the pseudolikelihood using the Berman-Turner computational approximation (BADDELEY, 2010). The quality of the fit was evaluated using an analysis of deviance table and visually by comparing the nonparametric and parametric estimates of the intensity function.

7.5 RESULTS

7.5.1 Elicited Expert Knowledge

According to the information gathered, the soil modellers had little experience with soil spatial modelling and this was the first time they were responsible for planning and making soil observations. Their expectation was to perform approximately $n = 500$ observations, given the available infrastructure, human resources and financial resources. The goal was to obtain a "satisfactory" coverage of both attribute and geographic spaces, with emphasis on the former. A mental stratification of the area was made to help achieve this goal. First, the area was divided into three strata according to geomorphological features: low elevation areas with flat and gently sloping terrain, steep slopes, and high elevation areas with gently sloping terrain. These strata represent three common physiographic regions of Southern Brazil called, respectively, Central (or Peripheral) Depression, Plateau Border and Plateau.

With their goal in mind, soil modellers started field campaigns in the Southern sector of the study area in 2008. This area represents the Central Depression, were soil modellers had less knowledge of soil/landscape relationships. Access to most sampling sites was granted by the absence of geographic barriers and presence of a dense road network. After a few field campaigns, the soil modellers noticed that available resources and infrastructure would not allow visiting $n = 500$ sites. Access restrictions were imposed by some landowners and mean access time to observation locations was larger than expected. Besides, a budget shortage imposed important restrictions to the continuity of the study. As a consequence, the goal and planning of field campaigns had to be changed. The most sensible change was the decrease of observation intensity, requiring the approximate location of observation sites to be predefined beforehand at approximately equally spaced distances to obtain a satisfactory spatial coverage. A final outcome of at least $n = 300$ soil observations became the new focal goal of the soil modellers.

Next field campaigns took place in the steepest areas of the study area. These areas represent the Plateau Border and possess severe access restrictions due to strong slopes and dense forest cover. Access restrictions were again imposed by some landowners. Following, field campaigns were carried out in the Northern and North-eastern sectors of the study area. These areas represent the Plateau, where the soil modellers had a better knowledge of soil-landscape relationships. Overall, there were only minor access issues due to geographic barriers and very few restrictions were imposed by landowners. More field campaigns were carried out in the areas of the Plateau Border but avoiding difficult-to-access areas (steep slopes and dense forests). In the last field campaigns, with a small amount of resources still available, the soil modellers performed a few more observations in some areas of the Plateau Border and Central Depression where their knowledge about the soil-landscape relationships was poorer. This yielded the final outcome of $n = 340$ soil observations.

7.5.2 Point Pattern Analysis

7.5.2.1 Observation intensity

There are two regions where soil observation was more intense. The first and largest of them is located in the South-western sector, while the second is located in the Middle-North-eastern sector, with the largest values found in the South-western sector. Differences in observation intensity resulted in the occurrence of patches with poor geographic coverage (Figure 7.2). The largest patches are located in the Middle and Middle-Eastern sectors.

The Stienen diagram (click here to download and open in Google Earth®) helped identifying areas where the observation pattern is approximately regular, such as the Northern sector where the circles are approximately aligned and have about the same size. In the Southern sector, where the size of the circles is variable, the observation process seem to be approximately random. The relation between observation intensity and environmental features is also very clear. There is a strong relation between nearest neighbour distance and topography. Overall,

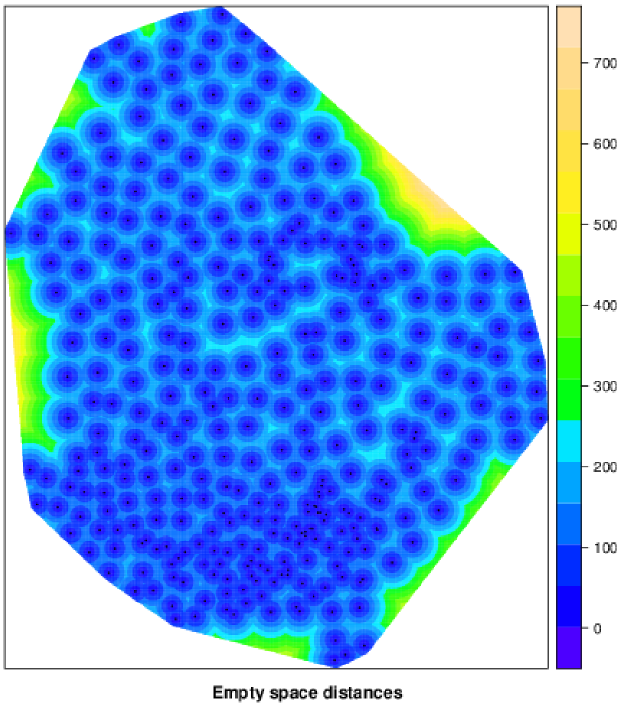


Figure 7.2: Estimated empty space distances. Values range from 1 to 709 m.

the smallest nearest neighbour distance values seem to occur in the Central Depression and increase towards the Plateau Border (which has a dense forest cover) and the Plateau.

Observation intensity is also related to the temporal order in which observations were made (Figure 7.3). First observations were made at short distances, resulting in a higher observation intensity. With time, nearest neighbour distance started to increase. This increase occurred until about the 150th soil sample, when the nearest neighbour distance reached about 250 m and remained approximately constant until the 300th sample. After the 300th observation, the nearest neighbour distance started to decrease and reached values around 50 m.

7.5.2.2 Spatial distribution

The estimated inhomogeneous G function of the spatial distribution of the point process is shown in Figure 7.4. The curve of the empirical point process (continuous black line) follows a different pattern than the theoretical curve (dashed red line) of a completely random spatial point process. This result supports the initial understanding that the point pattern under analysis can be the realization of a deterministic process. Figure 7.4 also shows the envelope of the estimated inhomogeneous G function built with $n = 99$ Monte Carlo simulations. The global statistical test confirms that observations with a nearest neighbour distance smaller than about 125 m have a random pattern of spatial distribution. At nearest neighbour distances above 125 m the empirical curve shows a strong deviation from the envelope, indicating that the spatial distribution of the soil observations is approximately regular. Because there is a significant difference between the empirical and theoretical curves, the point process under analysis can be regarded as not being a realization of complete spatial randomness, but of an yet unspecified point process (BADDELEY, 2010).

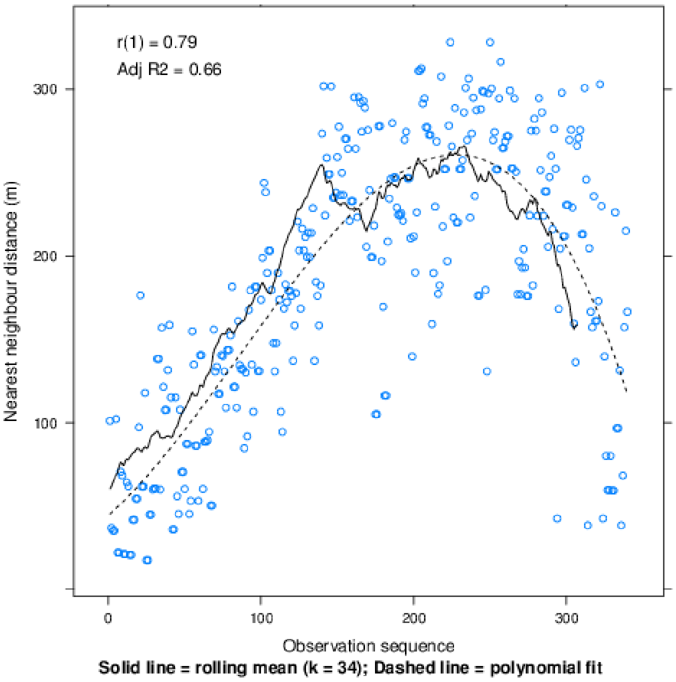


Figure 7.3: Estimated nearest neighbour distances (NND) as related to the temporal order in which soil observations were made. Twenty two field campaigns were carried out to make the $n = 340$ soil observations available. Deeper information about field campaigns is given in Figure 7.5. Function `rollmean` from `R`-package `zoo` was used to estimate the rolling mean using a window of $n = 34$ points (ZEILEIS; GROTHENDIECK, 2005).

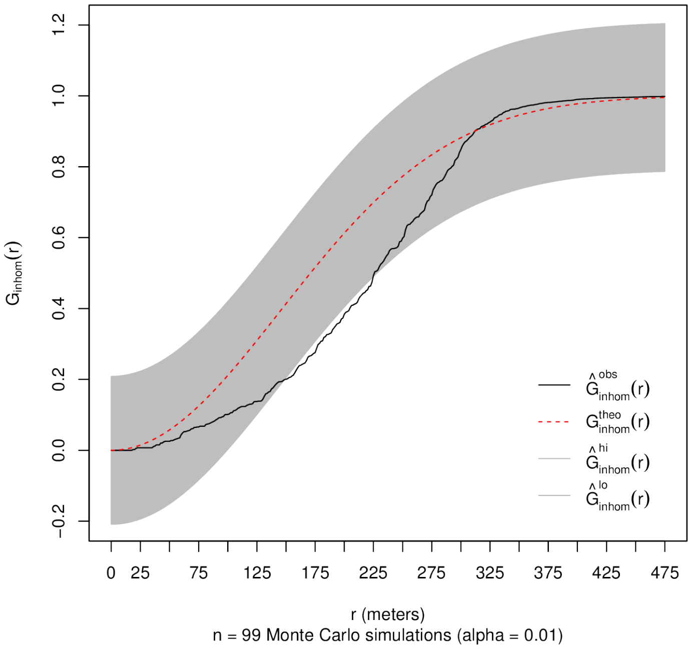


Figure 7.4: Spatial distribution of the planar point pattern estimated with the inhomogeneous G function and its global envelope. Note that at NND smaller than about 125 m the point pattern can be regarded as having a random pattern of spatial distribution. These observations were made in first and last field campaigns as shown in Figure 7.3.

7.5.2.3 Poisson point process model

Three potentially explanatory covariates were suggested by the judgements elicited from the experts to be included in the Poisson model to explain the point pattern: land use (seven classes) (SAMUEL-ROSA et al., 2011), terrain attributes derived from the sink-filled Shuttle Radar Topography Mission version 4 digital elevation model (REUTER et al., 2007), and field campaigns (indicator variables representing the 22 campaigns, respectively). The terrain attributes elevation, morphometric protection index and topographic position index were used to stratify the area into three physiographic strata, namely Central Depression, Plateau Border, and Plateau. A Dirichlet tessellation of the point pattern was computed to represent the field campaigns in the space domain.

Among all covariates available, land use, physiographic strata and field campaigns (Figure 7.5) are those which better explain the spatial distribution of the point process (Table 7.1). This corroborates the interpretation of the Stienen diagram (plotted in Google Earth®) and Figure 7.3 made above. Land use produces the largest deviance reduction, followed by field campaigns and physiographic strata. The interactions between predictors also reduce the deviance and were not included in the model to avoid increasing its complexity.

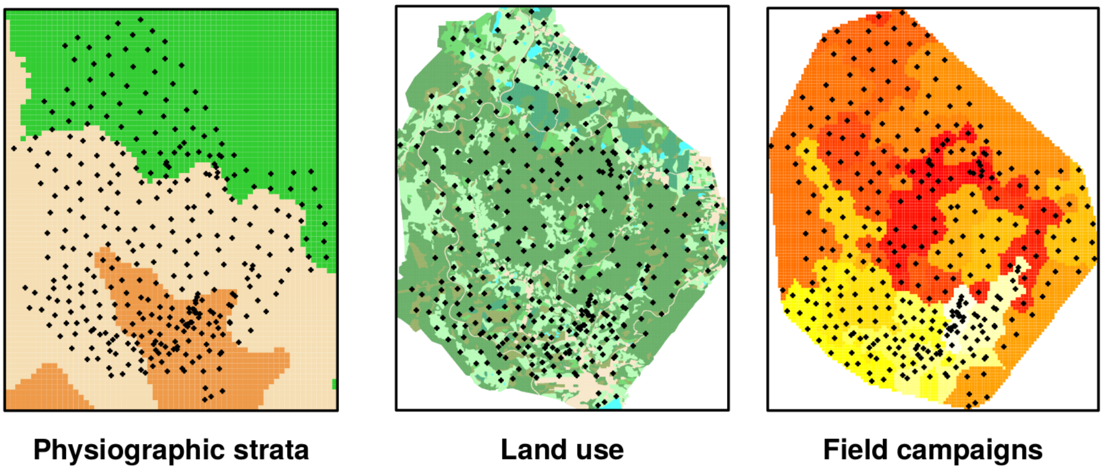


Figure 7.5: Covariates used to fit the non-stationary Poisson point process model superimposed with the planar point pattern. Physiographic strata include (South-North aligned): Central Depression, Plateau Border, and Plateau. Land use includes: native forest (more than 50 % of the area), shrubland, animal husbandry, crop agriculture, forestry, urban, and water. Field campaigns has 22 levels represented with increasing colour heat (white to red) from the first to the last campaign.

Estimated coefficients for predictor variables show that observation intensity has a decreasing tendency in the Plateau Border and Plateau (Table 7.2). The lowest observation intensity occurs in areas of native forests, urban land use and water bodies. All field observation campaigns were less intense than the first. Observation intensity reduction factor was similar between the second and seventh field campaigns (about -0.75 times). The largest reduction occurred in the 13th field campaign (-2.18 times), when observation intensity started to increase again until the last field campaign.

The fitted non-stationary Poisson point process model gives a fine representation of the observation intensity estimated with the isotropic Gaussian kernel (Figure 7.6). Both sectors of high observation intensity are correctly predicted. However, relative intensity values are strongly over-predicted in the South-western sector, while in border areas they are strongly

Table 7.1: Analysis of deviance for the fitted non-stationary Poisson point process model with two-tailed p-values for the chi-squared tests comparing the reduction in deviance due to the inclusion of each predictor variable. Terms were added sequentially to the model (first to last).

Predictor	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
Intercept only			3240	1943.05	
Physiographic strata	2	38.03	3238	1905.03	5.524e-09
Land use	6	120.91	3232	1784.12	< 2.2e-16
Field campaigns	21	63.28	3211	1720.84	4.027e-06

under-predicted. These under-predicted areas seem to be correlated to estimated empty space distances (Figure 7.2).

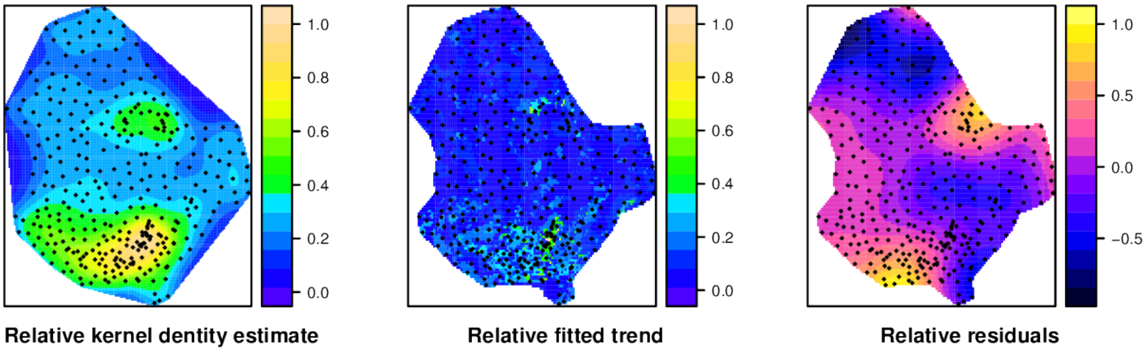


Figure 7.6: Relative empirical kernel density estimate (left), and relative trend (centre) and residuals (right) of the fitted non-stationary Poisson point process model. Data shown is relative to the largest absolute estimated value of intensity. Relative kernel density values range from 0.02 to 1.00, relative trend values range from 0.01 to 1.00, and relative residual values range from -1 to 0.81.

7.6 DISCUSSION

The spatial distribution of the point process has features resulting from the influence of three key factors. The first is *conceptual* and regards the knowledge of the soil modellers about soil-landscape relationships. The second factor is *operational* and relates to the available infrastructure, human resources and budget to make soil observations, as well to access restrictions imposed by landowners and geographic barriers. The last factor is *psychological*, which is also affected by the first two and is related to how the soil modellers perceive their surrounding environment and how the course of their motivation shifted during the soil observation process. The next three sections are devoted to better understand these factors.

7.6.1 Concentrating on Problem Areas

Soil observations were made during studies organized by young soil modellers. Both soil modellers had just started their careers and had more experience with well drained, deep and high iron oxide content soils derived from igneous rocks and occurring on gently sloping

Table 7.2: Estimated coefficients for the fitted non-stationary Poisson point process model and their lower and upper limits at the 95 % confidence interval.

Predictor^{a,b}	Estimate	Standard error	Z test	Lower limit	Upper limit
(Intercept)	-10.37	0.24	na	-10.85	-9.90
Plateau border	-0.03	0.20		-0.42	0.37
Plateau	-0.07	0.31		-0.67	0.53
Shrubland	1.17	0.18	***	0.82	1.52
Animal husbandry	0.95	0.14	***	0.67	1.23
Crop agriculture	1.59	0.19	***	1.22	1.96
Forestry	1.10	0.28	***	0.56	1.64
Urban	-1.12	0.52	*	-2.13	-0.11
Water	0.74	0.72		-0.68	2.16
Field campaign 2	-0.70	0.32	*	-1.32	-0.08
Field campaign 3	-0.68	0.31	*	-1.29	-0.07
Field campaign 4	-0.75	0.32	*	-1.37	-0.13
Field campaign 5	-0.85	0.39	*	-1.60	-0.09
Field campaign 6	-1.04	0.37	**	-1.77	-0.31
Field campaign 7	-0.74	0.43		-1.58	0.09
Field campaign 8	-1.36	0.34	***	-2.03	-0.69
Field campaign 9	-1.56	0.43	***	-2.40	-0.73
Field campaign 10	-1.49	0.38	***	-2.24	-0.75
Field campaign 11	-1.44	0.41	***	-2.24	-0.63
Field campaign 12	-1.39	0.36	***	-2.10	-0.68
Field campaign 13	-2.18	0.44	***	-3.04	-1.32
Field campaign 14	-1.64	0.45	***	-2.53	-0.75
Field campaign 15	-1.90	0.36	***	-2.61	-1.19
Field campaign 16	-1.93	0.44	***	-2.79	-1.07
Field campaign 17	-1.20	0.38	**	-1.94	-0.46
Field campaign 18	-1.25	0.41	**	-2.06	-0.45
Field campaign 19	-1.14	0.51	*	-2.15	-0.13
Field campaign 20	-1.33	0.41	**	-2.13	-0.53
Field campaign 21	-1.35	0.38	***	-2.10	-0.60
Field campaign 22	-0.28	0.42		-1.10	0.54

^{a} Physiographic strata has three levels: Central Depression, Plateau Border and Plateau. Land use has seven levels: native forest, shrubland, animal husbandry, crop agriculture, forestry, urban and water. Field campaigns has 22 levels represented with increasing number from the first to the last campaign.

^{b} Significance levels of the two-tailed Z test against the null hypothesis that each regression coefficient is equal to zero are given for p-values of 0 (***), 0.001 (**), 0.01 (*), 0.05 ( ) and 1 (na).

to sloping relief. This is the same type of soil and relief that prevails in the Northern and North-eastern sectors of the study area. Because the soil modellers started the study using the free survey method, locating soil observations to test hypotheses posited according to their mental model of soil-landscape relationships, more observations were made on the so called “problem areas” (ROSSITER, 2000). Problem areas are those areas for which the mental model of soil-landscape relationships is incomplete or possesses significant weaknesses, i.e. spatial soil variation is poorly predicted. Thus, it could be expected beforehand that the soil modellers would concentrate their efforts in Central Depression and Plateau Border areas. This is clearly evidenced by the higher observation intensity in the Central Depression and some areas of the

Plateau Border, while the Plateau has a lower observation intensity.

7.6.2 Managing Available Resources

Field work is the main component of any soil mapping effort (KEMPEN et al., 2012). Therefore, it demands an efficient planning of field campaigns to guarantee that the available infrastructure, human resources and budget are enough to accomplish the desired observation intensity. When the free survey method is used, planning of field campaigns strongly relies on the experience of the field soil modellers. Because the field soil modellers working in the study area had little field experience, it soon became clear to them that field campaigns were inefficiently planned. Overall, the field soil modellers underestimated the costs of variable components of field campaigns. One of these variable components is the access time to observation locations, which usually has a large impact on sampling costs (DOMBURG et al., 1997), especially in areas with many accessibility constraints.

Inefficient planning of field campaigns plus budget cuts and operational issues related to available infrastructure and human resources lead to the need for redefining the initially aimed total number of soil observations. Because the study started on problem areas, there is an *operational* bias towards employing stronger observation efforts in these areas, adding to the *conceptual* effect described above (See Section 7.6.1). However, operational issues also lead to the reduction of the observation intensity on poorly accessible areas, such as those with steep slopes and dense forest cover of the Plateau Border. In classical sampling theory an observation process strongly influenced by accessibility issues is called *convenience sampling* (DE GRUIJTER et al., 2006). The most extreme case occurs when soil observations are made only along the road network (CAMBULE et al., 2013). This is not the case of the observations made in the Plateau Border, but there is enough evidence to regard them as being the outcome of convenience sampling. Datasets obtained through convenience sampling usually present significant biases that affect the construction of robust soil mapping models (BRUS et al., 2011). Therefore, it can be expected beforehand that most soil mapping models built for the study area will have a poor performance in the Plateau Border areas.

7.6.3 Neurophysiological Responses

7.6.3.1 Spatial enclosure

When making soil observations, enclosed places can be biasedly oversampled due to the effect of the spatial enclosure on the way that field soil modellers perceive the distances in their surrounding environment. If this hypothesis is correct, then it can be expected that the two soil modellers would have located soil observations at shorter distances in the Central Depression than in the Plateau, resulting in a denser soil observation in the former. The same explanation is valid for the denser soil observation made in the last field campaigns carried out in a densely forested rugged terrain. Unfortunately the most expressive effect of the spatial enclosure occurs in the same places where the soil modellers had a poorer knowledge of soil-landscape relationships (Southern and Middle-Eastern sectors) and were not aware yet of the incompatibility between the available resources and their goals (Southern sector).

7.6.3.2 Motivation shifts

The U-shaped pattern of multi-goal pursuit shows a good fit to the soil observation process carried out by the field soil modellers in the study area. A strong motivation to follow well-known guidelines for soil observation existed in the first field campaigns. The soil modellers had the goal of obtaining a coverage of both attribute and geographic spaces to refine their knowledge of soil-landscape relationships and obtain a dataset that provided the means to make accurate predictions of the spatial variation of soil properties. Besides, this was one of the first studies under their complete responsibility, powering their initial motivational status. The possibility of understanding new soil-landscape relationships represented a pleasant challenge for the young soil modellers. With time the soil modellers faced accessibility constraints that started depleting their physiological and psychological resources. The situation became worst when the soil modellers realized that the available resources and the initial goal were incompatible because the costs were underestimated and due to budget cuts. They were forced to shift their focus to the outcome, i.e. obtaining a “satisfactory” coverage of the geographic space while keeping the costs of the study below the available amount of resources available. In other words, they had to relax their initial standards to save resources (physiological, psychological and financial). The result was the reduction of the observation intensity. A new motivation shift occurred when the end of the project became closer and the soil modellers perceived that their resources were not completely depleted. Problem areas were visited again and the free survey method used to make new soil observations in a very similar way of the first field campaigns.

7.6.4 Observation Model

The approach presented here helped formalizing a verbal representation of the mental model of the soil scientists who produced the set of $n = 340$ soil observations in the study area. Figure 7.7 shows one of the possible ways of depicting this model. Nearest neighbour distance is used as a quantitative indicator of the progress of the observation process.

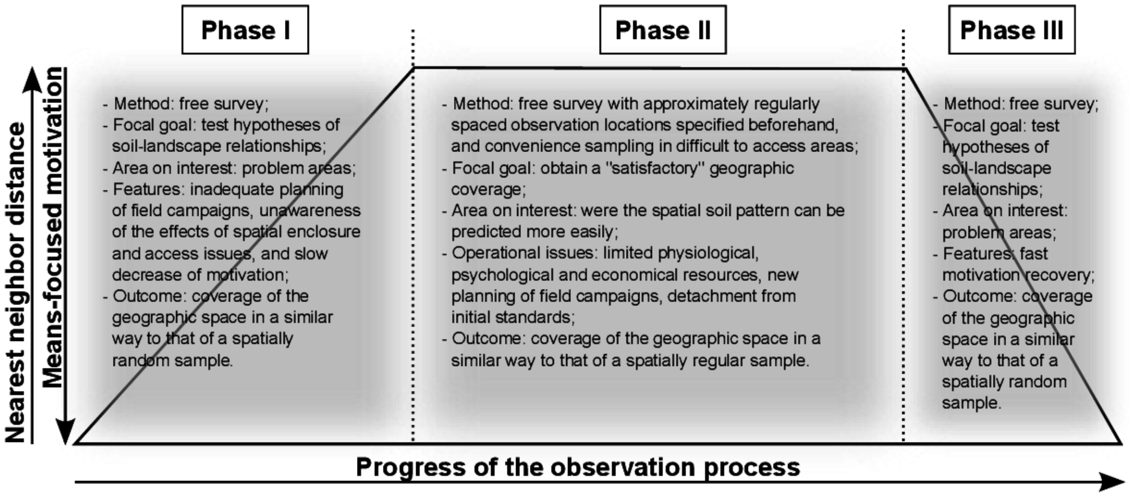

Phase I
Phase II
Phase III
Nearest neighbor distance
Means-focused motivation
- Method: free survey;
- Focal goal: test hypotheses of soil-landscape relationships;
- Area on interest: problem areas;
- Features: inadequate planning of field campaigns, unawareness of the effects of spatial enclosure and access issues, and slow decrease of motivation;
- Outcome: coverage of the geographic space in a similar way to that of a spatially random sample.
- Method: free survey with approximately regularly spaced observation locations specified beforehand, and convenience sampling in difficult to access areas;
- Focal goal: obtain a "satisfactory" geographic coverage;
- Area on interest: were the spatial soil pattern can be predicted more easily;
- Operational issues: limited physiological, psychological and economical resources, new planning of field campaigns, detachment from initial standards;
- Outcome: coverage of the geographic space in a similar way to that of a spatially regular sample.
- Method: free survey;
- Focal goal: test hypotheses of soil-landscape relationships;
- Area on interest: problem areas;
- Features: fast motivation recovery;
- Outcome: coverage of the geographic space in a similar way to that of a spatially random sample.
Progress of the observation process

Figure 7.7: Theoretical model of soil observation depicted using the proposed approach that includes elicitation of expert knowledge, point pattern analysis and articulation of psychological theories of perception and motivation. Phases I, II and III are guided by, respectively, means-, outcome- and means-focused motivation.

In the first phase the soil modellers employed the free survey method in its strict sense. The area was stratified into three primary observation units (physiographic strata) and many secondary observation units (land use type). The soil modellers were strongly motivated (means-focused motivation) and started observing the soil in problem areas. But they were unaware of the effects of spatial enclosure, access issues and of their inexperience which resulted in an inadequate planning of field campaigns. When the soil modellers became aware of these effects (after a few field campaigns had been already carried out), the main goal of the project had to be reviewed and planning of field campaigns reformulated. This marks the end of the first phase of the observation model, which had as outcome a point pattern covering the geographic space in a fashion similar to that of a spatially random sample.

The second phase of the observation model starts when the soil modellers are consistently less motivated than when they started the observation process. This decreased motivation comes along with a new focal goal: making a minimum number of observations to obtain a "satisfactory" geographic coverage of the area. This is achieved by reformulating the planning of field campaigns. The free survey still is the observation method used but with the location of soil observations made beforehand at approximately equally spaced intervals. In the field the soil modellers are free to change this location according to their mental model of soil-landscape relationships but respecting the previously established spacing between observations. They also move from problem areas to those were the spatial soil pattern can be predicted more easily. When strong access issues are faced, convenience sampling is employed. The objective is to save physiological, psychological and financial resources. When the soil modellers realized that they had reached the new focal goal and that their resources were not completely depleted yet, a new effort was employed to better understand problem areas. This marks the end of the second phase of the observation model, where the soil observation process was guided by an outcome-focused motivation. The outcome of this phase is a point pattern covering the geographic space in a similar way to that of a spatially regular sample.

The last phase of the observation model starts when the soil modellers have a renewed motivation to make soil observations in problem areas using the free survey method in its strict sense. Soil observations are made while there are resources available. Field campaigns are better planned now because the soil modellers have gained field experience. But the effects of spatial enclosure can still be present. The outcome of this phase is a point pattern which has similar features to that produced in the first phase, i.e. covers the geographic space in a fashion similar to that of a spatially random sample. The main difference is that the number observations made is smaller as was the amount of resources available.

7.7 CONCLUSIONS

Several factors influence how field soil spatial modellers decide upon where to place soil observation locations. These are of three types: conceptual, operational, and psychological. The first concerns the knowledge of the soil spatial modellers about soil-landscape relationships, and seems to be connected with the years of field experience. The second relates to the available resources (infrastructure, workforce, and budget) to make soil observations, as well with access restrictions imposed, for example, by landowners and geographic barriers. The third relates to how the soil modellers perceive their surrounding physical environment and how the course of their motivation shifts during the soil observation process. Point pattern analysis helped understanding that there is a trade-off between conceptual and operational factors, which determines how the motivation of field soil modellers shifts towards one or another focal goal. Depending on the focal goal, the resulting point pattern resembles a random (testing hypotheses of soil-landscape relationships – means-focused motivation) or a regular (maximizing the number of observations – outcome-focused motivation) spatial sample. Understanding the reasons behind the location of soil observations in free survey can help soil spatial modellers designing more efficient data-driven sampling strategies.

8 CHAPTER VII

OPTIMIZATION OF SAMPLE CONFIGURATIONS FOR SPATIAL TREND ESTIMATION FOR SOIL MAPPING

* This chapter is based on the study *spsann – optimization of sample patterns using spatial simulated annealing*, presented at the EGU General Assembly 2015 (SAMUEL-ROSA et al., 2015), and *Optimization of sample configurations for spatial trend estimation*, presented at Pedometrics 2015 (SAMUEL-ROSA et al., 2015). Also collaborated in the preparation of this document: Dick J. Brus (Alterra, Wageningen University and Research Centre, the Netherlands), Gerard B. M. Heuvelink (ISRIC – World Soil Information), Gustavo M. Vasques (Embrapa Soils, Brazil), and Lúcia Helena Cunha dos Anjos (Universidade Federal Rural do Rio de Janeiro, Brazil).

8.1 RESUMO

A tendência espacial é a parte de um modelo espacial do solo que explica, de maneira deterministica, a variação espacial do solo usando covariáveis. O método de amostragem mais comumente usado para identificar e estimar a tendência espacial é a *amostragem no hipercubo latino condicionado* (CLHS). Neste estudo nós propomos melhorias conceituais e algorítmicas no CLHS, as quais são avaliadas utilizando dados sintéticos derivados de um estudo de caso do mundo real, em Santa Maria, sul do Brasil. As melhorias incluem: 1) usar o r de Pearson somente quando todas as covariáveis são numéricas, e o V de Cramér quando algumas ou todas as covariáveis são fatores, 2) definir os estratos marginais de amostragem utilizando apenas os valores únicos dos quantis amostrais estimados com uma função descontínua, e 3) escalar as funções objetivo para o mesmo intervalo aproximado de valores usando a abordagem do limite superior-inferior com os valores máximo e mínimo de Pareto antes de agregá-las em uma única função de utilidade usando uma soma ponderada. Em comparação com os CLHS original, as modificação propostas resultaram em um algoritmo de amostragem com melhor comportamento numérico, mas isso não se traduz necessariamente em maior acurácia na predição. O tamanho da amostra tem uma influência maior na acurácia da predição do que o algoritmo de amostragem. No entanto, otimizar configurações amostrais visando a associação/correlação entre as covariáveis pode degradar a acurácia da predição.

Palavras-chave: Recozimento simulado. Otimização multi-objetivo. Pareto. Amostragem por hipercubo latino condicionado. Covariáveis.

8.2 ABSTRACT

The spatial trend is the part of a soil spatial model that deterministically explains the soil spatial variation using covariates. The sampling method most commonly used to identify and estimate the spatial trend is *conditioned Latin hypercube sampling* (CLHS). In this study we propose conceptual and algorithmic improvements on CLHS, which are evaluated using synthetic data derived from a real-world case study in Santa Maria, southern Brazil. The improvements include: 1) using the Pearson's r only when all covariates are numeric, and the Cramér's V when some or all covariates are factors, 2) defining marginal sampling strata using only the unique values of the sample quantiles estimated with a discontinuous function, and 3) scaling the objective functions to the same approximate range of values using the upper-lower bound approach with the Pareto maximum and minimum before aggregating them into a single utility function using a weighted sum. Compared to the original CLHS, our proposed modifications resulted in a sampling algorithm with an improved numerical behaviour, but this does not necessarily translates into improved prediction accuracy. Sample size has a larger influence on prediction accuracy than the sampling algorithm. However, optimizing sample configurations aiming at the association/correlation between covariates can degrade prediction accuracy.

Keywords: Simulated annealing. Multi-objective optimization. Pareto. Conditioned Latin hypercube sampling. Covariates.

8.3 INTRODUCTION

Modern soil mapping is based on using a model of spatial variation composed of two terms,

$$Y(\boldsymbol{s}) = m(\boldsymbol{s}) + e(\boldsymbol{s}). \tag{8.1}$$

The first term in the right-hand size in Equation 8.1 is the spatial trend, which corresponds to the spatial variation of the soil property $Y(\boldsymbol{s})$ that is explained deterministically using spatially exhaustive covariates; the remaining spatial variation of $Y(\boldsymbol{s})$ is explained stochastically with the second term (CRESSIE, 1993). Soil scientists devoted all their attention to $m(\boldsymbol{s})$ for more than a century (JENNY, 1961; FLORINSKY, 2012). Post-war technological developments in the fields of mathematics, statistics, and informatics, made many soil scientists turn their focus to $e(\boldsymbol{s})$ (WEBSTER; OLIVER, 1990). Recent developments in remote sensing and machine-learning algorithms made those soil scientists shift their attention back to $m(\boldsymbol{s})$ (MOORE et al., 1993) – but without forgetting $e(\boldsymbol{s})$ (ODEH et al., 1994) –, which now usually explains a considerably large proportion of the variation of $Y(\boldsymbol{s})$ compared to $e(\boldsymbol{s})$[24]. Besides, it is in $m(\boldsymbol{s})$ where we can incorporate most of our pedological knowledge (LARK, 2012).

Recent studies have shown that using more detailed covariates or more complex machine-learning algorithms can deliver more accurate soil maps, but the increase in prediction performance may be modest (SAMUEL-ROSA et al., 2015a) and largely depends on the calibration data (HEUNG et al., 2016). Limited to the currently available covariates and machine-learning algorithms, and to the existing pedological knowledge, one of the major operational issues that needs to be solved in any soil mapping project is how to design an efficient spatial sample to estimate $m(\boldsymbol{s})$. The sampling method most commonly used to solve this problem is the *conditioned Latin hypercube sampling* (CLHS). The CLHS was developed by Budiman Minasny and Alex McBratney at the University of Sydney in 2005, using an idea borrowed from the Latin hypercube sampling (MCKAY et al., 1979; MINASNY; MCBRATNEY, 2006). The popularity of the CLHS is due to its non-probabilistic nature, seen as a link with the sampling strategies used in "traditional soil survey", easiness to implement, and the high flexibility which makes the addition of new features simple (MINASNY; MCBRATNEY, 2010; ROUDIER et al., 2012; MULDER et al., 2013; CARVALHO JR et al., 2014; CLIFFORD et al., 2014).

The CLHS is a heuristic strategy of creating spatial samples that aim at three objectives: ($\mathcal{O}_1$) uniform coverage of the marginal distribution of numeric covariates (continuous and discrete data, e.g. elevation, slope, etc.), ($\mathcal{O}_2$) proportional sample sizes for the classes of factor covariates (binary, categorical, and ordinal data, e.g. geology, land use, etc.), and ($\mathcal{O}_3$) reproduction of the linear correlation of numeric covariates. The main idea was that if a spatial sample reproduces the marginal distribution of the numeric and factor covariates, as well as the correlation matrix of the numeric covariates, it will approximately cover the multivariate distribution of the covariates – this should put us closer to identifying the "true" spatial trend if we are (or assume to be) ignorant about its form.

Some critiques of the CLHS appeared in the literature since it was first published. Most of them focused on operational difficulties encountered in the field. For example, Cambule et al. (2013) argued that the CLHS is impractical in poorly-accessible areas, but Roudier et al. (2012) and Mulder et al. (2013) showed that this is just a matter of how the algorithm is implemented. Clifford et al. (2014) presented an algorithm for selecting an alternative sampling point when a CLHS sample point is inaccessible. Only recently soil scientists started paying

[24] Gerard Heuvelink shared the same opinion during his Richard Webster Medal speech at the conference of the Pedometrics Commission of the IUSS, which took place from 14–18 September 2015, in Córdoba, Spain.

more attention to the theoretical and algorithmic aspects of the CLHS. Minasny & McBratney (2010) demonstrated that, given an assumed known linear spatial trend, the CLHS is suboptimal. Clifford et al. (2014) questioned the importance of meeting the third objective ($\mathcal{O}_3$), as well as the mathematical approach used to find a solution for all three objectives jointly (see below). Finally, Brus (2015) proposed an alternative method for selecting Latin hypercube samples with known inclusion probabilities so that these samples can also be used for design-based inference.

Our objective is to propose conceptual and algorithmic improvements on the CLHS, all of which we describe in the next section. We then evaluate if the proposed improvements result in a more accurate representation of the feature space and in more accurate spatial predictions.

8.4 PROPOSED IMPROVEMENTS

8.4.1 Defining the Marginal Sampling Strata

Given a *numeric* covariate, CLHS uses the sample size n to define the number of marginal sampling strata c, i.e. $c = n$, and the interpolated sample quantiles to define the breakpoints of the c marginal sampling strata. The first objective of CLHS ($\mathcal{O}_1$) is to have exactly one sample point falling in each marginal sampling strata. However, depending on the level of discretization of the covariate values, CLHS may produce replicated breakpoints in the regions with a relatively high frequency of covariate values. For example, given a sample size of $n = 5$ and a covariate $\boldsymbol{a}$ with (ordered integer) values $\boldsymbol{a} = (1, 1, 1, 1, 2, 2, 3, 3, 4, 5, 8, 9, 9, 9, 9)$, the lower and upper boundaries of the $c = 5$ marginal sampling strata (mss) are $\boldsymbol{a}_{mss} = (1.0, 1.0, 2.6, 4.4, 9.0, 9.0)$. Because the marginal sampling strata in which a sample point b_i falls is evaluated using the indicator function

$$b_{sol_i} = \begin{cases} 1, & \text{if } a_{mss_j} \leq b_i \leq a_{mss_{j+1}} \text{ and } j = 1 \\ 1, & \text{if } a_{mss_j} < b_i \leq a_{mss_{j+1}} \text{ and } j > 1 \\ 0, & \text{otherwise} \end{cases}$$

where $i = 1, 2, \ldots, n$ refers to sample point candidates, and $j = 1, 2, \ldots, c$ to marginal sampling strata, the first and last marginal sampling strata (mms) of $\boldsymbol{a}$ will be empty, and the respective $n' = 2$ sample points will be allocated among the other three marginal sampling strata, with the set of allocation solutions $\boldsymbol{b}_{sol} = \{(0, 2, 1, 2, 0), (0, 1, 2, 2, 0), (0, 2, 2, 1, 0)\}$. Ergo, CLHS will be unable to find the globally optimum allocation solution $\boldsymbol{b}_{sol} = (1, 1, 1, 1, 1)$.

We propose defining the marginal sampling strata (mss) using only the unique values of the sample quantiles estimated with a discontinuous function (HYNDMAN; FAN, 1996). In our previous example, the strata boundaries would be $\boldsymbol{a}_{mss} = (1, 2, 4, 9)$. The number of sample points that should fall in each marginal sampling stratum is directly proportional to the number of sampling units (grids cells of a raster image) in that stratum of the covariate. For $\boldsymbol{a}$, this is $\boldsymbol{b}_{sol} = (2, 1, 2)$. The direct consequence of this modification is that, given a set of p covariates, each of them will potentially have a different number of (quasi-equal-size) marginal sampling strata, i.e. $c_i \leq n$, where $i = 1, 2, \ldots, p$. This will ultimately depend on the shape of their empirical frequency distribution, on the level of discretization of the covariate values, and on the sample size n.

8.4.2 Measuring the Association/Correlation Between Covariates

Two of the objectives of CLHS ($\mathcal{O}_1$ and $\mathcal{O}_3$) are concerned with *numeric* covariates, while only one ($\mathcal{O}_2$) focuses on *factor* covariates. $\mathcal{O}_1$ and $\mathcal{O}_2$ are mathematically equivalent – they aim at the coverage of the marginal distribution of the numeric and factor covariates, respectively –, and $\mathcal{O}_3$ measures the similarity between the population and sample correlation matrices of the numeric covariates as estimated with the Pearson's correlation coefficient r. The CLHS ignores the association among factor covariates, as well as among factor and numeric covariates. This means that CLHS gives more importance to numeric covariates. Such a bias cannot be corrected by simply attributing different *weights* to each objective (see below).

To address the problem above we propose to replace Pearson's r with Cramér's V

$$V = \sqrt{\frac{\chi^2/n}{min(n_{\text{col}} - 1, n_{\text{row}} - 1)}}, \tag{8.2}$$

where n_{row} and n_{col} are the number of rows and columns of the bivariate contingency table, n is the sample size, and χ^2 is the chi-squared statistic

$$\chi^2 = \sum_{i=1}^{n_{\text{row}}} \sum_{j=1}^{n_{\text{col}}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{8.3}$$

where O_{ij} and E_{ij} are the observed and expected frequencies, respectively, the marginal proportions of O being the maximum likelihood estimates of the marginal proportions of E (CRAMÉR, 1946; AGRESTI, 2002). The Cramér's V is a measure of association between factor covariates that ranges from 0 to $+1$: the closer to $+1$, the larger the association between two factor covariates. Accordingly, the only requirement for using the Cramér's V – instead of the Pearson's r – is that any numeric covariate be transformed into a factor covariate, with the factor levels defined using the marginal sampling strata. One could still use the Pearson's r when all covariates are numeric because computing the Cramér's V is more computationally demanding.

8.4.3 Aggregating the Objectives

Sampling for spatial trend estimation is a *multi-objective combinatorial optimization problem* (MOCOP): the spatial sample must meet a list of objectives among an almost infinite set of possible spatial samples. An important step for solving a MOCOP is to define each objective as a function, i.e. an *objective function* f_i (ARORA, 2011). An f_i associates a numerical value with each candidate spatial sample as a function only of the values of the p covariates used to describe the spatial domain – also known as *design variables* (ARORA, 2011) – at the n sample points. The lower the objective function value, the closer the spatial sample is to meeting the respective objective. Thus, when solving a MOCOP, one aims at minimizing the vector of k objective functions (ARORA, 2011)

$$\boldsymbol{f}(\boldsymbol{X}) = (f_1(\boldsymbol{X}), f_2(\boldsymbol{X}), \ldots, f_k(\boldsymbol{X})), \tag{8.4}$$

where $\boldsymbol{X}$ is the design matrix, a $n \times p$ matrix subject to the implicit constraints imposed by the finiteness of the spatial domain and discreteness of the p design variables. These implicit constraints define the set of values that can be assigned jointly to the design variables, i.e. the p-dimensional *feasible design space* $\mathcal{S}$, which, in turn, defines the set of numerical values that can be returned by the objective functions, i.e. the k-dimensional *feasible objective space* $\mathcal{Z}$ (MARLER; ARORA, 2004).

Ideally, there is a traceable unique *point cloud* $\boldsymbol{X}^*$ (i.e. a spatial sample with the values of the covariates at its sample points) that minimizes all objective functions simultaneously (MARLER; ARORA, 2009). However, in practice such a unique point cloud seldom exists, and if it exists it is hard to find. In most cases there is a large set of optima point clouds that map onto a set of optima points on $\mathcal{Z}$ because, for example, multiple point clouds can return the very same objective function value (ARORA, 2011). The set of optima point clouds is commonly defined using the concept of *Pareto optimality* (MARLER; ARORA, 2004): a point cloud $\boldsymbol{X}^*$ in $\mathcal{S}$ is Pareto optimum if and only if there is no other point cloud $\boldsymbol{X}$ in $\mathcal{S}$ that decreases the value of at least one objective function without increasing the value of another objective function.

A reasonable strategy to find a single optimum solution is to aggregate the objective functions into a single *utility function* U (MARLER; ARORA, 2005). The most common aggregation method is the *weighted sum* method, which is used in the CLHS. It employs weights to incorporate the *a priori* preferences of the user, their relative values reflecting the importance of each objective function (MARLER; ARORA, 2009). Thus, the MOCOP boils down to minimizing the *convex* combination of objective functions

$$U = \sum_{i=1}^{k} w_i f_i(\boldsymbol{X}), \tag{8.5}$$

where *convex* means that the weights w_i are constrained to $w_i > 0$ and $\sum_{i=1}^{k} w_i = 1$ (MARLER; ARORA, 2005; MARLER; ARORA, 2009). An important requirement of the weighted sum method is that the objective functions be scaled to the same approximate range of values so that any potential numerical dominance can be eliminated or minimized, and the weights can play the desired role (MARLER; ARORA, 2005; MARLER; ARORA, 2009).

There are several methods to scale the objective functions (MARLER; ARORA, 2005). The Fortran source code of CLHS shows that, although not mentioned in the original paper, CLHS scales $\mathcal{O}_1$ and $\mathcal{O}_3$ using the *upper-bound approach*, $f_i'' = f_i(\boldsymbol{X})/f_i^{max}$, where $f_{\mathcal{O}_1}^{max} = n \times p^{num}$ and $f_{\mathcal{O}_3}^{max} = 0.5{p^{num}}^2 + p^{num}$, p^{num} being the number of numerical covariates. $\boldsymbol{f}^{max}$ is a rough estimate of the single worst solution for $\mathcal{O}_1$ and $\mathcal{O}_3$, called the *nadir point cloud* (MARLER; ARORA, 2004). Thus, this transformation results in a non-dimensional objective function with an upper limit around 1, and its use is imposed due to the fact that, by definition, the three objective functions yield values of very different orders of magnitude: $\mathcal{O}_1 > \mathcal{O}_3 > \mathcal{O}_2$. This is because $\mathcal{O}_1$ uses the number of sample points per strata (0–n), while $\mathcal{O}_3$ uses the linear correlation coefficient (-1–1), and $\mathcal{O}_2$ uses the proportion of sample points per strata (0–1).

We believe that the *upper-bound approach* is insufficient for a proper scaling of the objective functions because $\boldsymbol{f}^{max}$ usually is unattainable – it does not correspond to any point cloud in $\mathcal{S}$ and/or is too far from the Pareto optimum set (MARLER; ARORA, 2004). Defining $\boldsymbol{f}^{max}$ as the median of the objective functions over multiple spatial samples generated by simple random sampling (CLIFFORD et al., 2014) is a suboptimal strategy because it only ensures that the objective functions will have similar orders of magnitude at the beginning of the optimization, which might have a negligible influence in the definition of $\mathcal{Z}$ (MARLER; ARORA, 2005). Besides, provided the optimization algorithm is well designed, the starting point should not influence the solution of the MOCOP (see below).

We propose using a more robust approach, i.e. the *upper-lower bound approach*,

$$f_i'' = \frac{f_i(\boldsymbol{X}) - f_i^\circ}{f_i^{max} - f_i^\circ}, \tag{8.6}$$

where f_i° is the *utopia point*, the single best solution for the ith objective function, and f_i'' is the ith non-dimensional, scaled objective function constrained between zero and one (MARLER; ARORA, 2005). Because of the above-mentioned problems regarding the definition of $\boldsymbol{f}^{max}$, it is more appropriate to use the *Pareto maximum*, $f_i^{max} = max_{1 \leq j \leq k} f_i(\boldsymbol{X}_j^*)$, where $\boldsymbol{X}_j^*$ is the point cloud that minimizes the jth objective function (MARLER; ARORA, 2005). In practice, we optimize a sample configuration regarding each of the k objective functions individually so that we end up with k optimized sample configurations. The objective function value of each of the k optimized sample configurations is recorded and set as the diagonal entries of the Pareto matrix $\varOmega$. Then, we take the ith optimized sample configuration and calculate the value of the jth objective functions, where $i \neq j$, and use the results as the off-diagonal entries of $\varOmega$. As such, $\varOmega$ is a $k \times k$ matrix with the objective function used to optimize the spatial sample in the columns, and the objective function used to calculate the objective function value in the rows. Finally, we lookup the columns of $\varOmega$ and record the largest absolute maximum observed in each of them, generally an off-diagonal entry. These are the Pareto maximum values of the k objective functions. In turn, f_i° is replaced with the Pareto minimum, i.e. the smallest absolute minimum value observed in each column of $\varOmega$, generally the diagonal entry. This is because, like $\boldsymbol{f}^{max}$, $\boldsymbol{f}^\circ$ exists in the objective space $\mathcal{Z}$, but usually is unattainable, i.e. it does not correspond to any point cloud in $\mathcal{S}$ (ARORA, 2011). The obvious drawback of this approach is the extra time needed to optimize each of the k objective functions individually.

8.4.4 Resulting Problem Definition

Given the proposed modifications, the problem of sampling for spatial trend estimation for soil mapping is redefined using two objective functions,

$$\text{CORR} = \sum_{i=1}^{p} \sum_{j=1}^{p} |\varphi_{ij} - \upsilon_{ij}|, \tag{8.7}$$

where φ_{ij} and υ_{ij} are the population and sample associations (or correlations in case all covariates are numeric) at the ith row and jth column of the p-dimensional population and sample association (or correlation) matrices, and

$$\text{DIST} = \sum_{i=1}^{p} \sum_{j=1}^{c_i} |\pi_{ij} - \gamma_{ij}|, \tag{8.8}$$

where π_{ij} and γ_{ij} are the proportion of sample and population points that fall in the jth class (or marginal sampling strata) of the ith covariate, c_i being the number of classes of the ith covariate. With these two objective functions, we define a utility function U as in Equation 8.5 aiming at a spatial sample that reproduces an **A**ssociation/**C**orrelation measure and the marginal **D**istribution of the **C**ovariates,

$$\text{ACDC} = w_1\text{CORR} + w_2\text{DIST}, \tag{8.9}$$

with $w_1 = w_2 = 0.5$ when we do not have *a priori* preferences towards the objective functions.

8.5 CASE STUDY

We developed a case study to evaluate the proposed improvements and compare them with the original CLHS. It was based on using synthetic data derived from a real-world case study (SAMUEL-ROSA et al., 2015a). The study site is a small catchment of about 2000 ha located on the southern edge of the Plateau of the Paraná Geologic Province, Rio Grande do Sul, Brazil. The real-world dataset contains $n = 350$ point soil observations of the topsoil, and includes several soil properties, but only bulk density data (BUDE, $\mathrm{Mg\,m^{-1}} \times 100$) was used ($n = 282$). The dataset also includes several covariates derived from area-class soil maps, digital elevation models, geological maps, land use maps, and satellite images. All processing steps used to derive the covariates were described by Samuel-Rosa et al. (2015a).

8.5.1 Soil Data Generating Process

In an ideal world, we would create $\mathcal{R} \geq 100$ spatial samples of $\mathcal{N} \geq 2$ sizes with each of the $\mathcal{A} \geq 2$ algorithms that we want to compare. Then we would go to the field, sample the soil, and measure a property to construct $\mathcal{D} = \mathcal{R} \times \mathcal{N} \times \mathcal{A}$ calibration datasets. The same property would be measured at a fixed set of probabilistically selected validation sites. Each calibration dataset would be used to calibrate a model, with which we would predict at the validation sites. The $\mathcal{A}$ sampling algorithms would then be compared on how well they performed, for each of the $\mathcal{N}$ sizes, using the confidence interval of a prediction error statistic over all $\mathcal{R}$ spatial samples. In such an experiment, the random selection of spatial samples would be the *source of variation* (DE GRUIJTER; TER BRAAK, 1990).

In the real world…Due to limited resources, we created only $\mathcal{R} = 1$ spatial sample with $\mathcal{N} = 3$ sizes with each of the $\mathcal{A} = 4$ sampling algorithms to be compared (CORR, DIST, ACDC, and CLHS). The variation in the experiment had to come from another source: we chose it to be the soil property data. We did so using unconditional sequential Gaussian simulation (GOOVAERTS, 2001; PEBESMA, 2004). To start, we defined a theoretical (or super-population) model, our *soil data generating process*, which should be as close to reality as possible. Thus, the soil data generating process was defined empirically by calibrating a (non)linear mixed model to the observed (real) BUDE data. The main calibration steps were as follows (BREIMAN, 2001; LIAW; WIENER, 2002; DIGGLE; RIBEIRO JR, 2007; LARK, 2012):

(I) Random forest regression: grow $n_{\text{trees}} = 500$ regression trees with a maximum terminal node size of $n_{\text{node size}} = 5$ points, each tree grown using $n = 282$ calibration points randomly selected with replacement (bootstrap sample) from the set of $n = 282$ point soil observations (about $n_{\text{in-bag}} = n \times \left(1 - \left(\frac{n-1}{n}\right)^n\right) = 178$ unique point soil observations), and $p_{\text{in-bag}} = 4$ covariates randomly selected at each split out of a set of $p = 12$ covariates (`SOIL_25c`, `SOIL_25h`, `SOIL_25j`, `LU2009a`, `LU2009b`, `LU2009d`, `GEO_50c`, `RED_30`, `ELEV_90`, `SLP_90_15`, `NOR_90_127`, `NOR_90_255`) selected and described as in Samuel-Rosa et al. (2015a).

(II) Out-of-bag predictions: use each of the $n_{\text{trees}} = 500$ regression trees from step (1) to predict BUDE at the point soil observations not included (out-of-bag) in the respective calibration dataset (about $n_{\text{out-of-bag}} = n - n_{\text{in-bag}} = 104$ point soil observations), and compute the average of the predicted BUDE at each point soil observation (about $n_{\widehat{BUDE}} = n_{\text{trees}} \times \left(\frac{n-1}{n}\right)^n = 184$ predicted values for each out-of-bag point).

(III) Linear mixed model: assume that the average of the out-of-bag predictions from step (2) are linearly related to BUDE and present insignificant conditional bias, and use them as a covariate in the fixed effects of a linear mixed model (LMM), the random effects modelled using the Whittle-Matérn model, all parameters being estimated by Gaussian restricted maximum likelihood (REML).

The parameters of the LMM are the coefficients β_0 and β_1 of the linear trend, which correct any linear bias in the random forest regression out-of-bag predictions (LIAW; WIENER, 2002), and the nugget (τ^2), sill (σ^2), and range (α) of the Whittle-Matérn model. The shape parameter (ν) of the Whittle-Matérn model was defined separately, by choosing from a set of discrete values $\nu = (0.5, 1.0, 2.0, 4.0, 8.0)$ based on the resulting profile likelihood for ν and maximized restricted log-likelihood, and on the computing time (STEIN, 1999; DIGGLE; RIBEIRO JR, 2007). The fitted LMM ($\beta_0 = 13.35\,\mathrm{Mg\,m^{-3}}$, $\beta_1 = 0.91$, $\tau^2 = 349.51\,\mathrm{Mg\,m^{-6}}$, $\sigma^2 = 97.24\,\mathrm{Mg\,m^{-6}}$, $\alpha = 210.99\,\mathrm{m}$, $\nu = 2.0$) explained 38 and 18 % of the sample variance of BUDE with $m(\boldsymbol{s})$ and $e(\boldsymbol{s})$, respectively (Figure 8.1).

With the random forest regression and the LMM at hand, we produced $\mathcal{R} = 1000$ equiprobable realizations of an isotropic Gaussian random field of BUDE. Each realization is constituted of a collection of BUDE values at a fine grid of $\sim 800\,000$ regularly spaced (5 m) points covering the entire study area. Because we used *unconditional* Gaussian simulation – a simulation that is only *globally* conditioned to the histogram (and variogram) of the data observed at the calibration locations (GOOVAERTS, 1997) – and a large number of realizations, we expected to find the whole set of possible BUDE values (i.e. the global histogram) at any point of the simulation grid, over the $\mathcal{R} = 1000$ realizations. Thus, the variance – our uncertainty – is the same everywhere. This is in contrast with *conditional* Gaussian simulation – a simulation that is *locally* conditioned to the data observed at the calibration locations –, which serves the purposes of honouring the data at the calibration locations, thus reducing the variance of the output realizations (GOOVAERTS, 1997). In our study, *conditional* Gaussian simulation is inappropriate as a model of uncertainty because the sampling algorithms that we evaluate are designed to situations where we know very little, everywhere, about the spatial structure of the soil data generating process.

Unconditional simulation algorithms approximate the globally conditioned distribution of a soil variable at a given point using neighbouring simulated values as local conditioning information (GOOVAERTS, 1997): the larger the neighbourhood, the better the approximation. A sensible criterion to select the nearest simulated values is the practical range of the variogram model. However, as the simulation proceeds, the number of simulated values within this neighbourhood becomes computationally prohibitive (GOOVAERTS, 1997; WEBSTER; OLIVER, 2007; PEBESMA, 2014). Thus, we used a fixed maximum number of simulated values $n_{max} = 100$ within the practical range of the variogram model (1132.657 m) that are closest to the point being simulated. Last, but not least, sequential simulation algorithms speed up computations by using the same random path in all simulations, allowing it to reuse the "expensive results": neighbourhood selection and solution to the kriging equations (GOOVAERTS, 1997; WEBSTER; OLIVER, 2007; PEBESMA, 2014). We believe that following the same random path to generate each of the $\mathcal{R} = 1000$ realizations can introduce a structural component in the approximation errors. To avoid that, simulations were carried out in five batches of $\mathcal{R} = 250$ realizations using a different seed for the pseudo-random number generators each time.

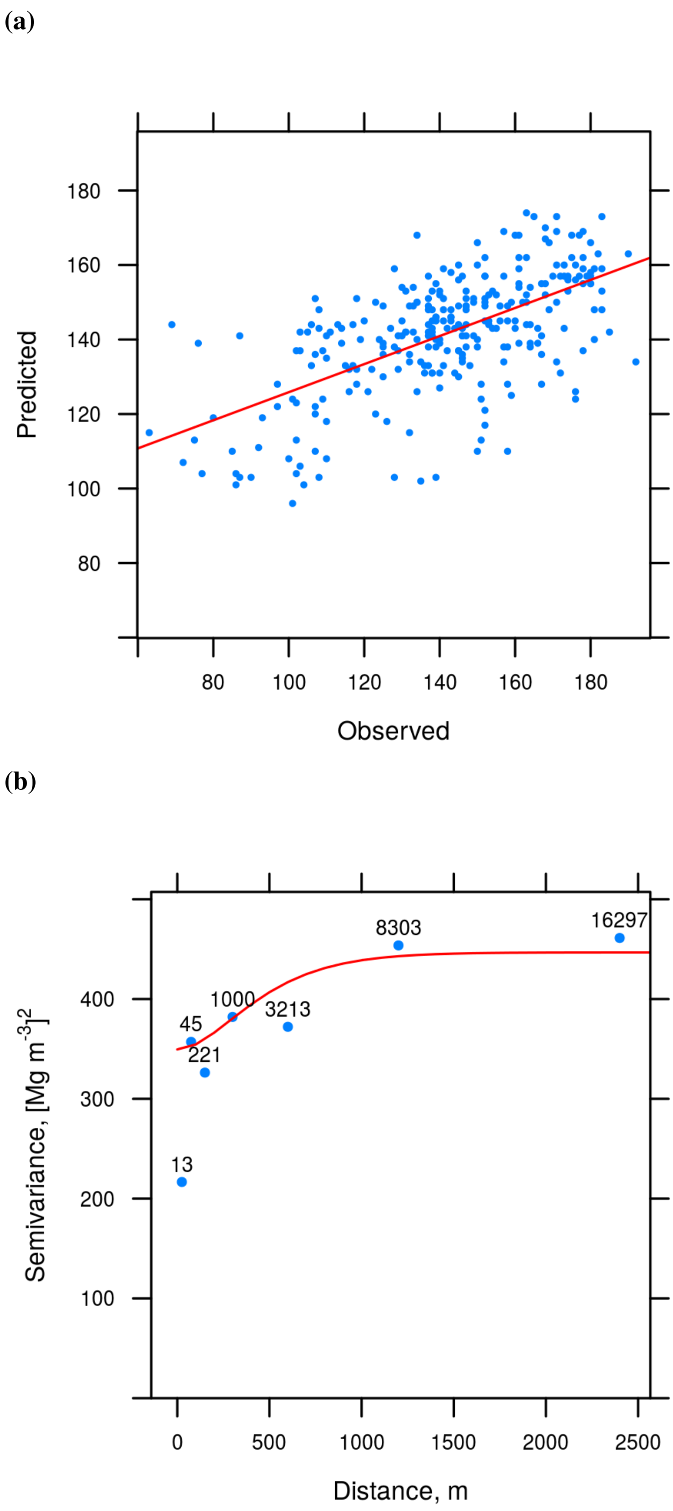


Figure 8.1: (Non)Linear mixed model fitted to the soil bulk density data (BUDE, Mg m^{-3} × 100) measured at $n = 282$ calibration locations. (a) shows the relation between out-of-bag random regression forest predictions and BUDE, here representing the deterministic component $m(\boldsymbol{s})$. (b) shows the restricted maximum likelihood fit of the variogram model to BUDE, i.e. the stochastic component $e(\boldsymbol{s})$. Exponentially spaced lags are used to depict the sample variogram along with the number of point-pairs in each lag.

8.5.2 Spatial Sampling

We compared $\mathcal{A} = 4$ criteria (CORR, DIST, ACDC, and CLHS) for optimizing sample configurations for spatial trend estimation for soil mapping using $\mathcal{N} = 3$ sample sizes $n = (100, 200, 400)$. These sample sizes correspond to the moderately high inspection density (1 sample point per 20, 10, and 5 ha, respectively) recommended for the production of soil maps published at a cartographic scale of 1:25 000 (ROSSITER, 2000). Spatial sample configurations were optimized regarding each of the four criteria using **sp**atial **s**imulated **ann**ealing as implemented in the `spsann` package for `R`, designed specifically for the purpose of this study and made available in The Comprehensive R Archive Network (CRAN).

8.5.2.1 Spatial simulated annealing

Simulated annealing is a popular method with widespread use to solve combinatorial optimization problems in the soil and geosciences such as stochastic simulation (DEUTSCH, 1992; GOOVAERTS, 2000), estimation of model parameters (LARK; PAPRITZ, 2003), resource allocation and land use planning (MUTTIAH et al., 1996; AERTS; HEUVELINK, 2002; DUH; BROWN, 2007), and spatial sampling and monitoring (VAN GROENIGEN et al., 1997; SIMBAHAN; DOBERMANN, 2006; BRUS; HEUVELINK, 2007; MARCHANT; LARK, 2006; MINASNY; MCBRATNEY, 2006; MELLES et al., 2011). This is mainly due to its robustness against local optima and easiness of implementation (METROPOLIS et al., 1953; KIRKPATRICK et al., 1983; ČERNÝ, 1985; AARTS; KORST, 1989; VAN GROENIGEN, 1999).

Simulated annealing is a heuristic algorithm that sequentially searches for the optimum solution for the problem at hand using the information “learned” from randomly trying solutions out of a large set of possible solutions, i.e. by trial and error. In spatial sampling, this means sequentially trying out randomly selected sample configurations $\boldsymbol{X}$ and checking how well they conform to the chosen criterion. Every time the newly selected sample configuration $\boldsymbol{X}_{i+1}$ returns an improved (lower) objective function value $\boldsymbol{f}(\boldsymbol{X}_{i+1})$ than the previously selected sample configuration $\boldsymbol{X}_i$, the latter is immediately discarded in favour of the former. The set of formal rules used to generate a new sample configuration $\boldsymbol{X}_{i+1}$ to be compared with $\boldsymbol{X}_i$ with respect to the chosen criterion is called the *generation mechanism* (VAN GROENIGEN, 1999; BRUS; HEUVELINK, 2007; WEBSTER; LARK, 2013).

The generation mechanism fundamentally works by means of randomly perturbing $\boldsymbol{X}_i$, specifically by adding random noise to the x- and y-coordinates of one of the sample points of $\boldsymbol{X}_i$. We call this process *jittering*. The main requirement is that the minimum (x_{min} and y_{min}) and maximum (x_{max} and y_{max}) quantity of random noise that can be added to the x- and y-coordinates of a sample point be specified. These will define the *neighbourhood* within which a sample point can be moved around. This neighbourhood corresponds to a rectangle centred at the sample point, with sides proportional to the sides of the rectangle that spans the sampling region. x_{max} and y_{max} should be large enough to enable all sample points visiting (almost) any place in the sampling region during the optimization such that the starting sample configuration – generally obtained by simple random sampling – has no direct influence on the final sample configuration.

Adding random noise to the x- and y-coordinates of a sample point corresponds to selecting a candidate location in the neighbourhood. This can only be done after the set of *effective* candidate locations has been identified, i.e. after the presence of non-sampling areas (e.g. buildings and water bodies), as well as the shape and finiteness of the sampling region have been taken into account. Using a *finite* set of candidate locations is an efficient way of achieving this because, by definition, the candidate location will always fall within the sampling region and out of non-sampling areas. A finite set of candidate locations is created by discretizing the sampling region beforehand, that is, by creating a fine grid of points that serve as candidate locations during the entire search for the optimum sample configuration. To minimize its disadvantages – such as the fact that not all locations in the sampling region can enter the sample – one can see the fine grid of points as the centre nodes of a finite set of grid cells and use a form of *two-stage random sampling* (WALVOORT et al., 2010): first, one of the candidate "grid cells" is selected with replacement in the neighbourhood, i.e. independently of already being occupied by another sample point; then, the candidate location for the sample point is selected within that "grid cell" by simple random sampling.

However, to be able to escape from apparently optima solutions that appear too early during the search – local optima solutions –, the newly selected sample configuration $\boldsymbol{X}_{i+1}$ can be accepted even if it returns an inferior objective function value, i.e. $\boldsymbol{f}(\boldsymbol{X}_{i+1}) > \boldsymbol{f}(\boldsymbol{X}_i)$. This means that the algorithm "understands" that taking a step back sometimes during the search can lead to better results. The decision whether to accept or not a sample configuration is described by the Metropolis criterion (METROPOLIS et al., 1953), used to compute the *acceptance probability* $P(\boldsymbol{X}_i \rightarrow \boldsymbol{X}_{i+1})$ as

$$P(\boldsymbol{X}_i \rightarrow \boldsymbol{X}_{i+1}) = \begin{cases} 1, & \text{if } \boldsymbol{f}(\boldsymbol{X}_{i+1}) \leq \boldsymbol{f}(\boldsymbol{X}_i), \\ exp\left(\frac{\boldsymbol{f}(\boldsymbol{X}_i)-\boldsymbol{f}(\boldsymbol{X}_{i+1})}{T}\right), & \text{if } \boldsymbol{f}(\boldsymbol{X}_{i+1}) > \boldsymbol{f}(\boldsymbol{X}_i), \end{cases} \tag{8.10}$$

where T is a positive control parameter that dictates how likely it is that an inferior sample configuration will be accepted.

Parameter T, traditionally called the *temperature* parameter, is at the core of simulated annealing. This is because simulated annealing tries to mimic the process used in metallurgy and materials science known as *annealing*, by which a molten metal is scheduled to gradually cool so that its solidifications results in such an arrangement of the atoms that form a perfect, minimum energy, mechanically resistant, crystalline structure (METROPOLIS et al., 1953; KIRKPATRICK et al., 1983; ČERNÝ, 1985). A closer example is the formation of extrusive (volcanic) igneous rocks by the cooling of effusive lava, where the cooling rate strongly determines the size of crystals in the resulting rock (HALDAR; TIŠLJAR, 2014): sudden cooling (seconds, minutes) of lava creates amorphous volcanic glass with vitreous texture such as pumice; slower cooling (hours, weeks) enables the growth of crystals, which can sometimes be visible to the naked eye, resulting in more mechanically resistant rocks such as basalt. Much larger crystals will be observed in intrusive (plutonic) igneous rocks such as granite, but their growth is due to a very slow cooling of magma, sometimes thousands or millions of years (HALDAR; TIŠLJAR, 2014).

As the concept of temperature says, what happens at the atomic level during the cooling process of a molten metal (or lava and magma) is the reduction of the average kinetic energy – the energy of motion – of the particles. When optimizing spatial samples using simulated annealing, high "kinetic energy" means that 1) considerably different sample configurations might be tried out every time, and 2) it is very likely that inferior sample configuration will be accepted. As the "kinetic energy" decreases, the difference between $\boldsymbol{f}(\boldsymbol{X}_{i+1})$ and $\boldsymbol{f}(\boldsymbol{X}_i)$ will be smaller, as well as the probability of accepting $\boldsymbol{f}(\boldsymbol{X}_{i+1})$ if it is inferior to $\boldsymbol{f}(\boldsymbol{X}_i)$. To control this in practice, one needs to set up an annealing schedule.

The *annealing schedule* corresponds to the set of formal rules that determine how the probability of accepting inferior sample configurations is decreased as the search for the globally optimum sample configuration evolves (AARTS; KORST, 1989; VAN GROENIGEN, 1999; WEBSTER; LARK, 2013). When optimizing the configuration of a spatial sample, we have to decide upon a feasible annealing schedule, i.e. a schedule that enables finding the globally optimum sample configuration (or a sample configuration very close to it) in a reasonable amount of time. For that end, a large initial value of T is chosen by trial and error so that almost all sample configurations tried out during the first temperature step are accepted. A temperature step, more precisely a *Markov chain*, corresponds to the series of tryouts made with a constant value of T. The number of tryouts made with a constant value of T corresponds to the length of a Markov chain $chain_{\text{length}}$, which should be long enough so that the "system" will tend to a "kinetic equilibrium".

In practice, the length of a Markov chain has to be limited, and the equilibrium can only be approximated (WEBSTER; LARK, 2013). The only requirement is that every point be jittered at least once as it would happen with the atoms in a physical system (METROPOLIS et al., 1953). To initiate a new Markov chain $j + 1$, one reduces T linearly by a control parameter δ_T, with $0 < \delta_T < 1$, so that $T_{j+1} = \delta_T T_j$ (AARTS; KORST, 1989). The same applies to the generation mechanism by using a decrement function to reduce the size of the neighbourhood as the search for the optimum sample configuration evolves. The reason for this is that, as the search evolves and approaches its end, it is likely that moving a sample point over a short distance contributes more to finding the global optimum than moving it over larger distances (VAN GROENIGEN; STEIN, 1998). The decrement function determines that the size of the neighbourhood is reduced linearly at the end of the jth chain. For the x-axis of the neighbourhood,

$$x_{\text{max}_{j+1}} = x_{\text{max}_{j=0}} - \frac{j}{n_{\text{chains}}} \times x_{\text{max}_{j=0}} - x_{\text{min}} + x_{\text{dim}}, \tag{8.11}$$

where $x_{\text{max}_{j+1}}$ is the dimension of the x-axis of the neighbourhood in the next chain, i.e. the maximum allowed shift in the x-coordinate, $x_{\text{max}_{j=0}}$ is the dimensions of the x-axis of the neighbourhood in the first chain, x_{min} is the minimum required shift in the x-coordinate, x_{dim} is the grid spacing in the x-coordinates, and n_{chains} is the maximum affordable total number of Markov chains (the very same applies to the y-axis of the neighbourhood). The maximum affordable total number of Markov chains n_{chains} has to be large enough such that the globally optimum sample configuration can be found. Finally, one or more criteria for stopping the search can be defined, such as the maximum affordable number of Markov chains completed without returning an improved sample configuration $n_{\text{chain stop}}$ (VAN GROENIGEN, 1999).

All parameters of the generation mechanism (x_{min}, y_{min}, x_{max} and y_{max}) and annealing schedule ($chain_{\text{length}}$, n_{chains}, and $n_{\text{chain stop}}$) have to be defined by trial and error, and based on the current knowledge and available resources (WEBSTER; LARK, 2013). A useful tool is a plot with the evolution of the objective function values as measured at the end of each Markov chain (LARK; PAPRITZ, 2003): objective function values should present large fluctuation during the first Markov chains, then gradually stabilize and decrease till a minimum is reached at which they remain for $n_{\text{chain stop}}$ Markov chains.

8.5.2.2 Experimental design

In this study, we used an annealing schedule with a maximum affordable total number of Markov chains $n_{\text{chains}} = 500$ without imposing any stopping criterion. The initial temperature T was set such that more than $95\,\%$ of the tryouts during the first Markov chain would be accepted. To initiate a new Markov chain $j + 1$, T was linearly decreased by a factor of $\delta_T = 0.95$. The length of each Markov chain was set to $chain_{\text{length}} = n$, i.e. the sample size $n = (100, 200, 400)$, with each sample point being jittered in turn.

The parameters of the generation mechanism were set such that the size of the neighbourhood in the first chain is equal to half the sides of the rectangle that spans the sampling region ($x_{\text{max}} = 2615\,\text{m}$ and $y_{\text{max}} = 2985\,\text{m}$), and that the minimum jitter is equal to zero ($x_{\text{min}} = y_{\text{min}} = 0$), i.e. that the grid cell where the sample point is located could be selected as well. With these settings, at the end of the search, the neighbourhood of a sample point was constrained to the set of nine grid cells composed of that in which the sample point falls and its eight surrounding grid cells.

Each spatial sample optimized with $\mathcal{N} = 3$ sample sizes using the $\mathcal{A} = 4$ sampling algorithms was used to sample from the $\mathcal{R} = 1000$ simulated realities of BUDE to constitute $\mathcal{D} = \mathcal{A} \times \mathcal{N} \times \mathcal{R} = 12\,000$ calibration datasets. Sampling was performed using nearest neighbour sampling because simulated data are available only at a finite set of regularly spaced (5 m) grid points. This corresponds to assuming that BUDE is the same everywhere inside the $5 \times 5\,\text{m}$ area surrounding every grid point.

8.5.3 Model Calibration

Each of the $\mathcal{D} = 12\,000$ calibration datasets was used to calibrate a (non)linear mixed model using the same steps employed for calibrating the soil data generating process as described in Section 8.5.1. Two important differences exist, though. First, the shape parameter, ν, of the Whittle-Matérn model was chosen from a set of only three values instead of five ($\nu = (0.5, 1.0, 2.0)$) following recommendations of Moyeed & Papritz (2002). This modification considerably reduced the computation time. Last, initial covariance parameters needed to solve the system of nonlinear equations for the REML fit of the (non)linear mixed model (KUENSCH et al., 2011) were chosen using a set of heuristics (`pedometrics::vgmICP`, Appendix B). These heuristics involve estimating the sample variogram at a sequence of seven exponentially spaced lag-distance classes up to half the diagonal of the rectangle that spans the data using Genton's robust Qn-estimator (GENTON, 1998). A minimum number of $n_{\text{pairs}} = 30$ point-pairs was required in a lag-distance class so that it be used for guessing the initial covariance parameters.

8.5.4 Evaluation of Sampling Algorithms

The influence of sampling design and sample size on the spatial prediction accuracy was assessed using $n = 1000$ validation points probabilistically selected from $g = 500$ quasi-equal-area compact geographical strata. Geographic stratification was obtained using the k-means algorithm as implemented in the `R`-package `spcosa` (WALVOORT et al., 2010). The algorithm was run using three different seeds for the pseudo-random number generator, the resulting stratification with the minimum value of the mean squared shortest distance (MSSD) being selected. Each stratum has $n = 2$ randomly selected validation points.

We used the realization-wise mean error (ME) and mean squared error (MSE) at the probabilistically selected validation points to evaluate the sampling designs. The ME shows the effect of the sampling design on the accuracy of the predictions, while the MSE indicates how the sampling design influence the calibration of the variogram model. As such, we can use the mean ratio of (empirical and theoretical) squared errors (MRSE) to see how good our estimate of the prediction errors are. A fourth validation measure was the amount of variance explained (AVE). At the end, for each sampling algorithm and sample size, we aggregated these measures over the $\mathcal{R} = 1000$ realizations using box-and-whisker plots to check visually how one sampling design compares to the others.

8.6 RESULTS AND DISCUSSION

Figure 8.2 shows the evolution of the objective function values during the optimization of sample configurations using all four sampling algorithms for all sample sizes. The observed pattern is that expected from a spatial simulated annealing algorithm (LARK; PAPRITZ, 2003): high and variable objective function values during the first Markov chains, and low and stable values as the optimization reaches its end. The remarkably larger variation of CLHS at the beginning of the optimization, compared to the other algorithms, is likely due to the fact that it is composed of three objective functions. Accordingly, the ACDC algorithm, which is a combination of the objective functions CORR and DIST, took more iterations to stabilize than the two functions individually. It is worth pointing that the optimum sample configuration was not found for any of the sampling algorithms after $n_{\text{chains}} = 500$ of size n as indicated by the descending behaviour of the objective function values at the end of the optimization (i.e. the objective function values did not level off). This is also the expected behaviour for finite Markov chains (WEBSTER; LARK, 2013).

Our expectation that the first objective of the CLHS, $\mathcal{O}_1$, which aims at the marginal distribution of the numeric covariates, would have a numerical dominance over the second and third objective functions, $\mathcal{O}_2$ and $\mathcal{O}_3$, which aim at the marginal distribution of the factor covariates and similarity between the population and sample correlation matrices of the numeric covariates, respectively, was confirmed as shown in Figure 8.3. Each panel in Figure 8.3 shows the region of the feasible objective space $\mathcal{Z}$ that has been jointly explored by the pairs of objective functions that compose CLHS and ACDC during the optimization of a sample configuration of size $n = 100$. Function values obtained during the first Markov chains are represented by the set of points located in the upper right corner of each panel, i.e. where a large variation is observed (see Figure 8.2). As the optimization proceeds, the variation in the objective function values decreases, till they reach a certain stability and concentrate in a specific region of the criterion space $\mathcal{Z}$ (left bottom corner of each panel). Optimally, the bivariate distribution of function values will follow the $45°$ line that crosses each panel from the top right corner to the bottom left corner. Deviations from the $45°$ line are evidence of numerical dominance of one objective function over the other.

Comparing $\mathcal{O}_1$ with $\mathcal{O}_2$ and $\mathcal{O}_3$, we see a strong deviation from the $45°$ line as evidenced by descendent vertical point cloud on the right side of both top panels. This means that the sample configuration is rapidly optimized with respect to $\mathcal{O}_2$ and $\mathcal{O}_3$ during the first Markov chains. The bottom left panel suggests that $\mathcal{O}_2$ and $\mathcal{O}_3$ are optimized jointly, with a small bias towards $\mathcal{O}_3$. Most of the optimization efforts are spent towards $\mathcal{O}_1$, as suggested by the high point density along the horizontal axis in both top panels. Thus, the bias in the CLHS can be ordered as $\mathcal{O}_1 > \mathcal{O}_3 > \mathcal{O}_2$. On the other hand, ACDC appears to be unbiased as suggested by the bottom right panel. The sample configuration is optimized with respect to CORR and DIST jointly, approximately following the $45°$ line, as expected. These results suggest that our proposed improvements on the CLHS produced a sampling algorithm with a better numerical behaviour. It is worth noting that the results presented in Figure 8.3 refer to the sample size of $n = 100$, but the same trends were observed with the other two sample sizes.

A snapshot of the optimized sample configurations is presented in Figure 8.4. We start by comparing the spatial samples produced using CORR and DIST regarding their geographic distribution. For all three sample sizes, DIST produces a spatial sample with a better geographic

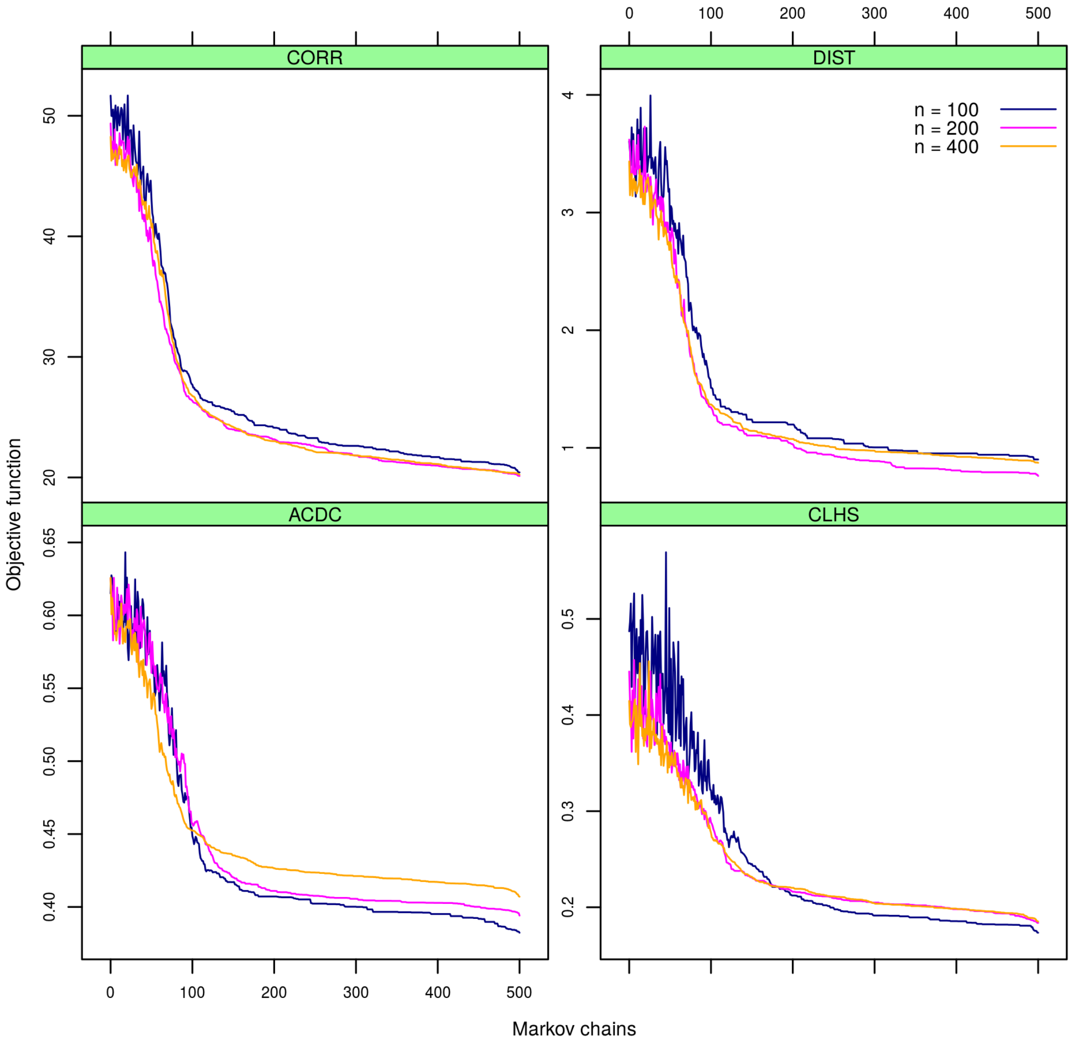


Figure 8.2: Objective function values during the optimization of sample configurations of sizes $n = (100, 200, 400)$ using sampling algorithms CORR, DIST, ACDC, and CLHS against the number of Markov chains of length n.

coverage, but the differences are smaller as the sample size increases. Since these algorithms do not aim at the coverage of the geographic space, leaving large spaces unsampled can have a negative effect on the prediction accuracy of calibrated models. CORR leaves larger empty spaces because it creates clusters of points, many of which resemble transects. The coverage of the geographic space of DIST samples is degraded when it is combined with CORR in ACDC. ACDC samples have more clusters of points than DIST, which again resemble transects. Both ACDC and CLHS samples present a similar level of coverage of the geographic space. The difference is that CLHS places clusters of samples in different regions than those targeted by ACDC and vice-versa. This is likely due to the numerical dominance of $\mathcal{O}_1$ in CLHS, which results in a bias towards numeric covariates.

Regarding prediction accuracy, as expected, increasing the sample size resulted in more accurate and precise predictions for all sampling algorithms (Figure 8.5). The spreads of the empirical distributions of all validation measures, which are approximately Gaussian, become narrower with larger sample sizes. The exception is the MRSE, whose distribution becomes severely skewed with $n = 400$. DIST presents the narrower distribution of MRSE values,

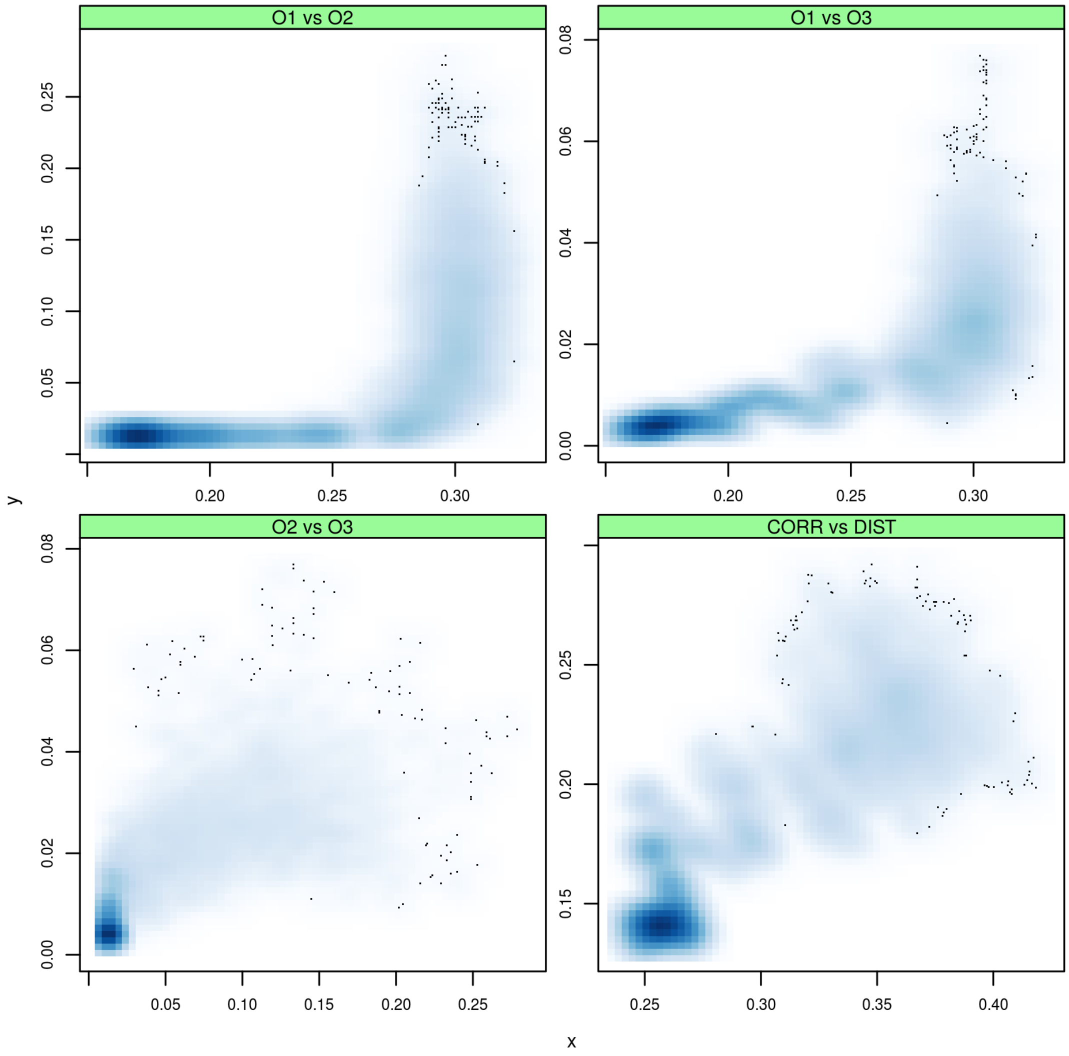


Figure 8.3: Region of the feasible objective space $\mathcal{Z}$ explored by pairs of objective functions (x vs y) that compose CLHS ($\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_3$) and ACDC (CORR and DIST) during the optimization of a sample configuration of size $n = 100$ using $n_{\text{chains}} = 500$ Markov chains of length n. Darker colours indicate a higher concentration of points.

but for all four sampling algorithms MRSE is above 1, which indicates that the prediction error variance was underestimated. This is likely because an incorrect variogram model was estimated with, for example, a very large nugget variance.

The CORR seems to have the largest spread of MSE values, specially for $n = 100$, suggesting that its predictions are the least precise. Models calibrated using spatial samples optimized using CORR consistently underpredicted BUDE for all sample sizes. Predictions of models calibrated with CLHS, ACDC and DIST spatial samples did not show a consistent bias. Together, these results suggest a degrading effect of taking the association/correlation between covariates into account when optimizing a sample configuration.

8.7 CONCLUSIONS

This study has shown that:

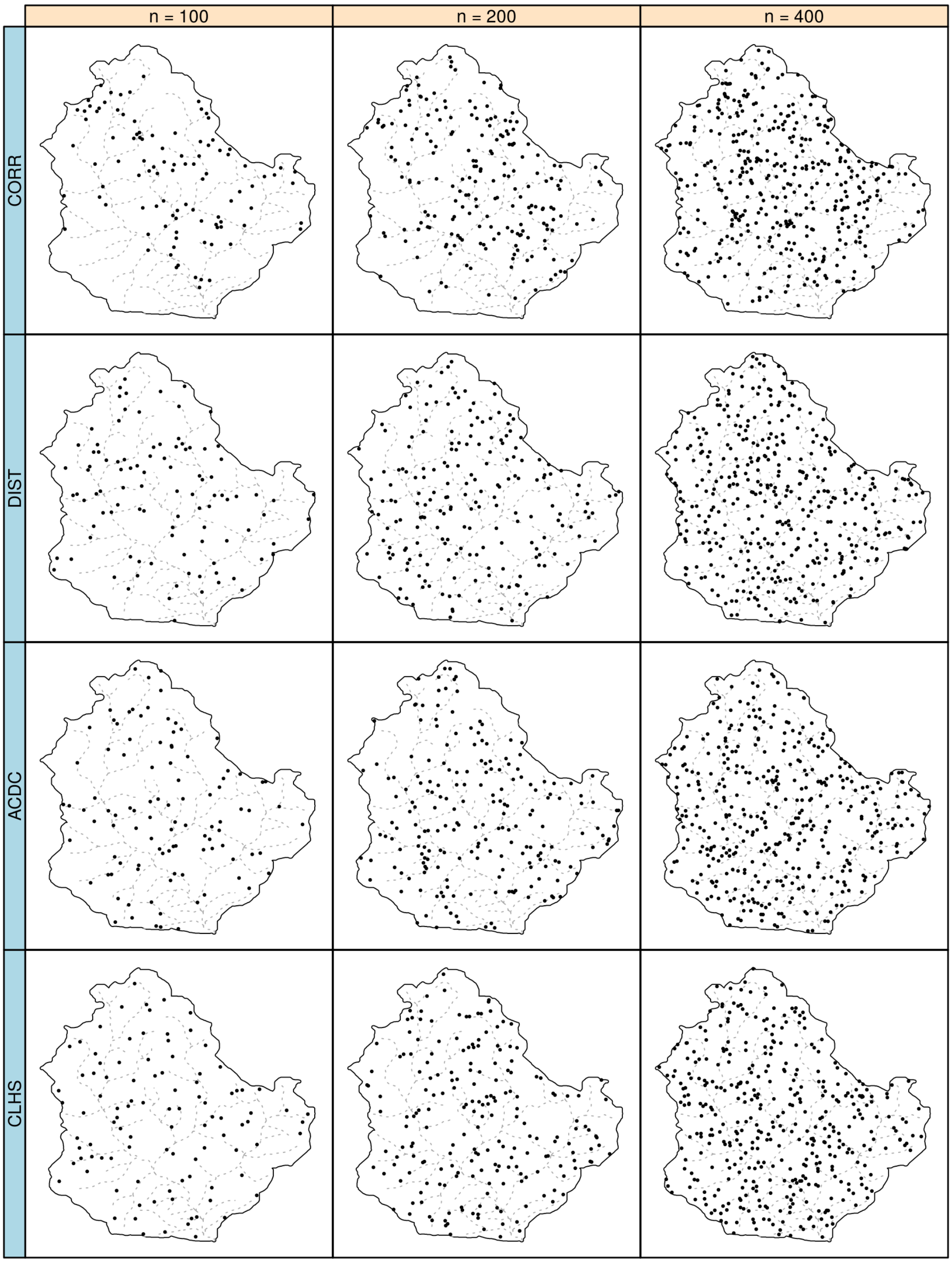


Figure 8.4: Sample configurations of size $n = (100, 200, 400)$ optimized using sampling algorithms CORR, DIST, ACDC, and CLHS superimposing the drainage network.

(I) The proposed modifications on the CLHS resulted in a sampling algorithm with an improved numerical behaviour, but this does not necessarily results in improved prediction accuracy;

(II) Larger sample sizes improve prediction quality irrespective of the sampling optimization algorithm that were compared;

(III) Sample configurations optimized aiming only at the association/correlation between co-

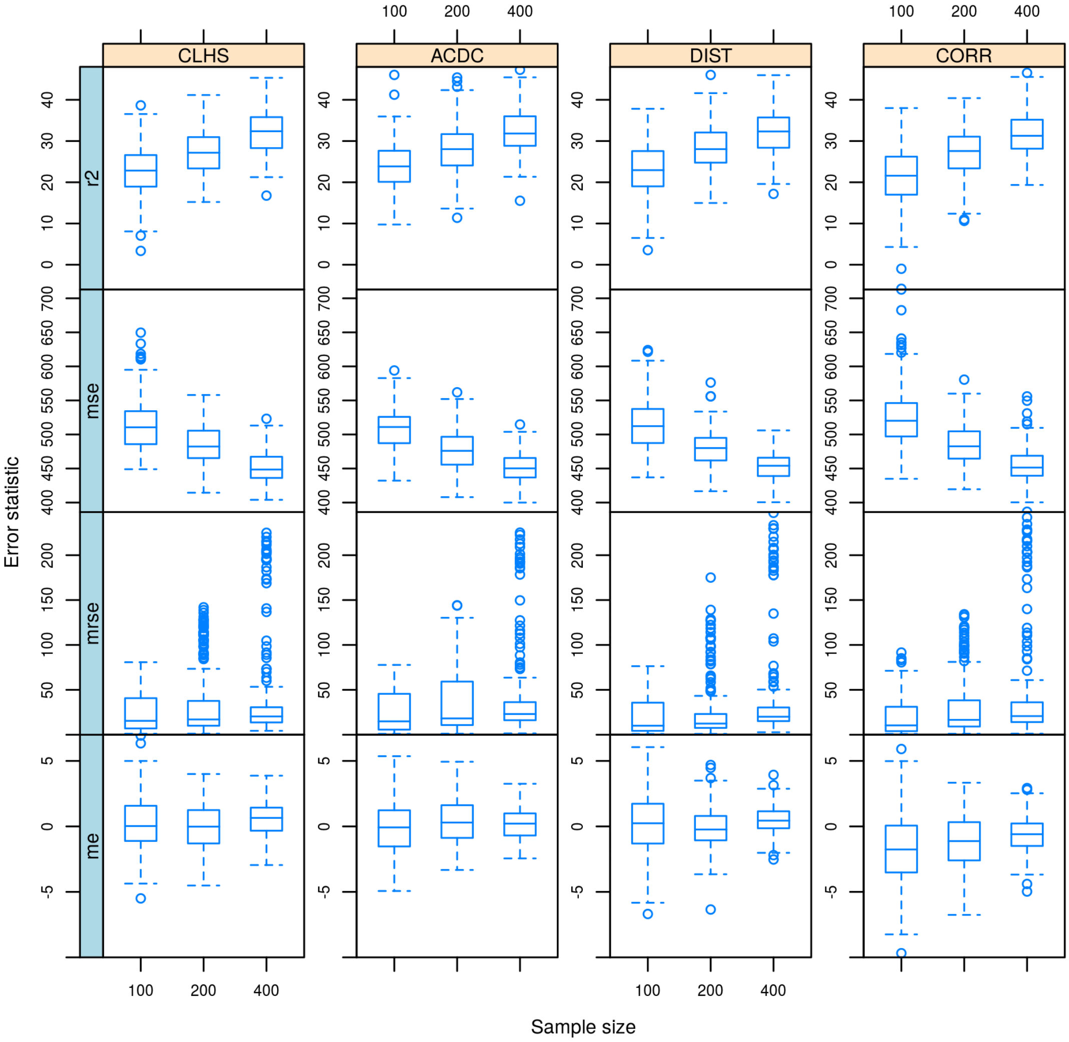


Figure 8.5: Statistics of the independent validation of the linear mixed models calibrated with simulated BUDE (Mg m^{-3} × 100) using sample configurations with sizes $n = (100, 200, 400)$ optimized using sampling algorithms CORR, DIST, ACDC, and CLHS. Statistics are the mean error (me), mean squared error (mse), mean ratio of squared errors (mrse), and amount of variance explained (r2).

variates can result in poorer predictions.

9 CHAPTER VIII

SAMPLING FOR SOIL MAPPING IN *TERRA INCOGNITA*

* This chapter is based on the study *spsann – optimization of sample patterns using spatial simulated annealing*, presented at the EGU General Assembly 2015 (SAMUEL-ROSA et al., 2015), and *Optimization of sample configurations for variogram estimation*, presented at Pedometrics 2015 (SAMUEL-ROSA et al., 2015b). Also collaborated in the preparation of this document: Gerard B. M. Heuvelink (ISRIC – World Soil Information), Dick J. Brus (Alterra, Wageningen University and Research Centre, the Netherlands), Gustavo M. Vasques (Embrapa Soils, Brazil), and Lúcia Helena Cunha dos Anjos (Universidade Federal Rural do Rio de Janeiro, Brazil).

9.1 RESUMO

Este estudo aborda um problema que muitos modeladores espaciais do solo enfrentam: como chegar à uma configuração amostral espacial eficiente para (I) estimar a tendência espacial, (II) estimar o variograma dos resíduos, e (III) fazer predições espaciais em situações onde nós sabemos muito pouco. A solução proposta consistem em formular um sólido problema de otimização multi-objetivo usando versões robustas de algoritmos de amostragem existentes. A amostra espacial visada deve reproduzir a distribuição marginal das covariáveis de modo que a tendência espacial possa ser estimada com acurácia. Ela também deve conter vários pequenos aglomerados dispersos por toda a região de amostragem para permitir fazer estimativa acurada do comportamento do variograma, especialmente próximo da origem. Finalmente, ela deve cobrir a região de amostragem da forma mais uniforme de tal modo que a média da variância do erro de predição é a menor possível. Esse problema de otimização multi-objetivo pode ser resolvido usando o recozimento simulado espacial conforme implementado no `R`-package `spann`.

Palavras-chave: Recozimento simulado. Otimização multi-objetivo. Estimativa do variogram. Predição espacial.

9.2 ABSTRACT

This study addresses a problem that many soil spatial modellers face: how to come up with an efficient spatial sample configuration to (I) estimate the spatial trend, (II) estimate the variogram of the residuals, and (III) make spatial predictions in situations where we know very little. The proposed solution is to formulate a sound multi-objective optimization problem using robust versions of existing sampling algorithms. The aimed spatial sample should reproduce the marginal distribution of the covariates such that the spatial trend can be accurately estimated. It should also contain several small clusters scattered throughout the sampling region to enable making an accurate estimate of the behaviour of the variogram, specially near the origin. Finally, it should cover the sampling region in the most uniform way such that the average prediction error variance is the least possible. This multi-objective optimization problem could be solved using spatial simulated annealing as implemented in the `R`-package `spsann`.

Keywords: Simulated annealing. Multi-objective optimization. Variogram estimation. Spatial prediction.

9.3 INTRODUCTION

The success of soil mapping largely depends on the sampling data, which are generally used to 1) estimate the spatial trend, 2) estimate the variogram of the residuals, and 3) make spatial predictions by calculating conditional distributions. A poor sampling strategy is likely to deliver a poor model and large prediction errors, resulting in a waste of financial resources, staff and time (VAN GROENIGEN et al., 1999; DE GRUIJTER et al., 2006; LAN; LIAN, 2010). This is undesirable because sampling already is the largest contributor to the costs of soil mapping (WEBSTER; OLIVER, 1990; VAN GROENIGEN et al., 1999; KEMPEN et al., 2012).

This study addresses a problem that many soil spatial modellers face: how to come up with a purposive spatial sample configuration that is effective and robust in situations where we know very little. We explore a scenario in which a) multiple soil properties have to be mapped, b) we are ignorant (or know very little) about the form of the model of spatial variation, and c) the operational constraints limit sampling to a single phase. The study starts with a review of the purposive sampling strategies employed by soil spatial modellers to meet one or more of the three objectives for which sampling data are used under the proposed scenario, i.e. to estimate the spatial trend and the variogram, and make spatial predictions. Based on theoretical and operational features, we indicate the purposive sampling strategies that we believe to be the most appropriate for each purpose and try to formulate a purposive sampling strategy that addresses all three objectives jointly. The chapter ends with a suggestion on how to test the performance of the proposed purposive sampling algorithms.

9.4 PURPOSIVE SAMPLING

Purposive sampling is the non-probability sampling mode by which the sampling locations are selected intentionally as to satisfy an *a priori* criterion. This criterion is commonly defined based on the *model* that will be used to infer the structure of spatial variation of a soil property $Y(\boldsymbol{s})$. Compared to probability sampling, purposive sampling generally is more efficient for *model-based inference* (DE GRUIJTER et al., 2006).

The criterion used to select the sampling locations can be defined based on the chosen statistical model (DE GRUIJTER et al., 2006; MÜLLER, 2007; WEBSTER; LARK, 2013). A set of mathematical and heuristic rules is then formalized in the form of a computer algorithm to find the sampling locations that minimize (or maximize) that criterion. The more we know about the structure of spatial variation of $Y(\boldsymbol{s})$, the more likely we are to obtain the optimum sample configuration given the chosen statistical model.

However, the statistical model is usually unknown before we sample. This is especially common when multiple soil properties have to be mapped and the available information is insufficient to decide on the structure of the spatial variation. Because we usually want to make the least possible number of assumptions about the model structure, the safest solution is to use a space filling design (HENGL et al., 2003; DE GRUIJTER et al., 2006; MÜLLER, 2007; WALVOORT et al., 2010): the locations are selected as to generate a sample that covers the geographic and/or feature space(s) as evenly as possible. In areas with very little information on the spatial variation of the soil properties of interest, referred to as *terra incognita* by Webster & Oliver (2007), where usually there are operational constraints that limit the sampling to a single phase, an efficient spatial sample configuration should be optimized to identify the correct model structure, estimate model parameters, and make spatial predictions.

9.4.1 Sampling for Spatial Trend Estimation

The spatial trend corresponds to the spatial variation of $Y(\boldsymbol{s})$ that is explained linearly or non-linearly by the covariates. For a linear spatial trend, the sample should cover the extremes of the distribution of the covariates (MÜLLER, 2007). For models with interactions and/or higher order terms there are the response surface designs (BOX; WILSON, 1951; LESCH et al., 1995). These approaches produce clusters of points and ignore the spatial autocorrelation of the residuals (BRUS; HEUVELINK, 2007; MÜLLER, 2007). Optimal sampling designs for neural nets, random forests, etc., are yet unknown.

A common solution for spatial trend estimation in *terra incognita* is to use a feature space filling sample. Hengl et al. (2003) sampled along the marginal distribution of the covariates using equal-range strata with weights proportional to the frequency distribution. Minasny et al. (2007) sampled equal-variance geographic strata created using the variance of the covariates retained in their first principal component.

A more elaborated method, formulated as a multi-objective optimization problem composed of three objective functions, was developed by Minasny & McBratney (2006) based on Latin hypercube sampling (LHS), a non purposive, probability sampling method (MCKAY et al., 1979). The method, known as conditioned Latin hypercube sampling (CLHS), is based on sampling along the marginal distribution of the numeric and factor covariates using equal-area strata (quantile sampling) and proportionally to the area occupied by each level, respectively, and reproducing the linear correlations among the numeric covariates (MINASNY; MCBRATNEY, 2006). CLHS is very flexible and can be easily extended, two important reasons for its popularity (MINASNY; MCBRATNEY, 2010; ROUDIER et al., 2012).

Recently, Samuel-Rosa et al. (2016) proposed conceptual and algorithmic improvements on the CLHS. Algorithmic improvements concern the definition of the marginal sampling strata, the measurement of the correlation between covariates, and the aggregation of the objective functions. These have resulted in a more numerically stable sampling algorithm which not necessarily translates into more accurate spatial predictions. Conceptually, the goal of the method was reformulated as aiming at a spatial sample that reproduces an association/correlation measure and the marginal distribution of the covariates (ACDC). This was presented as a more appropriate definition of the method, while the original denomination given by Minasny & McBratney (2006) was appointed as being misleading because it can lead one to think of CLHS as a (non purposive) probability sampling method Samuel-Rosa et al. (2016).

9.4.2 Sampling for Variogram Estimation

A variogram model explains the spatially correlated random part of the spatial variation of $Y(\boldsymbol{s})$. Several sampling methods exist to identify and/or estimate the variogram and its parameters (BRUS; DE GRUIJTER, 1994; DE GRUIJTER et al., 2006; MÜLLER, 2007; WEBSTER; LARK, 2013). Modern ones focus on maximum likelihood estimators (LARK, 2002; ZIMMERMAN, 2006; MÜLLER, 2007), their limitation being that a minimum knowledge about the form of the variogram is required. A Bayesian approach was suggested to account for the uncertainty of the estimated variogram (DIGGLE; LOPHAVEN, 2006; MARCHANT; LARK, 2006; ZHU; STEIN, 2006). But it is hard to implement for multiple variables simultaneously, and the uncertainty is likely to increase with the number of parameters that need to be estimated.

Sampling for variogram estimation should concentrate on relevant pairwise distances (MÜLLER; ZIMMERMAN, 1999; LARK, 2002). But how to do that when we are ignorant about the shape of the variogram? Bresler & Green (1982), Russo (1984), Warrick & Myers (1987) proposed a conservative solution: the points should be located as to match a uniform distribution of pairwise distances. Their claim was that the sample would be globally optimal for an infinite set of unknown variograms. This has not been proven mathematically nor corroborated by empirical evidence. The resulting sample usually is redundant (poorly informative), concentrating most of the points in a single large cluster, with a few scattered points – many of the point-pairs are computed using the same subset of points.

Another critique to the idea of Bresler & Green (1982), Russo (1984), Warrick & Myers (1987) is that it was rooted on the use of the method-of-moments to fit a continuous function to the binned empirical variogram. Nowadays we recognize that the estimates of the method-of-moments are affected by the correlation between the sequence of classes of pairwise distances and that more robust methods exist (DIGGLE; RIBEIRO JR, 2002). Maximum likelihood methods estimate model parameters using all data points, avoiding the need for an *ad hoc* definition of classes of pairwise distances that generally smooth out the structure of the spatial process (LARK, 2000b).

9.4.3 Sampling for Spatial Interpolation

Kriging is the best unbiased linear predictor of soil properties (LARK et al., 2006). Overall better prediction accuracy depends on spreading the sample points as uniformly as possible throughout the study area. This is because for a stationary isotropic random field the kriging variance is a function only of the distance between sample points (CRESSIE, 1993). Regular sampling grids are commonly used to obtain a uniform geographic coverage, although triangular equilateral grids are more efficient (WEBSTER; OLIVER, 2007). Their main weakness is the inefficient coverage of the geographic space when the sampling region is irregularly shaped and/or contains irregularly shaped non-sampling areas (WALVOORT et al., 2010).

The regression-kriging approach for soil mapping (HENGL et al., 2007) lead to the development of sampling methods that account for both feature and geographic spaces. Hengl et al. (2003) proposed sampling iteratively in the feature space and keeping the sample configuration with the best geographic coverage. Minasny & McBratney (2006) developed a sampling strategy for spatial trend estimation and claimed that the geographic space could be considered as well. Minasny et al. (2007) suggested that a geographic stratification based on the variance of the covariates would take into consideration the geographic coverage. These methods are suboptimal for spatial interpolation because they essentially operate in the feature space.

It has been suggested that efficient optimization of sample configurations for spatial interpolation depends upon minimizing a distance-based metric (ROYLE; NYCHKA, 1998). One such metric is the mean squared shortest distance (MSSD) between sample and prediction points, which is equivalent to finding, for each point in the prediction grid, the nearest neighbouring sample point (BRUS et al., 2006). Because the MSSD takes into account all points in the prediction grid, its minimization produces a spatial sample that uniformly covers the geographic space, irrespective of the sampling region being irregularly shaped and/or containing irregularly shaped non-sampling areas (WALVOORT et al., 2010). However, this renders the method computationally expensive because a large distance matrix has to be computed every time a candidate spatial sample is generated.

The problem of minimizing the MSSD can be speeded up reformulating it in terms of an unsupervised classification problem. The objects to be classified are the points of the prediction grid and the classification variables are the x- and y-coordinates, the cluster centers defining the sampling locations (WALVOORT et al., 2010). This classification problem can be solved using the k-means clustering algorithm, which is computationally fast, but sensitive to local optima solutions.

9.5 WHICH SAMPLING ALGORITHM?

9.5.1 Sampling for Spatial Trend Estimation

There are multiple methods for designing optimum spatial sample configurations for spatial trend estimation in *terra incognita*, each with different complexity levels. The method of Minasny & McBratney (2006) is very well suited, as indicated by its popularity. Because Samuel-Rosa et al. (2016) produced a more numerically stable sampling algorithm, we suggest that the improved version of the CLHS called ACDC should be used instead.

Different from CLHS, ACDC is a multi-objective optimization problem composed of only two objective functions,

$$\text{CORR} = \sum_{i=1}^{p}\sum_{j=1}^{p} |\varphi_{ij} - \upsilon_{ij}|, \tag{9.1}$$

where φ_{ij} and υ_{ij} are the population and sample associations (or correlations in case all covariates are numeric) at the ith row and jth column of the p-dimensional population and sample association (or correlation) matrices, and

$$\text{DIST} = \sum_{i=1}^{p}\sum_{j=1}^{c_i} |\pi_{ij} - \gamma_{ij}|, \tag{9.2}$$

where π_{ij} and γ_{ij} are the proportion of sample and population points that fall in the jth class (or marginal sampling strata) of the ith covariate, c_i being the number of classes of the ith covariate. As such, ACDC is defined as follows:

$$\text{ACDC} = w_1\text{CORR} + w_2\text{DIST}, \tag{9.3}$$

with weights $w_1 = w_2 = 0.5$ when we do not have *a priori* preferences towards the objective functions.

9.5.2 Sampling for Variogram Estimation

An efficient and robust method for variogram estimation in *terra incognita* seems to be missing. For that end, we propose that sampling should be based on placing several small clusters scattered throughout the spatial domain as to maximize the amount of information carried by the sample. The most relevant pairwise distances are those that 1) enable an accurate estimate of the behaviour of the variogram near the origin and 2) produce a low estimate of the nugget variance. Our main concern is with the fact that the shape of the variogram and the estimated nugget variance determine the smoothness of the spatial predictions (WEBSTER; OLIVER, 2007).

The design of our sampling algorithm starts with the method proposed by Warrick & Myers (1987). Instead of equidistant (Figure 9.1a), we use exponentially spaced lag-distance classes defined up to the circumradius r of the bounding box of the area as proposed by Truong et al. (2013). The exponential spacings are created sequentially from the largest to the smallest lag by halving the immediately preceding larger lag, resulting in narrower lags in the left side of the variogram (Figure 9.1b). It works as follows:

1. Find r. Use the result to define the upper bound of the first rightmost lag.
2. Halve r. Use the result to define the lower bound of the first rightmost lag.
3. Go to the next lag.
4. Set the lower bound of the last lag as the upper bound of the current lag.
5. Halve the upper bound. Use the result to define the lower bound of the current lag.
6. Proceed as in 3–5 until the upper and lower bounds of the leftmost lag have been defined.

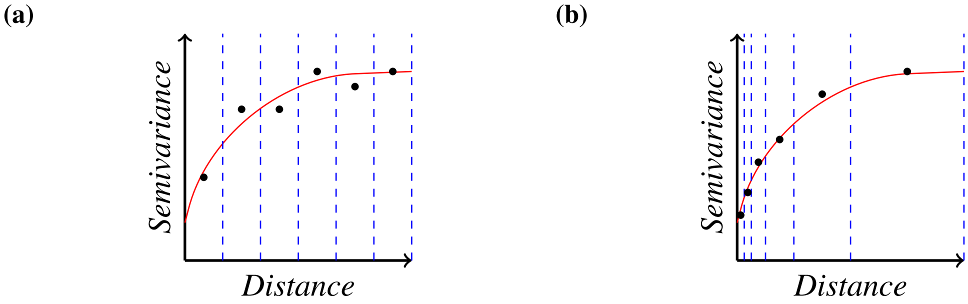


Figure 9.1: Representation of sample variograms with (a) equidistant and (b) exponentially spaced lag-distance classes. The dashed vertical lines represent the lower and upper bounds of the lag-distance classes.

The number of lag-distance classes depends on the wanted size of the smallest lag. For general applications, it seems appropriate to use a maximum of seven lags, where the size of the smallest lag will be approximately $2\,\%$ of the circumradius r of the bounding box of the area. Having defined the bounds of the lag-distance classes, our objective is to find a spatial sample configuration such that every point forms pairs with points that are separated by distances that fall in each of the (seven) lag-distance classes. In other words, we aim at each sample point contributing with at least one point-pair in each lag-distance class. For example, with seven lags and $n = 100$ sample points, the aimed solution is $\boldsymbol{l}^* = (100, 100, 100, 100, 100, 100, 100)$, i.e. a uniform distribution. When optimizing the sample configuration, the criterion to be minimized is the sum of differences between the vectors of the aimed $\boldsymbol{l}^*$ and observed $\boldsymbol{l}$ distributions of unique **P**oints **P**er **L**ag

$$\text{PPL} = \sum_{i=1}^{q} l_i^* - l_i, \tag{9.4}$$

where $\boldsymbol{l}^* = n$ for a uniform distribution, n being the number of sample points. This proposed modification differs from the original implementation of Warrick & Myers (1987) by the fact that the later aims at a uniform distribution of the number of *point-pairs* per lag-distance class.

9.5.3 Sampling for Spatial Interpolation

There are not many strategies for optimizing spatial samples for spatial interpolation in *terra incognita* as there are for spatial trend estimation. Among the existing options, the MSSD seems to be the most suited criterion. The main reasons for this are 1) its direct link with the need for minimizing the prediction error variance when making spatial predictions, and 2) its flexibility to deal with irregularly shaped sampling regions containing also irregularly shaped non-sampling areas.

The MSSD is given by

$$\text{MSSD} = \frac{1}{N}\sum_{i=1}^{N} min_j(D_{ij}^2), \tag{9.5}$$

where N is the number of points in the prediction grid, D_{ij}^2 is the squared Euclidean distance between the i point in the prediction grid and the j sampling point computed using the x- and y-coordinates, and min_j refers to taking the minimum over all j's for each i.

9.6 SAMPLING IN *TERRA INCOGNITA*

We propose a heuristic, general-purpose method to design sample configurations for soil mapping in *terra incognita*. Like sampling for spatial trend estimation (ACDC), it is based on solving a multi-objective combinatorial optimization problem (MOCOP) (Equation 8.4.3). A utility function is defined aggregating the three criteria described above using the weighted sum method (Equation 8.5) so that the sample points cover, extend over, spread over, SPAN the feature, variogram and geographic spaces,

$$\text{SPAN} = w_1\text{CORR} + w_2\text{DIST} + w_3\text{PPL} + w_4\text{MSSD}, \tag{9.6}$$

with $w_1 = w_2$ and $w_1 + w_2 = w_3 + w_4$ in the *terra incognita* setting. Before aggregation, each objective function is scaled to the same approximate range of values using the upper-lower bound approach with the Pareto minimum and maximum values (Equation 8.4.3) so that any potential numerical dominance can be eliminated or minimized, and the weights can play the desired role (MARLER; ARORA, 2005; MARLER; ARORA, 2009). The multi-objective optimization problem of sampling in *terra incognita* (SPAN) can be solved using spatial simulated annealing as implemented in the `R`-package `spsann` (Section 8.5.2.1 and Appendix A). Simulated annealing is a popular method with widespread use to solve combinatorial optimization problems due to its robustness against local optima and easiness of implementation (METROPOLIS et al., 1953; KIRKPATRICK et al., 1983; ČERNÝ, 1985; AARTS; KORST, 1989; VAN GROENIGEN, 1999).

9.7 CONCLUSIONS

This chapter presented sound strategies for optimizing spatial samples to estimate the spatial trend and the variogram, and make spatial predictions when we know very little about the soil spatial distribution. Overall, the main requirement is the formulation of a sound multi-objective combinatorial optimization problem (MOCOP) using robust versions of existing sampling algorithms. The aimed spatial sample should reproduce the marginal distribution of the covariates such that the spatial trend can be accurately estimated. This can be achieved using the recently improved version of the conditioned Latin hypercube sampling algorithm (ACDC).

The spatial sample should also contain several small clusters scattered throughout the sampling region, the reason being the need for an accurate estimate of the behaviour of the variogram near the origin. An efficient metric for achieving such objective is the number of unique points that form point-pairs in each of the exponentially spaced lag-distance classes of the sample variogram (PPL). Optimally, every point would form point-pairs in each lag class. Finally, the sampling region should be covered the most uniformly possible such that the average prediction error variance is the least possible. For that end, one can minimize the distance between every prediction point and its nearest neighbouring sampling point (MSSD).

We believe that the proposed general purpose sampling strategy need to be evaluated compared to (i) the single version of each objective function, (ii) popular sampling designs such as regular grids, and (iii) sampling algorithms that assume the model of soil spatial variation as known. The effect of sample size need to be addressed as well. Unfortunately, resource restrictions hinder the execution of such an evaluation using field data because sampling costs are high. A reasonable solution is to use synthetic data derived from a real-world case study. The first step would consist of defining soil data generating processes using existing point soil observations and spatially exhaustive covariates. At least two generating processes should be defined, possibly using different soil properties, such that the deterministic and random components of soil spatial variation have different forms. For example, linear and non-linear trends coupled with exponential and Gaussian variograms. Next, the generating processes would be used to produce multiple realizations of the soil data, from which we would sample using the spatial samples under comparison. Evaluation of sampling strategies would consist of measuring how well the spatial samples (i) capture the true form of the soil data generating process and (ii) make spatial predictions. The outcome of such an exercise should help us understanding on how to decide upon sampling strategies when we want to learn about the soil-landscape relationships and make accurate predictions.

10 GENERAL CONCLUSIONS

This thesis has made a pedological contribution with the development of a comprehensive description of the soil-forming factors and processes that determine the spatio-temporal distribution of soil properties in the Santa Maria case study area. The conceptual model of pedogenesis, presented in Chapter IV, showed that the spatial distribution of soil properties is highly variable, even when under the same land use. At coarse spatial scales, this spatial variation is determined by the geological and geomorphological diversity of the area, while at fine spatial scales, past and current (poor) agricultural practices seem to play a major role. Along with the conceptual model of pedogenesis, Chapter II and Chapter III constitute a technical contribution of this thesis. These chapters provide the basis for soil spatial modelling exercises in the study area.

Chapter V demonstrated that existing, freely available covariates are suitable for calibrating soil spatial models. It was shown that using more detailed covariates results in only a modest increase in the prediction accuracy of linear soil spatial models. The observed increase is comparable to the effect of incorporating spatial dependence in the soil spatial model, and may not outweigh the extra costs of using more detailed covariates. In general, a more detailed covariate has a greater potential to improve prediction accuracy when a soil property is poorly predicted by its less detailed version. However, the magnitude of the improvement may depend on which other covariates are included in the model. Choosing whether or not to invest in more detailed covariates depends on the strength of the relationship between the covariates and the soil property being modelled, and on the relative difference between the less, and more detailed versions of the covariates. It is likely better to substantially improve the detail of a less influential covariate than marginally increase the detail of the most influential covariate. However, one should always consider if more efficient means of increasing prediction accuracy exist (e.g. obtaining more soil observations).

Chapter VI showed that several factors influence how field soil spatial modellers decide upon where to place soil observation locations. These are of three types: conceptual, operational, and psychological. The first concerns the knowledge of the soil spatial modellers about soil-landscape relationships, and seems to be connected with the years of field experience. The second relates to the available resources (infrastructure, workforce, and budget) to make soil observations, as well as to access restrictions imposed by landowners and geographic barriers, for example. The third relates to how the soil modellers perceive their surrounding physical environment and how the course of their motivation shifts during the soil observation process. Point pattern analysis helped understanding that there is a trade-off between conceptual and operational factors, which determines how the motivation of field soil modellers shifts focus towards one or another immediate goal. Depending on the focal goal, the resulting sample configuration resembles a random (learning/verifying soil-landscape relationships – means-focused motivation) or a regular (maximizing the number of observations and geographic coverage – outcome-focused motivation) point pattern.

Chapter VII showed that the conditioned Latin hypercube sampling algorithm, a popular algorithm used to optimize spatial sample configurations for spatial trend estimation, can be considerably improved. Compared to the original CLHS, our proposed modifications resulted in a sampling algorithm with an improved numerical behaviour, but this does not necessarily translates into improved prediction accuracy. For instance, sample size has a larger influence on prediction accuracy than the sampling algorithm. However, aiming only at the association/correlation between covariates degrades prediction accuracy possibly because the coverage of the geographic space is poorer. As such, when optimizing a sample configuration for spatial trend estimation, it should suffice to aim only at reproducing the marginal distribution of the covariates. This should be done using only the non-empty marginal sampling strata.

Chapter VIII showed how to optimize sample configurations for spatial trend and variogram estimation, and spatial interpolation in situations where we know very little about the soil spatial distribution. The only requirement is that one formulates a sound multi-objective optimization problem using robust versions of existing sampling algorithms. The resulting spatial sample should reproduce the marginal distribution of the covariates such that the spatial trend can be accurately estimated. It should also contain several small clusters scattered throughout the spatial domain to enable making an accurate estimate of the behaviour of the variogram, specially near the origin. Finally, it should cover the sampling region in the most uniform way such that the average prediction error variance is the least possible.

This thesis has also contributed with two packages for the software environment for statistical computing and graphics `R`. The first package, called `pedometrics` (Appendix B), contains various functions for spatial exploratory data analysis and model calibration designed for the development of this thesis. The second package, called `spsann` (Appendix A), contains functions to optimize sample configurations to identify and estimate the variogram and spatial trend, and make spatial predictions. The latter was developed as part of Chapter VII and Chapter VIII. Both are freely available and can be obtained from The Comprehensive R Archive Network (CRAN).

Overall, this thesis showed that the complex interplay between soil and covariate data can have a large influence on the accuracy of soil maps. A single, universal, cost-effective recipe for reducing uncertainty in soil spatial modelling seems out of range. The case studies suggested that solutions are case specific and primarily depend on the existing soil and covariate data. Obtaining more soil samples showed to be an efficient strategy provided the available resources allow extra sampling. Otherwise, deciding upon cost-effective ways of reducing uncertainty requires, first, that we explore the full potential of existing soil and covariate data using robust spatial modelling techniques. Such an exercise requires a comprehensive knowledge of the soil-landscape relationships, as well as a thorough documentation of the soil and covariate data so that their weaknesses and strengths can be easily identified. Then, the decision of whether to invest on improving the quality of soil or covariate or both data sources will depend upon the trade-off between the increased data/prediction quality and the amount of resources required.

11 REFERÊNCIAS BIBLIOGRÁFICAS

AARTS, E. H. L.; KORST, J. H. M. Boltzmann machines for travelling salesman problems. *European Journal of Operational Research*, Elsevier BV, v. 39, n. 1, p. 79–95, Mar 1989. ISSN 0377-2217.

ABRÃO, P. U. R.; GIANLUPE, D.; AZOLIN, M. A. D. *Levantamento semi-detalhado dos solos da Estação Experimental de Silvicultura de Santa Maria*. Porto Alegre, 1988.

AERTS, J. C. J. H.; HEUVELINK, G. B. M. Using simulated annealing for resource allocation. *International Journal of Geographical Information Science*, v. 16, n. 6, p. 571–587, 2002. ISSN 13658816.

AGRESTI, A. *Categorical data analysis*. 2. ed. New York: Wiley-Interscience, 2002. 710 p. ISBN 0471360937. Disponível em: <http://www.stat.ufl.edu/~aa/cda2/cda.html>.

ANDERSEN, C. M.; BRO, R. Variable selection in regression – a tutorial. *Journal of Chemometrics*, v. 24, n. 11–12, p. 728–737, 2010.

ANTUNES, M. A. H.; SIQUEIRA, J. C. S. Características das imagens RapidEye para mapeamento e monitoramento agrícola e ambiental. In: EPIPHANIO, J. C. N.; GALVÃO, L. S. (Ed.). *Anais XVI Simpósio Brasileiro de Sensoriamento Remoto*. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2013. p. 547–554. Disponível em: <http://www.dsr.inpe.br/sbsr2013/files/p1253.pdf>.

ARORA, J. *Introduction to optimum design*. 3. ed. Waltham: Academic Press, 2011. 896 p. ISBN 978-0-12-381375-6.

AZOLIN, M. A. D. *Podologia das áreas marginais dos rios Ibicuí e Vacacaí*. Porto Alegre, 1977. 71 p.

AZOLIN, M. A. D.; MUTTI, L. S. M. *Solos da bacia hidrográfica do Vacacaí-Mirim*. Porto Alegre: DNOS-UFSM, 1988. 20 p. Disponível em: <http://1drv.ms/UAFIOK>.

BADDELEY, A. *Analysing spatial point patterns in R*. Canberra, 2010. 232 p. Disponível em: <http://www.coactivate.org/projects/plein-r/project-home/Baddeley_SPP-workshop_CSIRO_2008.pdf>.

BADDELEY, A. J.; MOLLER, J.; WAAGEPETERSEN, R. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerland*, Wiley-Blackwell, v. 54, n. 3, p. 329–350, Nov 2000. ISSN 1467-9574.

BARRERA-BASSOLS, N.; ZINCK, J. A. Ethnopedology: a worldwide view on the soil knowledge of local people. *Geoderma*, v. 111, p. 171–195, 2003. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S001670610200263X>.

BASHER, L. R. Is pedology dead and buried? *Australian Journal of Soil Research*, v. 35, p. 979–994, 1997. Disponível em: <http://www.publish.csiro.au/paper/S96110>.

BATLLE-BAYER, L.; BATJES, N. H.; BINDRABAN, P. S. Changes in organic carbon stocks upon land use conversion in the Brazilian Cerrado: A review. *Agriculture, Ecosystems & Environment*, Elsevier BV, v. 137, n. 1-2, p. 47–58, apr 2010.

BAZAGLIA FILHO, O.; RIZZO, R.; LEPSCH, I. F.; PRADO, H. do; GOMES, F. H.; MAZZA, J. A.; DEMATTÊ, J. A. M. Comparison between detailed digital and conventional soil maps of an area with complex geology. *Revista Brasileira de Ciência do Solo*, FapUNIFESP (SciELO), v. 37, n. 5, p. 1136–1148, 2013.

BEHRENS, T.; ZHU, A. X.; SCHMIDT, K.; SCHOLTEN, T. Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, v. 155, n. 3–4, p. 175–185, 2010. ISSN 0016-7061.

BIRKELAND, W. *Soils and geomorphology*. 3. ed. New York: Oxford University Press., 1999. 430 p.

BIVAND, R. S.; PEBESMA, E. J.; GÓMEZ-RUBIO, V. *Applied spatial data analysis with R*. 1. ed. New York: Springer, 2008. 374 p.

BIVAND, R. S.; PEBESMA, E. J.; GÓMEZ-RUBIO, V. *Applied spatial data analysis with R*. 2. ed. New York: Springer, 2013. 405 p.

BLANCO-CANQUI, H.; LAL, R. *Principles of soil conservation and management*. Dordrecht: Springer, 2008. 617 p. ISBN 9048185297.

BOCKHEIM, J. G.; GENNADIYEV, A. N. The role of soil-forming processes in the definition of taxa in soil taxonomy and the world soil reference base. *Geoderma*, v. 95, n. 1, p. 53–72, 2000.

BOCKHEIM, J. G.; GENNADIYEV, A. N. Soil-factorial models and earth-system science: a review. *Geoderma*, Elsevier BV, v. 159, n. 3–4, p. 243–251, Nov 2010. ISSN 0016-7061.

BONEZZI, A.; BRENDL, C. M.; ANGELIS, M. D. Stuck in the middle: the psychophysics of goal pursuit. *Psychological Science*, v. 22, n. 5, p. 607–612, 2011.

BONNES, M.; BONAIUTO, M. Environmental psychology: from spatial-physical environment to sustainable development. In: ____. *Handbook of environmental psychology*. New York: John Wiley & Sons, 2002. p. 28–54.

BORTOLUZZI, C. A. Contribuição à geologia da região de Santa Maria, Rio Grande do Sul, Brasil. *Pesquisas em Geociências*, v. 4, n. 1, p. 7–86, 1974. Disponível em: <http://seer.ufrgs.br/PesquisasemGeociencias/article/view/21834>.

BOX, G. E. P. Science and statistics. *Journal of the American Statistical Association*, v. 71, n. 356, p. 791–799, 1976. Disponível em: <http://www.jstor.org/stable/2286841>.

BOX, G. E. P.; WILSON, K. B. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society. Series B (Methodological)*, Wiley for the Royal Statistical Society, v. 13, n. 1, p. 1–45, 1951. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2983966>.

BRASIL. *Levantamento de reconhecimento dos solos do estado do Rio Grande do Sul*. Recife: Ministério da Agricultura. Departamento Nacional de Pesquisa Agropecuária. Divisão de Pesquisa Pedológica, 1973. 431 p. Escala 1:750 000. Disponível em: <http://library.wur.nl/isric/fulltext/isricu_i00003061_001.pdf>.

BRASIL. *Mapa geológico da folha Santa Maria*. Santa Maria, 1980. (1:50 000).

BRASIL. *Decreto nº 89.817, de 20 de junho de 1984. Estabelece as Instruções Reguladoras das Normas Técnicas da Cartografia Nacional.* Brasília: Diário Oficial da República Federativa do Brasil], 1984. 8884–8886 p. Disponível em: <http://www.concar.ibge.gov.br/detalheDocumentos.aspx?cod=8>.

BRASIL. *Geo catálogo do Ministério do Meio Ambiente – manual de uso*. 1.0. ed. Brasília, 2012. 35 p. Disponível em: <http://geocatalogo.mma.gov.br/>.

BREGT, A. K.; BOUMA, J.; JELLINEK, M. Comparison of thematic maps derived from a soil map and from kriging of point data. *Geoderma*, Elsevier BV, v. 39, n. 4, p. 281–291, may 1987. Disponível em: <http://dx.doi.org/10.1016/0016-7061(87)90048-6>.

BREIMAN, L. Random forests. *Machine Learning*, Springer Science + Business Media, v. 45, n. 1, p. 5–32, 2001. ISSN 0885-6125.

BRESLER, E.; GREEN, R. E. *Soil parameters and sampling scheme for characterizing soil hydraulic properties of a watershed*. Honolulu, 1982. 42 p. Technical Report 148. Disponível em: <http://hdl.handle.net/10125/1983>.

BREVIK, E. C.; HARTEMINK, A. E. Early soil knowledge and the birth and development of soil science. *Catena*, Elsevier BV, v. 83, n. 1, p. 23–33, oct 2010. Disponível em: <http://dx.doi.org/10.1016/j.catena.2010.06.011>.

BRUS, D. J. Balanced sampling: a versatile sampling approach for statistical soil surveys. *Geoderma*, v. 253–254, p. 111–121, 2015.

BRUS, D. J.; DE GRUIJTER, J. J. Estimation of non-ergodic variograms and their sampling variance by design-based sampling strategies. *Mathematical Geology*, Springer Science + Business Media, v. 26, n. 4, p. 437–454, May 1994. ISSN 1573-8868.

BRUS, D. J.; DE GRUIJTER, J. J.; VAN GROENIGEN, J. W. Designing spatial coverage samples using the k-means clustering algorithm. In: LAGACHERIE, A. M. P.; VOLTZ, M. (Ed.). *Digital soil mapping - an introductory perspective*. Amsterdam: Elsevier, 2006, (Developments in Soil Science, v. 31). cap. 14, p. 183–192.

BRUS, D. J.; HEUVELINK, G. B. M. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, v. 138, p. 86–95, 2007.

BRUS, D. J.; KEMPEN, B.; HEUVELINK, G. B. M. Sampling for validation of digital soil maps. *European Journal of Soil Science*, v. 62, n. 3, p. 394–407, 2011.

BURROUGH, P. A. Multiscale sources of spatial variation in soil. I. The application of fractal concepts to nested levels of soil variation. *Journal of Soil Science*, Wiley-Blackwell, v. 34, n. 3, p. 577–597, sep 1983.

CAMARGO, M. N.; JACOMINE, P. K. T.; OLMOS, J.; CARVALHO, A. P. Proposição preliminar de conceituação e distinção de Podzólico Vermelho-Escuro. In: *Conceituação sumária de algumas classes de solos recém-reconhecidas nos levantamentos e estudos de correlação do SNLCS*. Rio de Janeiro: Serviço Nacional de Levantamento e Conservação do Solo, 1982. p. 7–12. Circular técnica 1.

CAMBULE, A. H.; ROSSITER, D. G.; STOORVOGEL, J. J. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma*, v. 192, n. 0, p. 341–353, 2013.

CARRÉ, F.; MCBRATNEY, A. B.; MINASNY, B. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma*, Elsevier BV, v. 141, n. 1-2, p. 1–14, Sep 2007. ISSN 0016-7061.

CARRILLO, G. *vec2dtransf: 2D cartesian coordinate transformation*. [S.l.], 2012. 16 p. R package version 1.0. Disponível em: <http://CRAN.R-project.org/package=vec2dtransf>.

CARVALHO, A. P. Conceituação de terra Bruna Estruturada. In: *Conceituação sumária de algumas classes de solos recém-reconhecidas nos levantamentos e estudos de correlação do SNLCS*. Rio de Janeiro: Serviço Nacional de Levantamento e Conservação do Solo, 1982. p. 21–24. (Circular técnica 1).

CARVALHO, F. M.; MARCO, P. D.; FERREIRA, L. G. The Cerrado into-pieces: habitat fragmentation as a function of landscape use in the savannas of central Brazil. *Biological Conservation*, Elsevier BV, v. 142, n. 7, p. 1392–1403, jul 2009.

CARVALHO JR, W.; CHAGAS, C. S.; MUSELLI, A.; PINHEIRO, H. S. K.; PEREIRA, N. R.; BHERING, S. B. Conditioned Latin hypercube method for soil sampling in the presence of environmental covariates for digital soil mapping. *Revista Brasileira de Ciência do Solo*, v. 38, n. 2, p. 386–396, 2014. ISSN 0100-0683.

CAVAZZI, S.; CORSTANJE, R.; MAYR, T.; HANNAM, J.; FEALY, R. Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma*, v. 195-196, n. 0, p. 111–121, 2013. ISSN 0016-7061.

ČERNÝ, V. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications*, Springer Science + Business Media, v. 45, n. 1, p. 41–51, Jan 1985. ISSN 1573-2878.

CHATFIELD, C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, v. 158, n. 3, p. 419–466, 1995. Disponível em: <http://www.jstor.org/stable/2983440>.

CHRISTENSEN, O. F.; DIGGLE, P. J.; RIBEIRO JR, P. J. Analysing positive-valued spatial data: the transformed Gaussian model. In: MONESTIEZ, P.; ALLARD, D.; FROIDEVAUX, R. (Ed.). *Proceedings of the Third European Conference on Geostatistics for Environmental Applications*. Avignon: geoENVia Association, 2001. (Quantitative Geology and Geostatistics, v. 11), p. 287–298.

CHURCHMAN, G. J. The philosophical status of soil science. *Geoderma*, Elsevier BV, v. 157, n. 3–4, p. 214–221, jul 2010. Disponível em: <http://dx.doi.org/10.1016/j.geoderma.2010.04.018>.

CLAESSEN, M. E. C.; BARRETO, W. O.; PAULA, J. L.; DUARTE, M. N. *Manual de métodos de análise de solo*. 2. ed. Rio de Janeiro: Embrapa, 1997. 212 p.

CLIFFORD, D.; PAYNE, J. E.; PRINGLE, M.; SEARLE, R.; BUTLER, N. Pragmatic soil survey design using flexible Latin hypercube sampling. *Computers & Geosciences*, Elsevier BV, v. 67, p. 62–68, Jun 2014. ISSN 0098-3004.

COETERIER, J. Cues for the perception of the size of space in landscapes. *Journal of Environmental Management*, v. 42, n. 4, p. 333–347, 1994. ISSN 0301-4797. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0301479784710760>.

COMISSÃO PASTORAL DA TERRA. *Carta aberta à Sociedade Brasileira e à Presidência da República e ao Congresso Nacional sobre a destruição do Cerrado pelo MATOPIBA*. 2015. Eletronic. Carta final do I Encontro Regional dos Povos e Comunidades do Cerrado. Disponível em: <http://goo.gl/OgxdvS>.

COOKE, R. M. *Experts in uncertainty – opinion and subjective probability in science*. Oxford: Oxford University Press, 1991. 321 p.

CORREIA, J. R. *Pedologia e conhecimento local: proposta metodológica de interlocução entre saberes construídos por pedólogos e agricultores em área de Cerrado em Rio Pardo de Minas, MG*. 234 p. Tese (Doutorado) — Curso de Pós-graduação em Agronomia – Ciência do Solo, Universidade Federal Rural do Rio de Janeiro, 2005. Disponível em: <http://www.cpac.embrapa.br/quadro/87>.

CPRM. *Programa levantamentos geológicos básicos do Brasil - Agudo, Folha Sh.22-V-C-V, Estado do Rio Grande do Sul*. Brasília: CPRM (Serviço Geológico do Brasil), 2007. 97 p. (1:100 000).

CRAMÉR, H. *Mathematical methods of statistics*. Princeton: Princeton University Press, 1946. 575 p. ISBN 0-691-08004-6.

CRESSIE, N. The origins of kriging. *Mathematical Geology*, Springer Science + Business Media, v. 22, n. 3, p. 239–252, apr 1990.

CRESSIE, N. A. C. *Statistics for spatial data*. New York: John Wiley & Sons, 1993. 900 p. Disponível em: <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471002550.html>.

DALMOLIN, R. S. D. Faltam pedólogos no brasil. *Boletim Informativo da Sociedade Brasileira de Ciência do Solo*, v. 24, p. 13–15, 1999.

DALMOLIN, R. S. D.; GONÇALVES, C. N.; DICK, D. P.; KNICKER, H.; KLAMT, E.; KÖGEL-KNABNER, I. Organic matter characteristics and distribution in Ferralsol profiles of a climosequence in southern Brazil. *European Journal of Soil Science*, v. 57, p. 644–654, 2006.

DE GRUIJTER, J. J.; BRUS, D.; BIERKENS, M.; KNOTTERS, M. *Sampling for natural resource monitoring*. Berlin: Springer, 2006. 332 p. Disponível em: <http://www.springer.com/environment/environmental+toxicology/book/978-3-540-22486-0>.

DE GRUIJTER, J. J.; TER BRAAK, C. J. F. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, v. 22, p. 407–415, 1990.

DEUTSCH, C. V. *Annealing techniques applied to reservoir modeling and the integration of geological and engineering (well test) data*. 306 p. Tese (Doutorado) — Department of Applied Earth Sciences, Stanford University, 1992.

DIAS, J. R. *Aplication of the AGNPS2001 utilizing observed data in the Vacacaí-Mirim River watershed*. 118 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Engenharia Civi, Universidade Federal de Santa Maria, Santa Maria, 2003. Disponível em: <http://w3.ufsm.br/ppgec/wp-content/uploads/Janaina.pdf>.

DIGGLE, P.; LOPHAVEN, S. Bayesian geostatistical design. *Scandinavian Journal of Statistics*, Wiley-Blackwell, v. 33, n. 1, p. 53–64, Mar 2006. ISSN 1467-9469.

DIGGLE, P. J. A kernel method for smoothing point process data. *Applied Statistics (Journal of the Royal Statistical Society, Series C)*, v. 34, p. 138–147, 1985.

DIGGLE, P. J. *Statistical analysis of spatial point patterns*. 2. ed. New York: Oxford University Press, 2003.

DIGGLE, P. J.; RIBEIRO JR, P. J. Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling*, Informa UK Limited, v. 6, n. 2, p. 129–146, Nov 2002. ISSN 1469-8323.

DIGGLE, P. J.; RIBEIRO JR, P. J. *Model-based geostatistics*. 1. ed. New York: Springer, 2007. 228 p. Disponível em: <http://www.springer.com/earth+sciences+and+geography/book/978-0-387-32907-9>.

DILL, P. R. J.; PAIVA, E. M. C. D.; PAIVA, J. B. D.; ROCHA, J. S. M. Assoreamento do reservatório do Vacacaí-Mirim e sua relação com a deterioração da bacia hidrográfica contribuinte. *Revista Brasileira de Recursos Hídricos*, v. 9, p. 7–15, 2004. Disponível em: <http://jararaca.ufsm.br/websites/eloiza/download/Dill/RBRH-Dill.pdf>.

DOMBURG, P.; DE GRUIJTER, J. J.; VAN BEEK, P. Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. *Geoderma*, v. 75, n. 3-4, p. 183–201, 1997.

DRĂGUŢ, L.; SCHAUPPENLEHNER, T.; MUHAR, A.; STROBL, J.; BLASCHKE, T. Optimization of scale and parametrization for terrain segmentation: an application to soil-landscape modeling. *Computers & Geosciences*, Elsevier BV, v. 35, n. 9, p. 1875–1883, Sep 2009. ISSN 0098-3004.

DRAPER, N. R.; GUTTMAN, I.; KANEMASU, H. The distribution of certain regression statistics. *Biometrika*, v. 58, n. 2, p. 295–298, 1971. Disponível em: <http://www.jstor.org/stable/2334517>.

DRAPER, N. R.; SMITH, H. *Applied regression analyis*. 3. ed. Wiley, 1998. 736 p. (Probability and Statistics). ISBN 978-0-471-17082-2. Disponível em: <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471170828.html>.

DSG. *Camobi – SO. Folha SH.22-V-C-IV/2-SO*. Brasília: Ministério do Exército, Departamento de Engenharia e Comunicações, Diretoria do Serviço Geográfico do Exército, 1980. (1:25 000).

DSG. *Santa Maria – NE. Folha SH.22-V-C-IV-1-NE*. Brasília: Ministério do Exército, Departamento de Engenharia e Comunicações, Diretoria do Serviço Geográfico do Exército, 1992. (1:25 000).

DSG. *Santa Maria – SE. Folha SH.22-V-C-IV/1-SE*. Brasília: Ministério do Exército, Departamento de Engenharia e Comunicações, Diretoria do Serviço Geográfico do Exército, 1992. (1:25 000).

DUH, J.-D.; BROWN, D. G. Knowledge-informed pareto simulated annealing for multi-objective spatial allocation. *Computers, Environment and Urban Systems*, v. 31, n. 3, p. 253–281, 2007. ISSN 0198-9715.

DULLIUS, M. *Vegetação e solos de uma floresta estacional do Rio Grande do Sul*. 127 p. Dissertação (Mestrado) — Programa de Pós-Grauação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2012. Disponível em: <http://w3.ufsm.br/ppgcs/>.

DUNGAN, J. L.; PERRY, J. N.; DALE, M. R. T.; LEGENDRE, P.; CITRON-POUSTY, S.; FORTIN, M. J.; JAKOMULSKA, A.; MIRITI, M.; ROSENBERG, M. S. A balanced view of scale in spatial statistical analysis. *Ecography*, Wiley-Blackwell, v. 25, n. 5, p. 626–640, Oct 2002. ISSN 1600-0587.

EDIRISOORIYA, G. Stepwise regression is a problem, not a solution. In: *Annual Meeting of the Mid-South Educational Research Association*. Biloxi: Mid-South Educational Research Association, 1995. p. 16. Disponível em: <http://www.eric.ed.gov/>.

ELDEIRY, A. A.; GARCIA, L. A. Detecting soil salinity in alfalfa fields using spatial modeling and remote sensing. *Soil Science Society of America Journal*, Soil Science Society of America, v. 72, n. 1, p. 201–211, 2008. ISSN 1435-0661.

EPSTEIN, R.; KANWISHER, N. A cortical representation of the local visual environment. *Nature*, Macmillan Magazines Ltd., v. 392, n. 6676, p. 598–601, abr. 1998. ISSN 0028-0836. Disponível em: <http://dx.doi.org/10.1038/33402>.

ESPINDOLA, C. R. *Retrospectiva crítica sobre a pedologia – um repasse biliográfico*. 1. ed. Campinas: Editora da Unicamp, 2008. 397 p.

EVERITT, B. S. *The Cambridge dictionary of statistics*. 3. ed. Cambridge: Cambridge University Press, 2006. 432 p.

FAO. *The FAO voluntary guidelines for the right to food: lasting solutions against hunger*. Roma, 2005. 4 p. Disponível em: <http://www.fao.org/righttofood/KC/downloads/vl/docs/>.

FAO. *Guidelines for soil description*. 4. ed. Rome: FAO, 2006. 97 p. Disponível em: <ftp://ftp.fao.org/agl/agll/docs/guidel_soil_descr.pdf>.

FAO. *Pathways to success. Success stories in agricultural production and food security*. Rome, 2009. 34 p. Disponível em: <http://www.fao.org/fileadmin/user_upload/newsroom/docs/pathways.pdf>.

FAO. *State of food insecurity in the World: 2015*. Rome, 2015. 56 p. Disponível em: <http://reliefweb.int/sites/reliefweb.int/files/resources/a-i4646e.pdf>.

FARRAR, D. E.; GLAUBER, R. R. Multicollinearity in regression analysis: the problem revisited. *The Review of Econonomics and Statistics*, v. 49, p. 92–107, 1967. Disponível em: <http://hdl.handle.net/1721.1/48530>.

FERNANDES, B. M. Development models for the Brazilian countryside: paradigmatic and territorial disputes. *Latin American Perspectives*, SAGE Publications, v. 43, n. 2, p. 48–59, jan 2016.

FINKE, P. A. On digital soil assessment with models and the pedometrics agenda. *Geoderma*, v. 171-172, p. 3–15, 2012.

FISHER, P. F.; TATE, N. J. Causes and consequences of error in digital elevation models. *Progress in Physical Geography*, v. 30, n. 4, p. 467–489, 2006.

FLORINSKY, I. V. Accuracy of local topographic variables derived from digital elevation models. *International Journal of Geographical Information Science*, v. 12, p. 47–61, 1998.

FLORINSKY, I. V. The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication). *Eurasian Soil Science*, MAIK Nauka/Interperiodica distributed exclusively by Springer Science+Business Media LLC., v. 45, p. 445–451, 2012. ISSN 1064-2293.

FOX, J.; WEISBERG, S. *An R companion to applied regression*. 2. ed. Thousand Oaks: Sage, 2011. Disponível em: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>.

GASCH, C. K.; HENGL, T.; GRÄLER, B.; MEYER, H.; MAGNEY, T. S.; BROWN, D. J. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: the Cook Agronomy Farm data set. *Spatial Statistics*, Elsevier BV, v. 14, p. 70–90, nov 2015.

GASPARETTO, N. G. L.; MACIEL FILHO, C. L.; MEDEIROS, E. R.; MENEGOTTO, E.; SARTORI, P. L. P.; VEIGA, P. *Mapa geológico da Folha de Santa Maria*. Santa Maria, 1988. 1 p. (1:50 000).

GDAL DEVELOPMENT TEAM. *GDAL – Geospatial Data Abstraction Library*. [S.l.], 2013. (GDAL 1.10.0, released 2013/04/24). Disponível em: <http://www.gdal.org>.

GENTON, M. G. Highly robust variogram estimation. *Mathematical Geology*, Springer Science + Business Media, v. 30, n. 2, p. 213–221, 1998.

GESSLER, P. E.; MOORE, I. D.; MCKENZIE, N. J.; RYAN, P. J. Soil-landscape modelling and spatial prediction of soil attributes. *International Journal of Geographical Information Systems*, v. 9, n. 4, p. 421–432, 1995.

GOBIN, A.; CAMPLING, P.; FEYEN, J. Soil-landscape modelling to quantify spatial variability of soil texture. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, v. 26, n. 1, p. 41 – 45, 2001.

GOLDANI, J. Z. *Ocupação antrópica e sócio-ambiental na área de captação do reservatório do DNOS, Santa Maria-RS*. 104 p. Dissertação (Mestrado) — Graduate School in Geomatics, Universidade Federal de Santa Maria, Santa Maria, 2006. Disponível em: <http://cascavel.ufsm.br/tede/tde_busca/arquivo.php?codArquivo=120>.

GOOVAERTS, P. *Geostatistics for natural resources evaluation*. Oxford: Oxford University Press, 1997. 483 p. ISBN 0-19-511538-4.

GOOVAERTS, P. Estimation or simulation of soil properties? An optimization problem with conflicting criteria. *Geoderma*, v. 97, p. 165–186, 2000.

GOOVAERTS, P. Geostatistical modelling of uncertainty in soil science. *Geoderma*, v. 103, n. 1?2, p. 3–26, 2001.

GRUNWALD, S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, v. 152, n. 3-4, p. 195–207, 2009.

GRUNWALD, S. Current state of digital soil mapping and what is next. In: BOETTINGER, J.; HOWELL, D.; MOORE, A.; HARTEMINK, A.; KIENAST-BROWN, S. (Ed.). *Digital Soil Mapping*. Springer Netherlands, 2010, (Progress in Soil Science, v. 2). p. 3–12. ISBN 978-90-481-8862-8. Disponível em: <http://dx.doi.org/10.1007/978-90-481-8863-5_1>.

GRUNWALD, S.; THOMPSON, J. A.; BOETTINGER, J. L. Digital soil mapping and modeling at continental scales: finding solutions for global issues. *Soil Science Society of America Journal*, Soil Science Society of America, v. 75, n. 4, p. 1201–1213, 2011.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, v. 3, p. 1157–1182, 2003. Disponível em: <http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf>.

HACK, C.; LONGHI, S. J.; BOLIGON, A. A.; MURARI, A. B.; PAULESKI, D. T. Análise fitossociológica de um fragmento de floresta estacional decidual no município de Jaguari, RS. *Ciência Rural*, v. 35, p. 1083–1091, 2005.

HALDAR, S. K.; TIŠLJAR, J. Igneous rocks. In: ____. *Introduction to Mineralogy and Petrology*. 1. ed. Amsterdam: Elsevier, 2014. cap. 4, p. 93–120. ISBN 9780124167100.

HARRELL, F. E. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer, 2001. 571 p. (Springer Series in Statistics). ISBN 0387952322. Disponível em: <http://www.springer.com/mathematics/probability/book/978-0-387-95232-1>.

HARTEMINK, A. E. The depiction of soil profiles since the late 1700s. *Catena*, Elsevier BV, v. 79, n. 2, p. 113–127, nov 2009. Disponível em: <http://dx.doi.org/10.1016/j.catena.2009.06.002>.

HARTEMINK, A. E.; BOCKHEIM, J. G. Soil genesis and classification. *Catena*, Elsevier BV, v. 104, p. 251–256, may 2013. Disponível em: <http://dx.doi.org/10.1016/j.catena.2012.12.001>.

HARTEMINK, A. E.; MCBRATNEY, A. A soil science renaissance. *Geoderma*, v. 148, n. 2, p. 123–129, 2008.

HELDWEIN, A.; BURIOL, G.; STRECK, N. O clima de Santa Maria. *Ciência e Ambiente*, v. 38, p. 43–58, 2009.

HENGL, T. *Pedometric mapping – bridging the gaps between conventional and pedometric approaches*. 252 p. Tese (Doutorado) — Wageningen University, Wageningen, 2003. Disponível em: <http://library.wur.nl/WebQuery/edepot/121443>.

HENGL, T.; EVANS, I. S. Mathematical and digital models of the land surface. In: HENGL, T.; REUTER, H. I. (Ed.). *Geomorphometry – concepts, software, applications*. Amsterdam: Elsevier, 2009, (Developments in Soil Science, v. 33). cap. 2, p. 31–63.

HENGL, T.; HEUVELINK, G. B.; STEIN, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, v. 120, p. 75–93, 2004.

HENGL, T.; HEUVELINK, G. B. M.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; SHEPHERD, K. D.; SILA, A.; MACMILLAN, R. A.; JESUS, J. Mendes de; TAMENE, L.; AL. et. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. *PLOS ONE*, Public Library of Science (PLoS), v. 10, n. 6, p. e0125814, Jun 2015. ISSN 1932-6203. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0125814>.

HENGL, T.; HEUVELINK, G. B. M.; ROSSITER, D. G. About regression-kriging: from equations to case studies. *Computers & Geosciences*, v. 33, n. 10, p. 1301–1315, 2007. ISSN 0098-3004.

HENGL, T.; HUSNJAK, S. Evaluating adequacy and usability of soil maps in Croatia. *Soil Science Society of America Journal*, Soil Science Society of America, v. 70, n. 3, p. 920–929, 2006. ISSN 1435-0661.

HENGL, T.; JESUS, J. M. de; MACMILLAN, R. A.; BATJES, N. H.; HEUVELINK, G. B. M.; RIBEIRO, E.; SAMUEL-ROSA, A.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; AL. et. SoilGrids1km – global soil information based on automated mapping. *PLoS ONE*, Public Library of Science (PLoS), v. 9, n. 8, p. e105992, Aug 2014. ISSN 1932-6203.

HENGL, T.; NIKOLIC, M.; MACMILLAN, R. A. Mapping efficiency and information content. *International Journal of Applied Earth Observation and Geoinformation*, v. 22, p. 127–138, 2013.

HENGL, T.; ROSSITER, D. G.; STEIN, A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research*, v. 41, n. 8, p. 1403–1422, 2003.

HEUNG, B.; HO, H. C.; ZHANG, J.; KNUDBY, A.; BULMER, C. E.; SCHMIDT, M. G. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, Elsevier BV, v. 265, p. 62–77, mar 2016.

HEUVELINK, G. B. M. Identification of field attribute error under different models of spatial variation. *International journal of geographical information systems*, Informa UK Limited, v. 10, n. 8, p. 921–935, Dec 1996. ISSN 0269-3798.

HEUVELINK, G. B. M. *Error propagation in environmental modelling with GIS*. 1. ed. Boca Raton: Taylor and Francis, 1998. 127 p.

HEUVELINK, G. B. M. Propagation of error in spatial modelling with GIS. In: ____. *New developments in geographical information systems: principles, techniques, management and applications*. 2. ed. Wiley, 2005. cap. 14, p. 207–217. ISBN 978-0-471-73545-8. Disponível em: <http://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/>.

HEUVELINK, G. B. M.; BURROUGH, P. A.; STEIN, A. Propagation of errors in spatial modelling with GIS. *International journal of geographical information systems*, v. 3, n. 4, p. 303–322, 1989.

HEUVELINK, G. B. M.; BURROUGH, P. A.; STEIN, A. Propagation of errors in spatial modelling with GIS. In: ____. *Classics from IJGIS: twenty years of the International Journal of Geographical Information Science and Systems*. [S.l.]: CRC Press, 2006. v. 3, n. 4, p. 67–89.

HEUVELINK, G. B. M.; PEBESMA, E. J. Spatial aggregation and soil process modelling. *Geoderma*, v. 89, n. 1?2, p. 47–65, 1999. ISSN 0016-7061.

HEUVELINK, G. B. M.; WEBSTER, R. Modelling soil variation: past, present, and future. *Geoderma*, v. 100, n. 3-4, p. 269–301, 2001. ISSN 0016-7061.

HIRT, C.; FILMER, M.; FEATHERSTONE, W. Comparison and validation of recent freely-available ASTER-GDEM ver1, SRTM ver4.1 and GEODATA DEM-9S ver3 digital elevation models over Australia. *Australian Journal of Earth Sciences*, v. 57, n. 3, p. 337–347, 2010.

HOLTZ, M. *Do mar ao deserto: a evolução do Rio Grande do Sul no tempo geológico*. 2. ed. Porto Alegre: Editora da UFRGS, 2003. 144 p.

HUDSON, B. D. The soil survey as paradigm-based science. *Soil Science Society of America Journal*, v. 56, p. 836–841, 1992.

HULL, C. L. The goal-gradient hypothesis and maze learning. *Psychological Review*, v. 39, n. 1, p. 25–43, Jan 1932.

HUPY, C. M.; SCHAETZL, R. J.; MESSINA, J. P.; HUPY, J. P.; DELAMATER, P.; ENANDER, H.; HUGHEY, B. D.; BOEHM, R.; MITROKA, M. J.; FASHOWAY, M. T. Modeling the complexity of different, recently deglaciated soil landscapes as a function of map scale. *Geoderma*, Elsevier BV, v. 123, n. 1-2, p. 115–130, Nov 2004. ISSN 0016-7061.

HUTCHINSON, M. F. A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology*, v. 106, n. 3-4, p. 211–232, 1989. ISSN 0022-1694.

HYNDMAN, R. J.; FAN, Y. Sample quantiles in statistical packages. *The American Statistician*, Taylor & Francis, Ltd. on behalf of the American Statistical Association, v. 50, n. 4, p. 361–365, 1996. ISSN 00031305. Disponível em: <http://www.jstor.org/stable/2684934>.

IBGE. *Modelo de Ondulação Geoidal – MAPGEO2010*. 2010. Disponível em: <http://www.ibge.gov.br/home/geociencias/geodesia/modelo_geoidal.shtm>.

ISO. *ISO 7144:1986 Documentation – Presentation of theses and similar documents*. [S.l.], 1986. 10 p. Disponível em: <http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=13736>.

IUSS WORKING GROUP WRB. *World reference base for soil resources 2006 – a framework for international classification, correlation and communication, first update 2007*. Rome: Food and Agriculture Organization of the United Nations, 2007. 116 p. World Soil Resources Reports No. 103. Disponível em: <http://www.fao.org/fileadmin/templates/nr/images/resources/pdf_documents/wrb2007_red.pdf>.

JACKSON, D. A. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, v. 74, n. 8, p. 2204–2214, 1993. Disponível em: <http://www.jstor.org/stable/1939574>.

JANSSEN, P.; HEUBERGER, P. Calibration of process-oriented models. *Ecological Modelling*, Elsevier BV, v. 83, n. 1-2, p. 55–66, Dec 1995. ISSN 0304-3800.

JARVIS, A.; REUTER, H. I.; NELSON, A.; GUEVARA, E. *Hole-filled SRTM for the globe version 4*. [S.l.], 2008. Disponível em: <http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database-v4-1>.

JENNY, H. *Factors of soil formation – a system of quantitative pedology*. Toronto: Dover Publications, 1941. 281 p. ISBN 0-486-68128-9. Disponível em: <http://202.200.144.17/sykc/hjx/content/ckzl/6/2.pdf>.

JENNY, H. Derivation of state factor equations of soils and ecosystems. *Soil Science Society of America Journal*, Soil Science Society of America, v. 25, n. 5, p. 385–388, Sept 1961. ISSN 0361-5995.

JOLLIFFE, I. T. *Principal component analysis*. 2. ed. New York: Springer, 2002. 519 p.

KEMPEN, B. *Updating soil information with digital soil mapping*. 218 p. Tese (Doutorado) — Wageningen University, 2011. Disponível em: <http://edepot.wur.nl/187198>.

KEMPEN, B.; BRUS, D. J.; HEUVELINK, G. B. M.; STOORVOGEL, J. J. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. *Geoderma*, v. 151, p. 311–326, 2009.

KEMPEN, B.; BRUS, D. J.; STOORVOGEL, J. J.; HEUVELINK, G. B.; VRIES, F. de. Efficiency comparison of conventional and digital soil mapping for updating soil maps. *Soil Science Society of America Journal*, v. 76, n. 6, p. 2097–2115, 2012.

KEMPEN, B.; HEUVELINK, G. B. M.; BRUS, D. J.; STOORVOGEL, J. J. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. *European Journal of Soil Science*, Wiley Online Library, v. 61, n. 3, p. 333–347, 2010.

KER, J.; NOVAIS, R. Fundamentos para desenvolvimento da pedologia e da fertilidade do solo. In: *XXIX Congresso Brasileiro de Ciência do Solo, 2003, Ribeirão Preto, 2003*. [s.n.], 2003. p. 27. Disponível em: <http://jararaca.ufsm.br/websites/dalmolin/download/textospl/fundame.pdf>.

KER, J. C. Latossolos do Brasil: uma revisão. *Geônomos*, v. 5, p. 17–40, 1998. Disponível em: <http://goo.gl/vCMSl>.

KER, J. C. O futuro da pedologia no Brasil. *Boletim Informativo da Sociedade Brasileira de Ciência do Solo*, v. 24, p. 18–21, 1999.

KIM, J.; GRUNWALD, S.; RIVERO, R. G. Soil phosphorus and nitrogen predictions across spatial escalating scales in an aquatic ecosystem using remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, Institute of Electrical & Electronics Engineers (IEEE), v. 52, n. 10, p. 6724–6737, 2014. ISSN 1558-0644.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. *Science*, American Association for the Advancement of Science (AAAS), v. 220, n. 4598, p. 671–680, May 1983. ISSN 1095-9203.

KÖNIG, F. G.; BRUN, E. J.; SCHUMACHER, M. V.; LONGHI, S. J. Devolução de nutrientes via serapilheira em um fragmento de floresta estacional decidual no município de Santa Maria, RS. *Brasil Florestal*, v. 21, p. 45–52, 2002. Disponível em: <http://www.ibama.gov.br/ojs/index.php/braflor/article/viewArticle/110>.

KRASILNIKOV, P. V.; MARTÍ, J.-J. I.; ARNOLD, R.; SHOBA, S. *A handbook of soil terminology, correlation and classification*. 1. ed. London: Earthscan, 2009. 449 p.

KRIGE, D. G. *A statistical approach to some mine valuation and allied problems on the Witwatersrand*. 119–139 p. Dissertação (Mestrado) — University of the Witwatersrand, Johannesburg, 1951. Disponível em: <http://wiredspace.wits.ac.za/jspui/bitstream/10539/17975/1/Krige,%20D.%20G.%201951-001.pdf>.

KUENSCH, H. R.; PAPRITZ, A.; SCHWIERZ, C.; STAHEL, W. A. Robust estimation of the external drift and the variogram of spatial data. In: *Proceedings of the ISI 58th World Statistics Congress of the International Statistical Institute*. [s.n.], 2011. p. 1–8. Disponível em: <http://e-collection.library.ethz.ch/eserv/eth:7080/eth-7080-01.pdf>.

LAN, L.; LIAN, Z. Application of statistical power analysis - how to determine the right sample size in human health, comfort and productivity research. *Building and Environment*, v. 45, n. 5, p. 1202–1213, 2010.

LARK, R. M. A comparison of some robust estimators of the variogram for use in soil survey. *European Journal of Soil Science*, Wiley-Blackwell, v. 51, n. 1, p. 137–157, mar 2000.

LARK, R. M. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, v. 51, p. 717–728, 2000.

LARK, R. M. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, v. 105, p. 49–80, 2002.

LARK, R. M. Towards soil geostatistics. *Spatial Statistics*, Elsevier BV, v. 1, p. 92–99, May 2012. ISSN 2211-6753.

LARK, R. M.; BISHOP, T. F. A. Cokriging particle size fractions of the soil. *European Journal of Soil Science*, v. 58, p. 763–774, 2007.

LARK, R. M.; BISHOP, T. F. A.; WEBSTER, R. Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. *Geoderma*, v. 138, n. 1-2, p. 65–78, 2007.

LARK, R. M.; CULLIS, B. R. Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, v. 55, n. 4, p. 799–813, 2004. ISSN 1365-2389.

LARK, R. M.; CULLIS, B. R.; WELHAM, S. J. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science*, Wiley-Blackwell, v. 57, n. 6, p. 787–799, Dec 2006. ISSN 1365-2389.

LARK, R. M.; PAPRITZ, A. Fitting a linear model of coregionalization for soil properties using simulated annealing. *Geoderma*, Elsevier BV, v. 115, n. 3-4, p. 245–260, Aug 2003. ISSN 0016-7061.

LASLETT, G. M.; MCBRATNEY, A. B.; PAHL, P. J.; HUTCHINSON, M. F. Comparison of several spatial prediction methods for soil pH. *Journal of Soil Science*, Wiley-Blackwell, v. 38, n. 2, p. 325–341, Jun 1987. ISSN 0022-4588.

LEGROS, J.-P. *Mapping of the soil*. 1. ed. Enfield: Science Publishers, 2006. 411 p. Disponível em: <http://www.amazon.com/Mapping-Soil-Jean-Paul-Legros/dp/157808363X>.

LEMOINE, F. G.; KENYON, S. C.; FACTOR, J. K.; TRIMMER, R.; PAVLIS, N. K.; CHINN, D. S.; COX, C. M.; KLOSKO, S. M.; LUTHCKE, S. B.; TORRENCE, M. H.; WANG, Y. M.; WILLIAMSON, R. G.; PAVLIS, E. C.; RAPP, R. H.; OLSON, T. R. *The Development of the Joint NASA GSFC and NIMA Geopotential Model EGM96*. Greenbelt, 1998. Disponível em: <http://cddis.nasa.gov/926/egm96/egm96.html>.

LEMOS, R. C.; SANTOS, R. D. *Manual de descrição e coleta de solos no campo*. 2. ed. [S.l.]: Sociedade Brasileira de Ciência do Solo, 1982. 46 p.

LESCH, S. M.; STRAUSS, D. J.; RHOADES, J. D. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. *Water Resources Research*, Wiley-Blackwell, v. 31, n. 2, p. 387–398, Feb 1995. ISSN 0043-1397.

LIAW, A.; WIENER, M. Classification and regression by randomForest. *R News*, v. 2/3, p. 18–22, 2002. ISSN 1609-3631. Disponível em: <http://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf>.

LÓPEZ-GRANADOS, F.; JURADO-EXPÓSITO, M.; PEñA-BARRAGáN, J.; GARCÍA-TORRES, L. Using geostatistical and remote sensing approaches for mapping soil properties. *European Journal of Agronomy*, v. 23, p. 279–289, 2005.

MACARINI, J. P. A política econômica do governo Médici: 1970-1973. *Nova Economia*, FapUNIFESP (SciELO), v. 15, n. 3, p. 53–92, 2005. Disponível em: <http://dx.doi.org/10.1590/S0103-63512005000300003>.

MACHADO, J. L. F. Hidroestratigrafia química preliminar dos aquíferos na Região Central do Rio Grande do Sul. In: *X Congresso Brasileiro de Águas Subterrâneas*. São Paulo: [s.n.], 1998. Disponível em: <http://www.perfuradores.com.br>.

MACIEL FILHO, C. L. *Mapa geotécnico de Santa Maria*. Santa Maria, 1990. 21 p.

MACIEL FILHO, C. L.; GASPARETTO, N. V. L.; VEIGA, P.; SARTORI, P. L. P.; ALII et. *Mapa de formações superficiais e solos das folhas de Santa Maria e Camobi na escala* 1:50 000. Santa Maria, 1987. 1 p.

MACIEL FILHO, C. L.; GASPARETTO, N. V. L.; VEIGA, P.; SARTORI, P. L. P.; ALII et. *Mapa geológico das folhas de Santa Maria e Camobi, na escala de* 1:50 000. Santa Maria, 1987. 1 p.

MALUF, J. A new climatic classification for the state of Rio Grande do Sul, Brazil. *Revista Brasileira de Agrometeorologia*, v. 8, p. 141–150, 2000. Disponível em: <http://www.ufsm.br/rba/p14181.html>.

MARCHANT, B. P.; LARK, R. M. Adaptive sampling and reconnaissance surveys for geostatistical mapping of the soil. *European Journal of Soil Science*, v. 57, n. 6, p. 831–845, Dec 2006. ISSN 1365-2389.

MARINS, A. P. *Hidrologic simulation of the Vacacaí-Mirim reservoir, Santa Maria-RS, using the IPHS1 system*. 161 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Engenharia Civil, Universidade Federal de Santa Maria, Santa Maria, 2004.

MARLER, R. T.; ARORA, J. S. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, Springer Science + Business Media, v. 26, n. 6, p. 369–395, Apr 2004. ISSN 1615-1488.

MARLER, R. T.; ARORA, J. S. Function-transformation methods for multi-objective optimization. *Engineering Optimization*, Informa UK Limited, v. 37, n. 6, p. 551–570, Sep 2005. ISSN 1029-0273.

MARLER, R. T.; ARORA, J. S. The weighted sum method for multi-objective optimization: new insights. *Structural and Multidisciplinary Optimization*, Springer Science + Business Media, v. 41, n. 6, p. 853–862, Dec 2009. ISSN 1615-1488.

MARQUES, L. S.; ERNESTO, M. O magmatismo toleítico da bacia do Paraná. In: ____. *Geologia do continente sul-americano: evolução da obra de Fernando Flávio Marques de Almeida*. São Paulo: Beca, 2005. p. 245–263.

MARTINELLI, L. A.; NAYLOR, R.; VITOUSEK, P. M.; MOUTINHO, P. Agriculture in Brazil: impacts, costs, and opportunities for a sustainable future. *Current Opinion in Environmental Sustainability*, Elsevier BV, v. 2, n. 5-6, p. 431–438, dec 2010.

MASSY, W. F. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, American Statistical Association, v. 60, n. 309, p. 234–256, 1965. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2283149>.

MATÉRN, B. *Spatial variation: stochastic models and their application to some problems in forest surveys and other sampling investigations*. 144 p. Tese (Doutorado) — Statens Skogsforskningsinstitut, Stockholm, 1960. Disponível em: <http://pub.epsilon.slu.se/10033/1/medd_statens_skogsforskningsinst_049_05.pdf>.

MATHER, P. M. *Computer processing of remotely-sensed images - an introduction*. 3. ed. Chichester: John Wiley and Sons, Ltd, 2004. 324 p.

MATHERON, G. *Les variables régionalisées et leur estimation*. 305 p. Tese (Doutorado) — Faculte des Sciences, Universite de Paris, Paris, 1965. Disponível em: <http://cg.ensmp.fr/bibliotheque/public/MATHERON_Ouvrage_00083.pdf>.

MATHERON, G.; KLEINGELD, W. J. The evolution of geostatistics. In: *Proceedings of the Twentieth International Symposium on the Application of Computers and Mathematics in the Mineral Industries. Volume 3: Geostatistics*. Johannesburg: SAIMM, 1987. p. 9–12.

MAYNARD, J.; JOHNSON, M. Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: effects of grid resolution vs. neighborhood extent. *Geoderma*, Elsevier BV, v. 230-231, p. 29–40, Oct 2014. ISSN 0016-7061.

MAZOYER, M.; ROUDART, L. *História das agriculturas do mundo: do Neolítico à crise contemporânea*. São Paulo / Brasília: Editora UNESP / NEAD, 2008. 568 p.

MCBRATNEY, A.; MENDONÇA-SANTOS, M.; MINASNY, B. On digital soil mapping. *Geoderma*, v. 117, p. 3–52, 2003.

MCBRATNEY, A. B.; ODEH, I. O.; BISHOP, T. F.; DUNBAR, M. S.; SHATAR, T. M. An overview of pedometric techniques for use in soil survey. *Geoderma*, v. 97, p. 293–327, 2000.

MCKAY, M. D.; BECKMAN, R. J.; CONOVER, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, American Statistical Association and American Society for Quality, v. 21, n. 2, p. 239–245, 1979. ISSN 00401706.

MCKENZIE, N. J.; GALLANT, J. C. Digital soil mapping with improved environmental predictors and models of pedogenesis. In: LAGACHERIE, A. M. P.; VOLTZ, M. (Ed.). *Digital soil mapping - an introductory perspective*. Elsevier, 2006, (Developments in Soil Science, v. 31). cap. 24, p. 327 – 349. Disponível em: <http://www.sciencedirect.com/science/article/B7W58-4PT86XY-16/2/6e9c4fd649231554e34c769d9c1caa10>.

MCKENZIE, N. J.; RYAN, P. J. Spatial prediction of soil properties using environmental correlation. *Geoderma*, v. 89, p. 67–94, 1999.

MEBIUS, L. A rapid method for the determination of organic carbon in soil. *Analytica Chimica Acta*, v. 22, n. 0, p. 120 – 124, 1960. ISSN 0003-2670.

MEHL, H. U.; ELTZ, F. L. F.; REICHERT, J. M.; DIDONE, I. A. Caracterização de padrões de chuvas ocorrentes em Santa Maria (RS). *Revista Brasileira de Ciência do Solo*, v. 25, n. 2, p. 475–483, 2001. Disponível em: <http://sbcs.solos.ufv.br/solos/revistas/v25n2a23.pdf>.

MELLES, S. J.; HEUVELINK, G. B. M.; TWENHöFEL, C. J. W.; VAN DIJK, A.; HIEMSTRA, P. H.; BAUME, O.; STöHLKER, U. Optimizing the spatial pattern of networks for monitoring radioactive releases. *Computers & Geosciences*, v. 37, n. 3, p. 280–288, 2011. ISSN 0098-3004.

MENDONÇA-SANTOS, M. d. L.; SANTOS, H. G. *Mapeamento digital de classes e atributos de solos - métodos, paradigmas e novas técnicas*. Rio de Janeiro, 2003. 17 p. (Documento 55). Disponível em: <http://www.cnps.embrapa.br/publicacoes/pdfs/doc55_mapeamentodigital.pdf>.

MENEZES, F. P. *Humic substances in soils from different geomorphologic feature in the edge of Rio Grande do Sul Plateau*. 112 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2008. Disponível em: <http://w3.ufsm.br/ppgcs/>.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, AIP Publishing, v. 21, n. 6, p. 1087–1092, 1953.

MEYER, M. A.; BOOKER, J. M. *Eliciting and analyzing expert judgment: a practical guide*. London: ASA-SIAM Series on Statistics and Applied Probability, 2001. 459 p.

MIGUEL, P. *Pedological characterization, land use and modeling of the soil loss in hillslope areas the Plateau Border of RS*. 112 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2010. Disponível em: <http://w3.ufsm.br/ppgcs>.

MIGUEL, P. *Pedogeochemical and mineralogical variables in the identification of sources of sediments in a basin of hillside*. 98 p. Tese (Doutorado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2013. Disponível em: <http://w3.ufsm.br/ppgcs>.

MIGUEL, P.; DALMOLIN, R. S. D.; PEDRON, F. A.; SAMUEL-ROSA, A.; MEDEIROS, P. S. C.; MOURA-BUENO, J. M.; BALBINOT, A. Soil and land use dynamics in Plateau Border areas of Rio Grande do Sul. *Revista Brasileira de Agrociência*, v. 17, n. 4, p. 347–455, 2011. Disponível em: <http://www.ufpel.edu.br/faem/agrociencia/v17n4_arquivos/artigo05.htm>.

MIGUEL, P.; DALMOLIN, R. S. D.; PEDRON, F. d. A.; MOURA-BUENO, J. M.; TIECHER, T. Identificação de fontes de produção de sedimentos em uma bacia hidrográfica de encosta. *Revista Brasileira de Ciência do Solo*, FapUNIFESP (SciELO), v. 38, n. 2, p. 585–598, apr 2014. Disponível em: <http://dx.doi.org/10.1590/S0100-06832014000200023>.

MIKUTTA, R.; KLEBER, M.; KAISER, K.; JAHN, R. Review: organic matter removal from soils using hydrogen peroxide, sodium hypochlorite, and disodium peroxodisulfate. *Soil Science Society of America Journal*, v. 69, p. 120–135, 2005. Disponível em: <http://cat.inist.fr/?aModele=afficheN&cpsidt=16422287>.

MILANI, E. J. Comentários sobre a origem e a evolução tectônica da bacia do Paraná. In: ____. *Geologia do continente sul-americano: evolução da obra de Fernando Flávio Marques de Almeida*. São Paulo: Beca, 2005. p. 264–279.

MILLER, B.; SCHAETZL, R. The historical role of base maps in soil geography. *Geoderma*, Elsevier BV, v. 230–231, p. 329–339, May 2014. ISSN 0016-7061.

MILLER, B. A.; KOSZINSKI, S.; WEHRHAN, M.; SOMMER, M. Impact of multi-scale predictor selection for modeling soil properties. *Geoderma*, Elsevier BV, v. 239-240, p. 97–106, Feb 2015. ISSN 0016-7061.

MINASNY, B.; MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, Elsevier BV, v. 32, n. 9, p. 1378–1388, Nov 2006. ISSN 0098-3004.

MINASNY, B.; MCBRATNEY, A. B. Conditioned Latin Hypercube Sampling for calibrating soil sensor data to soil properties. In: ____. *Proximal Soil Sensing*. Amsterdam: Springer, 2010. (Progress in Soil Science), cap. 9, p. 111–119. ISBN http://id.crossref.org/isbn/978-90-481-8859-8.

MINASNY, B.; MCBRATNEY, A. B.; WALVOORT, D. J. The variance quadtree algorithm: use for spatial sampling design. *Computers & Geosciences*, v. 33, p. 383–392, 2007.

MOORE, I. D.; GESSLER, P. E.; NIELSEN, G. A.; PETERSON, G. A. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, v. 57, p. 443–452, 1993.

MOSER, J. M. Solos. In: ____. *Geografia do Brasil: Região Sul*. Rio de Janeiro: IBGE, 1990. p. 85–111.

MOURA-BUENO, J. M. *Soil erosion in the Plateau Border areas of the Rio Grande do Sul state*. Santa Maria: [s.n.], 2012. 37 p. Undergraduate Thesis. Disponível em: <http://1drv.ms/1jECraN>.

MOURA-BUENO, J. M.; SAMUEL-ROSA, A.; MIGUEL, P.; DALMOLIN, R. S. D.; FROSI, M. H.; DOTTO, A. C. Soil erosion in hillslope areas of sourthern brazil. In: *19th ISTRO Conference and IV SUCS Meeting*. Montevideu: [s.n.], 2012.

MOYEED, R. A.; PAPRITZ, A. An empirical comparison of kriging methods for nonlinear spatial point prediction. *Mathematical Geology*, Springer Science + Business Media, v. 34, n. 4, p. 365–386, 2002. ISSN 0882-8121.

MULDER, V. L.; DE BRUIN, S.; SCHAEPMAN, M. E. Representing major soil variability at regional scale by constrained Latin hypercube sampling of remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, Elsevier BV, v. 21, p. 301–310, Apr 2013. ISSN 0303-2434.

MULDER, V. L.; LACOSTE, M.; FORGES, A. C. R. de; MARTIN, M. P.; ARROUAYS, D. National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma*, v. 263, p. 16–34, 2016. ISSN 0016-7061.

MÜLLER, W. G. *Collecting spatial data - optimum design of experiments for random fields*. Berlin: Springer, 2007. 242 p. ISBN http://id.crossref.org/isbn/978-3-540-31174-4.

MÜLLER, W. G.; ZIMMERMAN, D. L. Optimal designs for variogram estimation. *Environmetrics*, v. 10, n. 1, p. 23–37, Jan 1999. ISSN 1099-095X.

MUTTIAH, R. S.; ENGEL, B. A.; JONES, D. D. Waste disposal site selection using GIS-based simulated annealing. *Computers & Geosciences*, Elsevier BV, v. 22, n. 9, p. 1013–1017, Nov 1996. ISSN 0098-3004.

NASCIMENTO, M. D.; PENNA E SOUZA, B. S. Mapeamento geomorfológico da área abrangida pela carta topográfica de Santa Maria – RS como subsídio ao planejamento ambiental. *Revista Brasileira de Geomorfologia*, v. 11, n. 2, p. 83–90, 2010. Disponível em: <http://www.lsie.unb.br/rbg/index.php?journal=rbg&page=article&op=view&path%5B%5D=155>.

NELSON, M. A.; BISHOP, T. F. A.; TRIANTAFILIS, J.; ODEH, I. O. A. An error budget for different sources of error in digital soil mapping. *European Journal of Soil Science*, Wiley-Blackwell, v. 62, n. 3, p. 417–430, Jun 2011. ISSN 1351-0754.

NUSSBAUM, M.; PAPRITZ, A.; BALTENSWEILER, A.; WALTHERT, L. Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging. *Geoscientific Model Development*, Copernicus GmbH, v. 7, n. 4, p. 1197–1210, 2014. ISSN 1991-962X.

ODEH, I. O. A.; MCBRATNEY, A. B.; CHITTLEBOROUGH, D. J. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, v. 63, n. 3?4, p. 197–214, 1994.

ODEH, I. O. A.; MCBRATNEY, A. B.; CHITTLEBOROUGH, D. J. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, v. 67, p. 215–226, 1995.

O'HAGAN, A.; BUCK, C.; DANESHKHAH, A.; EISER, J.; GARTHWAITE, P.; JENKINSON, D.; OAKLEY, J.; RAKOW, T. *Uncertain judgements: eliciting experts' probabilities*. Chichester: John Wiley and Sons, 2006. 321 p.

OLIPHANT, A. J.; SPRONKEN-SMITH, R. A.; STURMAN, A. P.; OWENS, I. F. Spatial variability of surface radiation fluxes in mountainous terrain. *Journal of Applied Meteorology*, v. 42, p. 113–128, 2003.

OLMOS, J.; CAMARGO, M. N. Concentração preliminar de Podzólico Bruno-Acinzentado tentativamente identificados no pais. In: *Conceituação sumária de algumas classes de solos recém-reconhecidas nos levantamentos e estudos de correlação do SNLCS*. Rio de Janeiro: Serviço Nacional de Levantamento e Conservação do Solo, 1982. p. 25–31. Circular técnica 1.

OMUTO, C.; NACHTERGAELE, F.; ROJAS, R. V. *State of the art report on global and regional soil information: Where are we? Where to go?* Rome: Food and Agriculture Organization of the United Nations, 2013. 69 p. ISBN 9251074496. Disponível em: <http://www.fao.org/3/a-i3161e.pdf>.

ORSI, F.; GENELETTI, D.; NEWTON, A. C. Towards a common set of criteria and indicators to identify forest restoration priorities: an expert panel-based approach. *Ecological Indicators*, v. 11, n. 2, p. 337 – 347, 2011. ISSN 1470-160X. Disponível em: <http://www.sciencedirect.com/science/article/B6W87-50G5H5F-2/2/1f174e1406d7629bbf6ceeb38c7384e6>.

PAIN, C. F.; OILIER, C. Inversion of relief - a component of landscape evolution. *Geomorphology*, v. 12, n. 2, p. 151 – 165, 1995. ISSN 0169-555X. Disponível em: <http://www.sciencedirect.com/science/article/pii/0169555X94000845>.

PAISANI, J.; GEREMIA, F. Evolution of hillslopes in the Basaltic Plateau based on the analysis of colluvium deposits - Middle Valley of Marrecas River - SW Paraná. *Geociências*, v. 29, n. 3, p. 321–334, 2010.

PAIVA, E. M. C. D.; PAIVA, J. B. D.; MOREIRA, A. P.; MAFFINI, G. F.; MELLER, A.; DILL, P. R. J. Evolução de processo erosivo acelerado em trecho do arroio Vacacaí-Mirim. *Revista Brasileira de Recursos Hídricos*, v. 6, p. 129–135, 2001. Disponível em: <http://www.abrh.org.br/>.

PCI GEOMATICS. *Geomatica® OrthoEngine® 10.1 user guide*. Richmond Hill, 2007. 174 p. (Version 10.1). Disponível em: <http://www.gis.unbc.ca/help/software/pci/orthoeng.pdf>.

PEBESMA, E. *gstat user's manual*. Utrecht, 2014. 108 p. Disponível em: <http://gstat.org/gstat.pdf>.

PEBESMA, E. J. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, v. 30, n. 7, p. 683–691, 2004. ISSN 0098-3004.

PEDRON, F. A. *Classification of land use potencial in the urban perimeter of Santa Maria - RS*. 65s p. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2005. Disponível em: <http://w3.ufsm.br/ppgcs/>.

PEDRON, F. A. *Mineralogia, morfologia e classificação de saprolitos e Neossolos derivados de rochas vulcânicas no Rio Grande do Sul*. 160 p. Tese (Doutorado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2007. Disponível em: <http://w3.ufsm.br/ppgcs/>.

PEDRON, F. A.; DALMOLIN, R. S. D.; AZEVEDO, A. C.; BOTELHO, M. R.; SAMUEL-ROSA, A. Spatial dynamic analysis of the land occupation and their conflicts of use in the urban perimiter of Santa Maria - RS (1975 – 2002). *Ciência Rural*, FapUNIFESP (SciELO), v. 36, n. 6, p. 1756–1764, Dec 2006. ISSN 0103-8478.

PEDRON, F. A.; SAMUEL-ROSA, A.; DALMOLIN, R. S. D. Variation in pedological characteristics and the taxonomic classification of Argissolos (Ultisols and Alfisols) derived from sedimentary rocks. *Revista Brasileira de Ciência do Solo*, v. 36, p. 1–9, 2012.

PERES-NETO, P. R.; JACKSON, D. A.; SOMERS, K. M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, v. 49, n. 4, p. 974–997, 2005.

PIERINI, C.; MIZUSAKI, A. M. P.; SCHERER, C. M.; ALVES, D. B. Integrated stratigraphic and geochemical study of the Santa Maria and Caturrita formations (Triassic of the Paraná Basin), Southern Brazil. *Journal of South American Earth Sciences*, v. 15, p. 669–681, 2002. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0895981102001141>.

PINHEIRO, R. J.; SOARES, J. M. Condicionantes geológicos-geotécnicos de movimentos de massa na encosta da Serra Geral - RS. *Teoria e Prática na Engenharia Civil*, v. 4, n. 4, p. 59–68, 2004. Disponível em: <http://www.editoradunas.com.br/revistatpec/Sumario_Numero4.htm>.

PINTO, J. S. *Estudo da condutividade hidráulica de solos para disposição de resíduos sólidos na região de Santa Maria*. 154 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Engenharia Civil, Universidade Federal de Santa Maria, Santa Maria, 2005.

POELKING, E. L. *Land suitability, evolution and land use conflicts in Itaara County, RS*. 67 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2007. Disponível em: <http://w3.ufsm.br/ppgcs/>.

POGGIO, L.; GIMONA, A. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation — an example from Scotland. *Geoderma*, Elsevier BV, v. 232–234, p. 284–299, Nov 2014. ISSN 0016-7061.

POGGIO, L.; GIMONA, A.; BREWER, M. J. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. *Geoderma*, Elsevier BV, v. 209-210, p. 1–14, nov 2013. Disponível em: <http://dx.doi.org/10.1016/j.geoderma.2013.05.029>.

PRUSCHA, H. *Statistical analysis of climate series*. Dordrecht: Springer Science + Business Media, 2013. 175 p. Disponível em: <http://dx.doi.org/10.1007/978-3-642-32084-2>.

QUANTUM GIS DEVELOPMENT TEAM. *Quantum GIS Geographic Information System*. [S.l.], 2013. Version 2.0.1-Dufour. Disponível em: <http://qgis.osgeo.org>.

RAMOS, D. P. Desafios da pedologia brasileira frente ao novo milênio. In: *Palestra proferida no XXIX Congresso Brasileiro de Ciência do Solo. Ribeirão Preto, SP, Julho 2003*. [s.n.], 2003. Disponível em: <http://jararaca.ufsm.br/websites/dalmolin/download/textospl/desafio.pdf>.

RAPIDEYE. *Satellite imagery product specifications*. 5. ed. Brandenburg an der Havel, 2013. 46 p. Disponível em: <http://www.rapideye.com/upload/RE_Product_Specifications_ENG.pdf>.

RATNER, B. Variable selection methods in regression: ignorable problem, outing notable solution. *Journal of Targeting, Measurement and Analysis for Marketing*, Nature Publishing Group, v. 18, n. 1, p. 65–75, mar 2010.

REFSGAARD, J. C.; SLUIJS, J. P. van der; BROWN, J.; KEUR, P. van der. A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources*, v. 29, n. 11, p. 1586–1597, 2006. ISSN 0309-1708. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0309170805002903>.

REUTER, H. I.; NELSON, A.; JARVIS, A. An evaluation of void-filling interpolation methods for SRTM data. *International Journal of Geographical Information Science*, v. 21, n. 9, p. 983–1008, 2007.

RIBEIRO, E.; BATJES, N. H.; LEENAARS, A. v. J. G. B.; JESUS, J. M. *Towards the standardization and harmonization of world soil data*. Wageningen, 2015. ISRIC Report 2015/03. Disponível em: <http://www.isric.org/sites/default/files/isric_report_2015_03.pdf>.

RIBEIRO JR, P. J.; DIGGLE, P. J. geoR: a package for geostatistical analysis. *R-NEWS*, v. 1, n. 2, p. 15–18, June 2001. Disponível em: <http://geodacenter.asu.edu/system/files/rnews1.2.15-18_0.pdf>.

RIPLEY, B. D.; RASSON, J. P. Finding the edge of a Poisson forest. *Journal of Applied Probability*, Cambridge University Press (CUP), v. 14, n. 3, p. 483, sep 1977.

RODRÍGUEZ, E.; MORRIS, C. S.; BELZ, J. E. A global assessment of the SRTM performance. *Photogrammetric Engineering and Remote Sensing*, v. 72, p. 249–260, 2006.

ROSSEL, R. A. V.; WEBSTER, R.; KIDD, D. Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. *Earth Surface Processes and Landforms*, Wiley-Blackwell, v. 39, n. 6, p. 735–748, Oct 2013. ISSN 0197-9337.

ROSSITER, D. G. *Methodology for soil resource inventories*. 2. ed. Enschede: Faculty of Geo-Information Science and Earth Observation, University of Twente, 2000. 132 p. Lecture Notes. Disponível em: <http://www.itc.nl/~rossiter/teach/ssm/SSM_LectureNotes2.pdf>.

ROUDIER, P.; BEAUDETTE, D. E.; HEWITT, A. E. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. In: MINASNY, B.; MALONE, B. P.; MCBRATNEY, A. B. (Ed.). *Digital soil assessments and beyond: proceedings of the 5th global workshop on digital soil mapping*. Sydney: CRC Press, 2012. p. 227–231.

ROWLINGSON, B.; DIGGLE, P. Splancs: Spatial point pattern analysis code in S-plus. *Computers & Geosciences*, Elsevier BV, v. 19, n. 5, p. 627–655, may 1993. Disponível em: <http://dx.doi.org/10.1016/0098-3004(93)90099-Q>.

ROYLE, J. A.; NYCHKA, D. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences*, Elsevier BV, v. 24, n. 5, p. 479–488, Jun 1998. ISSN 0098-3004.

RUSSO, D. Design of an optimal sampling network for estimating the variogram. *Soil Science Society of America Journal*, Soil Science Society of America, v. 48, n. 4, p. 708–716, 1984. ISSN 0361-5995.

SAMUEL-ROSA, A. *Uso da terra no Rebordo do Planalto do Rio Grande do Sul*. Santa Maria: [s.n.], 2009. 23 p.

SAMUEL-ROSA, A. *Spatial prediction functions of soil properties*. 201 p. Dissertação (Mestrado) — Post-Graduate Course in Soil Science, Universidade Federal de Santa Maria, Santa Maria, 2012. Disponível em: <http://w3.ufsm.br/ppgcs/>.

SAMUEL-ROSA, A.; ANJOS, L. H. C.; VASQUES, G. M. An approach to help formalizing the purposive sampling strategy of classical soil surveys. In: *Proceedings of the 20th World Congress of Soil Science*. Jeju: [s.n.], 2014. Disponível em: <https://www.researchgate.net/publication/264080772_An_approach_to_help_formalizing_the_purposive_sampling_strategy_of_classical_soil_surveys>.

SAMUEL-ROSA, A.; ANJOS, L. H. C.; VASQUES, G. M.; ANTUNES, M. A. H.; DALMOLIN, R. S. D. Identifying and correcting oblique striping in the Topodata digital elevation model. In: EPAGRI. *XXXIV Brazilian Congress of Soil Science*. Florianópolis: EPAGRI, 2013. Disponível em: <http://goo.gl/3zQmfq>.

SAMUEL-ROSA, A.; ANJOS, L. H. C.; VASQUES, G. M.; HEUVELINK, G. B. M. Evaluation of freely available ancillary data used for detailed soil mapping in Brazil. In: *Geophysical Research Abstracts – EGU General Assembly 2014*. Copernicus, 2014. v. 16, p. EGU2014–769–1. Disponível em: <http://meetingorganizer.copernicus.org/EGU2014/EGU2014-769-1.pdf>.

SAMUEL-ROSA, A.; ANJOS, L. H. C.; VASQUES, G. M.; HEUVELINK, G. B. M. *pedometrics - pedometric tools and techniques*. [S.l.], 2014. R package version 0.1-7. Disponível em: <https://r-forge.r-project.org/R/?group_id=1887>.

SAMUEL-ROSA, A.; BRUS, D. J.; VASQUES, G. M.; ANJOS, L. H. C. Optimization of sample configurations for spatial trend estimation. In: *Pedometrics 2015*. Córdoba: Universidad de Córdoba, 2015.

SAMUEL-ROSA, A.; BRUS, D. J.; VASQUES, G. M.; ANJOS, L. H. C. Optimization of sample configurations for spatial trend estimation for soil mapping. In preparation. 2016.

SAMUEL-ROSA, A.; DALMOLIN, R. S. D.; MIGUEL, P. Building predictive models of soil particle-size distribution. *Revista Brasileira de Ciência do Solo*, v. 37, n. 2, p. 422–430, 2013. ISSN 0100-0683.

SAMUEL-ROSA, A.; HEUVELINK, G.; VASQUES, G.; ANJOS, L. spsann – optimization of sample patterns using spatial simulated annealing. In: *Geophysical Research Abstracts – EGU General Assembly 2015*. Copernicus, 2015. v. 17, p. EGU2015–7780. Disponível em: <http://meetingorganizer.copernicus.org/EGU2015/EGU2015-7780.pdf>.

SAMUEL-ROSA, A.; HEUVELINK, G. B. M.; VASQUES, G. M.; ANJOS, L. H. C. Spatial point pattern analysis of soil survey sampling locations. In: *Proceedings of the 10th European Conference on Geostatistics for Environmental Applications*. Paris: [s.n.], 2014. Disponível em: <http://goo.gl/o9Hmky>.

SAMUEL-ROSA, A.; HEUVELINK, G. B. M.; VASQUES, G. M.; ANJOS, L. H. C. Do more detailed environmental covariates deliver more accurate soil maps? *Geoderma*, v. 243–244, p. 214–227, April 2015.

SAMUEL-ROSA, A.; HEUVELINK, G. B. M.; VASQUES, G. M.; ANJOS, L. H. C. Optimization of sample configurations for variogram estimation. In: *Pedometrics 2015*. Córdoba: Universidad de Córdoba, 2015.

SAMUEL-ROSA, A.; MIGUEL, P.; DALMOLIN, R. S. D.; PEDRON, F. A. Land use in the Plateau Border of the state of Rio Grande do Sul. *Ciência & Natura*, v. 33, p. 161–173, 2011. Disponível em: <http://cascavel.ufsm.br/revista_ccne/ojs/index.php/cienciaenatura/article/viewFile/542/404>.

SANCHEZ, P. A.; AHAMED, S.; CARRÉ, F.; HARTEMINK, A. E.; HEMPEL, J.; HUISING, J.; LAGACHERIE, P.; MCBRATNEY, A. B.; MCKENZIE, N. J.; MENDONÇA-SANTOS, M. L.; MINASNY, B.; MONTANARELLA, L.; OKOTH, P.; PALM, C. A.; SACHS, J. D.; SHEPHERD, K. D.; VAGEN, T.-G.; VANLAUWE, B.; WALSH, M. G.; WINOWIECKI, L. A.; ZHANG, G. lin. Digital soil map of the world. *Science*, v. 325, p. 680–681, 2009.

SANTOS, H. G.; JACOMINE, P. K. T.; ANJOS, L. H. C.; OLIVEIRA, V. A.; OLIVEIRA, J. B.; COELHO, M. R.; LUMBRERAS, J. F.; CUNHA, T. J. F. *Sistema Brasileiro de Classificação de Solos*. 2. ed. Rio de Janeiro: Embrapa Solos, 2006. 306 p. Disponível em: <http://200.20.158.8/blogs/sibcs/>.

SANTOS, H. G.; JACOMINE, P. K. T.; ANJOS, L. H. C.; OLIVEIRA, V. A.; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A.; CUNHA, T. J. F.; OLIVEIRA, J. B. *Sistema Brasileiro de Classificação de Solos*. 3. ed. Brasília: Embrapa, 2013. 353 p. Disponível em: <http://200.20.158.8/blogs/sibcs/>.

SANTOS, R. D.; LEMOS, R. C.; SANTOS, H. G.; KER, J. C.; ANJOS, L. H. C. *Manual of soil description and sampling in the field*. 5. ed. Viçosa: Sociedade Brasileira de Ciência do Solo, 2005. 92 p.

SANTOS, R. D.; LEMOS, R. C.; SANTOS, H. G.; KER, J. C.; ANJOS, L. H. C.; SHIMIZU, S. H. *Manual of soil description and sampling in the field*. 6. ed. Viçosa: Sociedade Brasileira de Ciência do Solo, 2013. 100 p.

SARTORI, P. L. P. Geology and geomorfology of Santa Maria. *Ciência e Ambiente*, v. 38, p. 19–42, 2009. Disponível em: <http://w3.ufsm.br/cienciaeambiente/resenha.php?IDResenha=397>.

SCHELLING, J. Soil genesis, soil classification and soil survey. *Geoderma*, Elsevier BV, v. 4, n. 3, p. 165–193, sep 1970. Disponível em: <http://dx.doi.org/10.1016/0016-7061(70)90002-9>.

SCHENEIDER, P. R.; GALVÃO, F.; LONGHI, S. J. Influência do pisoteio de bovinos em áreas florestais. *Revista Floresta*, v. 9, p. 19–23, 1978.

SCHNEIDER, M.; PERES, C. A. Environmental costs of government-sponsored agrarian settlements in Brazilian Amazônia. *PLOS ONE*, Public Library of Science (PLoS), v. 10, n. 8, p. e0134016, aug 2015. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0134016>.

SCHOWENGERDT, R. A. *Remote sensing: models and methods for image processing*. 3. ed. San Diego: Academic Press, 2007. 515 p.

SCHRAMA, M.; VEEN, G. F. C.; BAKKER, E. S. L.; RUIFROK, J. L.; BAKKER, J. P.; OLFF, H. An integrated perspective to explain nitrogen mineralization in grazed ecosystems. *Perspectives in Plant Ecology, Evolution and Systematics*, Elsevier BV, v. 15, n. 1, p. 32–44, Feb 2013. ISSN 1433-8319.

SCULL, P.; FRANKLIN, J.; CHADWICK, O.; MCARTHUR, D. Predictive soil mapping: a review. *Progress in Physical Geography*, v. 27, n. 2, p. 171–197, 2003.

SECOM. *Comissão debate compra de terra por estrangeiro*. 2015. 7 p. Eletronic. Disponível em: <http://www.camara.leg.br/internet/Jornal/JC20151215.pdf>.

SEMA/UFSM. *Relatório final do inventário florestal contínuo do Rio Grande do Sul*. Porto Alegre, 2001. 706 p. Disponível em: <http://w3.ufsm.br/ifcrs/frame.htm>.

SHI, X.; GIROD, L.; LONG, R.; DEKETT, R.; PHILIPPE, J.; BURKE, T. A comparison of LiDAR-based DEMs and USGS-sourced DEMs in terrain analysis for knowledge-based digital soil mapping. *Geoderma*, v. 170, n. 0, p. 217–226, 2012. ISSN 0016-7061.

SILVA, L. K. R. A migração dos trabalhadores gaúchos para a Amazônia Legal (1970-1985). II - A política de ocupação das fronteiras amazônicas. *Klepsidra – Revista Virtual de História*, v. 24, 2005. ISSN 1677-8944. Disponível em: <http://www.klepsidra.net/klepsidra24/agro-rs2.htm>.

SIMBAHAN, G. C.; DOBERMANN, A. Sampling optimization based on secondary information and its utilization in soil carbon mapping. *Geoderma*, v. 133, p. 345–362, 2006.

SMITH, G. D. *The Guy Smith interviews: rationale concepts in Soil Taxonomy*. 1. ed. New York: Soil Management Support Services. Soil Conservation Service. US Department of Agriculture, 1986. 260 p. SMSS technical monograph no. 11. ISBN 0-932865-05-4.

SOIL CONSERVATION SERVICE. *Soil survey investigations report*. Washington: United States Soil Conservation Service, 1972. 63 p. Disponível em: <http://catalog.hathitrust.org/Record/001720356>.

STAMPS, A. E. I.; KRISHNAN, V. Perceived enclosure of space, angle above observer, and distance to boundary. *Perceptual and Motor Skills*, v. 99, p. 1187–1192, 2004.

STEIN, A.; HOOGERWERF, M.; BOUMA, J. Use of soil-map delineations to improve (co-)kriging of point data on moisture deficits. *Geoderma*, Elsevier BV, v. 43, n. 2-3, p. 163–177, dec 1988. Disponível em: <http://dx.doi.org/10.1016/0016-7061(88)90041-9>.

STEIN, M. L. *Interpolation of spatial data: some theory for kriging*. New York: Springer, 1999. 247 p. ISBN 978-1-4612-1494-6. Disponível em: <http://www.springer.com/mathematics/probability/book/978-0-387-98629-6>.

STRECK, E. V.; KÄMPF, N.; DALMOLIN, R. S.; KLAMT, E.; NASCIMENTO, P. C.; SCHNEIDER P. GIASSON, E.; PINTO, L. F. *Solos do Rio Grande do Sul*. 2. ed. Porto Alegre: EMATER/RS, 2008. 222 p.

STÜRMER, S. L. K. *Water infiltration in Neossolos Regolíticos (Regossols) in the Plateau Edge of Rio Grande do Sul State*. 104 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2008. Disponível em: <http://w3.ufsm.br/ppgcs/>.

SUMFLETH, K.; DUTTMANN, R. Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. *Ecological Indicators*, v. 8, p. 485–501, 2008.

SUN, W.; MINASNY, B.; MCBRATNEY, A. Analysis and prediction of soil properties using local regression-kriging. *Geoderma*, v. 171-172, n. 0, p. 16–23, February 2012. Entering the Digital Era: Special Issue of Pedometrics 2009, Beijing.

SUTILI, F. J.; DURLO, M. A.; BRESSAN, D. A. Hidrografia de Santa Maria. *Ciência e Ambiente*, v. 38, p. 79–92, 2009.

SUZUKI, L. E. A. S.; REINERT, D. J.; KAISER, D. R.; KUNZ, M.; PELLEGRINI, A.; REICHERT, J. M.; ALBUQUERQUE, J. A. Areia total de solos sob diferentes tempos de agitação horizontal, tempo de contato do dispersante químico e dispersão mecânica. In: *Reunião Brasileira de Manejo e Conservação do Solo e da Água*. Santa Maria: Sociedade Brasileira de Ciência do Solo, 2004. p. 4. Disponível em: <http://www.fisicadosolo.ccr.ufsm.quoos.com.br/index.php?option=com_content&view=article&id=60&Itemid=89>.

SUZUKI, L. E. A. S.; REINERT, D. J.; KAISER, D. R.; KUNZ, M.; PELLEGRINI, A.; REICHERT, J. M.; ALBUQUERQUE, J. A. Teor de argila de solos sob diferentes tempos de agitação horizontal, tempo de contato do dispersante químico e dispersão mecânica. In: *Reunião Brasileira de Manejo e Conservação do Solo e da Água*. Santa Maria: Sociedade Brasileira de Ciência do Solo, 2004. p. 4. Disponível em: <http://www.fisicadosolo.ccr.ufsm.quoos.com.br/index.php?option=com_content&view=article&id=60&Itemid=89>.

TAYLOR, J. R. *An introduction to error analysis*. 2. ed. Sausalito: University Science Books, 1997. 327 p.

TEDESCO, M. J.; GIANELLO, C.; BISSANI, C. A.; BOHNEN, H.; VOLKWEISS, S. J. *Analysis of soil, plants and other materials*. 2. ed. [S.l.], 1995. 147 p.

TEN CATEN, A.; DALMOLIN, R. S. D.; PEDRON, F. A.; MENDONÇA-SANTOS, M. L. Principal components as predictor variables in digital mapping of soil classes. *Ciência Rural*, v. 41, p. 1170–1176, 2011.

TEN CATEN, A.; DALMOLIN, R. S. D.; PEDRON, F. A.; MENDONÇA-SANTOS, M. L. Spatial resolution of a digital elevation model defined by the wavelet function. *Pesquisa Agropecuária Brasileira*, v. 47, n. 3, p. 449–457, 2012.

TEN CATEN, A.; MINELLA, J. P. G.; MADRUGA, P. R. A. Disintensification of land use and its relation with soil erosion. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 16, n. 9, p. 1006–1014, 2012.

THOMPSON, J. A.; BELL, J. C.; BUTLER, C. A. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, v. 100, n. 1-2, p. 67–89, 2001.

TOURÉ-TILLERY, M.; FISHBACH, A. The course of motivation. *Journal of Consumer Psychology*, v. 21, n. 4, p. 414–423, 2011. ISSN 1057-7408. Special Issue on the Application of Behavioral Decision Theory. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1057740811000507>.

TOURÉ-TILLERY, M.; FISHBACH, A. The end justifies the means, but only in the middle. *Journal of Experimental Psychology: General*, v. 141, n. 3, p. 570–583, 2011.

TOUTIN, T. Evaluation of radargrammetric DEM from RADARSAT images in high relief areas. *IEEE Transactions on Geoscience and Remote Sensing*, v. 38, n. 2, p. 782–789, 2000.

TOUTIN, T. Geometric processing of remote sensing images: models, algorithms and methods. *International Journal of Remote Sensing*, v. 25, n. 10, p. 1893–1924, 2004. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/0143116031000101611>.

TRUONG, P. N.; HEUVELINK, G. B. M.; GOSLING, J. P. Web-based tool for expert elicitation of the variogram. *Computers & Geosciences*, v. 51, p. 390–399, Feb 2013. ISSN 0098-3004.

UFRRJ. *Manual de instruções para organização e apresentação de dissertações e teses na UFRRJ*. 3. ed. Seropédica, 2006. 25 p. Disponível em: <http://www.ufrrj.br/portal/modulo/dppg/Formularios_normas/manual_teses.pdf>.

VALERIANO, M. M.; ROSSETTI, D. F. Topodata: Brazilian full coverage refinement of srtm data. *Applied Geography*, v. 32, n. 2, p. 300–309, 2012.

VAN GROENIGEN, J.; SIDERIUS, W.; STEIN, A. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, v. 87, p. 239–259, 1999.

VAN GROENIGEN, J.; STEIN, A.; ZUURBIER, R. Optimization of environmental sampling using interactive gis. *Soil Technology*, v. 10, p. 83–97, 1997.

VAN GROENIGEN, J.-W. *Constrained optimisation of spatial sampling: a geostatistical approach*. 148 p. Tese (Doutorado) — Wageningen University, Wageningen, 1999. Disponível em: <http://edepot.wur.nl/192440>.

VAN GROENIGEN, J. W.; STEIN, A. Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, v. 27, n. 5, p. 1078–1086, 1998.

VENABLES, W. N.; RIPLEY, B. D. *Modern applied statistics with S*. 4. ed. New York: Springer, 2002. 504 p. ISBN 0-387-95457-0. Disponível em: <http://www.stats.ox.ac.uk/pub/MASS4>.

VERMOTE, E.; TANRE, D.; DEUZE, J. L.; HERMAN, M.; MORCETTE, J. J. Second Simulation of the Satellite Signal in the Solar Spectrum, 6S: an overview. *IEEE Transactions on Geoscience and Remote Sensing*, v. 35, n. 3, p. 675–686, 1997.

VOLTZ, M.; WEBSTER, R. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of Soil Science*, Wiley-Blackwell, v. 41, n. 3, p. 473–490, Sep 1990. ISSN 0022-4588.

WALVOORT, D. J. J.; BRUS, D. J.; DE GRUIJTER, J. J. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences*, v. 36, n. 10, p. 1261–1267, 2010. ISSN 0098-3004.

WARRICK, A. W.; MYERS, D. E. Optimization of sampling locations for variogram calculations. *Water Resources Research*, v. 23, n. 3, p. 496–500, Mar 1987. ISSN 0043-1397.

WEBSTER, R. Is soil variation random? *Geoderma*, Elsevier BV, v. 97, n. 3-4, p. 149–163, Sep 2000. ISSN 0016-7061.

WEBSTER, R. Let's re-write the scientific paper. *European Journal of Soil Science*, v. 54, p. 215–218, 2003.

WEBSTER, R.; LARK, R. M. *Field sampling for environmental science and management*. London: Routledge, 2013. 200 p.

WEBSTER, R.; OLIVER, M. A. *Statistical methods in soil and land resource survey*. Oxford: Oxford University Press, 1990. 316 p.

WEBSTER, R.; OLIVER, M. A. Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, Blackwell Publishing Ltd, v. 43, n. 1, p. 177–192, 1992. ISSN 1365-2389.

WEBSTER, R.; OLIVER, M. A. *Geostatistics for environmental scientists*. 2. ed. Chichester: John Wiley & Sons, 2007. 315 p.

WECHSLER, S. P. Perceptions of digital elevation model uncertainty by DEM users. *URISA Journal*, v. 15, n. 2, p. 57–64, 2003. Disponível em: <http://urisa.org/Journal/protect/Vol15No2/Wechsler.pdf>.

WEICHELT, H.; ROSSO, P.; MARX, A.; REIGBER, S.; DOUGLASS, K.; HEYNEN, M. *The RapidEye Red Edge Band*. [S.l.], 2013. 8 p. Disponível em: <http://blackbridge.com/rapideye/upload/Red_Edge_White_Paper.pdf>.

WERLANG, M. K.; GROSS, J. A.; PORTO, P.; RODRIGUES, P. G. Trabalho de campo em geomorfologia: visualização de formas de relevo, solos e dinâmica erosiva na topossequência desde a Depressão Periférica sul-rio-grandense até o Rebordo do Planalto (planaltos e chapadas da Bacia Sedimentar do Paraná) em Santa Maria-RS/Silveira Martins-RS. *Geografia Ensino e Pesquisa*, v. 14, n. 3, p. 18–26, 2010. Disponível em: <http://cascavel.ufsm.br/revistageografia/index.php/revistageografia/issue/view/63>.

WOOD, J. *The geomorphological characterisation of digital elevation models*. 185 p. Tese (Doutorado) — University of Leicester, Leicester, 1996. Disponível em: <http://www.soi.city.ac.uk/~jwo/phd/>.

YEOMANS, J. C.; BREMNER, J. M. A rapid and precise method for routine determination of organic carbon in soil. *Communications in Soil Science and Plant Analysis*, v. 19, n. 13, p. 1467–1476, 1988.

YFANTIS, E. A.; FLATMAN, G. T.; BEHAR, J. V. Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology*, Springer Science + Business Media, v. 19, n. 3, p. 183–205, apr 1987. Disponível em: <http://dx.doi.org/10.1007/BF00897746>.

ZALAMENA, J. *Impacto do uso da terra nos atributos químicos e físicos de solos do Rebordo do Planalto - RS*. 78 p. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência do Solo, Universidade Federal de Santa Maria, Santa Maria, 2008. Disponível em: <http://w3.ufsm.br/ppgcs/>.

ZEILEIS, A.; GROTHENDIECK, G. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, v. 14, n. 6, p. 1–27, 2005. Disponível em: <http://www.jstatsoft.org/v14/i06/>.

ZHU, A. X.; BURT, J. E.; SMITH, M.; WANG, R.; GAO, J. The impact of neighbourhood size on terrain derivatives and digital soil mapping. In: ZHOU, Q.; LEES, B.; TANG, G. (Ed.). *Advances in digital terrain analysis*. Berlin: Springer, 2008, (Lecture notes in geoinformation and cartography). p. 333–348. ISBN 978-3-540-77799-1.

ZHU, Z.; STEIN, M. L. Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, Springer Science + Business Media, v. 11, n. 1, p. 24–44, Mar 2006. ISSN 1537-2693.

ZIMMERMAN, D. L. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, Wiley-Blackwell, v. 17, n. 6, p. 635–652, 2006. ISSN 1099-095X.

12 APÊNDICE

APÊNDICE A – R-PACKAGE SPSANN: OPTIMIZATION OF SAMPLE CONFIGURATIONS USING SPATIAL SIMULATED ANNEALING

Package ‘spsann’

March 15, 2016

Type Package

Title Optimization of Sample Configurations using Spatial Simulated Annealing

Version 2.0-0

Date 2016-03-14

Description
Methods to optimize spatial sample configurations using spatial simulated annealing. Multiple objective functions are implemented for various purposes, such as variogram estimation, spatial trend
estimation, and spatial interpolation. A general purpose spatial simulated annealing function enables the
user to define his/her own objective function.

License GPL (>= 2)

Imports methods, pedometrics, Rcpp (>= 0.11.3), sp, SpatialTools

Suggests gstat, tcltk, knitr

LinkingTo Rcpp

Encoding UTF-8

RoxygenNote 5.0.1

VignetteBuilder knitr

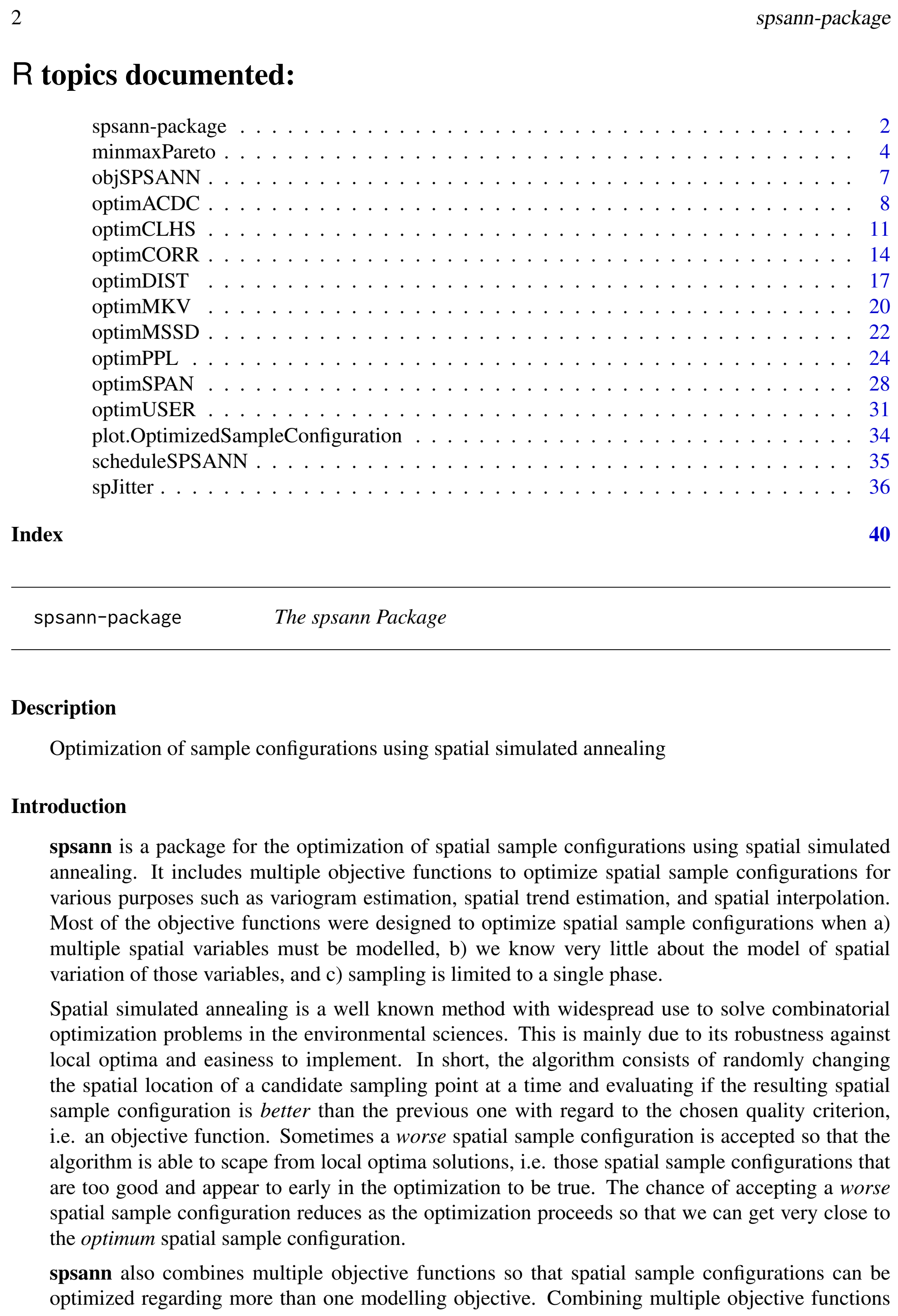
NeedsCompilation yes

Author Alessandro Samuel-Rosa [aut, cre],
Lucia Helena Cunha dos Anjos [ths],
Gustavo de Mattos Vasques [ths],
Gerard B M Heuvelink [ths],
Edzer Pebesma [ctb],
Jon Skoien [ctb],
Joshua French [ctb],
Pierre Roudier [ctb],
Dick Brus [ctb],
Murray Lark [ctb]

Maintainer Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

Repository CRAN

Date/Publication 2016-03-15 06:33:00



R topics documented:

spsann-package *The spsann Package*

Description

Optimization of sample configurations using spatial simulated annealing

Introduction

spsann is a package for the optimization of spatial sample configurations using spatial simulated annealing. It includes multiple objective functions to optimize spatial sample configurations for various purposes such as variogram estimation, spatial trend estimation, and spatial interpolation. Most of the objective functions were designed to optimize spatial sample configurations when a) multiple spatial variables must be modelled, b) we know very little about the model of spatial variation of those variables, and c) sampling is limited to a single phase.

Spatial simulated annealing is a well known method with widespread use to solve combinatorial optimization problems in the environmental sciences. This is mainly due to its robustness against local optima and easiness to implement. In short, the algorithm consists of randomly changing the spatial location of a candidate sampling point at a time and evaluating if the resulting spatial sample configuration is *better* than the previous one with regard to the chosen quality criterion, i.e. an objective function. Sometimes a *worse* spatial sample configuration is accepted so that the algorithm is able to scape from local optima solutions, i.e. those spatial sample configurations that are too good and appear to early in the optimization to be true. The chance of accepting a *worse* spatial sample configuration reduces as the optimization proceeds so that we can get very close to the *optimum* spatial sample configuration.

spsann also combines multiple objective functions so that spatial sample configurations can be optimized regarding more than one modelling objective. Combining multiple objective functions gives rise to a multi-objective combinatorial optimization problem (MOCOP). A MOCOP usually has multiple possible solutions. **spsann** finds a single solution by aggregating the objective functions using the weighted-sum method. With this method the relative importance of every objective function can be specified at the beginning of the optimization so that their relative influence on the resulting optimized spatial sample configuration can be different. But this requires the objective functions first to be scaled to the same approximate range of values. The upper-lower bound approach is used for that end. In this approach, every objective function is scaled using as reference the respective minimum and maximum attainable objective function values, also known as the Pareto minimum and maximum.

Package Structure

spsann has a very simple structure composed of three families of functions. The first is the family of `optim` functions. These are the functions that include the spatial simulated annealing algorithm, that is, the functions that perform the optimization regarding the chosen quality criterion (objective function). Every `optim` function is named after the objective function used as quality criterion. For example, the quality criterion used by `optimMSSD` is the *mean squared shortest distance* (MSSD) between sample and prediction points. As the example shows, the name of the `optim` functions is composed of the string `'optim'` followed by a suffix that indicates the respective objective function. In the example this is `'MSSD'`.

There currently are nine function in the `optim` family: `optimACDC`, `optimCLHS`, `optimCORR`, `optimDIST`, `optimMSSD`, `optimMKV`, `optimPPL`, `optimSPAN`, and `optimUSER`. The latter is a general purpose function that enables to user to define his/her own objective function and plug it in the spatial simulated annealing algorithm.

The second family of functions is the `obj` family. This family of functions is used to return the current objective function value of a spatial sample configuration. Like the family of `optim` functions, the name of the `obj` functions is composed of the string `'obj'` plus a suffix that indicates the objective function being used. For example, `objMSSD` computes the value of the mean squared shortest distance between sample and prediction points of any spatial sample configuration. Accordingly, there is a `obj` function for every `optim` function, except for `optimUSER`. A ninth `obj` function, `objSPSANN`, returns the objective function value at any point of the optimization, irrespective of the objective function used.

The third family of functions implemented in **spsann** corresponds to a set of auxiliary functions. These auxiliary functions can be used for several purposes, such as organizing the information needed to feed an `optim` function, retrieving information from an object of class `OptimizedSampleConfiguration`, i.e. an object containing an optimized sample configuration, generating plots of the spatial distribution an optimized sample configuration, and so on. These functions are named after the purpose for which they have been designed. For example: `countPPL`, `minmaxPareto`, `scheduleSPSANN`, `spJitter`, and `plot`.

Despite **spsann** functions are classified into three general family of functions defined according to the purpose for which they were designed, the documentation is constructed with regard to the respective objective functions. For example, every **spsann** function that uses as quality criterion the MSSD is documented in the same documentation page. The exception are the auxiliary functions, that generally are documented separately.

Support

spsann was initially developed as part of the PhD research project entitled 'Contribution to the

Construction of Models for Predicting Soil Properties', developed by Alessandro Samuel-Rosa under the supervision of Lúcia Helena Cunha dos Anjos <lanjos@ufrrj.br> (Universidade Federal Rural do Rio de Janeiro, Brazil), Gustavo de Mattos Vasques <gustavo.vasques@embrapa.br> (Embrapa Solos, Brazil), and Gerard B. M. Heuvelink <gerard.heuvelink@wur.nl> (ISRIC – World Soil Information, the Netherlands). The project was supported from March/2012 to February/2016 by the CAPES Foundation, Ministry of Education of Brazil, and the CNPq Foundation, Ministry of Science and Technology of Brazil.

Contributors

Some of the solutions used to build **spsann** were found in the source code of other R-packages and scripts developed and published by other researchers. For example, the original skeleton of the optimization functions was adopted from the **intamapInteractive** package with the approval of the package authors, Edzer Pebesma <edzer.pebesma@uni-muenster.de> and Jon Skoien <jon.skoien@gmail.com>. The current skeleton is based on the later adoption of several solutions implemented in the script developed and published by Murray Lark <mlark@bgs.ac.uk> as part of a short course ('Computational tools to optimize spatial sampling') offered for the first time at the 2015 EGU General Assembly in Vienna, Austria.

A few small solutions were adopted from the packages **SpatialTools**, authored by Joshua French <joshua.french@ucdenver.edu>, **clhs**, authored by Pierre Roudier <roudierp@landcareresearch.co.nz>, and **spcosa**, authored by Dennis Walvoort <dennis.Walvoort@wur.nl>, Dick Brus <dick.brus@wur.nl>, and Jaap de Gruijter <Jaap.degruijter@wur.nl>.

Major conceptual contributions were made by Gerard Heuvelink <gerard.heuvelink@wur.nl>, Dick Brus <dick.brus@wur.nl>, Murray Lark <mlark@bgs.ac.uk>, and Edzer Pebesma <edzer.pebesma@uni-muenster.

Author(s)

Author and Maintainer: Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>.

minmaxPareto *Pareto minimum and maximum values*

Description

Compute the minimum and maximum attainable values of the objective functions that compose a multi-objective combinatorial optimization problem.

Usage

```
minmaxPareto(osc, candi, covars)
```

Arguments

osc	A list of objects of class OptimizedSampleConfiguration (OSC). Each OSC of the list must be named after the objective function with which it has been optimized. For example, osc = list(CORR = osc_corr, DIST = osc_dist).

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`covars` Data frame or matrix with the covariates in the columns.

Details

Multi-objective combinatorial optimization problems: A method of solving a multi-objective combinatorial optimization problem (MOCOP) is to aggregate the objective functions into a single *utility function*. In **spsann**, the aggregation is performed using the *weighted sum method*, which incorporates in the weights the preferences of the user regarding the relative importance of each objective function.

The weighted sum method is affected by the relative magnitude of the different function values. The objective functions implemented in **spsann** have different units and orders of magnitude. The consequence is that the objective function with the largest values may have a numerical dominance during the optimization. In other words, the weights may not express the true preferences of the user, resulting that the meaning of the utility function becomes unclear because the optimization will favour the objective function which is numerically dominant.

A reasonable solution to avoid the numerical dominance of any objective function is to scale the objective functions so that they are constrained to the same approximate range of values. Several function-transformation methods can be used for this end and **spsann** has four of them available.

The *upper-lower-bound approach* requires the user to inform the maximum (nadir point) and minimum (utopia point) absolute function values. The resulting function values will always range between 0 and 1.

The *upper-bound approach* requires the user to inform only the nadir point, while the utopia point is set to zero. The upper-bound approach for transformation aims at equalizing only the upper bounds of the objective functions. The resulting function values will always be smaller than or equal to 1.

In most cases, the absolute maximum and minimum values of an objective function cannot be calculated exactly. If the user is uncomfortable with guessing the nadir and utopia points, there an option for using *numerical simulations*. It consists of computing the function value for many random system configurations. The mean function value obtained over multiple simulations is used to set the nadir point, while the the utopia point is set to zero. This approach is similar to the upper-bound approach, but the function values will have the same orders of magnitude only at the starting point of the optimization. Function values larger than one are likely to occur during the optimization. We recommend the user to avoid this approach whenever possible because the effect of the starting configuration on the optimization as a whole usually is insignificant or arbitrary.

The *upper-lower-bound approach* with the minimum and maximum *attainable* values of the objective functions that compose the MOCOP, also known as the *Pareto minimum and maximum values*, is the most elegant solution to scale the objective functions. However, it is the most time consuming. It works as follows:

1. Optimize a sample configuration with respect to each objective function that composes the MOCOP;
2. Compute the function value of every objective function that composes the MOCOP for every optimized sample configuration;
3. Record the minimum and maximum absolute function values attained for each objective function that composes the MOCOP – these are the Pareto minimum and maximum.

For example, consider **ACDC**, a MOCOP composed of two objective functions: **CORR** and **DIST**. The minimum absolute attainable value of **CORR** is obtained when the sample configuration is optimized with respect to only **CORR**, i.e. when the evaluator and generator objective functions are the same (see the intersection between the second line and second column in the table below). This is the Pareto minimum of **CORR**. It follows that the maximum absolute attainable value of **CORR** is obtained when the sample configuration is optimized with regard to only **DIST**, i.e. when the evaluator function is difference from the generator function (see the intersection between the first row and the second column in the table below). This is the Pareto maximum of **CORR**. The same logic applies for finding the Pareto minimum and maximum of **DIST**.

Evaluator	*Generator*	
	DIST	CORR
DIST	0.5	8.6
CORR	6.4	0.3

Value

A data frame with the Pareto minimum and maximum values.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Arora, J. *Introduction to optimum design*. Waltham: Academic Press, p. 896, 2011.

Marler, R. T.; Arora, J. S. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, v. 26, p. 369-395, 2004.

Marler, R. T.; Arora, J. S. Function-transformation methods for multi-objective optimization. *Engineering Optimization*, v. 37, p. 551-570, 2005.

Marler, R. T.; Arora, J. S. The weighted sum method for multi-objective optimization: new insights. *Structural and Multidisciplinary Optimization*, v. 41, p. 853-862, 2009.

See Also

`optimACDC`, `SPAN`

Examples

```
## Not run:
# This example takes more than 5 seconds
require(sp)
data(meuse.grid)
candi <- meuse.grid[, 1:2]
covars <- meuse.grid[, c(1, 2)]

# CORR
```

```
schedule <- scheduleSPSANN(initial.acceptance = 0.1, chains = 1,
                           x.max = 1540, y.max = 2060, x.min = 0,
                           y.min = 0, cellsize = 40)
set.seed(2001)
osc_corr <- optimCORR(points = 10, candi = candi, covars = covars,
                      schedule = schedule)

# DIST
set.seed(2001)
osc_dist <- optimDIST(points = 10, candi = candi, covars = covars,
                      schedule = schedule)

# PPL
set.seed(2001)
osc_ppl <- optimPPL(points = 10, candi = candi, schedule = schedule)

# MSSD
set.seed(2001)
osc_mssd <- optimMSSD(points = 10, candi = candi, schedule = schedule)

# Pareto
pareto <- minmaxPareto(osc = list(DIST = osc_dist, CORR = osc_corr,
                                  PPL = osc_ppl, MSSD = osc_mssd),
                       candi = candi, covars = covars)
pareto

## End(Not run)
```

objSPSANN *Auxiliary tools*

Description

Auxiliary tools used in the optimization of sample configurations using spatial simulated annealing.

Usage

```
objSPSANN(osc, at = "end", n = 1)
```

Arguments

osc	Object of class OptimizedSampleConfiguration.
at	Point of the optimization at which the energy state should be returned. Available options: "start", for the start, and "end", for the end of the optimization. Defaults to at = "end".
n	Number of instances that should be returned. Defaults to n = 1.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

optimACDC *Optimization of sample configurations for spatial trend identification and estimation (III)*

Description

Optimize a sample configuration for spatial trend identification and estimation. An utility function *U* is defined so that the sample reproduces the bivariate association/correlation between the covariates, as well as their marginal distribution (**ACDC**). The utility function is obtained aggregating two objective functions: **CORR** and **DIST**.

Usage

```
optimACDC(points, candi, covars, strata.type = "area", use.coords = FALSE,
  schedule = scheduleSPSANN(), plotit = FALSE, track = FALSE, boundary,
  progress = "txt", verbose = FALSE, weights = list(CORR = 0.5, DIST =
  0.5), nadir = list(sim = NULL, seeds = NULL, user = NULL, abs = NULL),
  utopia = list(user = NULL, abs = NULL))

objACDC(points, candi, covars, strata.type = "area", use.coords = FALSE,
  weights = list(CORR = 0.5, DIST = 0.5), nadir = list(sim = NULL, seeds =
  NULL, user = NULL, abs = NULL), utopia = list(user = NULL, abs = NULL))
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

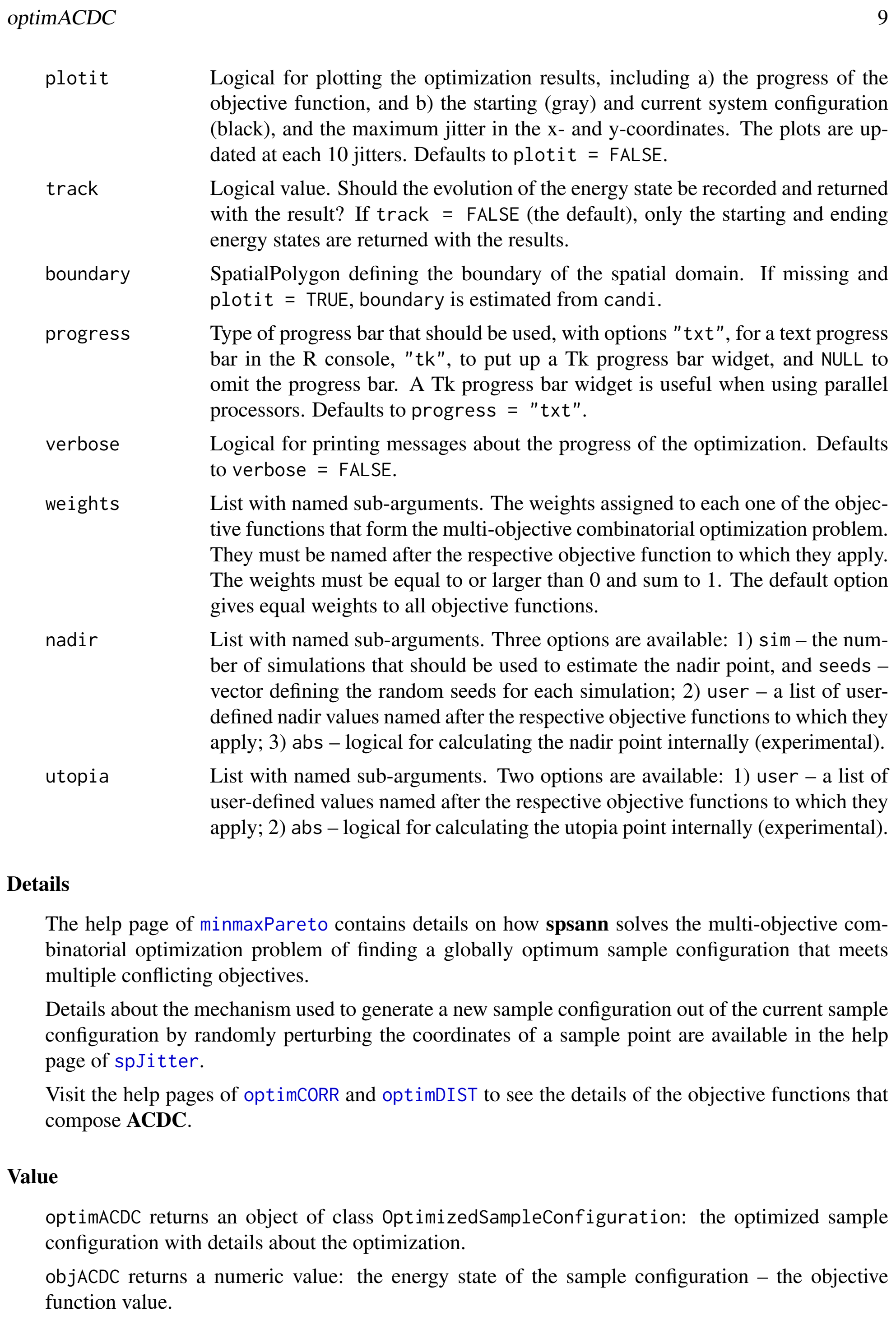
`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`covars` Data frame or matrix with the covariates in the columns.

`strata.type` Character value setting the type of stratification that should be used to create the marginal sampling strata (or factor levels) for the numeric covariates. Available options are `"area"`, for equal-area, and `"range"`, for equal-range. Defaults to `strata.type = "area"`.

`use.coords` Logical value. Should the geographic coordinates be used as covariates? Defaults to `use.coords = FALSE`.

`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.



`plotit` Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.

`track` Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.

`boundary` SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.

`progress` Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.

`verbose` Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

`weights` List with named sub-arguments. The weights assigned to each one of the objective functions that form the multi-objective combinatorial optimization problem. They must be named after the respective objective function to which they apply. The weights must be equal to or larger than 0 and sum to 1. The default option gives equal weights to all objective functions.

`nadir` List with named sub-arguments. Three options are available: 1) `sim` – the number of simulations that should be used to estimate the nadir point, and `seeds` – vector defining the random seeds for each simulation; 2) `user` – a list of user-defined nadir values named after the respective objective functions to which they apply; 3) `abs` – logical for calculating the nadir point internally (experimental).

`utopia` List with named sub-arguments. Two options are available: 1) `user` – a list of user-defined values named after the respective objective functions to which they apply; 2) `abs` – logical for calculating the utopia point internally (experimental).

Details

The help page of `minmaxPareto` contains details on how **spsann** solves the multi-objective combinatorial optimization problem of finding a globally optimum sample configuration that meets multiple conflicting objectives.

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Visit the help pages of `optimCORR` and `optimDIST` to see the details of the objective functions that compose **ACDC**.

Value

`optimACDC` returns an object of class `OptimizedSampleConfiguration`: the optimized sample configuration with details about the optimization.

`objACDC` returns a numeric value: the energy state of the sample configuration – the objective function value.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

This function was derived with modifications from the method known as the *conditioned Latin Hypercube sampling* originally proposed by Minasny and McBratney (2006), and implemented in the R-package **clhs** by Pierre Roudier.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Minasny, B.; McBratney, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, v. 32, p. 1378-1388, 2006.

Minasny, B.; McBratney, A. B. Conditioned Latin Hypercube Sampling for calibrating soil sensor data to soil properties. Chapter 9. Viscarra Rossel, R. A.; McBratney, A. B.; Minasny, B. (Eds.) *Proximal Soil Sensing*. Amsterdam: Springer, p. 111-119, 2010.

Roudier, P.; Beaudette, D.; Hewitt, A. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. *5th Global Workshop on Digital Soil Mapping*. Sydney, p. 227-231, 2012.

See Also

clhs, cramer

Examples

```
data(meuse.grid, package = "sp")
candi <- meuse.grid[1:1000, 1:2]
nadir <- list(sim = 10, seeds = 1:10)
utopia <- list(user = list(DIST = 0, CORR = 0))
covars <- meuse.grid[1:1000, 5]
schedule <- scheduleSPSANN(
  chains = 1, initial.temperature = 5, x.max = 1540, y.max = 2060,
  x.min = 0, y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimACDC(
  points = 10, candi = candi, covars = covars, nadir = nadir,
  use.coords = TRUE, utopia = utopia, schedule = schedule)
objSPSANN(res) - objACDC(
  points = res, candi = candi, covars = covars,
  use.coords = TRUE, nadir = nadir, utopia = utopia)
```

`optimCLHS` *Optimization of sample configurations for spatial trend identification and estimation (IV)*

Description

Optimize a sample configuration for spatial trend identification and estimation using the method proposed by Minasny and McBratney (2006), known as the conditioned Latin hypercube sampling. An utility function *U* is defined so that the sample reproduces the marginal distribution and correlation matrix of the numeric covariates, and the class proportions of the factor covariates (**CLHS**). The utility function is obtained aggregating three objective functions: **O1**, **O2**, and **O3**.

Usage

```
optimCLHS(points, candi, covars, use.coords = FALSE,
  schedule = scheduleSPSANN(), plotit = FALSE, track = FALSE, boundary,
  progress = "txt", verbose = FALSE, weights = list(O1 = 1/3, O2 = 1/3, O3
  = 1/3))

objCLHS(points, candi, covars, use.coords = FALSE, weights = list(O1 = 1/3,
  O2 = 1/3, O3 = 1/3))
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`covars` Data frame or matrix with the covariates in the columns.

`use.coords` Logical value. Should the geographic coordinates be used as covariates? Defaults to `use.coords = FALSE`.

`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.

`plotit` Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.

`track` Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.

`boundary` SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.

`progress` Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.

`verbose` Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

`weights` List with named sub-arguments. The weights assigned to each one of the objective functions that form the multi-objective combinatorial optimization problem. They must be named after the respective objective function to which they apply. The weights must be equal to or larger than 0 and sum to 1. The default option gives equal weights to all objective functions.

Details

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Marginal sampling strata: #' Reproducing the marginal distribution of the numeric covariates depends upon the definition of marginal sampling strata. *Equal-area* marginal sampling strata are defined using the sample quantiles estimated with `quantile` using a continuous function (`type = 7`), that is, a function that interpolates between existing covariate values to estimate the sample quantiles. This is the procedure implemented in the method of Minasny and McBratney (2006), which creates breakpoints that do not occur in the population of existing covariate values. Depending on the level of discretisation of the covariate values, that is, how many significant digits they have, this can create repeated breakpoints, resulting in empty marginal sampling strata. The number of empty marginal sampling strata will ultimately depend on the frequency distribution of the covariate and on the number of sampling points. The effect of these features on the spatial modelling outcome still is poorly understood.

Correlation between numeric covariates: The *correlation* between two numeric covariates is measured using the sample Pearson's *r*, a descriptive statistic that ranges from $-1$ to $+1$. This statistic is also known as the sample linear correlation coefficient. The effect of ignoring the correlation among factor covariates and between factor and numeric covariates on the spatial modelling outcome still is poorly understood.

Multi-objective combinatorial optimization: A method of solving a multi-objective combinatorial optimization problem (MOCOP) is to aggregate the objective functions into a single utility function *U*. In the **spsann** package, as in the original implementation of the CLHS by Minasny and McBratney (2006), the aggregation is performed using the *weighted sum method*, which uses weights to incorporate the preferences of the user about the relative importance of each objective function. When the user has no preference, the objective functions receive equal weights.

The weighted sum method is affected by the relative magnitude of the different objective function values. The objective functions implemented in `optimCLHS` have different units and orders of magnitude. The consequence is that the objective function with the largest values, generally **O1**, may have a numerical dominance during the optimization. In other words, the weights may not express the true preferences of the user, resulting that the meaning of the utility function becomes unclear because the optimization will favour the objective function which is numerically dominant.

An efficient solution to avoid numerical dominance is to scale the objective functions so that they are constrained to the same approximate range of values, at least in the end of the optimization. However, as in the original implementation of the CLHS by Minasny and McBratney (2006), `optimCLHS` uses the naive aggregation method, which ignores that the three objective functions have different units and orders of magnitude. The same aggregation procedure is implemented in the **clhs** package. The effect of ignoring the need to scale the objective functions on the spatial modelling outcome still is poorly understood.

Value

`optimCLHS` returns an object of class `OptimizedSampleConfiguration`: the optimized sample configuration with details about the optimization.

`objCLHS` returns a numeric value: the energy state of the sample configuration – the objective function value.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

The (only) difference of `optimCLHS` to the original Fortran implementation of Minasny and McBratney (2006), and to the `clhs` function implemented in the **clhs** package by Pierre Roudier, is the annealing schedule.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

References

Minasny, B.; McBratney, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, v. 32, p. 1378-1388, 2006.

Minasny, B.; McBratney, A. B. Conditioned Latin Hypercube Sampling for calibrating soil sensor data to soil properties. Chapter 9. Viscarra Rossel, R. A.; McBratney, A. B.; Minasny, B. (Eds.) *Proximal Soil Sensing*. Amsterdam: Springer, p. 111-119, 2010.

Roudier, P.; Beaudette, D.; Hewitt, A. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. *5th Global Workshop on Digital Soil Mapping*. Sydney, p. 227-231, 2012.

See Also

clhs, optimACDC

Examples

```
data(meuse.grid, package = "sp")
candi <- meuse.grid[1:1000, 1:2]
covars <- meuse.grid[1:1000, 5]
weights <- list(O1 = 0.5, O3 = 0.5)
schedule <- scheduleSPSANN(
  chains = 1, initial.temperature = 20, x.max = 1540, y.max = 2060,
  x.min = 0, y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimCLHS(
  points = 10, candi = candi, covars = covars, use.coords = TRUE,
  weights = weights, schedule = schedule)
objSPSANN(res) - objCLHS(
  points = res, candi = candi, covars = covars, use.coords = TRUE,
  weights = weights)
```

optimCORR — *Optimization of sample configurations for spatial trend identification and estimation (I)*

Description

Optimize a sample configuration for spatial trend identification and estimation. A criterion is defined so that the sample reproduces the bivariate association/correlation between the covariates (**CORR**).

Usage

```
optimCORR(points, candi, covars, strata.type = "area", use.coords = FALSE,
  schedule = scheduleSPSANN(), plotit = FALSE, track = FALSE, boundary,
  progress = "txt", verbose = FALSE)

objCORR(points, candi, covars, strata.type = "area", use.coords = FALSE)
```

Arguments

points	Integer value, integer vector, data frame or matrix. If points is an integer value, it defines the number of points that should be randomly sampled from candi to form the starting system configuration. If points is a vector of integer values, it contains the row indexes of candi that correspond to the points that form the starting system configuration. If points is a data frame or matrix, it must have three columns in the following order: [, "id"] the row indexes of candi that correspond to each point, [, "x"] the projected x-coordinates, and [, "y"] the projected y-coordinates. Note that in the later case, points must be a subset of candi.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`covars` Data frame or matrix with the covariates in the columns.

`strata.type` Character value setting the type of stratification that should be used to create the marginal sampling strata (or factor levels) for the numeric covariates. Available options are `"area"`, for equal-area, and `"range"`, for equal-range. Defaults to `strata.type = "area"`.

`use.coords` Logical value. Should the geographic coordinates be used as covariates? Defaults to `use.coords = FALSE`.

`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.

`plotit` Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.

`track` Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.

`boundary` SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.

`progress` Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.

`verbose` Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

Details

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Association/Correlation between covariates: The *correlation* between two numeric covariates is measured using the Pearson's r, a descriptive statistic that ranges from -1 to $+1$. This statistic is also known as the linear correlation coefficient.

When the set of covariates includes factor covariates, all numeric covariates are transformed into factor covariates. The factor levels are defined using the marginal sampling strata created from one of the two methods available (equal-area or equal-range strata).

The *association* between two factor covariates is measured using the Cramér's V, a descriptive statistic that ranges from 0 to $+1$. The closer to $+1$ the Cramér's V is, the stronger the association between two factor covariates.

The main weakness of using the Cramér's V is that, while the Pearson's r shows the degree and direction of the association between two covariates (negative or positive), the Cramér's V only measures the degree of association (weak or strong). The effect of replacing the Pearson's r with the Cramér's V on the spatial modelling outcome still is poorly understood.

Value

optimCORR returns an object of class OptimizedSampleConfiguration: the optimized sample configuration with details about the optimization.

objCORR returns a numeric value: the energy state of the sample configuration – the objective function value.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Cramér, H. *Mathematical methods of statistics*. Princeton: Princeton University Press, p. 575, 1946.

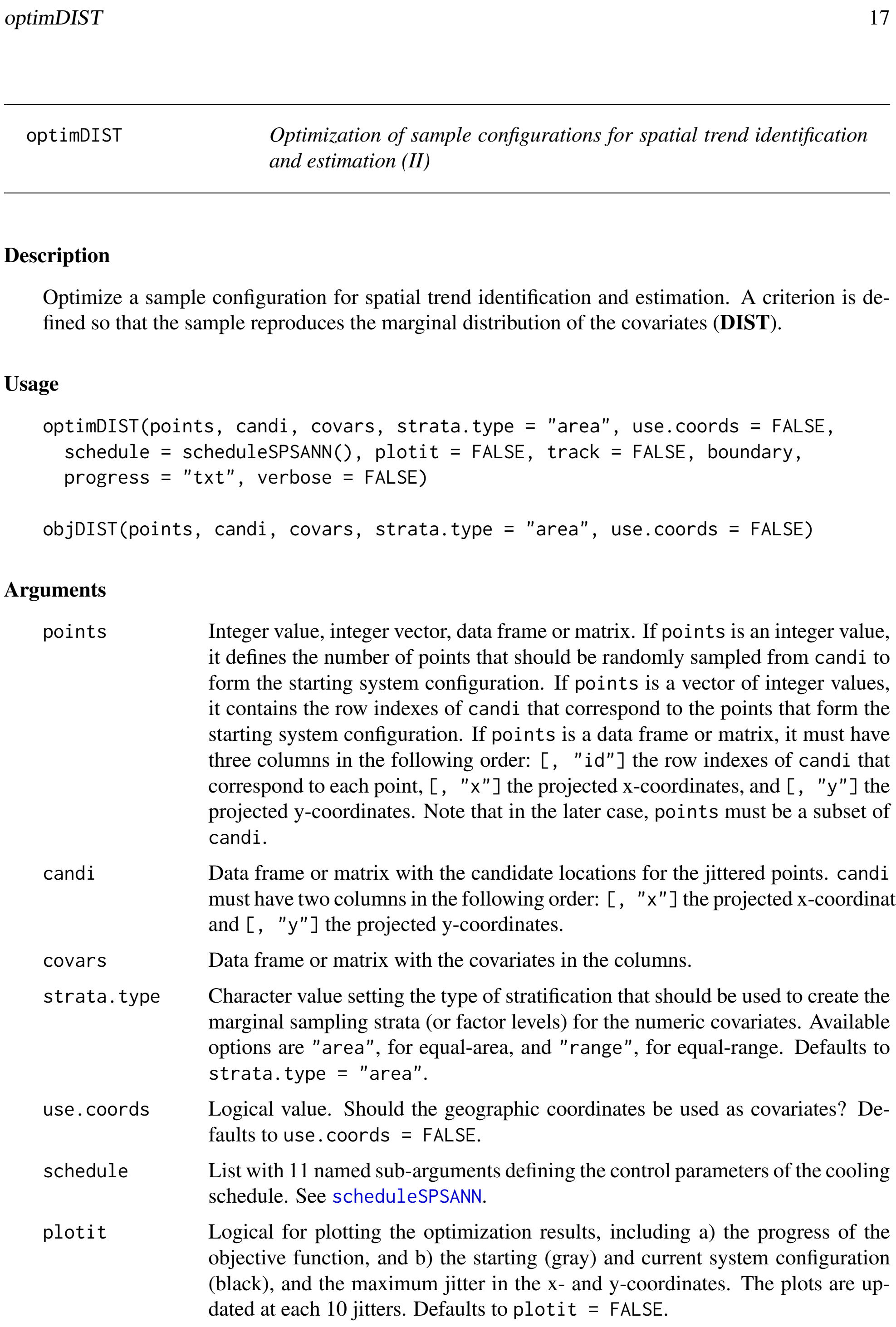
Everitt, B. S. *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press, p. 432, 2006.

See Also

clhs, cramer, optimACDC

Examples

```
data(meuse.grid, package = "sp")
candi <- meuse.grid[1:1000, 1:2]
covars <- meuse.grid[1:1000, 5]
schedule <- scheduleSPSANN(
  initial.temperature = 5, chains = 1, x.max = 1540, y.max = 2060,
  x.min = 0, y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimCORR(
  points = 10, candi = candi, covars = covars, use.coords = TRUE,
  schedule = schedule)
objSPSANN(res) - objCORR(
  points = res, candi = candi, covars = covars, use.coords = TRUE)
```



`optimDIST` *Optimization of sample configurations for spatial trend identification and estimation (II)*

Description

Optimize a sample configuration for spatial trend identification and estimation. A criterion is defined so that the sample reproduces the marginal distribution of the covariates (**DIST**).

Usage

```
optimDIST(points, candi, covars, strata.type = "area", use.coords = FALSE,
  schedule = scheduleSPSANN(), plotit = FALSE, track = FALSE, boundary,
  progress = "txt", verbose = FALSE)

objDIST(points, candi, covars, strata.type = "area", use.coords = FALSE)
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`covars` Data frame or matrix with the covariates in the columns.

`strata.type` Character value setting the type of stratification that should be used to create the marginal sampling strata (or factor levels) for the numeric covariates. Available options are `"area"`, for equal-area, and `"range"`, for equal-range. Defaults to `strata.type = "area"`.

`use.coords` Logical value. Should the geographic coordinates be used as covariates? Defaults to `use.coords = FALSE`.

`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.

`plotit` Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.

track Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.

boundary SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.

progress Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.

verbose Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

Details

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Marginal distribution of covariates: Reproducing the marginal distribution of the numeric covariates depends upon the definition of marginal sampling strata. These marginal sampling strata are also used to define the factor levels of all numeric covariates that are passed together with factor covariates. Two types of marginal sampling strata can be used: *equal-area* and *equal-range*.

Equal-area marginal sampling strata are defined using the sample quantiles estimated with `quantile` using a discontinuous function (`type = 3`). Using a discontinuous function avoids creating breakpoints that do not occur in the population of existing covariate values.

Depending on the level of discretization of the covariate values, `quantile` produces repeated breakpoints. A breakpoint will be repeated if that value has a relatively high frequency in the population of covariate values. The number of repeated breakpoints increases with the number of marginal sampling strata. Repeated breakpoints result in empty marginal sampling strata. To avoid this, only the unique breakpoints are used.

Equal-range marginal sampling strata are defined by breaking the range of covariate values into pieces of equal size. Depending on the level of discretization of the covariate values, this method creates breakpoints that do not occur in the population of existing covariate values. Such breakpoints are replaced with the nearest existing covariate value identified using Euclidean distances.

Like the equal-area method, the equal-range method can produce empty marginal sampling strata. The solution used here is to merge any empty marginal sampling strata with the closest non-empty marginal sampling strata. This is identified using Euclidean distances as well.

The approaches used to define the marginal sampling strata result in each numeric covariate having a different number of marginal sampling strata, some of them with different area/size. Because the goal is to have a sample that reproduces the marginal distribution of the covariate, each marginal sampling strata will have a different number of sample points. The wanted distribution of the number of sample points per marginal strata is estimated empirically as the proportion of points in the population of existing covariate values that fall in each marginal sampling strata.

Value

optimDIST returns an object of class OptimizedSampleConfiguration: the optimized sample configuration with details about the optimization.

objDIST returns a numeric value: the energy state of the sample configuration – the objective function value.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Hyndman, R. J.; Fan, Y. Sample quantiles in statistical packages. *The American Statistician*, v. 50, p. 361-365, 1996.

Everitt, B. S. *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press, p. 432, 2006.

See Also

clhs, optimACDC

Examples

```
require(sp)
data(meuse.grid)
candi <- meuse.grid[, 1:2]
covars <- meuse.grid[, 5]
schedule <- scheduleSPSANN(initial.temperature = 1, chains = 1,
                           x.max = 1540, y.max = 2060, x.min = 0,
                           y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimDIST(points = 10, candi = candi, covars = covars,
                 use.coords = TRUE, schedule = schedule)
objSPSANN(res) -
  objDIST(points = res, candi = candi, covars = covars, use.coords = TRUE)
```

optimMKV *Optimization of sample configurations for spatial interpolation (II)*

Description

Optimize a sample configuration for spatial interpolation with a known linear model. A criterion is defined so that the sample configuration minimizes the mean or maximum kriging variance (**MKV**).

Usage

```
optimMKV(points, candi, covars, eqn = z ~ 1, vgm, krige.stat = "mean", ...,
  schedule = scheduleSPSANN(), plotit = FALSE, track = FALSE, boundary,
  progress = "txt", verbose = FALSE)

objMKV(points, candi, covars, eqn = z ~ 1, vgm, krige.stat = "mean", ...)
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`covars` Data frame or matrix with the covariates in the columns.

`eqn` Formula string that defines the dependent variable `z` as a linear model of the independent variables contained in `covars`. Defaults to `eqn = z ~ 1`, that is, ordinary kriging. See the argument `formula` in the function `krige` for more information.

`vgm` Object of class `variogramModel`. See the argument `model` in the function `krige` for more information.

`krige.stat` Character value defining the statistic that should be used to summarize the kriging variance. Available options are `"mean"` and `"max"` for the mean and maximum kriging variance, respectively. Defaults to `krige.stat = "mean"`.

`...` further arguments passed to `krige`.

`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.

plotit	Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.
track	Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.
boundary	SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.
progress	Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.
verbose	Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

Details

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Value

`optimMKV` returns an object of class `OptimizedSampleConfiguration`: the optimized sample configuration with details about the optimization.

`objMKV` returns a numeric value: the energy state of the sample configuration – the objective function value.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

This function is based on the method originally proposed by Heuvelink, Brus and de Gruijter (2006) and implemented in the R-package **intamapInteractive** by Edzer Pebesma and Jon Skoien.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

References

Brus, D. J.; Heuvelink, G. B. M. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*. v. 138, p. 86-95, 2007.

Heuvelink, G. B. M.; Brus, D. J.; de Gruijter, J. J. Optimization of sample configurations for digital mapping of soil properties with universal kriging. In: Lagacherie, P.; McBratney, A. & Voltz, M. (Eds.) *Digital soil mapping - an introductory perspective*. Elsevier, v. 31, p. 137-151, 2006.

Examples

```
## Not run:
data(meuse.grid, package = "sp")
candi <- meuse.grid[1:1000, 1:2]
covars <- as.data.frame(meuse.grid)[1:1000, ]
vgm <- gstat::vgm(psill = 10, model = "Exp", range = 500, nugget = 8)
schedule <- scheduleSPSANN(
  initial.temperature = 10, chains = 1, x.max = 1540, y.max = 2060,
  x.min = 0,  y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimMKV(
  points = 10, candi = candi, covars = covars, eqn = z ~ dist,
  vgm = vgm, schedule = schedule)
objSPSANN(res) - objMKV(
  points = res, candi = candi, covars = covars,  eqn = z ~ dist,
  vgm = vgm)

## End(Not run)
```

optimMSSD *Optimization of sample configurations for spatial interpolation (I)*

Description

Optimize a sample configuration for spatial interpolation. The criterion used is the mean squared shortest distance (**MSSD**) between sample points and prediction points.

Usage

```
optimMSSD(points, candi, schedule = scheduleSPSANN(), plotit = FALSE,
  track = FALSE, boundary, progress = "txt", verbose = FALSE)

objMSSD(points, candi)
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.

`plotit` Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.

`track` Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.

`boundary` SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.

`progress` Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.

`verbose` Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

Details

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Spatial coverage sampling: Spatial coverage sampling is based on the knowledge that the kriging variance depends upon the distance between sample points. As such, the better the spread of the sample points in the spatial domain, the smaller the kriging variance. This is similar to using a regular grid of sample points. However, a regular grid usually is suboptimal for irregularly shaped areas.

Value

`optimMSSD` returns an object of class `OptimizedSampleConfiguration`: the optimized sample configuration with details about the optimization.

`objMSSD` returns a numeric value: the energy state of the sample configuration – the objective function value.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

This function was derived with modifications from the method known as *spatial coverage sampling* originally proposed by Brus, de Gruijter and van Groenigen (2006), and implemented in the R-package **spcosa** by Dennis Walvoort, Dick Brus and Jaap de Gruijter.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Brus, D. J.; de Gruijter, J. J.; van Groenigen, J.-W. Designing spatial coverage samples using the k-means clustering algorithm. In: P. Lagacherie,A. M.; Voltz, M. (Eds.) *Digital soil mapping – an introductory perspective*. Elsevier, v. 31, p. 183-192, 2006.

de Gruijter, J. J.; Brus, D.; Bierkens, M.; Knotters, M. *Sampling for natural resource monitoring*. Berlin: Springer, p. 332, 2006.

Walvoort, D. J. J.; Brus, D. J.; de Gruijter, J. J. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers and Geosciences*. v. 36, p. 1261-1267, 2010.

See Also

`distanceFromPoints`, `stratify`.

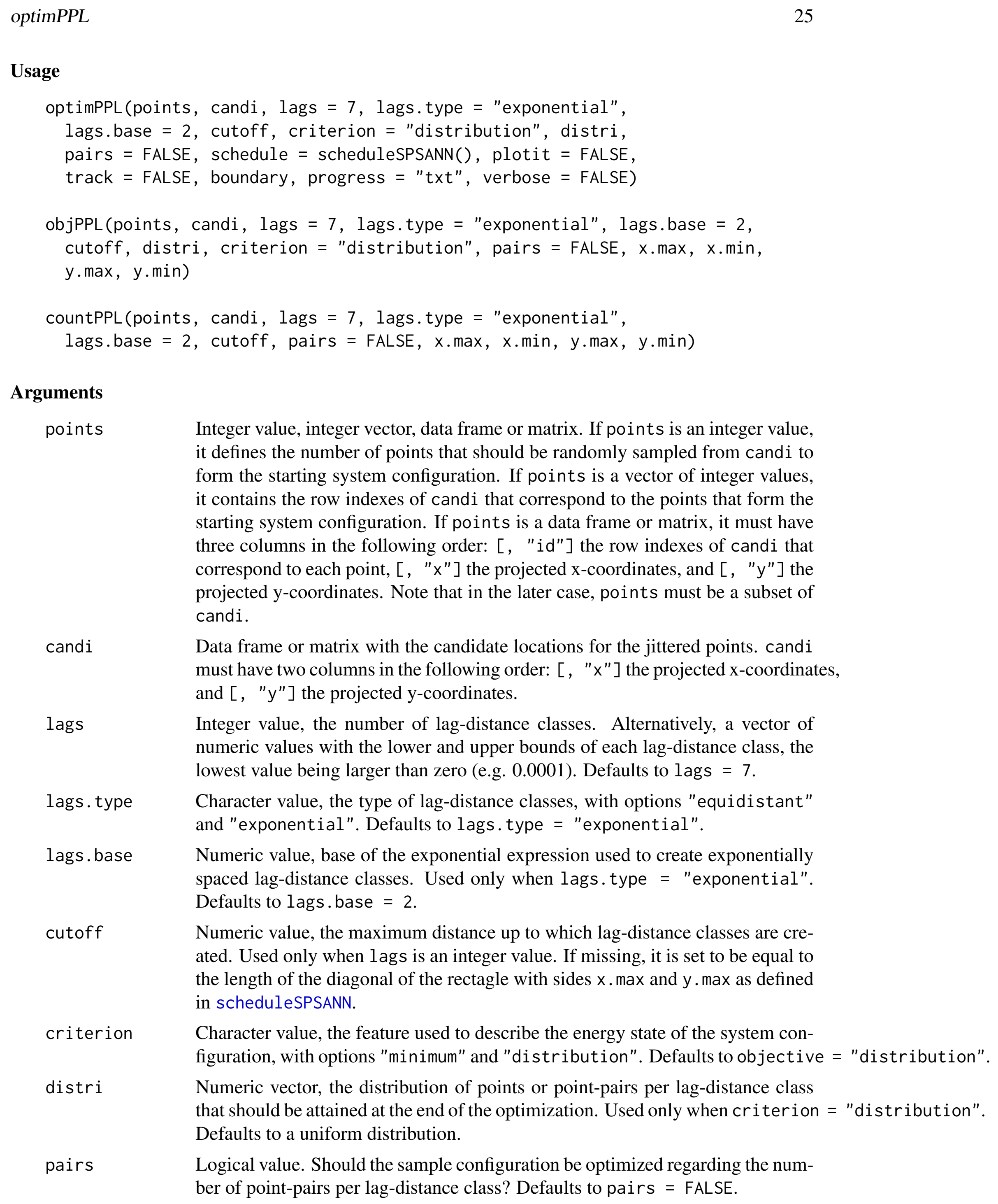
Examples

```
require(sp)
data(meuse.grid)
candi <- meuse.grid[, 1:2]
schedule <- scheduleSPSANN(chains = 1, initial.temperature = 5000000,
                           x.max = 1540, y.max = 2060, x.min = 0,
                           y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimMSSD(points = 10, candi = candi, schedule = schedule)
objSPSANN(res) - objMSSD(candi = candi, points = res)
```

`optimPPL` *Optimization of sample configurations for variogram identification and estimation*

Description

Optimize a sample configuration for variogram identification and estimation. A criterion is defined so that the optimized sample configuration has a given number of points or point-pairs contributing to each lag-distance class (**PPL**).



Usage

```
optimPPL(points, candi, lags = 7, lags.type = "exponential",
  lags.base = 2, cutoff, criterion = "distribution", distri,
  pairs = FALSE, schedule = scheduleSPSANN(), plotit = FALSE,
  track = FALSE, boundary, progress = "txt", verbose = FALSE)

objPPL(points, candi, lags = 7, lags.type = "exponential", lags.base = 2,
  cutoff, distri, criterion = "distribution", pairs = FALSE, x.max, x.min,
  y.max, y.min)

countPPL(points, candi, lags = 7, lags.type = "exponential",
  lags.base = 2, cutoff, pairs = FALSE, x.max, x.min, y.max, y.min)
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`lags` Integer value, the number of lag-distance classes. Alternatively, a vector of numeric values with the lower and upper bounds of each lag-distance class, the lowest value being larger than zero (e.g. 0.0001). Defaults to `lags = 7`.

`lags.type` Character value, the type of lag-distance classes, with options `"equidistant"` and `"exponential"`. Defaults to `lags.type = "exponential"`.

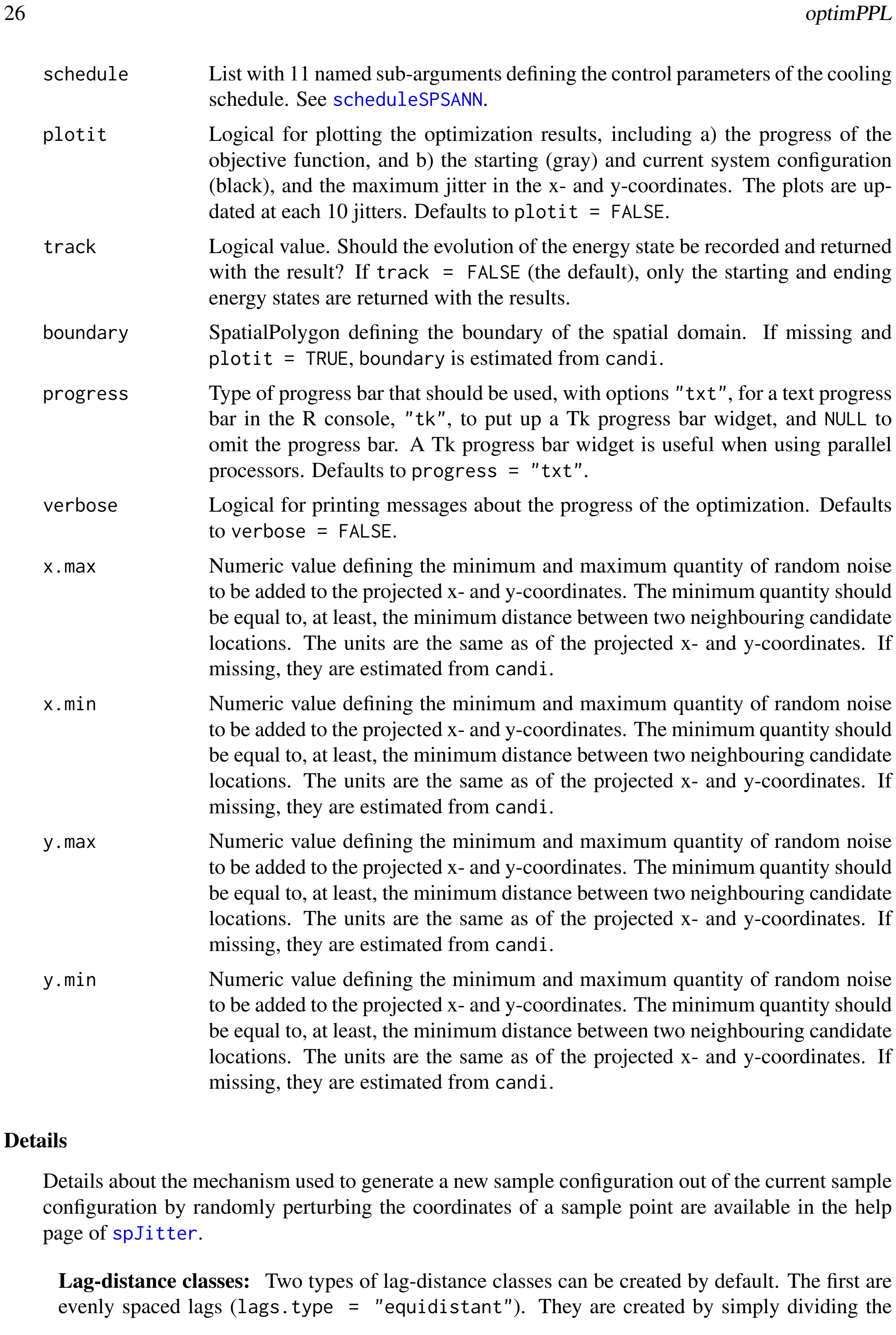
`lags.base` Numeric value, base of the exponential expression used to create exponentially spaced lag-distance classes. Used only when `lags.type = "exponential"`. Defaults to `lags.base = 2`.

`cutoff` Numeric value, the maximum distance up to which lag-distance classes are created. Used only when `lags` is an integer value. If missing, it is set to be equal to the length of the diagonal of the rectagle with sides `x.max` and `y.max` as defined in `scheduleSPSANN`.

`criterion` Character value, the feature used to describe the energy state of the system configuration, with options `"minimum"` and `"distribution"`. Defaults to `objective = "distribution"`.

`distri` Numeric vector, the distribution of points or point-pairs per lag-distance class that should be attained at the end of the optimization. Used only when `criterion = "distribution"`. Defaults to a uniform distribution.

`pairs` Logical value. Should the sample configuration be optimized regarding the number of point-pairs per lag-distance class? Defaults to `pairs = FALSE`.



`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.

`plotit` Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.

`track` Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.

`boundary` SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.

`progress` Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.

`verbose` Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

`x.max` Numeric value defining the minimum and maximum quantity of random noise to be added to the projected x- and y-coordinates. The minimum quantity should be equal to, at least, the minimum distance between two neighbouring candidate locations. The units are the same as of the projected x- and y-coordinates. If missing, they are estimated from `candi`.

`x.min` Numeric value defining the minimum and maximum quantity of random noise to be added to the projected x- and y-coordinates. The minimum quantity should be equal to, at least, the minimum distance between two neighbouring candidate locations. The units are the same as of the projected x- and y-coordinates. If missing, they are estimated from `candi`.

`y.max` Numeric value defining the minimum and maximum quantity of random noise to be added to the projected x- and y-coordinates. The minimum quantity should be equal to, at least, the minimum distance between two neighbouring candidate locations. The units are the same as of the projected x- and y-coordinates. If missing, they are estimated from `candi`.

`y.min` Numeric value defining the minimum and maximum quantity of random noise to be added to the projected x- and y-coordinates. The minimum quantity should be equal to, at least, the minimum distance between two neighbouring candidate locations. The units are the same as of the projected x- and y-coordinates. If missing, they are estimated from `candi`.

Details

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Lag-distance classes: Two types of lag-distance classes can be created by default. The first are evenly spaced lags (`lags.type = "equidistant"`). They are created by simply dividing the distance interval from 0.0001 to `cutoff` by the required number of lags. The minimum value of 0.0001 guarantees that a point does not form a pair with itself. The second type of lags is defined by exponential spacings (`lags.type = "exponential"`). The spacings are defined by the base b of the exponential expression b^n, where n is the required number of lags. The base is defined using the argument `lags.base`. See `vgmLags` for other details.

Using the default uniform distribution means that the number of point-pairs per lag-distance class (`pairs = TRUE`) is equal to $n \times (n-1)/(2 \times lag)$, where n is the total number of points and lag is the number of lags. If `pairs = FALSE`, then it means that the number of points per lag is equal to the total number of points. This is the same as expecting that each point contributes to every lag. Distributions other than the available options can be easily implemented changing the arguments `lags` and `distri`.

There are two optimizing criteria implemented. The first is called using `criterion = "distribution"` and is used to minimize the sum of the absolute differences between a pre-specified distribution and the observed distribution of points or point-pairs per lag-distance class. The second criterion is called using `criterion = "minimum"`. It corresponds to maximizing the minimum number of points or point-pairs observed over all lag-distance classes.

Value

`optimPPL` returns an object of class `OptimizedSampleConfiguration`: the optimized sample configuration with details about the optimization.

`objPPL` returns a numeric value: the energy state of the sample configuration – the objective function value.

`countPPL` returns a data.frame with three columns: a) the lower and b) upper limits of each lag-distance class, and c) the number of points or point-pairs per lag-distance class.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

References

Bresler, E.; Green, R. E. *Soil parameters and sampling scheme for characterizing soil hydraulic properties of a watershed*. Honolulu: University of Hawaii at Manoa, p. 42, 1982.

Pettitt, A. N.; McBratney, A. B. Sampling designs for estimating spatial variance components. *Applied Statistics*. v. 42, p. 185, 1993.

Russo, D. Design of an optimal sampling network for estimating the variogram. *Soil Science Society of America Journal*. v. 48, p. 708-716, 1984.

Truong, P. N.; Heuvelink, G. B. M.; Gosling, J. P. Web-based tool for expert elicitation of the variogram. *Computers and Geosciences*. v. 51, p. 390-399, 2013.

Warrick, A. W.; Myers, D. E. Optimization of sampling locations for variogram calculations. *Water Resources Research.* v. 23, p. 496-500, 1987.

Examples

```
## Not run:
# This example takes more than 5 seconds
require(sp)
data(meuse.grid)
candi <- meuse.grid[, 1:2]
schedule <- scheduleSPSANN(chains = 1, initial.temperature = 30,
                           x.max = 1540, y.max = 2060, x.min = 0,
                           y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimPPL(points = 10, candi = candi, schedule = schedule)
objSPSANN(res) - objPPL(points = res, candi = candi)
countPPL(points = res, candi = candi)

## End(Not run)
```

optimSPAN *Optimization of sample configurations for variogram and spatial trend identification and estimation, and for spatial interpolation*

Description

Optimize a sample configuration for variogram and spatial trend identification and estimation, and for spatial interpolation. An utility function U is defined so that the sample points cover, extend over, spread over, **SPAN** the feature, variogram and geographic spaces. The utility function is obtained aggregating four objective functions: **CORR**, **DIST**, **PPL**, and **MSSD**.

Usage

```
optimSPAN(points, candi, covars, strata.type = "area", use.coords = FALSE,
  lags = 7, lags.type = "exponential", lags.base = 2, cutoff,
  criterion = "distribution", distri, pairs = FALSE,
  schedule = scheduleSPSANN(), plotit = FALSE, track = FALSE, boundary,
  progress = "txt", verbose = FALSE, weights = list(CORR = 1/6, DIST =
  1/6, PPL = 1/3, MSSD = 1/3), nadir = list(sim = NULL, seeds = NULL, user =
  NULL, abs = NULL), utopia = list(user = NULL, abs = NULL))

objSPAN(points, candi, covars, strata.type = "area", use.coords = FALSE,
  lags = 7, lags.type = "exponential", lags.base = 2, cutoff,
  criterion = "distribution", distri, pairs = FALSE, x.max, x.min, y.max,
  y.min, weights = list(CORR = 1/6, DIST = 1/6, PPL = 1/3, MSSD = 1/3),
  nadir = list(sim = NULL, seeds = NULL, user = NULL, abs = NULL),
  utopia = list(user = NULL, abs = NULL))
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`covars` Data frame or matrix with the covariates in the columns.

`strata.type` Character value setting the type of stratification that should be used to create the marginal sampling strata (or factor levels) for the numeric covariates. Available options are `"area"`, for equal-area, and `"range"`, for equal-range. Defaults to `strata.type = "area"`.

`use.coords` Logical value. Should the geographic coordinates be used as covariates? Defaults to `use.coords = FALSE`.

`lags` Integer value, the number of lag-distance classes. Alternatively, a vector of numeric values with the lower and upper bounds of each lag-distance class, the lowest value being larger than zero (e.g. 0.0001). Defaults to `lags = 7`.

`lags.type` Character value, the type of lag-distance classes, with options `"equidistant"` and `"exponential"`. Defaults to `lags.type = "exponential"`.

`lags.base` Numeric value, base of the exponential expression used to create exponentially spaced lag-distance classes. Used only when `lags.type = "exponential"`. Defaults to `lags.base = 2`.

`cutoff` Numeric value, the maximum distance up to which lag-distance classes are created. Used only when `lags` is an integer value. If missing, it is set to be equal to the length of the diagonal of the rectagle with sides `x.max` and `y.max` as defined in `scheduleSPSANN`.

`criterion` Character value, the feature used to describe the energy state of the system configuration, with options `"minimum"` and `"distribution"`. Defaults to `objective = "distribution"`.

`distri` Numeric vector, the distribution of points or point-pairs per lag-distance class that should be attained at the end of the optimization. Used only when `criterion = "distribution"`. Defaults to a uniform distribution.

`pairs` Logical value. Should the sample configuration be optimized regarding the number of point-pairs per lag-distance class? Defaults to `pairs = FALSE`.

`schedule` List with 11 named sub-arguments defining the control parameters of the cooling schedule. See `scheduleSPSANN`.

`plotit` Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to `plotit = FALSE`.

`track` Logical value. Should the evolution of the energy state be recorded and returned with the result? If `track = FALSE` (the default), only the starting and ending energy states are returned with the results.

`boundary` SpatialPolygon defining the boundary of the spatial domain. If missing and `plotit = TRUE`, `boundary` is estimated from `candi`.

`progress` Type of progress bar that should be used, with options `"txt"`, for a text progress bar in the R console, `"tk"`, to put up a Tk progress bar widget, and `NULL` to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to `progress = "txt"`.

`verbose` Logical for printing messages about the progress of the optimization. Defaults to `verbose = FALSE`.

`weights` List with named sub-arguments. The weights assigned to each one of the objective functions that form the multi-objective combinatorial optimization problem. They must be named after the respective objective function to which they apply. The weights must be equal to or larger than 0 and sum to 1. The default option gives equal weights to all objective functions.

`nadir` List with named sub-arguments. Three options are available: 1) `sim` – the number of simulations that should be used to estimate the nadir point, and `seeds` – vector defining the random seeds for each simulation; 2) `user` – a list of user-defined nadir values named after the respective objective functions to which they apply; 3) `abs` – logical for calculating the nadir point internally (experimental).

`utopia` List with named sub-arguments. Two options are available: 1) `user` – a list of user-defined values named after the respective objective functions to which they apply; 2) `abs` – logical for calculating the utopia point internally (experimental).

`x.max, x.min, y.max, y.min`

Numeric value defining the minimum and maximum quantity of random noise to be added to the projected `x`- and `y`-coordinates. The minimum quantity should be equal to, at least, the minimum distance between two neighbouring candidate locations. The units are the same as of the projected `x`- and `y`-coordinates. If missing, they are estimated from `candi`.

Details

The help page of `minmaxPareto` contains details on how **spsann** solves the multi-objective combinatorial optimization problem of finding a globally optimum sample configuration that meets multiple conflicting objectives.

Details about the mechanism used to generate a new sample configuration out of the current sample configuration by randomly perturbing the coordinates of a sample point are available in the help page of `spJitter`.

Visit the help pages of `optimCORR`, `optimDIST`, `optimPPL`, and `optimMSSD` to see the details of the objective functions that compose **SPAN**.

Value

`optimSPAN` returns an object of class `OptimizedSampleConfiguration`: the optimized sample configuration with details about the optimization.

objSPAN returns a numeric value: the energy state of the sample configuration – the objective function value.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

See Also

optimCORR, optimDIST, optimPPL, optimMSSD

Examples

```
## Not run:
# This example takes more than 5 seconds to run!
require(sp)
data(meuse.grid)
candi <- meuse.grid[, 1:2]
nadir <- list(sim = 10, seeds = 1:10)
utopia <- list(user = list(DIST = 0, CORR = 0, PPL = 0, MSSD = 0))
covars <- meuse.grid[, 5]
schedule <- scheduleSPSANN(chains = 1, initial.temperature = 1,
                           x.max = 1540, y.max = 2060, x.min = 0,
                           y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimSPAN(points = 10, candi = candi, covars = covars, nadir = nadir,
                 use.coords = TRUE, utopia = utopia, schedule = schedule)
objSPSANN(res) -
  objSPAN(points = res, candi = candi, covars = covars, nadir = nadir,
          use.coords = TRUE, utopia = utopia)

## End(Not run)
```

---

optimUSER *Optimization of sample configurations using a user-defined objective function*

Description

Optimize a sample configuration using a user-defined objective function.

Usage

```
optimUSER(points, candi, fun, ..., schedule = scheduleSPSANN(),
  plotit = FALSE, track = FALSE, boundary, progress = "txt",
  verbose = FALSE)
```

Arguments

points	Integer value, integer vector, data frame or matrix. If points is an integer value, it defines the number of points that should be randomly sampled from candi to form the starting system configuration. If points is a vector of integer values, it contains the row indexes of candi that correspond to the points that form the starting system configuration. If points is a data frame or matrix, it must have three columns in the following order: [, "id"] the row indexes of candi that correspond to each point, [, "x"] the projected x-coordinates, and [, "y"] the projected y-coordinates. Note that in the later case, points must be a subset of candi.
candi	Data frame or matrix with the candidate locations for the jittered points. candi must have two columns in the following order: [, "x"] the projected x-coordinates, and [, "y"] the projected y-coordinates.
fun	A function defining the objective function that should be used to evaluate the energy state of the system configuration at each random perturbation of a candidate sample point. See 'Details' for more information.
...	Other arguments passed to the objective function. See 'Details' for more information.
schedule	List with 11 named sub-arguments defining the control parameters of the cooling schedule. See scheduleSPSANN.
plotit	Logical for plotting the optimization results, including a) the progress of the objective function, and b) the starting (gray) and current system configuration (black), and the maximum jitter in the x- and y-coordinates. The plots are updated at each 10 jitters. Defaults to plotit = FALSE.
track	Logical value. Should the evolution of the energy state be recorded and returned with the result? If track = FALSE (the default), only the starting and ending energy states are returned with the results.
boundary	SpatialPolygon defining the boundary of the spatial domain. If missing and plotit = TRUE, boundary is estimated from candi.
progress	Type of progress bar that should be used, with options "txt", for a text progress bar in the R console, "tk", to put up a Tk progress bar widget, and NULL to omit the progress bar. A Tk progress bar widget is useful when using parallel processors. Defaults to progress = "txt".
verbose	Logical for printing messages about the progress of the optimization. Defaults to verbose = FALSE.

Details

The user-defined objective function fun must be an object of class function and include the argument points. The argument points is defined in optimUSER as a matrix with three columns:

`[, 1]` the identification of each sample point given by the respective row indexes of `candi`, `[, 2]` the x-coordinates, and `[, 3]` the y-coordinates. The identification is useful to retrieve information from any data matrix used by the objective function defined by the user.

Value

`optimUSER` returns an object of class `OptimizedSampleConfiguration`: the optimized sample configuration with details about the optimization.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

Examples

```
## Not run:
# This example takes more than 5 seconds
require(sp)
require(SpatialTools)
data(meuse.grid)
candi <- meuse.grid[, 1:2]
schedule <- scheduleSPSANN(chains = 1, initial.temperature = 30,
                           x.max = 1540, y.max = 2060, x.min = 0,
                           y.min = 0, cellsize = 40)

# Define the objective function - number of points per lag distance class
objUSER <-
  function (points, lags, n_lags, n_pts) {
    dm <- SpatialTools::dist1(points[, 2:3])
    ppl <- vector()
    for (i in 1:n_lags) {
      n <- which(dm > lags[i] & dm <= lags[i + 1], arr.ind = TRUE)
      ppl[i] <- length(unique(c(n)))
    }
    distri <- rep(n_pts, n_lags)
    res <- sum(distri - ppl)
  }
lags <- seq(1, 1000, length.out = 10)

# Run the optimization using the user-defined objective function
set.seed(2001)
timeUSER <- Sys.time()
resUSER <- optimUSER(points = 10, fun = objUSER, lags = lags, n_lags = 9,
                     n_pts = 10, candi = candi, schedule = schedule)
timeUSER <- Sys.time() - timeUSER

# Run the optimization using the respective function implemented in spsann
set.seed(2001)
timePPL <- Sys.time()
resPPL <- optimPPL(points = 10, candi = candi, lags = lags,
                   schedule = schedule)
timePPL <- Sys.time() - timePPL
```

```
# Compare results
timeUSER
timePPL
lapply(list(resUSER, resPPL), countPPL, candi = candi, lags = lags)
objSPSANN(resUSER) - objSPSANN(resPPL)

## End(Not run)
```

plot.OptimizedSampleConfiguration

Plot an optimized sample configuration

Description

Plot the evolution of the energy state and the optimized sample configuration

Usage

```
## S3 method for class 'OptimizedSampleConfiguration'
plot(x, which = 1:2, boundary, ...)
```

Arguments

x	Object of class OptimizedSampleConfiguration returned by one of the optim-functions.
which	Which plot should be produced: evolution of the energy state (1), optimized sample configuration (2), or both (1:2)? Defaults to which = 1:2.
boundary	Object of class Spatial defining the boundary of the sampling region.
...	Other options passed to plot.

Examples

```
require(sp)
data(meuse.grid)
candi <- meuse.grid[, 1:2]
covars <- meuse.grid[, 5]
schedule <- scheduleSPSANN(initial.temperature = 5, chains = 1,
                           x.max = 1540, y.max = 2060, x.min = 0,
                           y.min = 0, cellsize = 40)
set.seed(2001)
res <- optimCORR(points = 10, candi = candi, covars = covars,
                 use.coords = TRUE, schedule = schedule)
plot(res)
```

scheduleSPSANN **spsann** *annealing schedule*

Description

Set the control parameters for the annealing schedule of **spsann** functions.

Usage

```
scheduleSPSANN(initial.acceptance = 0.95, initial.temperature = 0.001,
  temperature.decrease = 0.95, chains = 500, chain.length = 1,
  stopping = 10, x.max, x.min = 0, y.max, y.min = 0, cellsize)
```

Arguments

`initial.acceptance`
: Numeric value between 0 and 1 defining the initial acceptance probability, i.e. the proportion of proposed system configurations that should be accepted in the first chain. The optimization is stopped and a warning is issued if this value is not attained. Defaults to `initial.acceptance = 0.95`.

`initial.temperature`
: Numeric value larger than 0 defining the initial temperature of the system. A low `initial.temperature`, combined with a low `initial.acceptance` result in the algorithm to behave as a greedy algorithm, i.e. only better system configurations are accepted. Defaults to `initial.temperature = 0.001`.

`temperature.decrease`
: Numeric value between 0 and 1 used as a multiplying factor to decrease the temperature at the end of each Markov chain. Defaults to `temperature.decrease = 0.95`.

`chains`
: Integer value defining the maximum number of chains, i.e. the number of cycles of jitters at which the temperature and the size of the neighbourhood should be kept constant. Defaults to `chains = 500`.

`chain.length`
: Integer value defining the length of each Markov chain relative to the number of sample points. Defaults to `chain.length = 1`, i.e. one time the number of sample points.

`stopping`
: Integer value defining the maximum allowable number of Markov chains without improvement of the objective function value. Defaults to `stopping = 10`.

`x.max, x.min, y.max, y.min`
: Numeric value defining the minimum and maximum quantity of random noise to be added to the projected x- and y-coordinates. The units are the same as of the projected x- and y-coordinates. If missing, they are estimated from `candi`, `x.min` and `y.min` being set to zero, and `x.max` and `y.max` being set to half the maximum distance in the x- and y-coordinates, respectively.

`cellsize`
: Vector with two numeric values defining the horizontal (x) and vertical (y) spacing between the candidate locations in `candi`. A single value can be used if the spacing in the x- and y-coordinates is the same.

Value

A list with a set of control parameters of the annealing schedule.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Aarts, E. H. L.; Korst, J. H. M. Boltzmann machines for travelling salesman problems. *European Journal of Operational Research*, v. 39, p. 79-95, 1989.

Černý, V. Thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications*, v. 45, p. 41-51, 1985.

Brus, D. J.; Heuvelink, G. B. M. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, v. 138, p. 86-95, 2007.

Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by simulated annealing. *Science*, v. 220, p. 671-680, 1983.

Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, v. 21, p. 1087-1092, 1953.

van Groenigen, J.-W.; Stein, A. Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*. v. 27, p. 1078-1086, 1998.

Webster, R.; Lark, R. M. *Field sampling for environmental science and management*. London: Routledge, p. 200, 2013.

See Also

optimACDC, optimCORR, optimDIST, optimMKV, optimMSSD, optimPPL, optimSPAN, optimUSER.

Examples

```
schedule <- scheduleSPSANN()
```

spJitter *Random perturbation of spatial points*

Description

Randomly perturb (‘jitter’) the coordinates of spatial points.

Usage

```
spJitter(points, candi, x.max, x.min, y.max, y.min, which.point, cellsize)
```

Arguments

`points` Integer value, integer vector, data frame or matrix. If `points` is an integer value, it defines the number of points that should be randomly sampled from `candi` to form the starting system configuration. If `points` is a vector of integer values, it contains the row indexes of `candi` that correspond to the points that form the starting system configuration. If `points` is a data frame or matrix, it must have three columns in the following order: `[, "id"]` the row indexes of `candi` that correspond to each point, `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates. Note that in the later case, `points` must be a subset of `candi`.

`candi` Data frame or matrix with the candidate locations for the jittered points. `candi` must have two columns in the following order: `[, "x"]` the projected x-coordinates, and `[, "y"]` the projected y-coordinates.

`x.max, x.min, y.max, y.min`
Numeric value defining the minimum and maximum quantity of random noise to be added to the projected x- and y-coordinates. The minimum quantity should be equal to, at least, the minimum distance between two neighbouring candidate locations. The units are the same as of the projected x- and y-coordinates. If missing, they are estimated from `candi`.

`which.point` Integer values defining which point should be perturbed.

`cellsize` Vector with two numeric values defining the horizontal (x) and vertical (y) spacing between the candidate locations in `candi`. A single value can be used if the spacing in the x- and y-coordinates is the same.

Details

Jittering methods: There are multiple mechanism to generate a new sample configuration out of the current sample configuration. The main step consists of randomly perturbing the coordinates of a sample point, a process known as 'jittering'. These mechanisms can be classified based on how the set of candidate locations is defined. For example, one could use an *infinite* set of candidate locations, that is, any location in the sampling region can be selected as the new location of a jittered point. All that is needed is a polygon indicating the boundary of the sampling region. This method is the most computationally demanding because every time a point is jittered, it is necessary to check if the point falls in sampling region.

Another approach consists of using a *finite* set of candidate locations for the jittered points. A finite set of candidate locations is created by discretising the sampling region, that is, creating a fine grid of points that serve as candidate locations for the jittered point. This is the least computationally demanding jittering method because, by definition, the jittered point will always fall in the sampling region.

Using a finite set of candidate locations has two important inconveniences. First, not all locations in the sampling region can be selected as the new location for a jittered point. Second, when a point is jittered, it may be that the new location already is occupied by another point. If this happens, another location has to be iteratively sought for, say, as many times as the number of points in the sample. In general, the more points there are in the sample, the more likely it is that the new location already is occupied by another point. If a solution is not found in a reasonable time, the point selected to be jittered is kept in its original location. Such a procedure clearly is suboptimal.

spsann uses a more elegant method which is based on using a finite set of candidate locations coupled with a form of *two-stage random sampling* as implemented in spsample. Because the candidate locations are placed on a finite regular grid, they can be seen as the centre nodes of a finite set of grid cells (or pixels of a raster image). In the first stage, one of the “grid cells” is selected with replacement, i.e. independently of already being occupied by another sample point. The new location for the point chosen to be jittered is selected within that “grid cell” by simple random sampling. This method guarantees that virtually any location in the sampling region can be selected. It also discards the need to check if the new location already is occupied by another point, speeding up the computations when compared to the first two approaches.

Value

A matrix with the jittered projected coordinates of the points.

Note

The distance between two points is computed as the Euclidean distance between them. This computation assumes that the optimization is operating in the two-dimensional Euclidean space, i.e. the coordinates of the sample points and candidate locations should not be provided as latitude/longitude. **spsann** has no mechanism to check if the coordinates are projected: the user is responsible for making sure that this requirement is attained.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Edzer Pebesma, Jon Skoien with contributions from Olivier Baume, A. Chorti, D.T. Hristopulos, S.J. Melles and G. Spiliopoulos (2013). *intamapInteractive: procedures for automated interpolation - methods only to be used interactively, not included in* intamap *package.* R package version 1.1-10.

van Groenigen, J.-W. *Constrained optimization of spatial sampling: a geostatistical approach.* Wageningen: Wageningen University, p. 148, 1999.

Walvoort, D. J. J.; Brus, D. J.; de Gruijter, J. J. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences*. v. 36, p. 1261-1267, 2010.

See Also

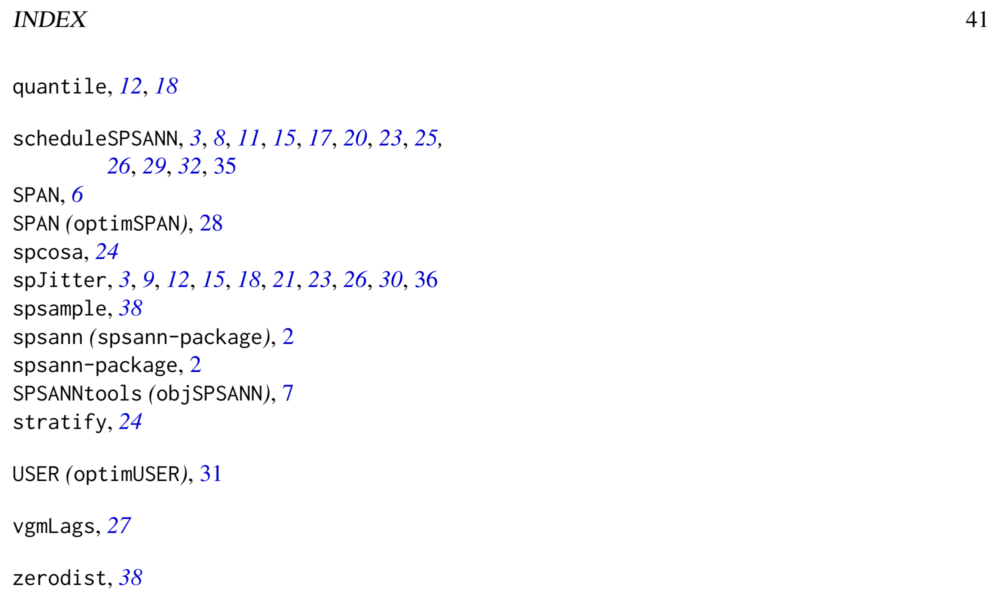
ssaOptim, zerodist, jitter, jitter2d.

Examples

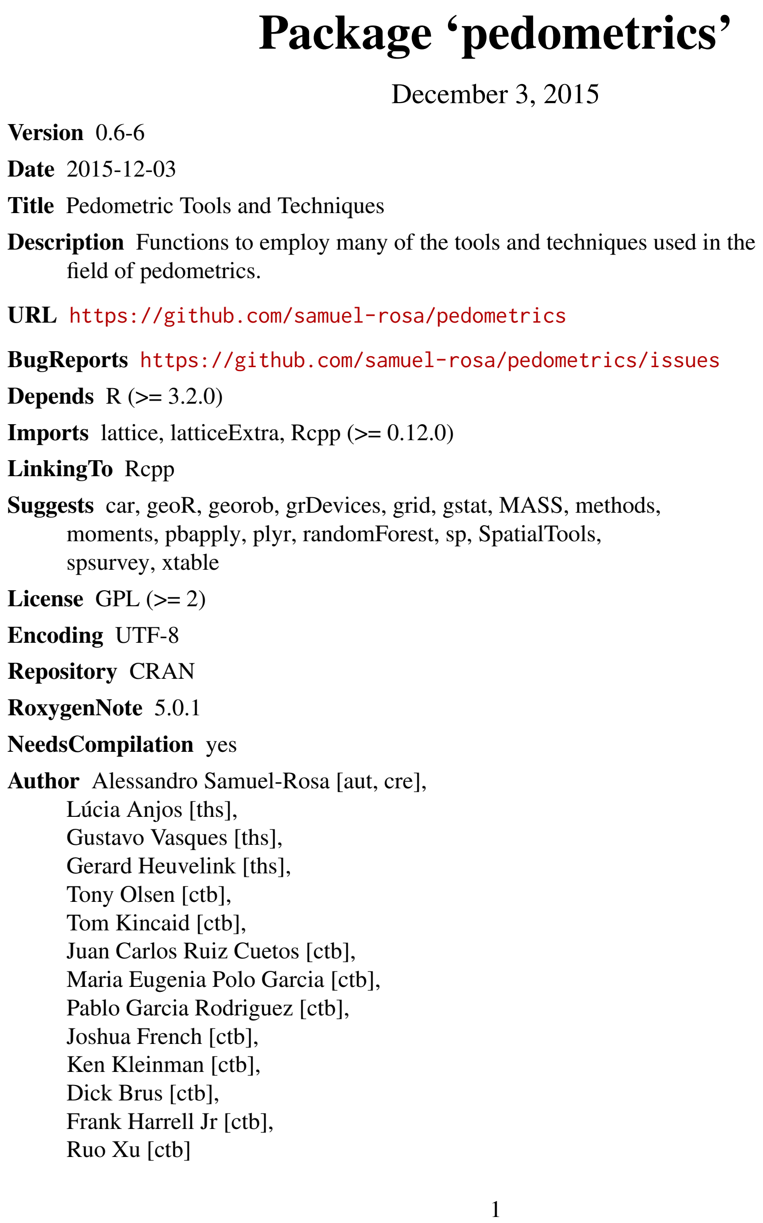
```
require(sp)
data(meuse.grid)
meuse.grid <- as.matrix(meuse.grid[, 1:2])
meuse.grid <- matrix(cbind(1:dim(meuse.grid)[1], meuse.grid), ncol = 3)
pts1 <- sample(c(1:dim(meuse.grid)[1]), 155)
pts2 <- meuse.grid[pts1, ]
```

```
pts3 <- spJitter(points = pts2, candi = meuse.grid, x.min = 40,
                 x.max = 100, y.min = 40, y.max = 100,
                 which.point = 10, cellsize = 40)
plot(meuse.grid[, 2:3], asp = 1, pch = 15, col = "gray")
points(pts2[, 2:3], col = "red", cex = 0.5)
points(pts3[, 2:3], pch = 19, col = "blue", cex = 0.5)

# Cluster of points
pts1 <- c(1:55)
pts2 <- meuse.grid[pts1, ]
pts3 <- spJitter(points = pts2, candi = meuse.grid, x.min = 40,
                x.max = 80, y.min = 40, y.max = 80,
                which.point = 1, cellsize = 40)
plot(meuse.grid[, 2:3], asp = 1, pch = 15, col = "gray")
points(pts2[, 2:3], col = "red", cex = 0.5)
points(pts3[, 2:3], pch = 19, col = "blue", cex = 0.5)
```

Index



APÊNDICE B – R-PACKAGE PEDOMETRICS: PEDOMETRIC TOOLS AND TECHNIQUES



Package ‘pedometrics’

December 3, 2015

Version 0.6-6

Date 2015-12-03

Title Pedometric Tools and Techniques

Description Functions to employ many of the tools and techniques used in the field of pedometrics.

URL `https://github.com/samuel-rosa/pedometrics`

BugReports `https://github.com/samuel-rosa/pedometrics/issues`

Depends R (>= 3.2.0)

Imports lattice, latticeExtra, Rcpp (>= 0.12.0)

LinkingTo Rcpp

Suggests car, geoR, georob, grDevices, grid, gstat, MASS, methods, moments, pbapply, plyr, randomForest, sp, SpatialTools, spsurvey, xtable

License GPL (>= 2)

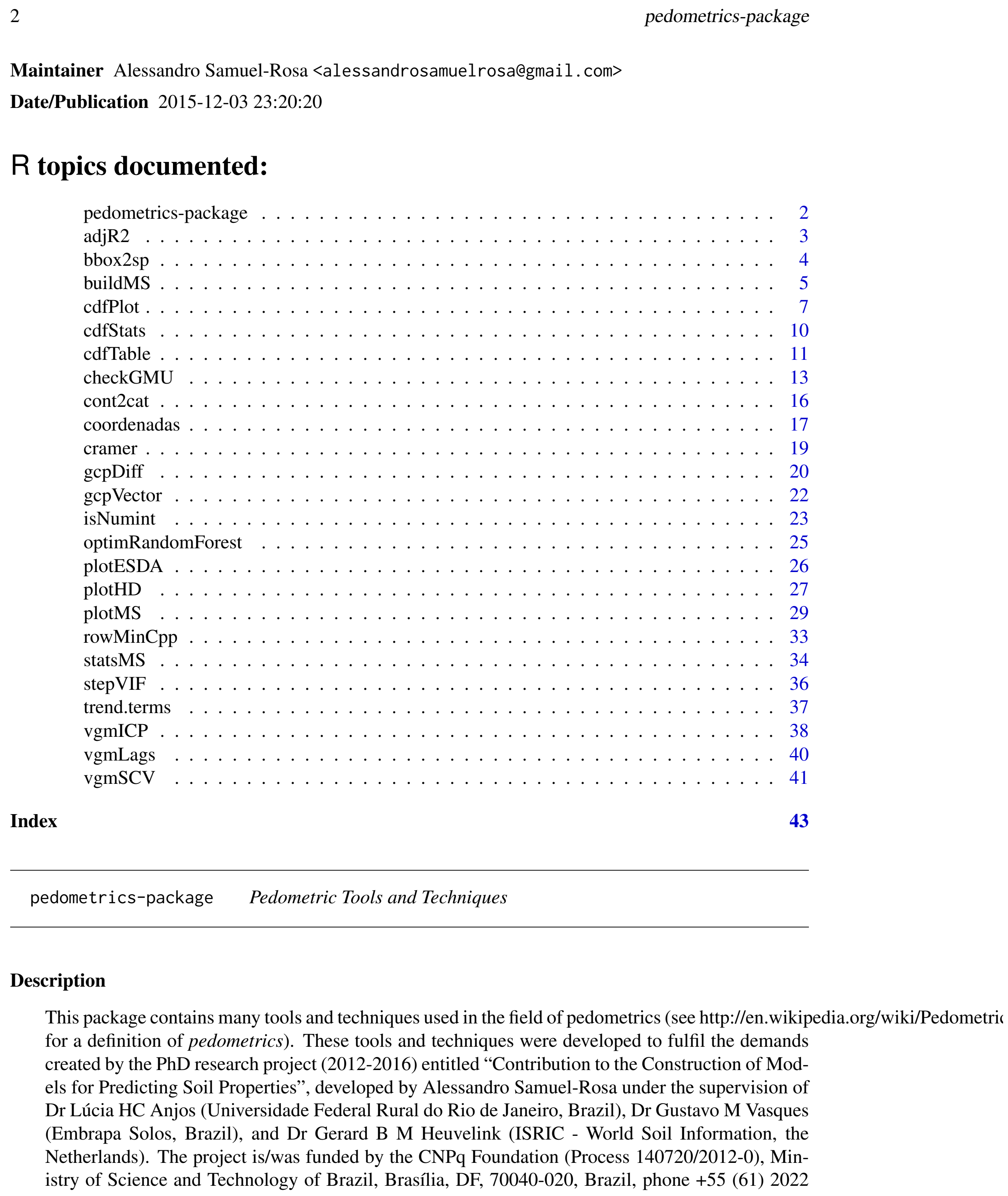
Encoding UTF-8

Repository CRAN

RoxygenNote 5.0.1

NeedsCompilation yes

Author Alessandro Samuel-Rosa [aut, cre],
Lúcia Anjos [ths],
Gustavo Vasques [ths],
Gerard Heuvelink [ths],
Tony Olsen [ctb],
Tom Kincaid [ctb],
Juan Carlos Ruiz Cuetos [ctb],
Maria Eugenia Polo Garcia [ctb],
Pablo Garcia Rodriguez [ctb],
Joshua French [ctb],
Ken Kleinman [ctb],
Dick Brus [ctb],
Frank Harrell Jr [ctb],
Ruo Xu [ctb]



Maintainer Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

Date/Publication 2015-12-03 23:20:20

R topics documented:

`pedometrics-package` *Pedometric Tools and Techniques*

Description

This package contains many tools and techniques used in the field of pedometrics (see http://en.wikipedia.org/wiki/Pedometric for a definition of *pedometrics*). These tools and techniques were developed to fulfil the demands created by the PhD research project (2012-2016) entitled "Contribution to the Construction of Models for Predicting Soil Properties", developed by Alessandro Samuel-Rosa under the supervision of Dr Lúcia HC Anjos (Universidade Federal Rural do Rio de Janeiro, Brazil), Dr Gustavo M Vasques (Embrapa Solos, Brazil), and Dr Gerard B M Heuvelink (ISRIC - World Soil Information, the Netherlands). The project is/was funded by the CNPq Foundation (Process 140720/2012-0), Ministry of Science and Technology of Brazil, Brasília, DF, 70040-020, Brazil, phone +55 (61) 2022

6002, and the CAPES Foundation (Process ID BEX 11677/13-9), Ministry of Education of Brazil, Brasília, DF, 70040-020, Brazil, phone: +55 (61) 2022 6210.

Details

Several functions simply extend the functionalities of other functions commonly used for the analysis of pedometric data. It should be noted that changes are likely to occur quite often and the use of this package as a dependency for other packages is strongly discouraged.

Package:	pedometrics
Type:	Package
Version:	0.6-6
Date:	2015-12-03
License:	GPL (>= 2)

Author(s)

Author and Maintainer: Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

`adjR2` *Adjusted coefficient of determination*

Description

Calculates the adjusted coefficient of determination of a multiple linear regression model.

Usage

```
adjR2(r2, n, p)
```

Arguments

`r2`	Numeric vector with the coefficient of determination to be adjusted.
`n`	Numeric vector providing the number of observations used to fit the multiple linear regression model.
`p`	Numeric vector providing the number of parameters included in the multiple linear regression model.

Details

Details will be added later.

Value

A numeric vector with the adjusted coefficient of determination.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

References

Coefficient of determination. Wikipedia, The Free Encyclopedia. Available at `http://en.wikipedia.org/wiki/Coefficient_of_determination`. [Online; accessed 31-July-2014].

Examples

```
adjR2(r2 = 0.95, n = 100, p = 80)
```

`bbox2sp` *Create Spatial object from a bounding box*

Description

This function takes the bounding box of a Spatial* object and creates a SpatialPoints* or SpatialPolygons* object from it.

Usage

```
bbox2sp(obj, sp = "SpatialPolygons", keep.crs = TRUE)
```

Arguments

`obj`	Object of class Spatial*.
`sp`	Class of the resulting object. Available options are `"SpatialPoints"`, `"SpatialPointsDataFrame"`, `"SpatialPolygons"` and `"SpatialPolygonsDataFrame"`.
`keep.crs`	Logical for assigning the same coordinate reference system to the resulting Spatial* object.

Value

An object of class SpatialPoints* or SpatialPolygons*.

Note

Some of the solutions used to build this function were found in the source code of the R-package **intamapInteractive**. As such, the authors of that package, Edzer Pebesma <<`edzer.pebesma@uni-muenster.de`>> and Jon Skoien <<`jon.skoien@gmail.com`>>, are entitled ‘contributors’ to the R-package **pedometrics**.

Author(s)

Alessandro Samuel-Rosa <<`alessandrosamuelrosa@gmail.com`>>

References

Edzer Pebesma, Jon Skoien with contributions from Olivier Baume, A. Chorti, D.T. Hristopulos, S.J. Melles and G. Spiliopoulos (2013). *intamapInteractive: procedures for automated interpolation - methods only to be used interactively, not included in intamap package.* R package version 1.1-10. http://CRAN.R-project.org/package=intamapInteractive

Examples

```
require(sp)
data(meuse)
coordinates(meuse) <- ~ x + y
bbox2sp(meuse, keep.crs = FALSE)
```

buildMS *Build a series of linear models using automated variable selection*

Description

This function allows building a series of linear models (`lm`) using one or more automated variable selection implemented in function `stepVIF` and `stepAIC`.

Usage

```
buildMS(formula, data, vif = FALSE, vif.threshold = 10,
  vif.verbose = FALSE, aic = FALSE, aic.direction = "both",
  aic.trace = FALSE, aic.steps = 5000, ...)
```

Arguments

`formula` A list containing one or several model formulas (a symbolic description of the model to be fitted).

`data` Data frame containing the variables in the model formulas.

`vif` Logical for performing backward variable selection using the Variance-Inflation Factor (VIF). Defaults to `VIF = FALSE`.

`vif.threshold` Numeric value setting the maximum acceptable VIF value. Defaults to `vif.threshold = 10`.

`vif.verbose` Logical for printing iteration results of backward variable selection using the VIF. Defaults to `vif.verbose = FALSE`.

`aic` Logical for performing variable selection using Akaike Information Criterion (AIC). Defaults to `aic = FALSE`.

`aic.direction` Character string setting the direction of variable selection when using AIC. Available options are `"both"`, `"forward"`, and `"backward"`. Defaults to `aic.direction = "both"`.

`aic.trace` Logical for printing iteration results of variable selection using the AIC. Defaults to `aic.trace = FALSE`.

`aic.steps` Integer value setting the maximum number of steps to be considered for variable selection using the AIC. Defaults to `aic.steps = 5000`.

`...` Further arguments passed to the function `stepAIC`.

Details

This function was devised to deal with a list of linear model formulas. The main objective is to bring together several functions commonly used when building linear models, such as automated variable selection. In the current implementation, variable selection can be done using `stepVIF` or `stepAIC` or both. `stepVIF` is a backward variable selection procedure, while `stepAIC` supports backward, forward, and bidirectional variable selection. For more information about these functions, please visit their respective help pages.

An important feature of `buildMS` is that it records the initial number of candidate predictor variables and observations offered to the model, and adds this information as an attribute to the final selected model. Such feature was included because variable selection procedures result biased linear models (too optimistic), and the effective number of degrees of freedom is close to the number of candidate predictor variables initially offered to the model (Harrell, 2001). With the initial number of candidate predictor variables and observations offered to the model, one can calculate penalized or adjusted measures of model performance. For models built using `builtMS`, this can be done using `statsMS`.

Some important details should be clear when using `buildMS`:

1. this function was originally devised to deal with a list of formulas, but can also be used with a single formula;
2. in the current implementation, `stepVIF` runs before `stepAIC`;
3. function arguments imported from `stepAIC` and `stepVIF` were named as in the original functions, and received a prefix (`aic` or `vif`) to help the user identifying which function is affected by a given argument without having to go check the documentation.

Value

A list containing the fitted linear models.

TODO

Add option to set the order in which `stepAIC` and `stepVIF` are run.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

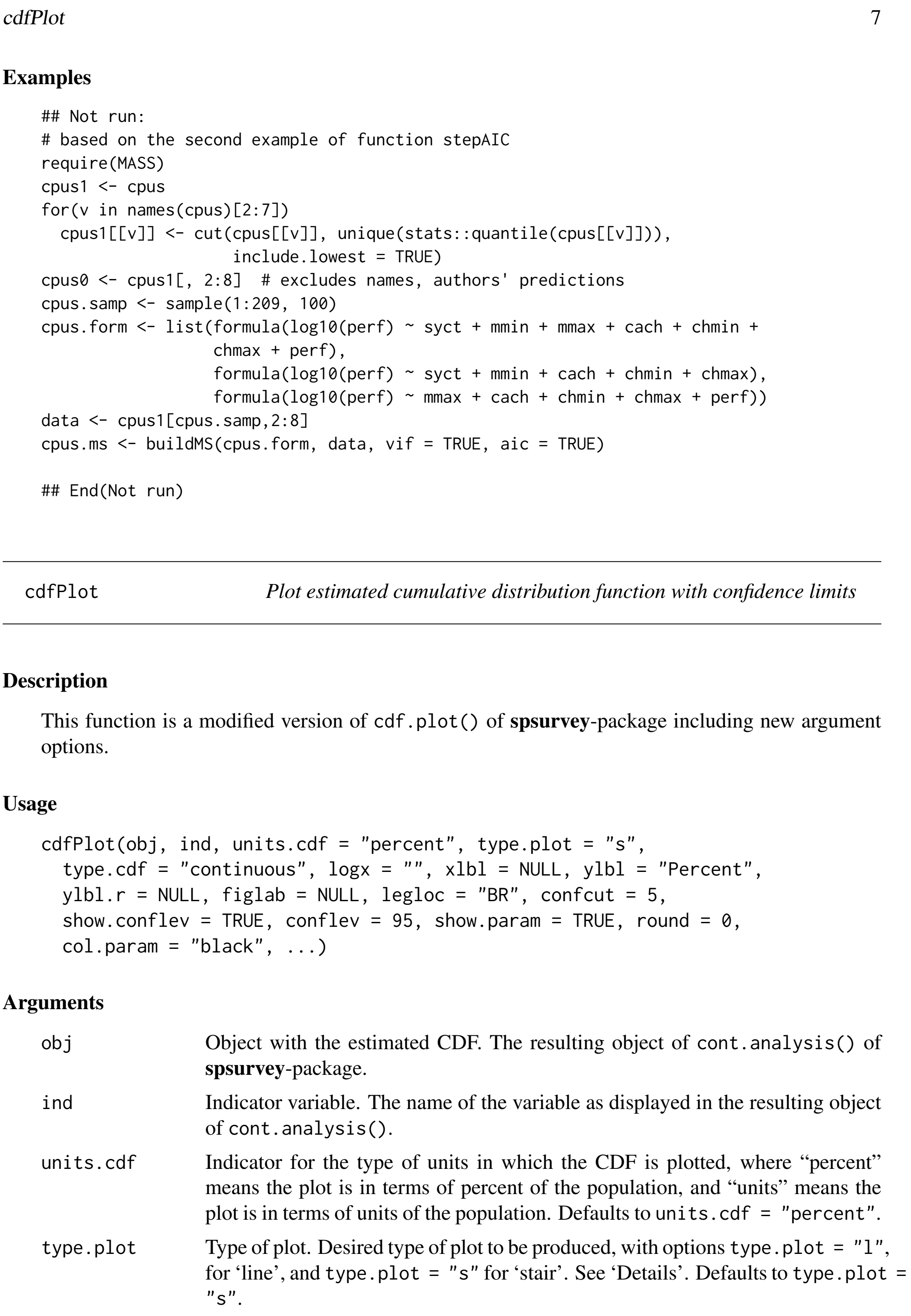
References

Harrell, F. E. (2001) *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis.* First edition. New York: Springer.

Venables, W. N. and Ripley, B. D. (2002) *Modern applied statistics with S.* Fourth edition. New York: Springer.

See Also

`stepAIC`, `stepVIF`, `statsMS`.



Examples

```
## Not run:
# based on the second example of function stepAIC
require(MASS)
cpus1 <- cpus
for(v in names(cpus)[2:7])
  cpus1[[v]] <- cut(cpus[[v]], unique(stats::quantile(cpus[[v]])),
                    include.lowest = TRUE)
cpus0 <- cpus1[, 2:8]  # excludes names, authors' predictions
cpus.samp <- sample(1:209, 100)
cpus.form <- list(formula(log10(perf) ~ syct + mmin + mmax + cach + chmin +
                  chmax + perf),
                  formula(log10(perf) ~ syct + mmin + cach + chmin + chmax),
                  formula(log10(perf) ~ mmax + cach + chmin + chmax + perf))
data <- cpus1[cpus.samp,2:8]
cpus.ms <- buildMS(cpus.form, data, vif = TRUE, aic = TRUE)

## End(Not run)
```

`cdfPlot` *Plot estimated cumulative distribution function with confidence limits*

Description

This function is a modified version of `cdf.plot()` of **spsurvey**-package including new argument options.

Usage

```
cdfPlot(obj, ind, units.cdf = "percent", type.plot = "s",
  type.cdf = "continuous", logx = "", xlbl = NULL, ylbl = "Percent",
  ylbl.r = NULL, figlab = NULL, legloc = "BR", confcut = 5,
  show.conflev = TRUE, conflev = 95, show.param = TRUE, round = 0,
  col.param = "black", ...)
```

Arguments

`obj`	Object with the estimated CDF. The resulting object of `cont.analysis()` of **spsurvey**-package.
`ind`	Indicator variable. The name of the variable as displayed in the resulting object of `cont.analysis()`.
`units.cdf`	Indicator for the type of units in which the CDF is plotted, where "percent" means the plot is in terms of percent of the population, and "units" means the plot is in terms of units of the population. Defaults to `units.cdf = "percent"`.
`type.plot`	Type of plot. Desired type of plot to be produced, with options `type.plot = "l"`, for 'line', and `type.plot = "s"` for 'stair'. See 'Details'. Defaults to `type.plot = "s"`.

`type.cdf` Character string consisting of the value “continuous” or “ordinal” that controls the type of CDF plot for each indicator. Defaults to `type.cdf = "continuous"`.

`logx` Character string consisting of the value `""` or `"x"` that controls whether the x axis uses the original scale (`""`) or the base 10 logarithmic scale (`"x"`). Defaults to `logx = ""`.

`xlbl` Character string providing the x-axis label. If this argument equals `NULL`, then the indicator name is used as the label. Defaults to `xlbl = NULL`.

`ylbl` Character string providing the the y-axis label. Defaults to `ylbl = "Percent"`.

`ylbl.r` Character string providing the label for the right side y-axis, where `ylbl.r = NULL` means a label is not created, and `ylbl.r = "Same"` means the label is the same as the left side label (i.e., argument `ylbl`). Defaults to `ylbl.r = NULL`.

`figlab` Character string providing the plot title. Defaults to `figlab = NULL`.

`legloc` Indicator for location of the plot legend, where `legloc = "BR"` means bottom right, `legloc = "BL"` means bottom left, `legloc = "TR"` means top right, and `legloc = "TL"` means top left. Defaults to `legloc = "BR"`.

`confcut` Numeric value that controls plotting confidence limits at the CDF extremes. Confidence limits for CDF values (percent scale) less than `confcut` or greater than 100 minus `confcut` are not plotted. A value of zero means confidence limits are plotted for the complete range of the CDF. Defaults to `confcut = 5`.

`show.conflev` Logical for showing the confidence limits of the CDF. Defaults to `show.conflev = TRUE`.

`conflev` Numeric value of the confidence level used for confidence limits. Defaults to `conflev = 95`.

`show.param` Logical for showing the parameters of the CDF. Available parameters are the mean, the median, and a percentile defined by the argument `conflev`. The legend displays de actual values of all three parameters, including the standard deviation of the mean. The percentile value is calculated using `spsurvey::interp.cdf()`.

`round` Numeric to set the rounding level of the parameters of the CDF.

`col.param` Color of the lines showing the parameters of the CDF. Defaults to `col.param = "black"`.

`...` Additional arguments passed to `plot()`. See ‘Details’.

Details

Parameter `type.plot` is used only when `type.cdf = "Continuous"`.

Care should be taken with possible conflicts between the arguments of the original function `cdf.plot` and those passed to `plot()` using `...`. The existence of conflicts between these two functions was one of the reasons for creating this new implementation.

Value

A plot of the estimated cumulative distribution function with confidence limits.

Note

Most of the source code that constitutes this function was originaly published in the **spsurvey**-package, version 2.6 (2013-09-20). The authors were asked to include a few new functionalities, but did not seem to be interested in doing so, since no reply was obtained. This implementation is a way of including such functionalities. When using this function, credit should be given to the authors of the original implementation in the **spsurvey**-package.

Author(s)

Tony Olsen <Olsen.Tony@epa.gov>
Tom Kincaid <Kincaid.Tom@epa.gov>
Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Brus, D. J., Kempen, B. and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, v. 62, p. 394-407.

Diaz-Ramos, S., D.L. Stevens, Jr., and A.R. Olsen. (1996). *EMAP Statistical Methods Manual*. EPA/620/R-96/XXX. Corvallis, OR: U.S. Environmental Protection Agency, Office of Research and Development, National Health Effects and Environmental Research Laboratory, Western Ecology Division.

Kincaid, T. M. and Olsen, A. R. (2013) *spsurvey: Spatial Survey Design and Analysis*. R package version 2.6. URL: http://www.epa.gov/nheerl/arm/.

See Also

cdf.plot.

Examples

```
## Not run:
## Estimate the CDF
my.cdf <- spsurvey::cont.analysis(spsurvey.obj = my.spsurvey)

## See indicator levels in the resulting object
levels(my.cdf$Pct$Indicator)

## Plot CDF
cdfPlot(obj = my.cdf, ind = "dz", figlab = "",
   xlbl = "Difference (m)", xlim = c(-30, 10), type.plot = "s")

## End(Not run)
```

`cdfStats` *Descriptive statistics of the cumulative distribution function of a continuous variable*

Description

This function returns summary statistics of the cumulative distribution function of a continuous variable estimated with **spsurvey**-package.

Usage

```
cdfStats(obj, ind, all = TRUE)
```

Arguments

`obj`	Object containing the estimated cumulative distribution function of the continuous variable. The resulting object of `cont.analysis()` of **spsurvey**-package.
`ind`	Indicator variable. The name of the continuous variable as displayed in the resulting object of `cont.analysis()`.
`all`	Summary statistics to be returned. The default option (`all = TRUE`) returns all summary statistics available. If `all = FALSE`, then only estimated population mean and standard deviation are returned. See 'Details'.

Details

The function `cont.analysis()` of **spsurvey**-package estimates the population total, mean, variance, and standard deviation of a continuous variable. It also estimates the standard error and confidence bounds of these population estimates. In some cases it may be interesting to see all estimates, for which one uses `all = TRUE`. However, in other circumstances there might be interest only in taking a look at the estimated population mean and standard deviation. Then the argument `all` has to be set to `FALSE`.

Value

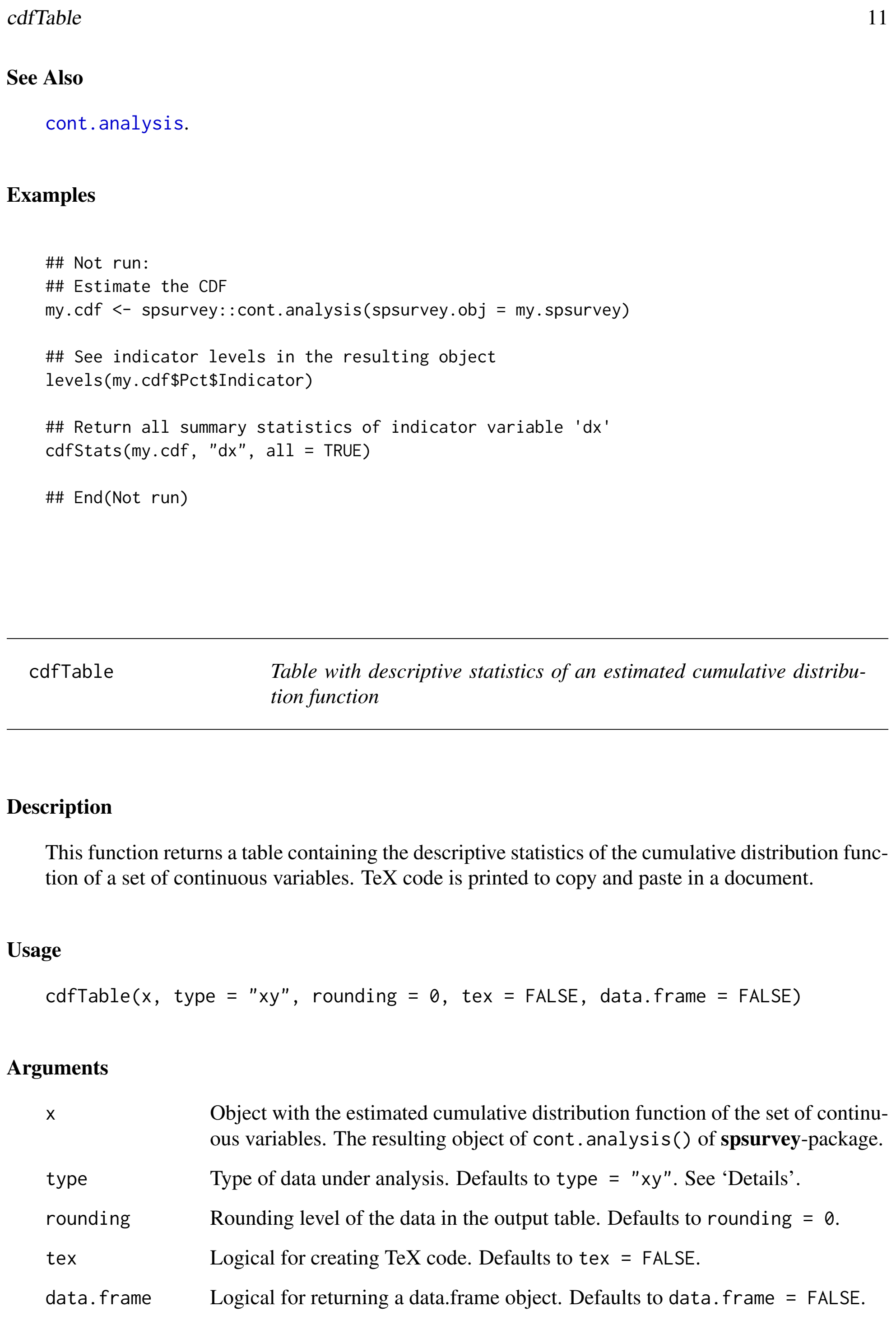
A `data.frame` containing summary statistics of the cumulative distribution function of a continuous variable.

Author(s)

Alessandro Samuel-Rosa <<`alessandrosamuelrosa@gmail.com`>>

References

Kincaid, T. M. and Olsen, A. R. (2013). spsurvey: Spatial Survey Design and Analysis. R package version 2.6. URL: <http://www.epa.gov/nheerl/arm/>.



See Also

cont.analysis.

Examples

```
## Not run:
## Estimate the CDF
my.cdf <- spsurvey::cont.analysis(spsurvey.obj = my.spsurvey)

## See indicator levels in the resulting object
levels(my.cdf$Pct$Indicator)

## Return all summary statistics of indicator variable 'dx'
cdfStats(my.cdf, "dx", all = TRUE)

## End(Not run)
```

---

cdfTable — *Table with descriptive statistics of an estimated cumulative distribution function*

Description

This function returns a table containing the descriptive statistics of the cumulative distribution function of a set of continuous variables. TeX code is printed to copy and paste in a document.

Usage

```
cdfTable(x, type = "xy", rounding = 0, tex = FALSE, data.frame = FALSE)
```

Arguments

`x`	Object with the estimated cumulative distribution function of the set of continuous variables. The resulting object of `cont.analysis()` of **spsurvey**-package.
`type`	Type of data under analysis. Defaults to `type = "xy"`. See 'Details'.
`rounding`	Rounding level of the data in the output table. Defaults to `rounding = 0`.
`tex`	Logical for creating TeX code. Defaults to `tex = FALSE`.
`data.frame`	Logical for returning a data.frame object. Defaults to `data.frame = FALSE`.

Details

Summary statistics included in the table (estimated population mean and standard deviation) are obtained from the resulting object of `cont.analysis()` by internally using the function `cdfStats()`.

There are two types of data that can be submitted to function `cdfTable()`. The first (`type = "xy"`) is composed by two instances ('x' and 'y') and is produced during horizontal (positional) validation exercises (validation in the geographic space). Thus, 'x' and 'y' represent, respectively, the horizontal displacement (error) in 'x' and 'y' coordinates.

The second type of data (`type = "z"`) is composed by only one instance ('z') and is generated by vertical validation exercises (validation in the attribute space). Thus, 'z' represents the vertical displacement (error) of the attribute 'z' being measured.

Value

Returned value depends on how arguments `type` and `tex` are set.

`list("type")` If `type = "xy"`, then the function returns a table with estimated population mean and standard deviation of error statistics for 'x' and 'y' coordinates. These error statistics include the mean error, mean absolute error, and mean square error. It also returns the estimated mean and mean square error vector (module), and the estimated mean azimuth. The number of ground control points used to make the estimates is printed by default.

If `type = "z"`, then the function returns a table with estimated population mean and standard deviation of error statistics for 'z', the attribute under analysis. These error statistics include the mean error, mean absolute error, and mean square error. The number of ground control points used to make the estimates is printed by default.

`list("tex")` If `tex = TRUE`, them the function prints the TeX code for the table defined by the argument `type`. Otherwise the TeX code is not generated.

Author(s)

Alessandro Samuel-Rosa <<`alessandrosamuelrosa@gmail.com`>>

References

Kincaid, T. M. and Olsen, A. R. (2013). spsurvey: Spatial Survey Design and Analysis. R package version 2.6. URL: <`http://www.epa.gov/nheerl/arm/`>.

See Also

`cdfStats`, `cont.analysis`.

Examples

```
## Not run:
## Estimate the CDF
my.cdf <- cont.analysis(spsurvey.obj = my.spsurvey)
```

```
## Print table and TeX code
cdfTable(my.cdf)

## End(Not run)
```

---

checkGMU *Evaluation of geostatistical models of uncertainty*

Description

Evaluate the local quality of a geostatistical model of uncertainty (GMU) using summary measures and graphical displays.

Usage

```
checkGMU(observed, simulated, pi = seq(0.01, 0.99, 0.01), symmetric = TRUE,
  plotit = TRUE)
```

Arguments

`observed`	Vector of observed values at the validation points. See 'Details' for more information.
`simulated`	Data frame or matrix with simulated values (columns) for each validation point (rows). See 'Details' for more information.
`pi`	Vector defining the width of the series of probability intervals. Defaults to `pi = seq(0.01, 0.99, 0.01)`. See 'Details' for more information.
`symmetric`	Logical for choosing the type of probability interval. Defaults to `symmetric = TRUE`. See 'Details' for more information.
`plotit`	Logical for plotting the results. Defaults to `plotit = TRUE`.

Details

There is no standard way of evaluating the local quality of a GMU. The collection of summary measures and graphical displays presented here is far from being comprehensive. A few definitions are given bellow.

Error statistics: Error statistics measure how well the GMU predicts the measured values at the validation points. Four error statistics are presented:

Mean error (ME) Measures the bias of the predictions of the GMU, being defined as the mean of the differences between the average of the simulated values and the observed values, i.e. the average of all simulations is taken as the predicted value.

Mean squared error (MSE) Measures the accuracy of the predictions of the GMU, being defined as the mean of the squared differences between the average of the simulated values and the observed values.

Scaled root mean squared error (SRMSE) Measures how well the GMU estimate of the prediction error variance (PEV) approximates the observed prediction error variance, where the first is given by the variance of the simulated values, while the second is given by the squared differences between the average of the simulated values, i.e. the squared error (SE). The SRMSE is computed as the average of SE / PEV, where SRMSE > 1 indicates underestimation, while SRMSE < 1 indicates overestimation.

Pearson correlation coefficient Measures how close the GMU predictions are to the observed values. A scatter plot of the observed values versus the average of the simulated values can be used to check for possible unwanted outliers and non-linearities. The square of the Pearson correlation coefficient measures the fraction of the overall spread of observed values that is explained by the GMU, that is, the amount of variance explained (AVE), also known as coefficient of determination or ratio of scatter.

Coverage probabilities: The coverage probability of an interval is given by the number of times that that interval contains its parameter over several replications of an experiment. For example, consider the interquartile range $IQR = Q3 - Q1$ of a Gaussian distributed variable with mean equal to zero and variance equal to one. The nominal coverage probability of the IQR is 0.5, i.e. two quarters of the data fall within the IQR. Suppose we generate a Gaussian distributed *random* variable with the same mean and variance and count the number of values that fall within the IQR defined above: about 0.5 of its values will fall within the IQR. If we continue generating Gaussian distributed *random* variables with the same mean and variance, on average, 0.5 of the values will fall in that interval.

Coverage probabilities are very useful to evaluate the local quality of a GMU: the closer the observed coverage probabilities of a sequence of probability intervals (PI) are to the nominal coverage probabilities of those PIs, the better the modelling of the local uncertainty.

Two types of PIs can be used here: symmetric, median-centred PIs, and left-bounded PIs. Papritz & Dubois (1999) recommend using left-bounded PIs because they are better at evidencing deviations for both large and small PIs. The authors also point that the coverage probabilities of the symmetric, median-centred PIs can be read from the coverage probability plots produced using left-bounded PIs.

In both cases, the PIs are computed at each validation location using the quantiles of the conditional cumulative distribution function (ccdf) defined by the set of realizations at that validation location. For a sequence of PIs of increasing width, we check which of them contains the observed value at all validation locations. We then average the results over all validation locations to compute the proportion of PIs (with the same width) that contains the observed value: this gives the coverage probability of the PIs.

Deutsch (1997) proposed three summary measures of the coverage probabilities to assess the local *goodness* of a GMU: accuracy (A), precision (P), and goodness (G). According to Deutsch (1997), a GMU can be considered “good” if it is both accurate and precise. Although easy to compute, these measures seem not to have been explored by many geostatisticians, except for the studies developed by Pierre Goovaerts and his later software implementation (Goovaerts, 2009). Richmond (2001) suggests that they should not be used as the only measures of the local quality of a GMU.

Accuracy An accurate GMU is that for which the proportion p^* of true values falling within the p PI is equal to or larger than the nominal probability p, that is, when $p^* \geq p$. In the coverage probability plot, a GMU will be more accurate when all points are on or above the 1:1 line. The range of A goes from 0 (lest accurate) to 1 (most accurate).

Precision The *precision*, \$P\$, is defined only for an accurate GMU, and measures how close p^* is to \$p\$. The range of \$P\$ goes from 0 (lest precise) to 1 (most precise). Thus, a GMU will be more accurate when all points in the PI-width plot are on or above the 1:1 line.

Goodness The *goodness*, \$G\$, is a measure of the departure of the points from the 1:1 line in the coverage probability plot. \$G\$ ranges from 0 (minimum goodness) to 1 (maximum goodness), the maximum \$G\$ being achieved when $p^* = p$, that is, all points in both coverage probability and interval width plots are exactly on the 1:1 line.

It is worth noting that the coverage probability and PI-width plots are relevant mainly to GMU created using *conditional simulations*, that is, simulations that are locally conditioned to the data observed at the validation locations. Conditioning the simulations locally serves the purposes of honouring the available data and reducing the variance of the output realizations. This is why one would like to find the points falling above the 1:1 line in both coverage probability and PI-width plots. For *unconditional simulations*, that is, simulations that are only globally conditioned to the histogram (and variogram) of the data observed at the validation locations, one would expect to find that, over a large number of simulations, the whole set of possible values (i.e. the global histogram) can be generated at any node of the simulation grid. In other words, it is expected to find all points on the 1:1 line in both coverage probability and PI-width plots. Deviations from the 1:1 line could then be used as evidence of problems in the simulation.

Note

Comments by Pierre Goovaerts <pierre.goovaerts@biomedware.com> were important to describe how to use the coverage probability and PI-width plots when a GMU is created using unconditional simulations.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Deutsch, C. Direct assessment of local accuracy and precision. Baafi, E. Y. & Schofield, N. A. (Eds.) *Geostatistics Wollongong '96*. Dordrecht: Kinwer Academic Publishers, v. I, p. 115-125, 1997.

Papritz, A. & Dubois, J. R. Mapping heavy metals in soil by (non-)linear kriging: an empirical validation. Gómez-Hernández, J.; Soares, A. & Froidevaux, R. (Eds.) *geoENV II – Geostatistics for Environmental Applications*. Springer, p. 429-440, 1999.

Goovaerts, P. Geostatistical modelling of uncertainty in soil science. *Geoderma*. v. 103, p. 3 - 26, 2001.

Goovaerts, P. AUTO-IK: a 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. *Computers & Geosciences*. v. 35, p. 1255-1270, 2009.

Richmond, A. J. Maximum profitability with minimum risk and effort. Xie, H.; Wang, Y. & Jiang, Y. (Eds.) *Proceedings 29th APCOM*. Lisse: A. A. Balkema, p. 45-50, 2001.

Ripley, B. D. *Stochastic simulation*. New York: John Wiley & Sons, p. 237, 1987.

Examples

```
## Not run:
set.seed(2001)
observed <- round(rnorm(100), 3)
simulated <- t(
  sapply(1:length(observed), function (i) round(rnorm(100), 3)))
resa <- checkGMU(observed, simulated, symmetric = T)
resb <- checkGMU(observed, simulated, symmetric = F)
resa$error;resb$error
resa$goodness;resb$goodness

## End(Not run)
```

cont2cat *Stratification and categorization of continuous data*

Description

Compute break points and marginal strata proportions, stratify and convert continuous data (numeric) into categorical data (factor or integer).

Usage

```
cont2cat(x, breaks, integer = FALSE)

breakPoints(x, n, type = "area", prop = FALSE)

stratify(x, n, type = "area", integer = FALSE)
```

Arguments

x	Vector, data frame or matrix; the continuous data to be processed.
breaks	Vector or list; the lower and upper limits that should be used to break the continuous data into categories. See 'Details' for more information.
integer	Logical value; should the categorical data be returned as integers? Defaults to integer = FALSE.
n	Integer value; the number of strata that should be created.
type	Character value; the type of strata, with options "area", for equal-area, and "range", for equal-range strata. Defaults to type = "area".
prop	Logical value; should the marginal strata proportions be returned? Defaults to prop = FALSE.

Details

Breaks must be a vector if x is a vector, but a list if x is a data frame or matrix. Using a list allows breaking the data into a different number of classes.

Value

A vector, data frame, or matrix, depending on the class of x.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

See Also

cut2

Examples

```
## Compute the break points of marginal strata
x <- data.frame(x = round(rnorm(10), 1), y = round(rlnorm(10), 1))
x <- breakPoints(x = x, n = 4, type = "area", prop = TRUE)
x

## Convert continuous data into categorical data
# Matrix
x <- y <- c(1:10)
x <- cbind(x, y)
breaks <- list(c(1, 2, 4, 8, 10), c(1, 5, 10))
y <- cont2cat(x, breaks)
y
# Data frame
x <- y <- c(1:10)
x <- data.frame(x, y)
breaks <- list(c(1, 2, 4, 8, 10), c(1, 5, 10))
y <- cont2cat(x, breaks, integer = TRUE)
y
# Vector
x <- c(1:10)
breaks <- c(1, 2, 4, 8, 10)
y <- cont2cat(x, breaks, integer = TRUE)
y

## Stratification
x <- data.frame(x = round(rlnorm(10), 1), y = round(rnorm(10), 1))
x <- stratify(x = x, n = 4, type = "area", integer = TRUE)
x
```

coordenadas *Prepare object for argument* design *of* spsurvey.analysis()

Description

This function returns an object to feed the argument design when creating an object of class spsurvey.analysis.

Usage

```
coordenadas(x)
```

Arguments

x	Object of class SpatialPointsDataFrame from which site ID and XY coordinates are to be returned.

Details

The argument design used to create object of class spsurvey.analysis requires a series of inputs. However, it can be fed with data about site ID and coordinates. coordenadas() returns a data frame that provides this information, assuming that all other design variables are provided manually in the arguments list.

Value

An object of class data.frame containing three columns with names siteID, xcoord, and ycoord.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Kincaid, T. M. and Olsen, A. R. (2013). spsurvey: Spatial Survey Design and Analysis. R package version 2.6. URL: <http://www.epa.gov/nheerl/arm/>.

See Also

gcpDiff, cont.analysis.

Examples

```
## Not run:
## Create an spsurvey.analysis object
my.spsurvey <-
  spsurvey.analysis(design = coordenadas(my.data),
                    data.cont = delta(ref.data, my.data),
                    popcorrect = TRUE, pcfsize = length(my.data$id),
                    support = rep(1, length(my.data$id)),
                    wgt = rep(1, length(my.data$id)), vartype = "SRS")

## End(Not run)
```

`cramer` *Association between categorical variables*

Description

Compute the Cramer's V, a descriptive statistic that measures the association between categorical variables.

Usage

```
cramer(x)
```

Arguments

`x`	Data frame or matrix with a set of categorical variables.

Details

Any integer variable is internally converted to a factor.

Value

A matrix with the Cramer's V between the categorical variables.

Note

The original code is available at `http://sas-and-r.blogspot.nl/`, Example 8.39: calculating Cramer's V, posted by Ken Kleinman on Friday, June 3, 2011. As such, Ken Kleinman <`<Ken_Kleinman@hms.harvard.edu>`> is entitled a 'contributor' to the R-package **pedometrics**.

The function `bigtabulate` used to compute the chi-squared test is the main bottleneck in the current version of `cramer`. Ideally it will be implemented in C++.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

References

Cramér, H. *Mathematical methods of statistics*. Princeton: Princeton University Press, p. 575, 1946.

Everitt, B. S. *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press, p. 432, 2006.

See Also

`assocstats`

Examples

```
## Not run:
data <- read.csv("http://www.math.smith.edu/r/data/help.csv")
data <- data[, c("female", "homeless", "racegrp")]
str(data)
test <- cramer(data)
test

## End(Not run)
```

gcpDiff *Difference on xyz coordinates between ground control points*

Description

This function estimates the difference, absolute difference, and squared difference on x, y and z coordinates of two sets of ground control points (GCP). It also estimates the module (difference vector), its square and azimuth. The result is a data frame ready to be used to define a object of class spsurvey.object.

Usage

```
gcpDiff(measured, predicted, type = "xy", aggregate = FALSE, rounding = 0)
```

Arguments

measured	Object of class SpatialPointsDataFrame with the reference GCP. A column named ‘siteID’ giving case names is mandatory. See ‘Details’, item ‘Type of data’.
predicted	An object of class SpatialPointsDataFrame with the point data being validated. A column named ‘siteID’ giving case names is mandatory. See ‘Details’, item ‘Type of data’.
type	Type of data under analysis. Defaults to type = "xy". ‘Details’, item ‘Type of data’.
aggregate	Logical for aggregating the data when it comes from cluster sampling. Used only when type = "z". Defaults to aggregate = FALSE. See ‘Details’, item ‘Data aggregation’.
rounding	Rounding level of the data in the output data frame.

Details

Type of data: Two types of validation data that can be submitted to function gcpDiff(): those coming from horizontal (positional) validation exercises (type = "xy"), and those coming from vertical validation exercises (type = "z").

Horizontal (positional) validation exercises compare the position of measured point data with the position of predicted point data. Horizontal displacement (error) is measured in both ‘x’ and 'y' coordinates, and is used to calculate the error vector (module) and its azimuth. Both objects `measured` and `predicted` used with function `gcpDiff()` must be of class `SpatialPointsDataFrame`. They must have at least one column named 'siteID' giving the identification of every case. Matching of case IDs is mandatory. Other columns are discarded.

Vertical validation exercises are interested in comparing the `measured` value of a variable at a given location with that `predicted` by some model. In this case, error statistics are calculated only for the the vertical displacement (error) in the 'z' coordinate. Both objects `measured` and `predicted` used with function `gcpDiff()` must be of class `SpatialPointsDataFrame`. They also must have a column named 'siteID' giving the identification of evary case. Again, matching of case IDs is mandatory. However, both objects must have a column named 'z' which contains the values of the 'z' coordinate. Other columns are discarded.

Data aggregation: Validation is sometimes performed using cluster or transect sampling. Before estimation of error statistics, the data needs to be aggregated by cluster or transect. The function `gcpDiff()` aggregates validation data of `type = "z"` calculating the mean value per cluster. Thus, aggregation can only be properly done if the 'siteID' column of both objects `measured` and `predicted` provides the identification of clusters. Setting `aggregate = TRUE` will return aggregated estimates of error statistics. If the data has been aggregated beforehand, the parameter `aggregate` can be set to `FALSE`.

Case matching: There are circumstances in which the number of cases in the object `measured` is larger than that in the object `predicted`. The function `gcpDiff()` compares the number of cases in both objects and automatically drops those cases of object `measured` that do not match the cases of object `predicted`. However, case matching can only be done if case IDs are exactly the same for both objects. Otherwise, estimated error statistics will have no meaning at all.

Value

An object of class `data.frame` ready to be used to feed the argument `data.cont` when creating a `spsurvey.analysis` object.

Note

Data of `type = "xy"` cannot be submitted to cluster aggregation in the present version.

Author(s)

Alessandro Samuel-Rosa <<`alessandrosamuelrosa@gmail.com`>>

References

Kincaid, T. M. and Olsen, A. R. (2013). spsurvey: Spatial Survey Design and Analysis. R package version 2.6. URL: `http://www.epa.gov/nheerl/arm/`.

See Also

`coordenadas`, `gcpVector`, `spsurvey.analysis`.

Examples

```
## Not run:
## Create an spsurvey.analysis object
my.spsurvey <-
  spsurvey.analysis(design = coordenadas(my.data),
                    data.cont = delta(ref.data, my.data),
                    popcorrect = TRUE, pcfsize = length(my.data$id),
                    support = rep(1, length(my.data$id)),
                    wgt = rep(1, length(my.data$id)), vartype = "SRS")

## End(Not run)
```

gcpVector *Calculate module and azimuth*

Description

This function calculates the module and azimuth of the difference on x and y coordinates between two sets of ground control points (GCP).

Usage

```
gcpVector(dx, dy)
```

Arguments

dx	Numeric vector containing the difference on the 'x' coordinate between two sets of GCP.
dy	Numeric vector containing the difference on the 'y' coordinate between two sets of GCP.

Details

This function is suited to perform calculations for topographical coordinates only. The origin is set in the y coordinate, and rotation performed clockwise.

Value

An object of the class `data.frame` containing the module, its square and azimuth. These three columns are named 'module', 'sq.module' and 'azimuth'.

Note

This function was addapted from `LoadData`.

Author(s)

Juan Carlos Ruiz Cuetos <bilba_t@hotmail.com>
Maria Eugenia Polo Garcia <mepolo@unex.es>
Pablo Garcia Rodriguez <pablogr@unex.es>
Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Ruiz-Cuetos J.C., Polo M.E. and Rodriguez P.G. (2012). *VecStatGraphs2D: Vector analysis using graphical and analytical methods in 2D*. R package version 1.6. http://CRAN.R-project.org/package=VecStatGraphs2D

See Also

LoadData, gcpDiff

Examples

```
## Not run:
gcpVector(dx = rnorm(3, 5, 10), dy = rnorm(3, 5, 10))

## End(Not run)
```

---

isNumint *Tests for data types*

Description

Evaluate the data type contained in an object.

Usage

```
isNumint(x)

allNumint(x)

anyNumint(x)

whichNumint(x)

allInteger(x)

anyInteger(x)

whichInteger(x)
```

```
allFactor(x)

anyFactor(x)

whichFactor(x)

allNumeric(x)

anyNumeric(x)

whichNumeric(x)

uniqueClass(x)
```

Arguments

x	Object to be tested.

Value

TRUE or FALSE depending on whether x contains a given data type.

Author(s)

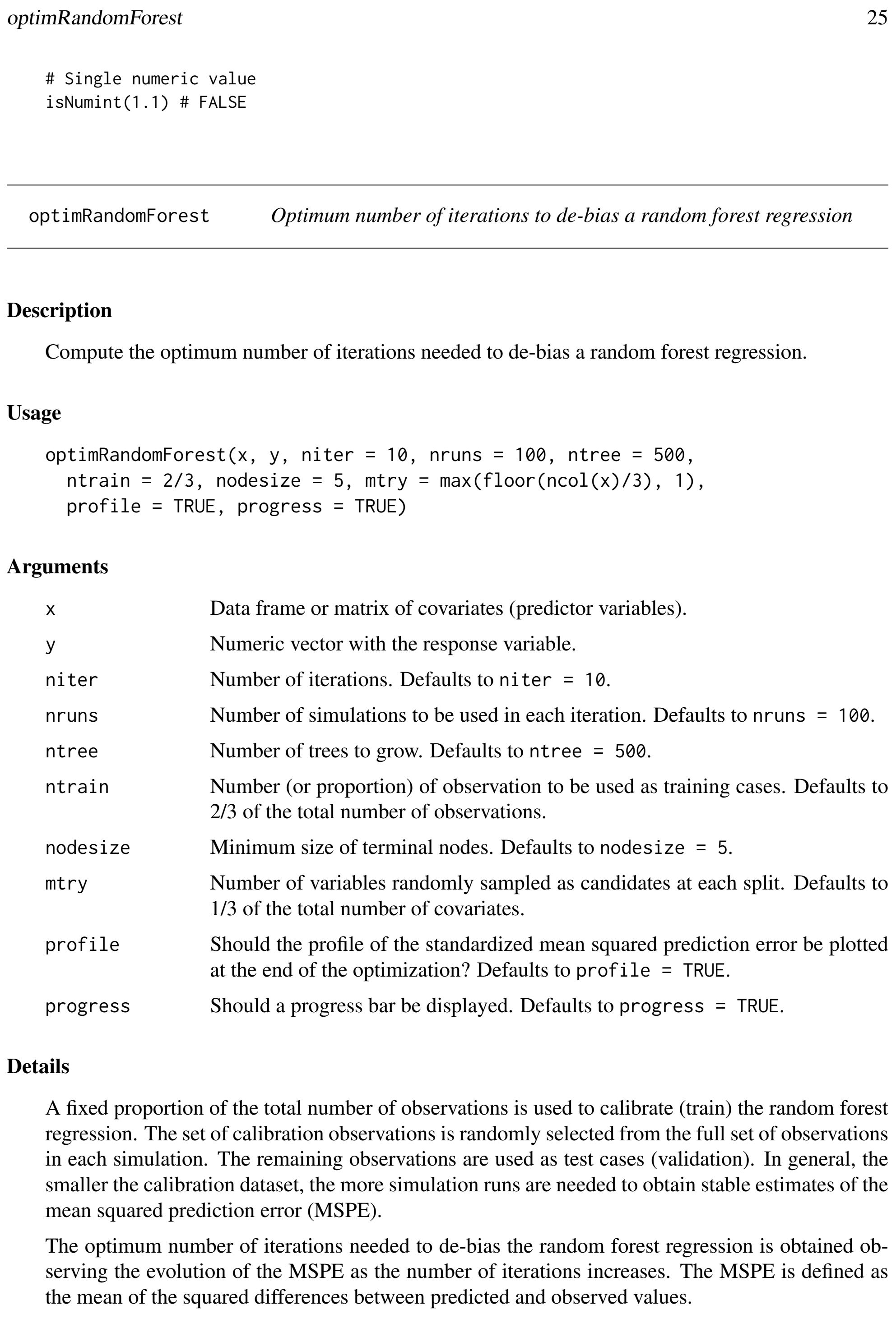
Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

See Also

is.numeric, is.integer, is.factor.

Examples

```
# Vector of integers
x <- 1:10
isNumint(x) # FALSE

# Vector of numeric integers
x <- as.numeric(x)
isNumint(x) # TRUE

# Vector of numeric values
x <- c(1.1, 1, 1, 1, 2)
isNumint(x) # FALSE
allNumint(x) # FALSE
anyNumint(x) # TRUE
whichNumint(x)

# Single numeric integer
isNumint(1) # TRUE
```



```
# Single numeric value
isNumint(1.1) # FALSE
```

optimRandomForest *Optimum number of iterations to de-bias a random forest regression*

Description

Compute the optimum number of iterations needed to de-bias a random forest regression.

Usage

```
optimRandomForest(x, y, niter = 10, nruns = 100, ntree = 500,
  ntrain = 2/3, nodesize = 5, mtry = max(floor(ncol(x)/3), 1),
  profile = TRUE, progress = TRUE)
```

Arguments

`x`	Data frame or matrix of covariates (predictor variables).
`y`	Numeric vector with the response variable.
`niter`	Number of iterations. Defaults to `niter = 10`.
`nruns`	Number of simulations to be used in each iteration. Defaults to `nruns = 100`.
`ntree`	Number of trees to grow. Defaults to `ntree = 500`.
`ntrain`	Number (or proportion) of observation to be used as training cases. Defaults to 2/3 of the total number of observations.
`nodesize`	Minimum size of terminal nodes. Defaults to `nodesize = 5`.
`mtry`	Number of variables randomly sampled as candidates at each split. Defaults to 1/3 of the total number of covariates.
`profile`	Should the profile of the standardized mean squared prediction error be plotted at the end of the optimization? Defaults to `profile = TRUE`.
`progress`	Should a progress bar be displayed. Defaults to `progress = TRUE`.

Details

A fixed proportion of the total number of observations is used to calibrate (train) the random forest regression. The set of calibration observations is randomly selected from the full set of observations in each simulation. The remaining observations are used as test cases (validation). In general, the smaller the calibration dataset, the more simulation runs are needed to obtain stable estimates of the mean squared prediction error (MSPE).

The optimum number of iterations needed to de-bias the random forest regression is obtained observing the evolution of the MSPE as the number of iterations increases. The MSPE is defined as the mean of the squared differences between predicted and observed values.

Note

The original function was published as part of the dissertation of Ruo Xu, which was developed under the supervision of Daniel S Nettleton <dnett@iastate.edu> and Daniel J Nordman <dnordman@iastate.edu>.

Author(s)

Ruo Xu <xuruo.isu@gmail.com>, with improvements by Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Breiman, L. Random forests. *Machine Learning*. v. 45, p. 5-32, 2001.

Breiman, L. *Using adaptive bagging to debias regressions*. Berkeley: University of California, p. 16, 1999.

Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News*. v. 2/3, p. 18-22, 2002.

Xu, R. *Improvements to random forest methodology*. Ames, Iowa: Iowa State University, p. 87, 2013.

See Also

randomForest

plotESDA *Plots for exploratory spatial data analysis (ESDA)*

Description

This function creates four plots for exploratory spatial data analysis (ESDA): histogram + density plot, bubble plot, variogram plot, and variogram map.

Usage

```
plotESDA(z, lat, lon, lags, cutoff, width = c(cutoff/20))
```

Arguments

z	Vector of numeric values of the variable for with ESDA plots should be created.
lat	Vector of numeric values containing the y coordinate (latitude) of the point locations where the z variable was observed.
lon	Vector of numeric values containing the x coordinate (longitude) of the point locations where the z variable was observed.
lags	Numerical vector; upper boundaries of lag-distance classes. See argument boundaries of variogram for more info.

`cutoff` Integer value defining the spatial separation distance up to which point pairs are included in semi-variance estimates. Defaults to the length of the diagonal of the box spanning the data divided by three.

`width` Integer value specifying the width of subsequent distance intervals into which data point pairs are grouped for semi-variance estimates. Defaults to `width = cutoff / 20`.

Details

The user should visit the help pages of `variogram`, `plotHD`, `bubble` and `spplot` to obtain more details about the main functions used to built `plotESDA`.

Value

Four plots: histogram and density plot, bubble plot, empirical variogram, and variogram map.

Author(s)

Alessandro Samuel-Rosa `<alessandrosamuelrosa@gmail.com>`

References

Cressie, N.A.C. (1993) *Statistics for Spatial Data*. New York: John Wiley \& Sons, p.900, 1993.

Pebesma, E.J. (2004) Multivariable geostatistics in S: the gstat package. *Computers \& Geosciences*, 30:683-691, 2004.

Webster, R. \& Oliver, M.A. *Geostatistics for environmental scientists*. Chichester: John Wiley \& Sons, p.315, 2007.

See Also

`variogram`, `plotHD`, `bubble`, `spplot`.

Examples

```
# require(gstat)
# data(meuse)
# plotESDA(z = meuse$zinc, lat = meuse$y, lon = meuse$x)
```

`plotHD` *Histogram and density plot*

Description

This function plots a histogram and a density plot of a single variable using the R-package **lattice**.

Usage

```
plotHD(x, HD = "over", nint = 20, digits = 2, stats = TRUE,
  BoxCox = FALSE, col = c("lightgray", "black"), lwd = c(1, 1),
  lty = "dashed", xlim, ylim, ...)
```

Arguments

`x` Vector of numeric values of the variable for which the histogram nd density plot should be created.

`HD` Character value indicating the type of plot to be created. Available options are `"over"`, to create a histogram superimposed by the theoretical density plot of a normally distributed variable, and `"stack"`, to create a histogram and an empirical density plot in separated panels. Defaults to `HD = "over"`.

`nint` Integer specifying the number of histogram bins. Defaults to `nint = 20`.

`digits` Integer indicating the number of decimal places to be used when printing the statistics of the variable `x`. Defaults to `digits = 2`.

`stats` Logical to indicate if descriptive statistics of the variable `x` should be added to the plot. Available only when `HD = "over"`. The function tries to automatically find the best location to put the descriptive statistics given the shape of the histogram. Defaults to `stats = TRUE`.

`BoxCox` Logical to indicate if the variable `x` should be transformed using the Box-Cox family of power transformations. The estimated lambda value of the Box-Cox transform is printed in the console. It is set to zero when negative. Defaults to `BoxCox = FALSE`.

`col` Vector of two elements, the first indicating the colour of the histogram, the second indicating the colour of the density plot. Defaults to `col = c("lightgray", "black")`.

`lwd` Vector of two elements, the first indicating the line width of the histogram, the second indicating the line width of the density plot. Defaults to `lwd = c(1, 1)`.

`lty` Character value indicating the line type for the density plot. Defaults to `lty = "dashed"`.

`xlim` Vector of two elements defining the limits of the x axis. The function automatically optimizes `xlim` based on the density plot.

`ylim` Vector of two elements defining the limits of the y axis. The function automatically optimizes `ylim` based both histogram and density plot.

`...` Other arguments that can be passed to **lattice** functions. There is no guarantee that they will work.

Details

The user should visit the help pages of `histogram`, `densityplot`, `panel.mathdensity`, `powerTransform` and `bcPower` to obtain more details about the main functions used to built `plotHD`.

Value

An object of class `"trellis"`. The `update.trellis` method can be used to update components of the object and the `print.trellis` print method (usually called by default) will plot it on an appropriate plotting device.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Sarkar, Deepayan (2008) *Lattice: Multivariate Data Visualization with R*, Springer. http://lmdvr.r-forge.r-project.org/

See Also

histogram, densityplot, panel.mathdensity, powerTransform, bcPower.

Examples

```
x <- rnorm(100, 10, 2)
plotHD(x, HD = "stack")
plotHD(x, HD = "over")
```

---

plotMS *Model series plot*

Description

This function produces a graphical output that allows the examination of the effect of using different model specifications (design) on the predictive performance of these models (a model series). It generally is used to access the results of functions buildMS and statsMS, but can be easily adapted to work with any model structure and performance measure.

Usage

```
plotMS(obj, grid, line, ind, type = c("b", "g"), pch = c(20, 2),
  size = 0.5, arrange = "desc", color = NULL, xlim = NULL,
  ylab = NULL, xlab = NULL, at = NULL, ...)
```

Arguments

obj	Object of class data.frame, generally returned by statsMS, containing a 1) series of performance statistics of several models, and 2) the design information of each model. See 'Details' for more information.
grid	Vector of integer values or character strings indicating the columns of the data.frame containing the design data which will be gridded using the function levelplot. See 'Details' for more information.
line	Character string or integer value indicating which of the performance statistics (usually calculated by statsMS) should be plotted using the function xyplot. See 'Details' for more information.

`ind` Integer value indicating for which group of models the mean rank is to be calculated. See ‘Details’ for more information.

`type` Vector of character strings indicating some of the effects to be used when plotting the performance statistics using `xyplot`. Defaults to `type = c("b", "g")`. See `panel.xyplot` for more information on how to set this argument.

`pch` Vector with two integer values specifying the symbols to be used to plot points. The first sets the symbol used to plot the performance statistic, while the second sets the symbol used to plot the mean rank of the indicator set using argument `ind`. Defaults to `pch = c(20, 2)`. See `points` for possible values and their interpretation.

`size` Numeric value specifying the size of the symbols used for plotting the mean rank of the indicator set using argument `ind`. Defaults to `size = 0.5`. See `grid.points` for more information.

`arrange` Character string indicating how the model series should be arranged, which can be in ascending (`asc`) or descending (`desc`) order. Defaults to `arrange = "desc"`. See `arrange` for more information.

`color` Vector defining the colors to be used in the grid produced by function `levelplot`. If `NULL`, defaults to `color = cm.colors(n)`, where `n` is the number of unique values in the columns defined by argument `grid`. See `cm.colors` to see how to use other color palettes.

`xlim` Numeric vector of length 2, giving the x coordinates range. If `NULL` (which is the recommended value), defaults to `xlim = c(0.5, dim(obj)[1] + 0.5)`. This is, so far, the optimum range for adequate plotting.

`ylab` Character vector of length 2, giving the y-axis labels. When `obj` is a `data.frame` returned by `statsMS`, and the performance statistic passed to argument `line` is one of those calculated by `statsMS` (`"candidates"`, `"df"`, `"aic"`, `"rmse"`, `"nrmse"`, `"r2"`, `"adj_r2"` or `"ADJ_r2"`), the function tries to automatically identify the correct `ylab`.

`xlab` Character vector of length 1, giving the x-axis labels. Defaults to `xlab = "Model ranking"`.

`at` Numeric vector indicating the location of tick marks along the x axis (in native coordinates).

`...` Other arguments for plotting, although most of these have no been tested. Argument `asp`, for example, is not effective since the function automatically identifies the best aspect for plotting based on the dimensions of the design data.

Details

This section gives more details about arguments `obj`, `grid`, `line`, `arrange`, and `ind`.

obj: The argument `obj` usually constitutes a `data.frame` returned by `statsMS`. However, the user can use any `data.frame` object as far as it contains the two basic units of information needed:

1. design data passed with argument `grid`
2. performance statistic passed with argument `line`

grid: The argument `grid` indicates the *design* data which is used to produce the grid output in the top of the model series plot. By *design* we mean the data that specify the structure of each model and how they differ from each other. Suppose that eight linear models were fit using three types of predictor variables (`a`, `b`, and `c`). Each of these predictor variables is available in two versions that differ by their accuracy, where `0` means a less accurate predictor variable, while `1` means a more accurate predictor variable. This yields 2^3 = 8 total possible combinations. The *design* data would be of the following form:

```
> design
  a b c
1 0 0 0
2 0 0 1
3 0 1 0
4 1 0 0
5 0 1 1
6 1 0 1
7 1 1 0
8 1 1 1
```

line: The argument `line` corresponds to the performance statistic that is used to arrange the models in ascending or descending order, and to produce the line output in the bottom of the model series plot. For example, it can be a series of values of adjusted coefficient of determination, one for each model:

```
adj_r2 <- c(0.87, 0.74, 0.81, 0.85, 0.54, 0.86, 0.90, 0.89)
```

arrange: The argument `arrange` automatically arranges the model series according to the performance statistics selected with argument `line`. If `obj` is a `data.frame` returned by `statsMS()`, then the function uses standard arranging approaches. For most performance statistics, the models are arranged in descending order. The exception is when `"r2"`, `"adj_r2"` or `"ADJ_r2"` are used, in which case the models are arranged in ascending order. This means that the model with lowest value appears in the leftmost side of the model series plot, while the models with the highest value appears in the rightmost side of the plot.

```
> arrange(obj, adj_r2)
  id a b c adj_r2
1  5 1 0 1  0.54
2  2 0 0 1  0.74
3  3 1 0 0  0.81
4  4 0 1 0  0.85
5  6 0 1 1  0.86
6  1 0 0 0  0.87
7  8 1 1 1  0.89
8  7 1 1 0  0.90
```

This results suggest that the best performing model is that of `id = 7`, while the model of `id = 5` is the poorest one.

ind: The model series plot allows to see how the design influences model performance. This is achieved mainly through the use of different colours in the grid output, where each unique value in the *design* data is represented by a different colour. For the example given above, one could try to see if the models built with the more accurate versions of the predictor variables have a better performance by identifying their relative distribution in the model series plot. The models placed at the rightmost side of the plot are those with the best performance.

The argument `ind` provides another tool to help identifying how the design, more specifically how each variable in the *design* data, influences model performance. This is done by simply calculating the mean ranking of the models that were built using the updated version of each predictor variable. This very same mean ranking is also used to rank the predictor variables and thus identify which of them is the most important.

After arranging the `design` data described above using the adjusted coefficient of determination, the following mean rank is obtained for each predictor variable:

```
> rank_center
     a    b    c
1 5.75 6.25 5.25
```

This result suggests that the best model performance is obtained when using the updated version of the predictor variable `b`. In the model series plot, the predictor variable `b` appears in the top row, while the predictor variable `c` appears in the bottom row.

Value

An object of class `"trellis"` consisting of a model series plot.

Warning

Use the original functions `xyplot` and `levelplot` for higher customization.

Note

Some of the solutions used to build this function were found in the source code of the R-package **mvtsplot**. As such, the author of that package, Roger D. Peng <<`rpeng@jhsph.edu`>>, is entitled 'contributors' to the R-package **pedometrics**.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

References

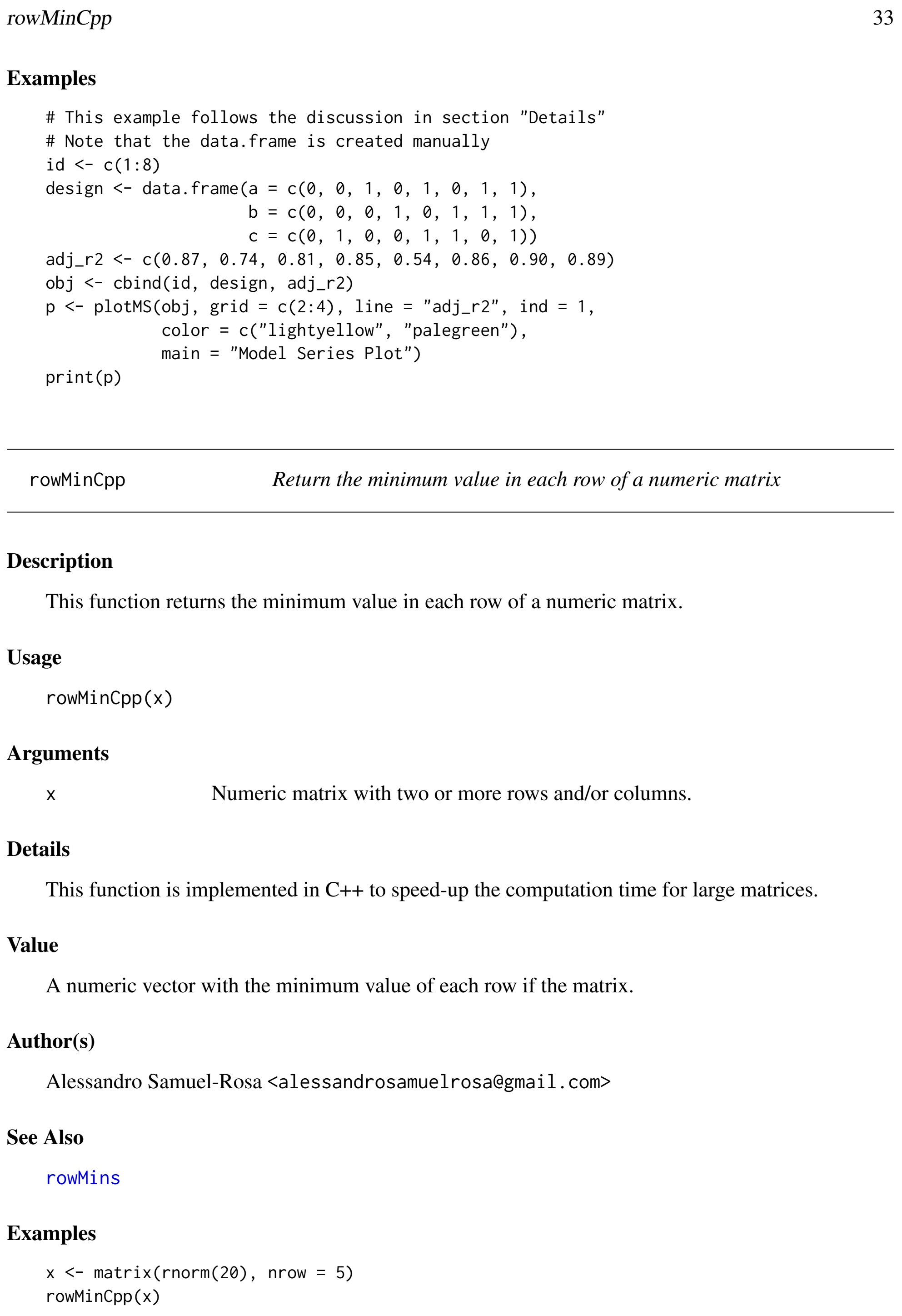
Deepayan Sarkar (2008). *Lattice: Multivariate Data Visualization with R.* Springer, New York. ISBN 978-0-387-75968-5.

Roger D. Peng (2008). *A method for visualizing multivariate time series data.* Journal of Statistical Software. v. 25 (Code Snippet), p. 1-17.

Roger D. Peng (2012). *mvtsplot: Multivariate Time Series Plot.* R package version 1.0-1. `http://CRAN.R-project.org/package=mvtsplot`.

See Also

`levelplot`, `xyplot`, `mvtsplot`.



rowMinCpp 33

Examples

```
# This example follows the discussion in section "Details"
# Note that the data.frame is created manually
id <- c(1:8)
design <- data.frame(a = c(0, 0, 1, 0, 1, 0, 1, 1),
                     b = c(0, 0, 0, 1, 0, 1, 1, 1),
                     c = c(0, 1, 0, 0, 1, 1, 0, 1))
adj_r2 <- c(0.87, 0.74, 0.81, 0.85, 0.54, 0.86, 0.90, 0.89)
obj <- cbind(id, design, adj_r2)
p <- plotMS(obj, grid = c(2:4), line = "adj_r2", ind = 1,
            color = c("lightyellow", "palegreen"),
            main = "Model Series Plot")
print(p)
```

rowMinCpp *Return the minimum value in each row of a numeric matrix*

Description

This function returns the minimum value in each row of a numeric matrix.

Usage

```
rowMinCpp(x)
```

Arguments

x	Numeric matrix with two or more rows and/or columns.

Details

This function is implemented in C++ to speed-up the computation time for large matrices.

Value

A numeric vector with the minimum value of each row if the matrix.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

See Also

rowMins

Examples

```
x <- matrix(rnorm(20), nrow = 5)
rowMinCpp(x)
```

statsMS *Obtain performance statistics of a series of linear models*

Description

This function returns several statistics measuring the performance of a series of linear models built using the function buildMS, with an option to rank the models based on one of the returned performance statistics.

Usage

```
statsMS(model, design.info, arrange.by, digits)
```

Arguments

model	A list of linear models returned by buildMS.
design.info	Extra information about the linear models in the series.
arrange.by	Character string defining if the table with the performance statistics of the linear models should be arranged, and which column should be used. Available options are "candidates", "df", "aic", "rmse", "nrmse", "r2", "adj_r2", and "ADJ_r2". Descending order is used by default and cannot be changed in the current implementation. See 'Value' for more information.
digits	Integer or vector with six integers indicating the number of decimal places to be used to round the performance statistics. If a vector is passed to the function, the number of decimal places should be in the following order: c("aic", "rmse", "nrmse", "r2", "adj_r2", "ADJ_r2").

Details

This function was devised to deal with a list of linear models generated by the function buildMS. The main objective is to compare several linear models using several performance statistics. Such statistics can then be used to rank the linear models and identify, for example, the best performing model, given the selected performance statistics.

An important feature of statsMS is that it uses the information about the initial number of candidate predictor variables offered to the build the model to calculate penalized or adjusted measures of model performance. Such information is recorded as an attribute of the final model selected by buildMS. This feature was included in statsMS because data-driven variable selection results biased linear models (too optimistic), and the effective number of degrees of freedom is close to the number of candidate predictor variables initially offered to the model (Harrell, 2001).

Value

A data frame with several performance statistics:

id Identification of the model.

candidates Number of candidate predictor variables initially offered to the model.

df Number of degrees of freedom of the final selected model.

aic Akaike's Information Criterion (AIC). Obtained using `extractAIC`.

rmse Root-mean squared error, calculated based on the number of candidate predictor variables initially offered to the model.

nrmse Normalized Root-mean squared error, calculated as the ratio between the RMSE and the standard deviation of the observed values of the dependent variable.

r2 Multiple coefficient of determination.

adj_r2 Adjusted multiple coefficient of determination.

ADJ_r2 Adjusted multiple coefficient of determination. Calculations are done based on the number of candidate predictor variables initially offered to the model.

TODO

1. Include other performance statistics such as: PRESS, BIC, Mallow's Cp, max(VIF);
2. Add option to select which performance statistics should be returned.

Author(s)

Alessandro Samuel-Rosa <alessandrosamuelrosa@gmail.com>

References

Harrell, F. E. (2001) *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis.* First edition. New York: Springer.

Venables, W. N. and Ripley, B. D. (2002) *Modern applied statistics with S.* Fourth edition. New York: Springer.

See Also

`buildMS`, `plotMS`.

Examples

```
## Not run:
# based on the second example of function stepAIC
require(MASS)
cpus1 <- cpus
for(v in names(cpus)[2:7])
  cpus1[[v]] <- cut(cpus[[v]], unique(quantile(cpus[[v]])),
                    include.lowest = TRUE)
cpus0 <- cpus1[, 2:8]  # excludes names, authors' predictions
cpus.samp <- sample(1:209, 100)
cpus.form <- list(formula(log10(perf) ~ syct + mmin + mmax + cach + chmin +
                  chmax + perf),
                  formula(log10(perf) ~ syct + mmin + cach + chmin + chmax),
                  formula(log10(perf) ~ mmax + cach + chmin + chmax + perf))
data <- cpus1[cpus.samp,2:8]
cpus.ms <- buildMS(cpus.form, data, vif = TRUE, aic = TRUE)
cpus.des <- data.frame(a = c(0, 1, 0), b = c(1, 0, 1), c = c(1, 1, 0))
stats <- statsMS(cpus.ms, design.info = cpus.des, arrange.by = "aic")

## End(Not run)
```

stepVIF *Variable selection using the variance-inflation factor*

Description

This function takes a linear model and selects the subset of predictor variables that meet a user-specific collinearity threshold measured by the variance-inflation factor (VIF).

Usage

```
stepVIF(model, threshold = 10, verbose = FALSE)
```

Arguments

`model` Linear model (object of class 'lm') containing collinear predictor variables.

`threshold` Positive number defining the maximum allowed VIF. Defaults to `threshold = 10`.

`verbose` Logical for indicating if iteration results should be printed. Defaults to `verbose = FALSE`.

Details

`stepVIF` starts computing the VIF of all predictor variables in the linear model. Because some predictor variables can have more than one degree of freedom, such as categorical variables, generalized variance-inflation factors (Fox and Monette, 1992) are calculated instead using `vif`. Generalized variance-inflation factors (GVIF) consist of VIF corrected to the number of degrees of freedom (df) of the predictor variable:

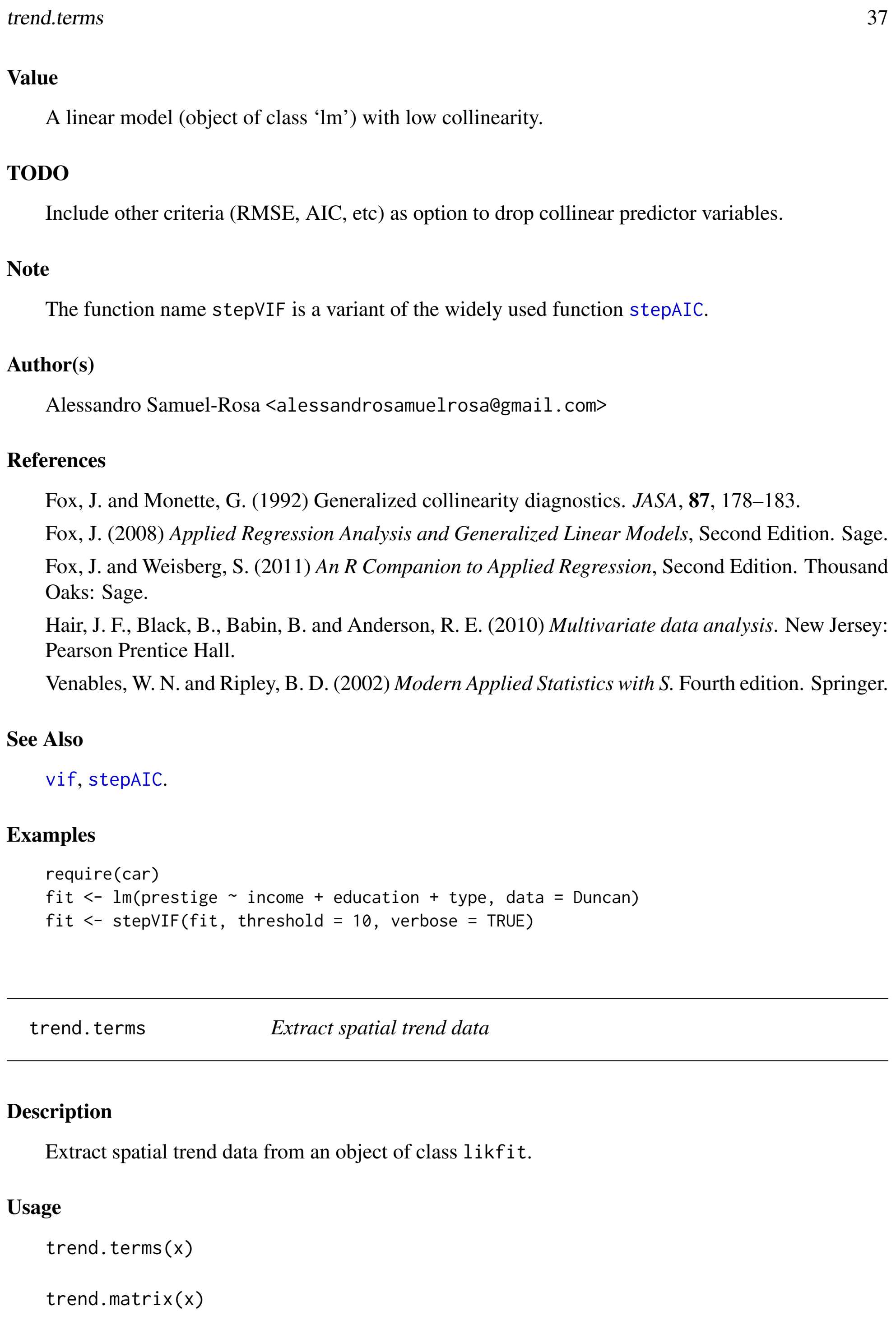
$GVIF = VIF^{1/(2\times df)}$

GVIF are interpretable as the inflation in size of the confidence ellipse or ellipsoid for the coefficients of the predictor variable in comparison with what would be obtained for orthogonal data (Fox and Weisberg, 2011).

The next step is to evaluate if any of the predictor variables has a VIF larger than the specified threshold. Because `stepVIF` estimates GVIF and the threshold corresponds to a VIF value, the last is transformed to the scale of GVIF by taking its square root. If there is only one predictor variable that does not meet the VIF threshold, it is automatically removed from the model and no further processing occurs. When there are two or more predictor variables that do not meet the VIF threshold, `stepVIF` fits a linear model between each of them and the dependent variable. The predictor variable with the lowest adjusted coefficient of determination is dropped from the model and new coefficients are calculated, resulting in a new linear model.

This process lasts until all predictor variables included in the new model meet the VIF threshold.

Nothing is done if all predictor variables have a VIF value inferior to the threshold, and `stepVIF` returns the original linear model.



Value

A linear model (object of class ‘lm’) with low collinearity.

TODO

Include other criteria (RMSE, AIC, etc) as option to drop collinear predictor variables.

Note

The function name `stepVIF` is a variant of the widely used function `stepAIC`.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

References

Fox, J. and Monette, G. (1992) Generalized collinearity diagnostics. *JASA*, **87**, 178–183.

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition. Thousand Oaks: Sage.

Hair, J. F., Black, B., Babin, B. and Anderson, R. E. (2010) *Multivariate data analysis*. New Jersey: Pearson Prentice Hall.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.

See Also

`vif`, `stepAIC`.

Examples

```
require(car)
fit <- lm(prestige ~ income + education + type, data = Duncan)
fit <- stepVIF(fit, threshold = 10, verbose = TRUE)
```

`trend.terms` *Extract spatial trend data*

Description

Extract spatial trend data from an object of class `likfit`.

Usage

```
trend.terms(x)

trend.matrix(x)
```

Arguments

`x`	Object of class `likfit`.

Details

`trend.terms` is similar to `terms`.

`trend.matrix` is similar to `model.frame`.

Author(s)

Alessandro Samuel-Rosa <`alessandrosamuelrosa@gmail.com`>

See Also

`likfit`

vgmICP *Initial covariance parameters (ICP)*

Description

Guess the initial values for the covariance parameters required to fit a variogram model.

Usage

```
vgmICP(z, coords, lags, cutoff = 0.5, method = "a", min.npairs = 30,
  model = "matern", nu = 0.5, estimator = "qn", plotit = FALSE)
```

Arguments

`z`	Numeric vector with the values of the response variable for which the initial values for the covariance parameters should be guessed.
`coords`	Data frame or matrix with the projected x- and y-coordinates.
`lags`	Numeric scalar defining the width of the lag-distance classes, or a numeric vector with the lower and upper bounds of the lag-distance classes. If missing, the lag-distance classes are computed using `vgmLags`. See 'Details' for more information.
`cutoff`	Numeric value defining the fraction of the diagonal of the rectangle that spans the data (bounding box) that should be used to set the maximum distance up to which lag-distance classes should be computed. Defaults to `cutoff = 0.5`, i.e. half the diagonal of the bounding box.
`method`	Character keyword defining the method used for guessing the initial covariance parameters. Defauls to `method = "a"`. See 'Details' for more information.
`min.npairs`	Positive integer defining the minimum number of point-pairs required so that a lag-distance class is used for guessing the initial covariance parameters. Defaults to `min.npairs = 30`.

`model` Character keyword defining the variogram model that will be fitted to the data. Currently, most basic variogram models are accepted. See `cov.spatial` for more information. Defaults to `model = "matern"`.

`nu` numerical value for the additional smoothness parameter ν of the correlation function. See `RMmodel` and argument kappa of `cov.spatial` for more information.

`estimator` Character keyword defining the estimator for computing the sample variogram, with options `"qn"`, `"mad"`, `"matheron"`, and `"ch"`. Defaults to `estimator = "qn"`. See `sample.variogram` for more details.

`plotit` Should the guessed initial covariance parameters be plotted along with the sample variogram? Defaults to `plotit = FALSE`.

Details

There are five methods two guess the initial covariance parameters (ICP). Two of them (`"a"` and `"b"`) rely a sample variogram with exponentially spaced lag-distance classes, while the other three (`"b"`, `"d"`, and `"e"`) use equidistant lag-distance classes (see `vgmLags`). All of them are heuristic.

Method `"a"` was developed in-house, and is the most elaborated of them, specially for guessing the nugget variance. Method `"c"` is implemented in the **automap**-package and was developed by Hiemstra et al. (2009).

Method `"b"` was proposed by Jian et al. (1996) and is implemented in SAS/STAT(R) 9.22. Method `"d"` was developed by Desassis & Renard (2012). Method `"e"` was proposed by Larrondo et al. (2003) and is implemented in the VARFIT module of GSLIB.

Value

A vector of numeric values: the guesses for the covariance parameters nugget, partial sill, and range.

Author(s)

Alessandro Samuel-Rosa <<`alessandrosamuelrosa@gmail.com`>>

References

Desassis, N. & Renard, D. Automatic variogram modelling by iterative least squares: univariate and multivariate cases. *Mathematical Geosciences*. Springer Science + Business Media, v. 45, p. 453-470, 2012.

Hiemstra, P. H.; Pebesma, E. J.; Twenhöfel, C. J. & Heuvelink, G. B. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Computers & Geosciences*. Elsevier BV, v. 35, p. 1711-1721, 2009.

Jian, X.; Olea, R. A. & Yu, Y.-S. Semivariogram modelling by weighted least squares. *Computers & Geosciences*. Elsevier BV, v. 22, p. 387-397, 1996.

Larrondo, P. F.; Neufeld, C. T. & Deutsch, C. V. *VARFIT: a program for semi-automatic variogram modelling*. Edmonton: Department of Civil and Environmental Engineering, University of Alberta, p. 17, 2003.

See Also

vgmLags, sample.variogram, autofitVariogram

Examples

```
data(meuse, package = "sp")
icp <- vgmICP(z = log(meuse$copper), coords = meuse[, 1:2])
```

vgmLags *Lag-distance classes for variogram estimation*

Description

Computation of lag-distance classes for variogram estimation.

Usage

```
vgmLags(coords, n.lags = 7, type = "exp", cutoff = 0.5, base = 2,
  zero = 0.001, count = "pairs")
```

Arguments

coords	Data frame or matrix with the projected x- and y-coordinates.
n.lags	Integer value defining the number of lag-distance classes that should be computed. Defaults to `n = 7`.
type	Character value defining the type of lag-distance classes that should be computed, with options `"equi"` (equidistant) and `"exp"` (exponential). Defaults to `type = "exp"`.
cutoff	Numeric value defining the fraction of the diagonal of the rectangle that spans the data (bounding box) that should be used to set the maximum distance up to which lag-distance classes should be computed. Defaults to `cutoff = 0.5`, i.e. half the diagonal of the bounding box.
base	Numeric value defining the base of the exponential expression used to create exponentially spaced lag-distance classes. Used only when `type = "exp"`. Defaults to `base = 2`, i.e. the width of the rightmost lag-distance classes is equal to half the diagonal of `cutoff`, and so on.
zero	Numeric value setting the minimum pair-wise separation distance that should be used to compute the lag-distance classes. Defaults to `zero = 0.0001`.
count	Should the number of points (`"points"`) or point-pairs (`"pairs"`) per lag-distance class be computed? Defaults to `count = "pairs"`.

Value

Vector of numeric values with the lower and upper boundaries of the lag-distance classes. The number of points or point-pairs per lag-distance class is returned as an attribute.

Author(s)

Alessandro Samuel-Rosa <<alessandrosamuelrosa@gmail.com>>

References

Truong, P. N.; Heuvelink, G. B. M.; Gosling, J. P. Web-based tool for expert elicitation of the variogram. *Computers and Geosciences*. v. 51, p. 390-399, 2013.

See Also

optimPPL

Examples

```
data(meuse, package = "sp")
lags_points <- vgmLags(coords = meuse[, 1:2], count = "points")
lags_pairs <- vgmLags(coords = meuse[, 1:2], count = "pairs")
```

vgmSCV *Spatially correlated variance (SCV)*

Description

Compute the proportion of the variance that is spatially correlated.

Usage

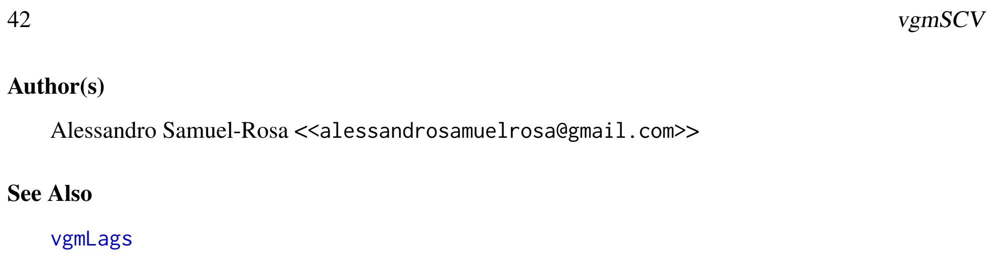
```
## S3 method for class 'variomodel'
vgmSCV(obj, digits = 4)

## S3 method for class 'variogramModel'
vgmSCV(obj, digits = 4)

## S3 method for class 'georob'
vgmSCV(obj, digits = 4)
```

Arguments

obj	Variogram model fitted with available function in geostatistical packages such as **gstat**, **geoR**, and **georob**.
digits	Integer indicating the number of decimal places to be used.

Value

Numeric value indicating the proportion of the variance that is spatially correlated.



Author(s)

Alessandro Samuel-Rosa <<`alessandrosamuelrosa@gmail.com`>>

See Also

`vgmLags`

Index

APÊNDICE C – INTRODUÇÃO GERAL

A modelagem espacial do solo moderna é baseada na utilização de modelos estatísticos para explorar a relação empírica entre as condições ambientais e as propriedades do solo. Estes modelos espaciais do solo, como qualquer outro modelo, não são nada mais do que uma simplificação da realidade. A menos que observemos o solo em todos os lugares – o que destruiria o solo e tornaria as observações inúteis –, não importa quão grande seja o volume de dados, ou quão abrangente o nosso conhecimento for, *nunca* será possível construir um modelo que explique toda a complexidade do solo. Assim, o resultado de um modelo espacial do solo, ou seja, um mapa do solo, *sempre* desviará da "verdade" – esse desvio da "verdade" é o que chamamos de *error*. O que um mapa do solo transmite é o que esperamos que o solo seja, reconhecendo que há *incerteza* sobre ele.

Dado que modeladores espaciais do solo procuram utilizar os recursos disponíveis para produzir a representação mais precisa do solo, um programa de pesquisa sensível é investigar as principais causas para os mapas do solo serem mais ou menos *incertos*. Existem muitas fontes de incerteza na modelagem espacial do solo, tais como os erros que resultam da utilização de um modelo estatístico deficiente ou de fazer interpolações e extrapolações para predizer as propriedades do solo em locais não visitados. Outra importante fonte de incerteza é os dados utilizados para avaliar a relação empírica entre as condições ambientais e as propriedades do solo: dados de covariáveis e solo.

O objetivo geral dessa tese é avaliar importantes fontes de incerteza na modelagem espacial do solo com ênfase em dados de solo e de covariáveis. Este objetivo geral pode ser dividido em objetivos específicos e suas respectivas questões de pesquisa:

(I) Determinar a aptidão de covariáveis disponíveis gratuitamente para calibrar modelos espaciais do solo.

  (a) Será que o uso de covariáveis mais detalhadas resulta em mapas consideravelmente mais precisos do solo?

  (b) Como é que a incorporação de dependência espacial em um modelo espacial do solo se compara ao ganho na acurácia de predição obtido pelo uso de covariáveis mais detalhadas?

  (c) As respostas a essas questões de pesquisa são consistentes entre propriedades do solo?

(II) Identificar os fatores que determinam como modeladores espaciais de campo do solo selecionam os locais de observação do solo.

  (a) Quais fatores são considerados para decidir sobre a localização das observações do solo? Eles têm uma origem pedológica?

  (b) Será que os fatores desempenham o mesmo papel ao longo do curso do processo de observação do solo?

  (c) Pode a análise de padrão pontual ajudar a compreender a estratégia de amostragem intencional tradicionalmente empregada por modeladores espaciais de campo do solo?

(III) Identificar tamanhos e delineamentos amostrais de calibração apropriados para a modelagem espacial do solo.

   (a) Pode o algoritmo de amostragem hipercubo Latino condicionado ser melhorado? Será que essa melhoria resulta em pedições espaciais do solo mais acuradas?

   (b) Quais são os algoritmos de amostragem teoricamente mais sólidos para estimativa da espacial tendência, estimativa do variograma, e predição espacial quando sabemos muito pouco sobre a variação espacial do solo?

   (c) Podem esses algoritmos de amostragem ser usados para construir um algoritmo genérico de amostragem intencional?

A tese é composta por oito capítulos, onde cada um dos objetivos mencionados acima são cumpridos. Embora haja uma sequência lógica na sua apresentação, todos os capítulos foram planejadas para que pudessem ser lidos separadamente. Isto significa que existe uma certa sobreposição entre eles, isto é, informações repetidas. As referências a seções específicas de outros capítulos utilizando hiperlinks coloridos (azul) são comuns.

Capítulo I é uma revisão comentada da literatura sobre modelagem espacial do solo e suas principais fontes de incerteza. A revisão começa com uma discussão sobre os esforços envidados pelos modeladores espaciais do solo para aumentar a conscientização sobre a importância da informação espacial do solo. Esses esforços parecem ter impulsionado uma demanda científica global por informação espacial do solo atualizada e em alta resolução. O capítulo continua com uma descrição da modelagem espacial do solo ao longo da história humana, sugerindo que o objetivo de produzir mapas do solo permanece mais ou menos o mesmo desde a Revolução Neolítica (ca. 10 000 anos). O capítulo termina com as principais fontes de incerteza.

Capítulo II descreve os dados do solo incluídos no *conjunto de dados de Santa Maria*, que foi usado para desenvolver os estudos de caso apresentados nesta tese. O conjunto de dados de Santa Maria é composto de $n = 410$ observações do solo compiladas de estudos realizados entre 2004 e 2013. Esses estudos visavam a geração de mapas semi-detalhados do solo e uso da terra, e a modelagem do estoque de carbono na camada superficial do solo e da vulnerabilidade à erosão. Uma descrição detalhada dos dados de covariáveis incluídos no conjunto de dados de Santa Maria, e seu processamento, é dada no Capítulo III. Capítulo IV apresenta o modelo conceitual de pedogênese (em português), que consiste numa descrição da área de estudo que inclui uma descrição explícita dos fatores de formação do solo (clima, geologia, geomorfologia, hidrologia, uso da terra e vegetação) e processos que determinam a distribuição espaço-temporal do solo. Além de descrever os dados utilizados na tese, o objetivo desses capítulos, juntamente com o modelo conceitual de pedogênese, é fornecer a base para futuros exercícios de modelagem espacial do solo na área de estudo, e servir como exemplo para novas estudos de modelagem espacial do solo desenvolvidos em outros lugares.

Baseado em um artigo publicado na revista avaliada por pares Geoderma, Capítulo V serve o propósito de atingir o primeiro objetivo da tese e responder a suas respectivas questões de pesquisa. O desempenho preditivo de modelos lineares espaciais do solo calibrados utilizando covariáveis (mapas do solo área de classe, mapas de uso da terra, mapas geológicos, modelos digitais de elevação, e imagens de satélite) disponíveis em dois níveis de detalhe é avaliado. A influência de levar em conta a dependência espacial dos resíduos também é avaliada.

Capítulo VI apresenta uma abordagem que visa ajudar a compreender a estratégia de amostragem intencional tradicionalmente empregada por modeladores de campo do solo, ou seja, caminhamento livre. Isso é importante porque muitos projetos de modelagem espacial do solo dependem de dados legados, ou seja, dados do solo produzidos previamente e disponibilizados (publicamente ou não), cujos locais de observação foram selecionados intencionalmente por modeladores espaciais do solo usando regras tácitas mal documentadas. O capítulo foi concebido para responder às questões de pesquisa do segundo objetivo da tese. Análise de padrão pontual é utilizada para caracterizar a configuração espacial da amostra, enquanto teorias emprestadas da Psicologia são usados para elaborar sobre os fatores subjetivos envolvidos na seleção de locais de observação do solo.

O objetivo 3 e suas questões de pesquisa são abordados no Capítulo VII e Capítulo VIII. No Capítulo VII, três algoritmos de amostragem aperfeiçoados são comparados com o algoritmo original de amostragem hipercubo Latino condicionado em como eles afetam a cobertura geográfica, os parâmetros estimados do modelo e a acurácia preditiva. A influência do tamanho da amostra também é discutida. Capítulo VIII apresenta as estratégias de amostragem mais eficientes para estimativa da tendência espacial, estimativa do variograma, e predição espacial quando sabemos muito pouco sobre a variação espacial do solo. O capítulo termina com um novo algoritmo genérico de amostragem que visa os três objetivos conjuntamente.

A sequência de oito capítulos é encerrada com Conclusões Gerais onde destaco os principais resultados da pesquisa e contribuições do estudo. Em seguida, há dois apêndices, ambos dedicados à descrição dos dois pacotes para `R` desenvolvidos para apoiar a tese: `spsann` (Apêndice A) e `pedometria` (Apêndice B). O primeiro foi projetado para a otimização de configurações amostrais usando o recozimento simulado espacial. O segundo inclui funções auxiliares que foram colocadas juntas para facilidade de uso. Todas as referências da literatura são apresentadas sob uma lista única de Referências Bibliográficas no final da tese.

APÊNDICE D – CONCLUSÃO GERAL

Esta tese fez uma contribuição pedológica com o desenvolvimento de uma descrição abrangente dos fatores e processos de formação do solo que determinam a distribuição espaço-temporal das propriedades do solo na área de estudo de Santa Maria. O modelo conceitual de pedogênese, apresentado em Capítulo IV, mostrou que a distribuição espacial das propriedades do solo é muito variável, mesmo quando sob o mesmo uso da terra. Em escalas espaciais maiores, essa variação espacial é determinada pela diversidade geológica e geomorfológica da área, enquanto que em escalas espaciais menores, as práticas agrícolas usadas no passado e no presente parecem desempenhar papel preponderante. Juntamente com o modelo conceitual de pedogênese, Capítulo II e Capítulo III constituem uma contribuição técnica dessa tese. Esses capítulos fornecem a base para exercícios de modelagem espacial do solo na área de estudo.

Capítulo V demonstrou que as covariáveis existentes, gratuitamente disponíveis, são adequadas para calibrar modelos espaciais do solo. Foi demonstrado que a utilização de covariáveis mais detalhadas resulta apenas em pequeno aumento na acurácia da predição de modelos lineares espaciais do solo. O aumento observado é comparável ao efeito da incorporação de dependência espacial no modelo espacial do solo, e pode não compensar os custos adicionais de usar covariáveis mais detalhados. Em geral, uma covariável mais detalhada tem maior potencial de melhorar a acurácia da predição quando uma propriedade do solo é pobremente predita pela sua versão menos detalhada. No entanto, a magnitude da melhoria pode depender das outras covariáveis incluídas no modelo. Decidir se deve-se ou não investir em covariáveis mais detalhados depende da força da relação entre as covariáveis e a propriedade do solo que está sendo modelada, e da diferença relativa entre as versões menos e mais detalhadas das covariáveis. É provavelmente melhor aumentar substancialmente o detalhe de uma covariável menos influente do que aumentar marginalmente o detalhe da covariável mais influente. No entanto, deve-se sempre considerar se meios mais eficientes de aumentar a acurácia da predição existem (por exemplo, a obtenção de mais observações do solo).

Capítulo VI mostrou que vários fatores influenciam o modo como modeladores espaciais de campo do solo escolhem os locais de observação do solo. Esses fatores são de três tipos: conceituais, operacionais e psicológicos. O primeiro diz respeito ao conhecimento dos modeladores espacial do solo sobre as relações solo-paisagem, e parece estar relacionado com os anos de experiência de campo. O segundo refere-se aos recursos disponíveis (infra-estrutura, força de trabalho e do orçamento) para fazer observações do solo, bem como a restrições ao acesso impostas por proprietários de terras e barreiras geográficas, por exemplo. O terceiro refere-se à forma como os modeladores do solo percebem seu ambiente físico circundante e como o curso de sua motivação muda durante o processo de observação do solo. A análise de padrão pontual ajudou a compreender que existe um balanço entre os fatores conceituais e operacionais, o qual determina como a motivação dos modeladores de campo do solo muda o foco para um ou outro objetivo imediato. Dependendo do objetivo focal, a configuração amostral resultante se assemelha a um padrão pontual aleatório (aprendizagem/verificação das relações solo-paisagem – motivação focada nos meios) ou regular (maximizar o número de observações e a cobertura geográfica – motivação focada no resultado).

Capítulo VII mostrou que o algoritmo de amostragem no hipercubo latino condicionado (CLHS), um algoritmo popular usado para otimizar configurações amostrais espaciais para a estimativa da tendência espacial, pode ser consideravelmente aperfeiçoado. Em comparação com o CLHS original, as modificações propostas resultaram em um algoritmo de amostragem com um comportamento numérico superior, mas isso não se traduz necessariamente em maior acurácia da predição. Por exemplo, o tamanho da amostra tem uma influência maior na acurácia da predição do que o algoritmo de amostragem. No entanto, ter em vista a associação/correlação entre covariáveis degrada a acurácia da predição, possivelmente porque a cobertura do espaço geográfico é mais pobre. Como tal, ao otimizar uma configuração amostral para a estimativa de tendência espacial, deveria ser suficiente visar apenas a reprodução da distribuição marginal das covariáveis. Isso deve ser feito usando apenas os estratos marginais de amostragem não vazios.

Capítulo VIII mostrou como otimizar configurações amostrais para estimativa da tendência espacial e do variograma, e interpolação espacial em situações em que sabemos muito pouco sobre a distribuição espacial do solo. A única exigência é que seja formulado um sólido problema de otimização multi-objetivo usando versões robustas de algoritmos de amostragem existentes. A amostra espacial resultante deve reproduzir a distribuição marginal das covariáveis de modo que a tendência espacial possa ser estimada com acurácia. Ela também deve conter vários pequenos aglomerados dispersos por todo o domínio espacial para permitir fazer uma estimativa acurada do comportamento do variograma, especialmente próximo da origem. Finalmente, ela deve cobrir a região de amostragem da forma mais uniforme de tal modo que a média da variância do erro de predição é a menor possível.

Essa tese também contribuiu com dois pacotes para o ambiente de computação estatística e gráficos `R`. O primeiro pacote, chamado `pedometrics` (Apêndice B), contém várias funções para a análise exploratória de dados espaciais e calibração de modelos projetadas para o desenvolvimento dessa tese. O segundo pacote, chamado `spsann` (Apêndice A), contém funções para otimizar configurações amostrais para identificar e estimar o variograma e a tendência espacial, e fazer predições espaciais. O último foi desenvolvido como parte do Capítulo VII e Capítulo VIII. Ambos estão disponíveis gratuitamente e podem ser obtidos no The Comprehensive R Archive Network (CRAN).

No geral, essa tese mostrou que o complexa interação entre o solo e os dados covariáveis pode ter uma grande influência sobre a acurácia dos mapas do solo. Uma receita única, universal, de baixo custo para reduzir a incerteza na modelagem espacial do solo parece fora de alcance. Os estudos de caso sugeriram que soluções são específicas para cada caso e dependem principalmente dos dados do solo e covariáveis existentes. A obtenção de mais amostras do solo mostrou ser uma estratégia eficiente, desde que os recursos disponíveis permitam a amostragem extra. Caso contrário, decidir sobre formas rentáveis de reduzir a incerteza requer, em primeiro lugar, que exploremos todo o potencial dos dados do solo e covariáveis existentes usando técnicas robustas de modelagem espaciais. Tal exercício exige um conhecimento abrangente das relações solo-paisagem, bem como uma minuciosa documentação dos dados do solo e covariáveis para que os seus pontos fracos e fortes possam ser facilmente identificados. Então, a decisão sobre investir na melhoria da qualidade dos dados do solo ou covariáveis ou ambos dependerá do balanço entre o aumento da qualidade dos dados/predição e da quantidade de recursos necessários.