Propositions

1. Decent soil data for land use analyses are impossible to obtain without the expertise of a soil scientist. (this thesis)

2. Digital soil mapping requires an internationally acknowledged standard protocol, similar to the soil survey manual for conventional soil mapping. (this thesis)

3. Yield gaps will decrease in Sub-Saharan Africa when people choose farming as a profession instead of farming as a need to survive.

4. Adoption of climate smart agricultural practices does not keep up with climate change.

5. Field surveys are essential to put the results of a land use analysis in proper context.

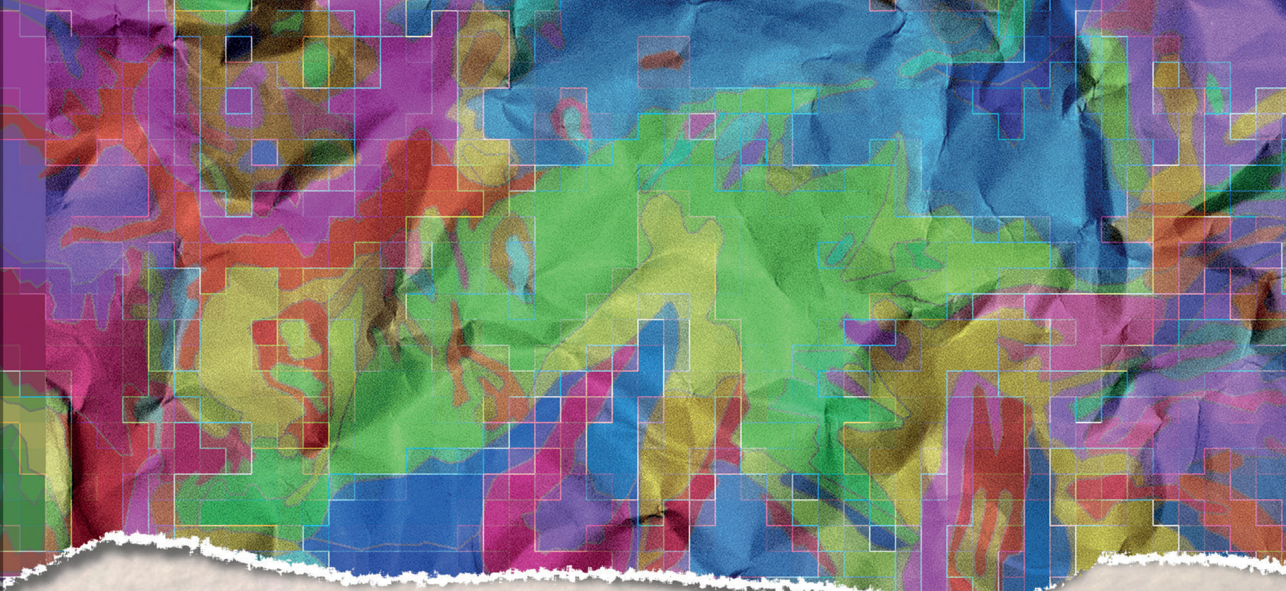
6. The hospitality of stakeholders during research activities should be rewarded by sharing research results.

7. A PhD is like competitive swimming, you can only reach the goal with encouragement and good trainers.

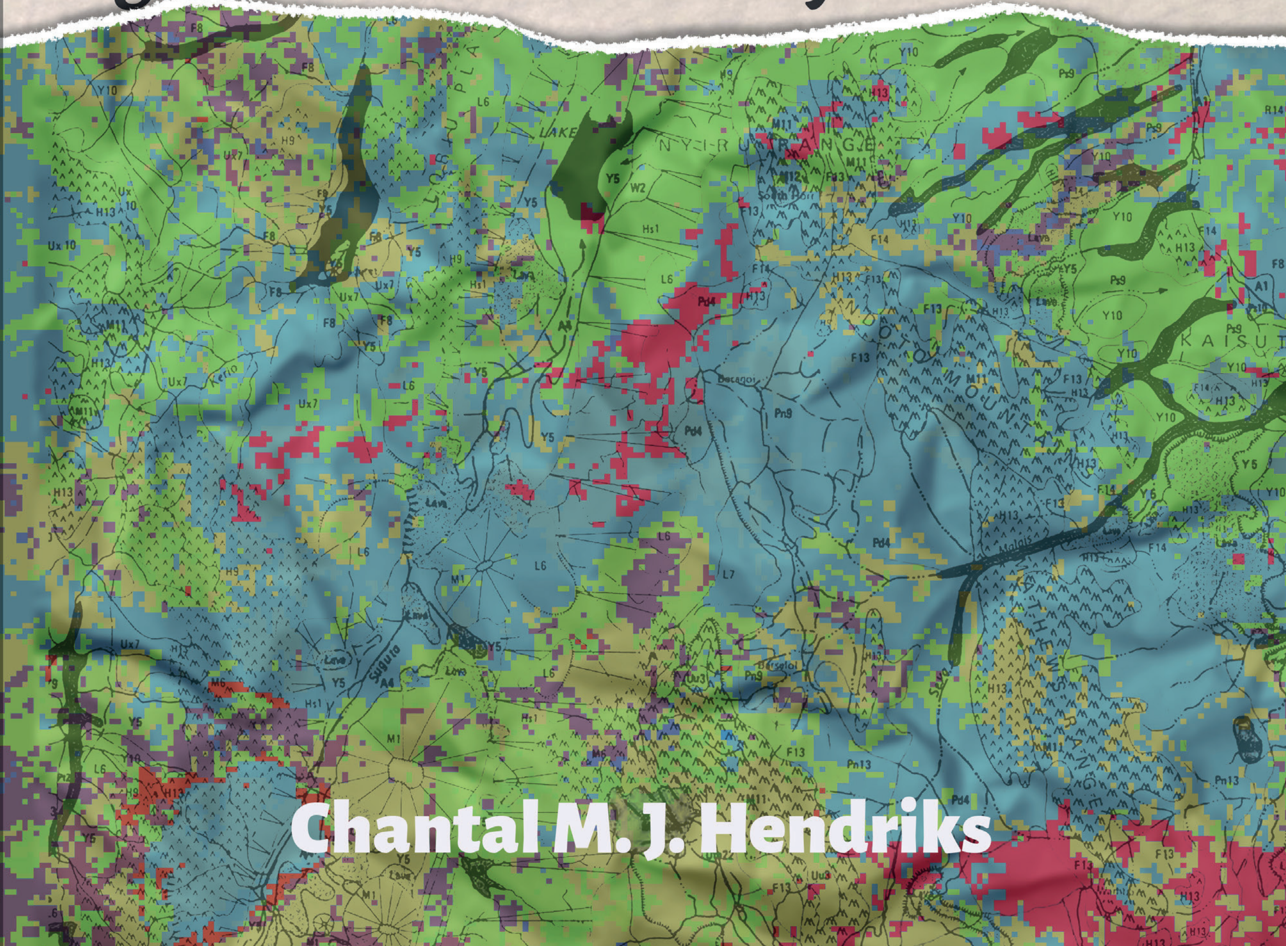
Propositions belonging to the thesis, entitled

'Bridging the gap between the available and required soil data for regional land use analysis'.

Chantal Hendriks
Wageningen, 5 April 2018.



Bridging the gap between the available and required soil data for regional land use analysis



Chantal M. J. Hendriks

Bridging the gap between the available and required soil data for regional land use analysis

Chantal Hendriks

Thesis committee

Promotor

Prof. Dr J. Wallinga
Professor of Soil Geography and Landscape
Wageningen University & Research

Co-promotors

Dr J.J. Stoorvogel
Associate professor, Soil Geography and Landscape Group
Wageningen University & Research

Dr L.F.G. Claessens
Assistant professor, Soil Geography and Landscape Group
Wageningen University & Research

Other members

Prof. Dr M.K. van Ittersum, Wageningen University & Research
Dr B. Jansen, University of Amsterdam
Dr P.T. Gicheru, Kenya Agricultural and Livestock Research Organisation, Embu, Kenya
Dr K. Vancampenhout, KU Leuven, Belgium

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC).

Bridging the gap between the available and required soil data for regional land use analysis

Chantal M.J. Hendriks

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Thursday 05 April 2018
at 11 a.m. in the Aula.

Chantal M. J. Hendriks
Bridging the gap between the available and required soil data for regional land use analysis,
198 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)
With references, with summaries in English and Dutch

ISBN 978-94-6343-737-0
DOI https://doi.org/10.18174/431522

Table of contents

Acknowledgements 7

Chapter 1: Introduction 11

Chapter 2: Exploring the challenges with soil data in regional land use analysis ... 31

Chapter 3: Combining new and available soil data to enrich local to regional land use analyses 67

Chapter 4: A mechanistic model for digital soil mapping to predict the soil organic matter content in nature areas 85

Chapter 5: How to obtain soil data for regional land use analyses? 111

Chapter 6: Synthesis 135

References 155

Summary 179

Samenvatting 185

About the author 193

Acknowledgements

My PhD went by like a roller-coaster ride. At the start I needed to get familiar with the research topic by gaining knowledge from reading literature and talking to other researchers and experts. This took quite some time and effort, but my vision and bird's eye view grew every day. Since the day my first fieldwork campaign started, my PhD has been a furious journey. It was a joyful, rich and sometimes loopy ride. I realize that without the collaboration and support of some persons, this roller-coaster ride would have been much less fun. Therefore, I would like to express my gratefulness to these persons.

First of all, I would like to thank my supervisors Jetse and Lieven. Your enthusiasm and energy for this PhD research worked contagiously. I hope we keep collaborating in the future, because we still have some great ideas to elaborate. Our professional and personal talks and discussions helped me to put things into perspective. Besides, I really enjoyed our chitchats at the coffee machine, at Doppio, in the office, at your homes or in the pub. I have learnt very much from you, perhaps even more than you think. You helped me to become an independent scientist that is able to set up a fieldwork campaign, analyze data, write papers and communicate about research findings. Besides that, I also learnt much about myself and about the way I approach the loops that I had to face during my roller-coaster ride. Besides my supervisors, I would also like to thank my promotor Jakob for all his help, advise and support. Your door was always open and I really appreciated your adequate response when necessary. I am also very grateful that, due to your help, I got the opportunity to get involved in the new strategy and action plan for the Environmental Science Group of Wageningen University & Research.

My fieldwork campaigns took place in Kenya and Spain. I would like to thank the International Crops Research Institute for Semi-Arid Tropics (ICRISAT), the

WorldAgroforestry Centre ICRAF and the Environmental Hydraulics Institute IH Cantabria for providing logistics, transport, housing, soil sample analyses, good drivers and translators. These institutes guaranteed that I could do my job efficiently. A special thanks to Dr. Moses Siambi for hosting me at ICRISAT in Nairobi and Lynette Bwire for helping me with issues during the fieldwork campaign. The first fieldwork campaign went to Machakos and Makueni counties in Kenya. I would like to express my thanks to all District Agricultural Officers, the people working at the KARI weather station in Kambi ya Mawe, the drivers, translators and, most importantly, the farmers. The hospitality and willingness to help succeed this research is highly appreciated. Femke, I can barely express how thankful I am that you joined me during this fieldwork campaign as a master student. You called me a 'boss', but your support and sense of humor made this fieldwork very pleasant and resulted in a very dear friendship. We have experienced many adventures during our fieldwork, which I will never forget. The second fieldwork campaign went to the Lower Nyando basin. Many thanks to Jusper Kiplimo and Wiebke Foerch for sharing data and pictures of the study area and bringing me in contact with persons that were working in the study area. I would like to thank the site manager of ICRAF Kisumu, Georges Aertssen, and ICRAF researcher Joash Mango for their support to find accommodation, translators and a fantastic driver Raymond Otieno. Raymond, you were not just a driver to me. You took the time to understand our research, showed interest in the research topic and translated it in such a way to the farmers that they understood it as well. Besides driving us around, you fixed the car several times, translated the interviews, introduced us to the cultural differences between tribes and digged and augered even as hard as we did. During this fieldwork I felt again very blessed that a great student joint me. Renee, it was a pleasure to do fieldwork with you. You worked very hard and never complained. Thanks for all your positive energy that kept the spirit high during some long days in the field. I would like to thank CropNuts for analyzing all the soil samples of Kenya with high accuracy and precision. For the last fieldwork campaign I went to Spain. I would like to thank Pepe Barquín and José Álvarez Martínez for our close collaboration. I am also very thankful to Sara Alcalde

Aparicio and Ignacio Pérez Silos who conducted most of the fieldwork and who did the laboratory analyses.

Besides the fieldwork campaigns, I spent most of my time at the office in Wageningen. My roomies Maricke, Luc, Kasia, Marcos, Rocky, Rafael and Cindy are the best. You knew exactly when I needed a serious talk, a nonsense talk, a walk, a pingpong table or a good joke. Because of you I had to laugh at least one time a day so loud that everybody in the corridor could enjoy. I also want to thank all my other PhD fellows, my colleagues at the Soil Geography and Landscape group, at ISRIC-World Soil Information, Alterra and the colleagues I met through the Dutch Soil Association.

Last but not least, I would like to thank my family and friends for their personal support and interest. I feel very lucky that I am surrounded by many good friends. You know who you are, so I will keep it short by thanking you for our friendship. A special thanks to Maricke and Femke for being my paranymphs, and to Maarten who helped designing my front cover. Mum, dad, Jolanda and Swen, thank you very much for showing interest in understanding my research and for not getting bored asking about it every time again. Mum, it is impossible to say how worried I was when you became ill during my first fieldwork campaign and how blessed we all are that you recovered. I would like to thank my godmother Til, my godfather John and their families for their interest and sympathy. I also thank my extended family, Marjan, Rob, Teun and Bas, for their heartwarming support and for the many times we could join you for dinner when we were working till late. During my roller-coaster ride, there was always one person that was sitting next to me. Without you Erik, the cabin would have been unstable. I cannot express how thankful I am for your support and love.

Introduction

Chapter 1

1.1 Problem statement

Concerns about future human well-being gained increased attention over recent decades. This resulted in several global initiatives, strategic reports and programs on sustainable development. The UN proposal for 17 Sustainable Development Goals (SDGs) was accepted in 2015 (UN-SDSN, 2014) and gave a clear focus and direction on how to achieve sustainable development. Integrated impact assessment studies are required to help achieving sustainable development (Bond et al., 2001; Keesstra et al., 2016). Regional land use analyses (RLUA), defined in Textbox 1.1, are essential for many of these studies. In the past, RLUA mainly focussed on qualitative land evaluation and land use planning. However, the increased focus on integrated impact assessment studies resulted in more quantitative RLUA that make increasingly use of simulation models. The change in focus had consequences for the required input data, including soil data. Soil data are essential input data for RLUA. In the past, required soil data were obtained from conventional soil surveys and sampling. For current RLUA, conventional soil surveys are often too qualitative and resources to collect new soil data are limited. Consequently, a gap developed between the available soil data (defined in Textbox 1.2) and the required soil data. Soil science can contribute to a wide variety of RLUA that help achieving the SDGs

Textbox 1.1. Definition of regional land use analyses

Regional land use analyses study the assessment of land potential and identify ways to attain these potentials in order to develop adequate and sustainable land use plans for an area or region (Fresco et al., 1990). These plans aim to satisfy changing human needs while maintaining or improving the quality of the environment and conserving natural resources (TAC, 1988), and involve those who are concerned (e.g., population, policy makers, planners, scientists).

Textbox 1.2. Definition of available soil data

Available soil data, i.e., legacy soil data, include all soil data that are available prior to the study.

(Fig.1.1) (Keesstra et al., 2016), but current sustainable development programs often make limited use of soil science expertise (Bouma et al., 2014). To make soil science expertise contribute more effectively in sustainable development programs, the gap needs to be bridged.

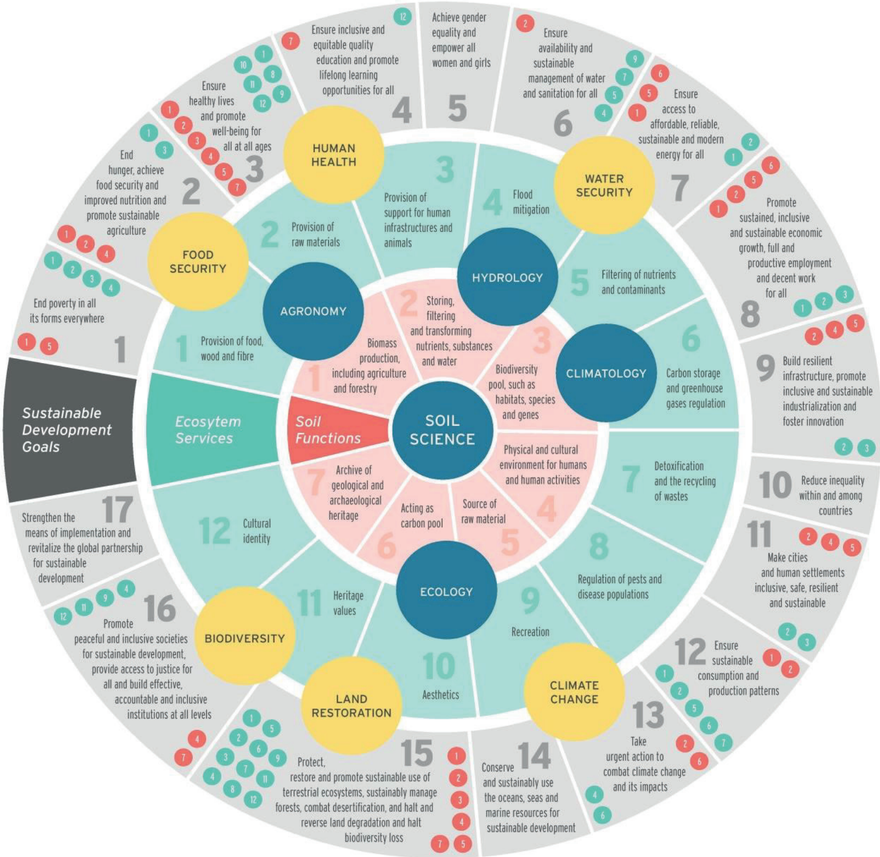


Figure 1.1 *The contribution that soil science can have in achieving the United Nations Sustainable Development Goals (SDGs), from Keesstra et al. (2016).*

In this introduction, a background on land use analysis and available soil data is given. This background helps to understand how the gap could develop. The focus and boundary conditions of this thesis are defined in section 1.5: ‘Scope of the thesis’. In section 1.6 the aim of the research, the research questions, and the hypothesis are formulated. Finally, an ouline of all chapters is given to guide the reader through this thesis.

1.2 Regional land use analysis

1.2.1 Diversity in regional land use analysis

Diversity in RLUA is mainly caused by: (i) pursuing different goals, (ii) including different levels of complexity and computation, and (iii) operating at different spatial and temporal scale.

i. **Pursuing different goals.** The goals RLUA focus on differ and this makes the required soil properties study-specific. Agronomic studies that aim to analyse, for example, plant species, yield gaps, factors that cause yield reduction, the effect of manuring practices and crop growth monitoring, the design of farming systems and regional land use systems (Van Ittersum et al., 2003), require different soil data than e.g., hydrological studies. These studies aim to analyse, for example, the amount of runoff and discharge, water quality, flood and drought risk, effects of water conservation measures, climate change impact, stream flow and stream velocity (Cornelissen et al., 2013). Besides differences in required soil properties, the required level of accuracy differs as well. The required level of accuracy depends on the aim of RLUA. The '4 per 1000 Soils for Food Security and Climate'-initiative requires, for example, a high level of accuracy, because a prospected increase in global soil organic matter stocks of 0.4% needs to be measured. The accuracy that RLUA require has influence on the desired accuracy of the soil data.

ii. **Including different levels of complexity and computation.** The level of complexity and computation of RLUA can be categorized by the diagram of Hoosbeek and Bryant (1992), which was adapted by Bouma and Hoosbeek (1996) (Fig.1.2). The figure illustrates the degree of complexity ranging from empirical to mechanistic and the degree of computation ranging from qualitative to quantitative. Five knowledge levels are distinguished by Bouma and Hoosbeek (1996). Analyses that use user's expertise or expert knowledge are defined as K1 and K2 respectively. These analyses include, for example, qualitative land evaluation and land use planning. Knowledge derived from simple 'black box' models are categorized as K3. This knowledge level includes linear programming techniques and simple modelling (Bouma, 1997). Increasingly, knowledge is obtained from comprehensive models covering entire systems (K4) or very detailed, specialized models covering parts of the system (K5).

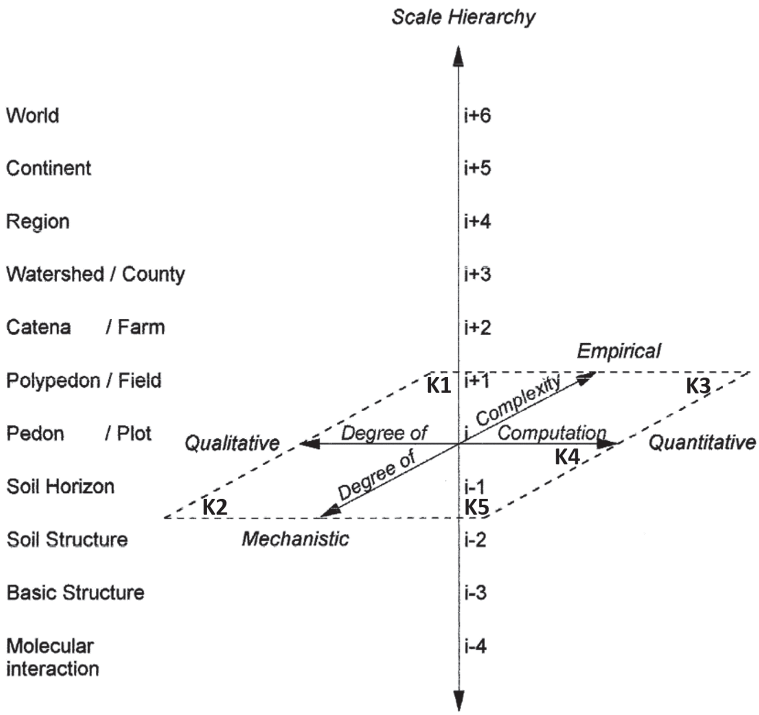


Figure 1.2. *The diagram of Hoosbeek and Bryant (1992) after slight adaptations by Bouma and Hoosbeek (1996). The diagram illustrates the classification of modelling approaches based on hierarchic scale levels, degrees of computation and degrees of complexity. Five knowledge (K) levels are distinguished.*

iii. **Operating at different spatial and temporal scale.** RLUA operate at spatial scales ranging from molecular interaction to global (Fig. 1.2.). Figure 1.2 does not include a time dimension. However, this is implicitly present when modelling at K3, K4 or K5 level (Bouma, 1997). Different processes dominate at different scales and it depends on the modelling approach of a study at which level of detail the spatial and temporal variation need to be described.

1.2.2 Change in focus of regional land use analysis

Over recent decades RLUA changed. These changes caused that available soil data often do not meet the data requirements anymore. In Table 1.1 ten land use analyses were selected. These studies described the limitations that were faced using the selected soil dataset for the land use analysis. Some of these limitations developed over recent decades due to a change in focus of RLUA. A brief background on the change in focus of RLUA will provide insight in the change in soil data requirements.

Land use analyses result from the necessity to evaluate land use and not just soils. Early land use analyses aimed to analyse land management practices and to decide most suitable crops and management practices for a given soil (Soil Science Division Staff, 2017). Soils played a central role in these early land use analyses. The role soils and soil science play in RLUA changed over recent decades. The attention for sustainable development increased, especially after the World Commission on Environment and Development presented a "Global agenda for change" in 1976. A number of policy debates, conferences and reports on sustainable development followed. Most well-known are the United Nations Conference on the Human Environment (UNCHE) in 1972, the Brundtland Report in 1987 (Keeble, 1988), the Nations General Assembly Special Session on Sustainable Development in 1992, the World Summit on Sustainable Development in 2002, the United Nations Conference on Sustainable Development (UNCSD) in 2012, the Sustainable Development Goals (DSGs) presented at the Sustainable Development Summit of 2015 (UN-SDSN, 2014) and the '4 per mille Soils for Food Security and Climate'-initiative signed at the 21st Conference of the Parties (COP21) to the United Nations Framework Convention on Climate Change in 2015 (Minasny et al., 2017). In the past, land use analyses focussed dominantly on what to cultivate where. In this stage, it was important to describe the distribution of soils in the landscape. In a later stage, it became more important to analyse how to practice land use and land management in a sustainable way. For the latter, the understanding of soil genesis became important. Besides soil data, other land characteristics (e.g., slope, rainfall, vegetation), and land qualities (e.g., moisture availability, erosion resistance, nutritive value) were required for land suitability

Table 1.1. *Ten examples of land use analyses studies.*

Source	Contributes to SDG	Aim of the study	Used soil data	Limitations of the soil data
Van Ittersum et al., 2013	2. Zero Hunger	Estimating yield gaps using a bottom-up approach.	AfSIS-GYGA (Leenaars et al., 2015).	Depth weighted average of the soil profile is taken, while the simulation models require multiple soil layers.
Liu et al., 2012	2. Zero Hunger	Analyse the limitations of maize productivity in a semi-arid environment.	Chinese Soil Taxonomy (CSSD, 2015). Physical and chemical soil data collected.	Soil map does not correspond with field data. Only upper soil layer (0-20cm) is sampled, while data on the rooting zone are required.
Hoffmann et al., 2016	2. Zero Hunger	Estimating the effect of soil and climate input data aggregation on regional yield simulations.	Bodenkarte 1:50.000. FIS SToBO database (Geological Service NRW, 2004).	Not all required soil data are available. One representative soil profile for each mapping unit is included for the analysis.
Uda et al., 2017	3. Climate Action	Illustrating potential solutions for sustainable management of Indonesian tropical peatlands.	Indonesia Peatland Map Scale 1:250,000 (Ritung et al., 2011).	Lack of data on peat depth. Maps differ in peatland coverage and distribution.

Source	Contributes to SDG	Aim of the study	Used soil data	Limitations of the soil data
Tiktak et al., 2004	6. Clean water and sanitation	Assess the pesticide leaching risk at Pan-European level.	1:1,000,000 soil map of the European Union with the Soil Profile Analytical Database of Europe (Jamagne et al., 1995).	Austria, Sweden and Finland are not included in the analyses, because there was insufficient soil profile data. The maximum resolution is restricted by the soil map of Europe (at 1:1,000,000 scale). Soil data for preferential flow models (e.g., quantitative soil structure information) are not yet available at the Pan-European scale. Two soil data sources cannot be linked. 25% of the agricultural area cannot be assigned to a soil profile.
Persson and Kværnø, 2017	13. Climate Action	Determine the impact of projected mid-21st century climate on regional spring wheat yields.	1:5000 national soil survey database of the Norwegian Forest and Landscape Institute (2014).	The number of soil profiles is not sufficiently large to result in marked differences between profiles in soil hydraulic characteristics. The required input data are not available.

Source	Contributes to SDG	Aim of the study	Used soil data	Limitations of the soil data
Fraga et al., 2016	13. Climate Action	Modelling climate change impacts on viticultural yield, phenology and stress conditions in Europe.	Harmonized World Soil Database (HWSD) (Fischer et al., 2008).	Variation over depth is described by two soil layers (0-30cm and 30-100 cm). Not all required soil data are available.
Cornelissen et al., 2013	13. Climate Action	Estimating the effect of climate and land use on discharges.	Soil map 1:200,000. Soil properties of Sintondji (2005) and Hiepe (2008).	One representative soil profile for each mapping unit is included for the analysis. Hydraulic conductivity values are a source of uncertainty that produces problems in the correct representation of infiltration characteristics.
Wanders and Wada, 2015	13. Climate Action	Analyse the human and climate impacts on the hydrological drought.	FAO-UNESCO Digital Soil Map of the World (1988).	Data on two soil layers are required (direct runoff and interflow), not on soil horizon. Not all required soil data are available.
Bindraban et al., 2000	15. Life on land	Indicate land qualities for sustainable land management.	FAO-UNESCO Digital Soil Map of the World (1988). Batjes (1997).	Not all required soil data were available.

assessment (FAO, 1985). Different land evaluation reports were published to assess land suitability for, e.g., irrigated agriculture (FAO, 1985), agriculture in the tropics (Verdoodt and Van Ranst, 2003) and forestry (FAO, 1987). In current land use analyses, the impact of certain land use and land management practices on human well-being needs to be studied. This increased the importance of integrated impact assessment studies that use an interdisciplinary approach. The quantification of pedogenetic processes is crucial for the disciplines that are integrated in these assessments (Hoosbeek and Bryant, 1992).

1.3 Available soil data

1.3.1 Diversity in available soil data

There are different types of soil data available: conventional soil surveys, point data that come along with e.g., agronomic experiments, digital soil mapping and remotely sensed soil data. Conventional soil surveys examine, describe, classify and map soils according to standardized surveying (e.g., the Soil Survey Manual of the Soil Science Division Staff, 2017) and classification systems (e.g., the World Reference Base of the FAO, 2015). The soil maps resulting from conventional soil surveys are vector based and describe one or multiple soil types within discrete mapping units that are based on geology, landforms, topography, climate and natural vegetation (FAO, 2015). Each soil type is represented by a soil profile description including chemical and physical analyses. Digital soil mapping (DSM), i.e. predictive soil mapping, predicts soil characteristics spatially exhaustive by deriving relationships between observed soil characteristics and spatially exhaustive auxiliary data (described in Textbox 1.3) that represent the five soil forming factors defined by Jenny (1941); climate, organisms, relief, parent material and time.

Textbox 1.3. Definition of auxiliary data

Auxiliary data are all other data that are required for processing a product which is not part of the primary measurement data. In soil science, there is an increased demand for spatially exhaustive auxiliary data that can be used for soil mapping.

Different models can be used to derive the relationships between observed soil characteristics and auxiliary data. For example, regression models, classification and regression trees, neural networks, fuzzy systems and geo-statistical models. The soil maps resulting from DSM are gridded maps. Studies on DSM operate often at local scale, because DSM relies on soil observations which are costly to collect. For remotely sensed soil data, soil properties are derived from a range of sensing platforms and sensor types. The complexity of soil components and soil spectra makes it difficult to derive soil properties from sensing platforms and sensor types. This type of soil data is therefore not widely available (Ge et al., 2011). Available soil data differ in (i) scale, (ii) availability and (iii) quality:

i. Scale: the scale of vector soil maps (e.g., conventional soil maps) can be divided in detailed soil maps (1:10,000 or 1:25,000), detailed reconnaissance soil maps (1:50,000 to 1:125:000), reconnaissance soil maps (1:125,000 to 1:250,000) and schematic soil maps (1:500,000 or smaller) (Canada Department of Agriculture, 1970). The mapping units of vector soil maps have a minimum size delineation, which is often considered to be 0.5 cm^2 (Soil Survey Division Staff, 2017). Each mapping unit is described by one soil type (in the case of a soil consociation), or more soil types (in the case of a soil complex or soil association) (Soil Science Division Staff, 2017). Often, the type of mapping unit that is chosen depends on the scale of the soil map. About 31% of the global land surface is covered by conventional soil maps at 1:1M scale or finer (Nachtergaele and Van Ranst, 2003). For developed countries more often detailed soil maps are available compared to less developed countries (Nachtergaele and Van Ranst, 2003; Hartemink and Sonneveld, 2013). Gridded soil maps (e.g., digital soil maps) differ in map extent, resolution and support. The resolution of digital soil maps is increasing, because auxiliary data become available at finer resolution. For example, SoilGrids 250m resolution (Hengl et al., 2017) recently replaced SoilGrids 1km resolution (Hengl et al., 2014). However, this does not automatically result in a soil map of higher quality (Samuel-Rosa et al., 2015; Geza and McCray, 2008).

ii. Availability: the soil data density is unequally distributed over the global land surface. For 69% of the global land surface no or only schematic conventional soil maps are available (Nachtergaele and Van Ranst, 2003). Areas that are covered by conventional soil maps are based on one soil observation per 1 cm^2 to 4 cm^2 on the map and each soil type comes along a representative soil profile description (Soil Science Division Staff, 2017). For DSM, the number of soil observations depends on the minimum required accuracy of the map. Different sampling schemes can be used to collect soil data, e.g., simple random sampling, stratified sampling, systematic and grid sampling, ranked set sampling or adaptive cluster sampling (EPA, 2002), depending on the aim of the study and the available auxiliary data. Available soil data are harmonized and stored in databases such as the Harmonized World Soil Database (HWSD) (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) and the WISE Soil Property database (Batjes, 2009).

iii. Quality: the quality of available soil data differs per dataset and is strongly correlated to the scale of the map. Some soil datasets are not accompanied with a quality assessment. The quality of soil maps can be influenced by the used mapping technique, the sampling density, the spatial variation, the quality and resolution of auxiliary data and the quality of the laboratory. Different methods are available to indicate the quality of a soil map. For conventional soil surveys a purity was suggested as quality indicator. The purity gives the maximum percentage of other soil types permissible in a mapping unit (Soil Science Division Staff, 2017). The quality of digital soil maps can be quantified by the spatial cumulative distribution function of the prediction errors (Brus et al., 2011). Digital soil maps can be validated using data-splitting or cross-validation methods, or using additional probability sampling. The prediction errors of data-splitting and cross-validation methods can be spatially auto-correlated (Brus et al., 2011) and therefore additional probability sampling is the most favoured, but also the most costly, validation technique. Digital soil maps that come along a model efficiency only provide the quality of the model fit and not the quality of the digital soil map.

1.3.2 Changes in collecting soil data

The methods used to collect soil data changed over recent decades. The rapid increase in computing technology and the availability of high resolution auxiliary data stimulated this change. Communicating about soils started at the end of the 19th century, when a journal dedicated to soil science was founded (Sibirtsev, 1900). At that time, information on soils was needed to help increase the agricultural production that was required to feed the growing population. Soils were considered as the weathering products of parent material and soil texture was seen as the key property for soil productivity (Arnold, 2016). At the beginning of the 20th century, soils were studied independent of geology and soils became a concept of 'mature soil'. This means that the soil has passed through the major development phases, which were climate and vegetation at that time. Dokuchaev was the first person who described different soil types by explaining differences in the soil forming factors for soil formation (Bockheim et al., 2005). This resulted in the first soil classification system. The concept of Dokuchaev was further elaborated by Jenny (1941), who published the concept of soils as a function of parent material, climate, organisms, relief and time. This concept is still fundamental in soil science and soil classification systems are still used to communicate about soils. There are some global classification systems available, e.g., Keys to Soil Taxonomy (Soil Survey Staff, 2014) and the World Reference Base for soil resources (FAO, 2014), but there are also many national soil classification systems, e.g., Bakker and Schelling, 1966 and Isbell, 2016.

For several decades, conventional soils surveys were collected to obtain soil data that are appropriate for general purpose interpretations (Zinck, 1995). Standardized soil surveying systems warrant a certain consistency among conventional soil surveys. The surveys include soil maps, map unit descriptions, soil profile descriptions, soil classifications, and interpretations for the use and management of soils (Soil Science Division Staff, 2017). When the need for quantitative soil data increased, the soil profile descriptions were compiled and harmonized in soil databases such as the HWSD. Soil profile descriptions were collected at locations that were representative for the soil type. Therefore, the variation within a soil type is often unknown. Besides that, the exact location of a soil type within a mapping unit is unknown when the mapping unit consists of more than one soil type, because each soil type within a mapping unit is presented as a proportion. When the use of simulation models for land use analyses increased, the need for quantitative, spatially exhaustive soil profile data increased. There was a need for new soil data that better met the data requirements for these simulation models. The development of new mapping techniques was stimulated by the availability of mapping tools such as geographic information systems (GIS), GPS, and remote and proximal sensors. However, many studies that use these new mapping techniques only provide soil data of the topsoil. The need for soil profile data resulted in the development of complex digital soil mapping techniques. For example, three-dimensional (3-D) DSM was introduced to include variation over depth (e.g. Kempen et al., 2011; Gasch et al., 2015). Another example is the use of Structural Equation Modelling in DSM (e.g. Angelini et al., 2016), where relationships between soil properties are integrated in the model. Artificial neural networks can be used in DSM as well. It selects a large number of soil properties in the artificial neural networks and only the most relevant soil variables are selected by machine-learning. However, these DSM techniques require a large number of soil observations and the quality of the resulting soil maps is often poor (e.g. Kempen et al., 2011; Angelini et al., 2016).

1.4 Gap between the available and required soil data

There are two main reasons that make the gap between available and required soil data for regional land use analysis difficult to trace. The first reason is that the change in focus and the wide diversity in RLUA make it difficult to decide which soil data to collect, because the gap differs per RLUA. The second reason is that available soil data often do not meet the soil data requirements and the resources to collect new soil data are limited. Limitations of available soil data that are most frequently mentioned by studies on land use analyses are: lack in required soil properties, lack in quantitative soil data and lack in data on spatial soil variability (Table 1.1). Issues with soil data covering large extents and soil-cover complexity, and the representation of short-distance variability carried over from conventional soil surveys to DSM (Lagacherie and McBratney, 2006).

1.5 Scope of the thesis

The gap between the available and required soil data differs per RLUA and therefore many different solutions on bridging the gap can be developed. It is therefore essential to define the focus and boundary conditions of this thesis. This thesis focusses on providing a series of solutions on bridging the gap. Therefore, the thesis includes different case studies that focus on different aims and study areas. Each case study is linked to an ongoing land use analysis, which makes the process to identifying the gap and searching for solutions interactive. The solutions are placed in a broader context, so they can be used for other RLUA as well. The case studies focus on agronomy and the regional scale, because solutions for bridging the gap are of major importance in these studies. At regional scale, available soil data are often too coarse and digital soil maps are not available or of poor quality (e.g., Kempen et al., 2011; Angelini et al., 2016). Soil data at regional scale are important to answer questions for regional or national policy and at this scale interdisciplinary approaches that involve people and institutions are required (Stoorvogel and Antle, 2001).

The thesis is elaborated in collaboration with the Consultative Group on International Agricultural Research (CGIAR) research program on Climate Change, Agriculture and Food Security (CCAFS). This research program operates at different scales and in different regions across the globe. The projects of CCAFS contribute to achieving the SDGs and are therefore interesting to select for this research.

1.6 Bridging the gap

The gap between the available and required soil data need to be bridged to better meet the soil data requirements for RLUA. RLUA have a certain soil data demand, which changed over time. In some cases, available soil data are still suitable for RLUA, but in many cases available soil data do not meet the required soil data anymore. Missing soil data need to be complemented either through collecting and/or processing new soil data or through transforming available soil data. This study makes use of 'smart analysis' to complement the missing soil data. These analyses are called smart, because they make efficient use of available soil data, project resources, auxiliary data, mapping tools and techniques and pedological knowledge. A flowchart illustrates the options on how the gap can be addressed (Fig. 1.3).

The aim of this study is:

Bridging the gap between the available and required soil data for regional land use analyses.

To reach this aim I will try to answer the following research questions (RQ):

RQ.1. Does it matter which available soil data are used for a regional land use analysis?

RQ.2. What complementary data are needed to meet the required soil data demand for regional land use analysis?

RQ.3. How to obtain the required soil data for regional land use analyses in an effective manner?

Collecting new soil data is expensive, but to obtain the required soil data for RLUA collecting new data is sometimes unavoidable. This thesis hypothesises that the need for new soil data can be minimized by making 'smart' use of available soil data.

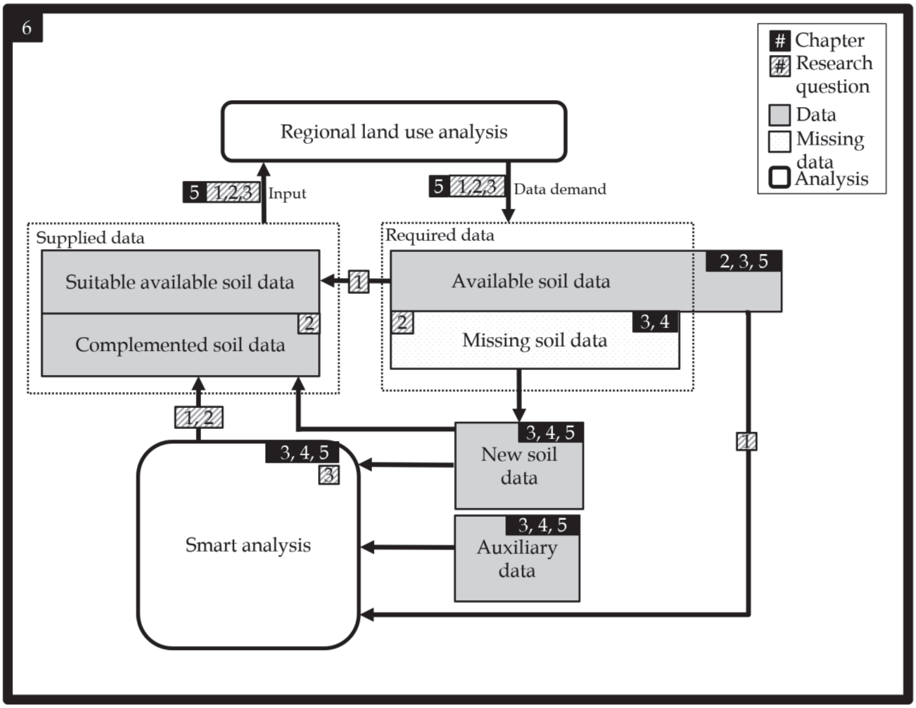


Figure 1.3. *A flowchart on how the gap between the required and available soil data for regional land use analyses (RLUA) can be addressed. Missing soil data, i.e. the gap, can be bridged by collecting and/or processing new soil data or by transforming available soil data using 'smart analysis'. The soil data that result from 'smart analysis' or from the collection of new soil data, complement the suitable available soil data for RLUA. In this case, the supplied soil data that serve as input data for the RLUA meet the required soil data that RLUA demand.*

1.7 Outline

The outline of this thesis is illustrated in Figure 1.3. An attempt to answer the research questions is presented in Chapters 2 to 5 and a synthesis on the different studies is provided in Chapter 6. In Chapter 2 the effect of using different soil datasets for a RLUA is analysed. The soil properties of different spatially exhaustive soil datasets are compared selecting random locations. The effects that were made when a soil dataset was established, were analysed collecting new soil data. The spatially exhaustive soil datasets were applied in a crop-growth simulation model to analyse the effect different soil datasets have on a RLUA.

To bridge the gap, smart analyses can be applied to obtain complementary soil data (Chapter 3 and Chapter 4). Chapter 3 analyses the potential of combining available soil data and new soil data. A literature study was carried out to analyse how land use analyses obtain the required soil data. The potential of combining available soil data and new soil data was illustrated by carrying out two case studies, one at local scale and one at regional scale. In Chapter 4, the potential of incorporating pedological knowledge in a model for DSM is analysed. This study analyses whether the soil organic matter content can be predicted using a mechanistic model for DSM.

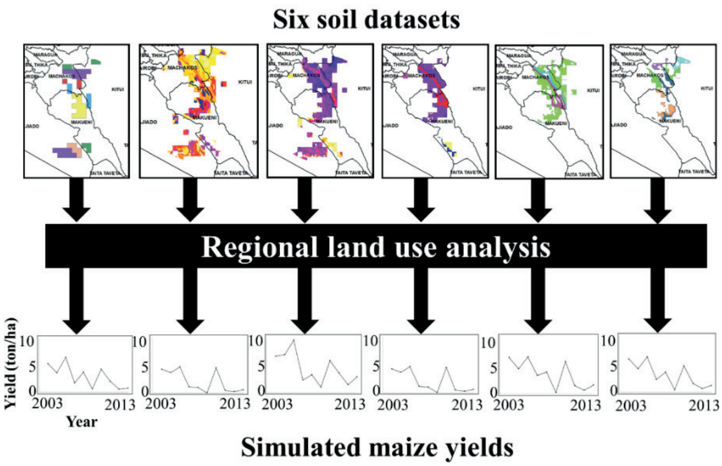
The question on how to obtain the required soil data for RLUA is tried to be answered in Chapter 5. Soil data are nowadays often obtained without aiming to meet the data requirements for RLUA or they are obtained using highly complex techniques. These assumptions are analysed by carrying out three case studies that require soil data on the spatial variation at different levels of detail. The synthesis (Chapter 6) provides a flowchart that helps studies that use RLUA obtaining the required soil data. Besides that, recommendations towards the soil science community and the people that are involved in RLUA are provided to make soil science expertise contribute more effectively in sustainable development programs.

Chapter 2

Exploring the challenges with soil data in regional land use analysis

Highlights:

- Six soil datasets are shown to differ strongly due to the use of different data sources, assumptions and processing methods.
- Field tests showed that assumptions made to derive soil datasets are not always valid.
- The selection of the soil dataset for a regional land use analysis largely influences the results.
- The quality of soil datasets is often unknown hampering their use and requiring validation.



Slightly modified: Hendriks, C.M.J., Stoorvogel, J.J. and Claessens, L.
Agricultural Systems 144 (2016)

2.1 Introduction

There is an increased pressure on our natural resources due to e.g., population growth, economic growth and climate change. Globally, the increase in agricultural production does not keep up with population growth resulting in a decline in food security (Van Ittersum et al., 2013). This increases the need to study the interactions between our natural resources and land use. To study these interactions, regional land use analyses (RLUA) were adapted. In the past, RLUA mainly focused on qualitative land evaluation. The change in RLUA in combination with the increased information technology and data availability opened the possibility to use quantitative simulation models for RLUA (e.g., models simulating crop growth, soil erosion, water quality, land use change) (McBratney et al., 2000). These developments coincided with changing data requirements. In general, the simulation models need quantitative, high resolution and spatially exhaustive data. As many research programmes lack the resources to collect new data, most RLUA rely on available data. However, available soil data often do not match with the data requirements resulting in a gap between the available and required soil data for RLUA. This gap may lead to operational problems in RLUA. This study aims to identify the main challenges with soil data in RLUA by exploring and analysing the effect different soil datasets have on RLUA.

In general, we distinguish four types of soil data:

1. Conventional soil survey (CSS). The CSS is originally established for qualitative land evaluation and is the most common type of soil data. The spatial variation of soils is represented by discrete mapping units. Each mapping unit is described by one (in the case of a consociation) or more (in the case of a soil complex or association) soil types. The boundaries of the mapping units in CSS are abrupt (Cambule et al., 2013 ; Heuvelink and Webster, 2001). The compound mapping units are described by multiple soil types for which often relative area coverages are provided. Less abundant soil types are sometimes left out. Soil types are

characterized by soil morphology and, chemical and physical analyses of representative soil profiles, before they are classified using e.g., Soil Taxonomy (Soil Survey Staff, 2014) or World Reference Base (IUSS Working Group WRB, 2014). By providing representative soil profile descriptions for the soil types, their internal variation is often ignored, i.e., the soil types are considered to be homogeneous. Nowadays, 31% of the global land surface is mapped by CSS at 1:1 million scale or larger (Nachtergaele and Van Ranst, 2003). The reconnaissance survey of the Kapenguna area (Gelens et al., 1976) is a good example of a CSS in Kenya. Those conventional exploratory maps and more general maps like the 1:2M scale provisional soil map of East Africa (Milne et al., 1936) formed the basis for the Exploratory Soil Map of Kenya (Sombroek et al., 1982).

2. Point data. These data are available from a wide range of sources. They can accompany the CSS as representative soil profiles, but they can also be provided along with e.g., agronomic experiments. Point data can be qualitative or quantitative. For example, the Fertilizer Use Recommendation Project (FURP) in Kenya carried out a large number of agronomic experiments in different agro-ecological zones in Kenya (FURP, 1987 ; FURP, 1994). Each experiment was accompanied by a soil profile description including chemical and physical soil characteristics.

3. Digital soil maps. Digital soil mapping (DSM) spatially predicts soil characteristics by deriving statistical relationships between observed soil characteristics and auxiliary information representing the soil forming factors (e.g., digital elevation models representing topography and satellite imagery representing vegetation) (McBratney et al., 2003). The quality of digital soil maps depends on the quality and sampling density of the soil data, on the quality of the auxiliary information, and on the used mapping techniques. An example of DSM in Kenya is presented by Mora-Vallejo et al. (2008).

4. Remotely sensed soil data. These soil data are derived from a broad range of sensing platforms and sensor types. This technique is a relatively new inventory technique. Ge et al. (2011) and Mulder et al. (2011) provide an overview of the various techniques that are available. Most remote sensing studies so far have been

performed locally (e.g. Palacios-Orueta and Ustin, 1998) and no standardized remote sensing based methodology for soil inventory has been established yet (Mulder et al., 2011).

Each soil data type describes soil variability in its own specific way and presents opportunities, but also drawbacks for its use in RLUA. For example, the CSS gives spatially exhaustive data and quantitative data come from representative soil profiles. However, CSS does not describe the soil variability within a soil type and the scale of CSS is often not detailed enough for RLUA (Nachtergaele and Van Ranst, 2003). Point data provide quantitative data, but the data are not spatially exhaustive. Digital soil maps provide quantitative, spatial exhaustive data. However, the soil characteristic maps resulting from digital soil mapping are often established independently. In comparison to conventional soil surveys, digital soil mapping has no unified (soil classification) system. Different digital soil maps of the same area can therefore vary depending on which source data are used, which assumptions are made and how the data are processed.

Our study focuses on Machakos and Makueni counties (Kenya), a semi-arid area where agriculture and food security plays an important role. For this area, six soil datasets are compiled from available soil data sources. The study consists of three steps. In the first step, the six soil datasets are compared. In the second step, we verify assumptions that are made to establish soil dataset using a field survey. In the third step, the effect of selecting a soil dataset for a study on RLUA is analysed. The Global Yield Gap Atlas (GYGA) project[1] was taken as a case study. GYGA assesses yield gaps to study food security and guide potential investments in agricultural research and development (Van Ittersum et al., 2013).

2.2 Materials and methods

2.2.1 Study area

The major staple food crop in Kenya is maize. The total harvested maize area is estimated at 2.16 million ha with an average maize yield of 1.8 tons/ha (FAO Statistics Division, 2015), which is far below the average water-limited maize yield potential of approximately 7.1 tons/ha[1]. Main causes for this large yield gap are i) nutrient depleted soils, ii) low application of mineral fertilizer, iii) scarcity in manure, iv) variable rainfall patterns, and v) lack of resources to improve degraded soils (Claessens et al., 2012). Narrowing the gap between the actual yield and the potential yield is at the top of the agenda of Kenyan governmental agencies. Problems faced by the Ministry of Agriculture and Ministry of Livestock and Fisheries Development (2004) are for example: lack of resilience during droughts and floods, low and declining fertility of land, crop diseases, and lack of coherent land policies.

An important maize cropping area in Kenya, which is also selected as a study site by the GYGA project, is located in Eastern Province and includes Machakos and Makueni counties (Fig. 2.1). The counties are 1.35 million ha and half of that area is under agriculture (Mora-Vallejo et al., 2008). The area is hilly with elevations varying between 418 m and 2053 m above sea level. It has a semi-arid climate with low and highly variable rainfall distributed over two seasons. Average rainfall for each season ranges from 100 mm to 350 mm and the mean annual temperature varies between 15 °C and 25 °C. The main geological parent material originates from the Basement System and contains old intrusive and metamorphic rocks. Deep and friable soils

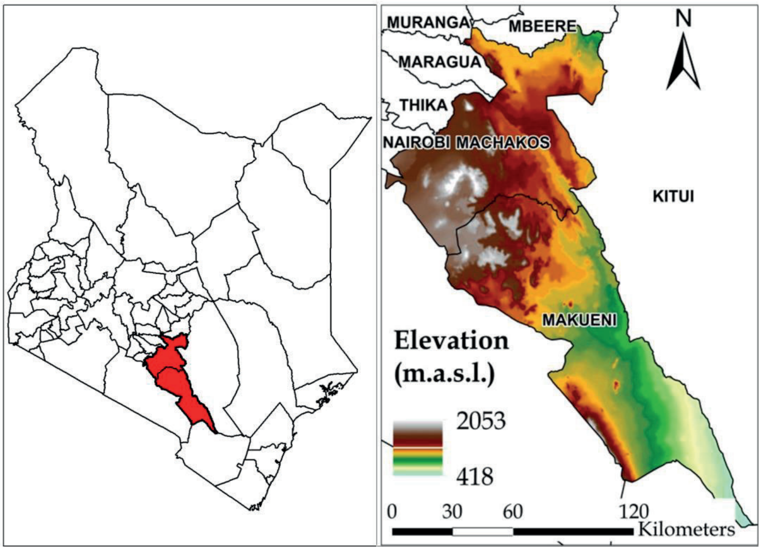


Figure. 2.1. *Machakos and Makueni study area in Eastern Province of Kenya.*

developed in this parent material. The soils are inherently poor in nutrients with the exception of some volcanic areas. The textures range from clay to sandy clay and the soils generally have good drainage. According to the Kenya Soils and Terrain Database (KenSOTER) (Batjes and Gicheru, 2004), the most dominant soil types in Machakos and Makueni counties are Rhodic Ferrasols, Chromic Cambisols, Eutric Vertisols, Haplic Lixisols and Chromic Luvisols.

The study area has several seasonal rivers and the permanent Athi River in the East. Due to fast runoff in seasonal rivers and steep topography around the permanent river, the possibilities for irrigation are limited. Maize is often intercropped with beans, legumes and sorghum. Other cultivated crops are vegetables, fruits and roots. Mixed smallholder farming systems are prevalent in the area. Due to increased agricultural activities in the early 1930s, caused by population growth, soil erosion took place (Tiffen et al., 1994). Governmental enforcement in erosion control, e.g. by terracing agricultural fields and reforestation of highly degraded areas and steep areas, slowed down the land degradation. Despite these measures and the willingness of people to voluntarily maintain the terraces (Tiffen et al., 1994; De Jager et al., 2005), the yields are low. Nowadays, still 59.6% of the population in Machakos and 64.1% in Makueni fall below the poverty line of 1 US$/person/day (Commission on Revenue Allocation, 2011). These numbers underline the need for RLUA.

2.2.2 Soil datasets

For the study area, six soil datasets were compiled from available soil data sources. The datasets are spatially exhaustive, but differ in extent, scale/resolution and spatial variation (Table 2.1). Only two soil datasets collected field data to establish the dataset. The other datasets are derivatives of available soil datasets, whether or not combined with collected field data.

1.The ISRIC-WISE Derived Soil Properties dataset (Batjes, 2012) is a global 5 by 5 arc minutes gridded map. Data sources behind this dataset are the Digital Soil Map

Table 2.1. *Description of six soil datasets available for Machakos and Makueni counties (Kenya).*

	Project	Extent	Scale/resolution	Spatial variation	Source
1	ISRIC-WISE	Global	5 arc minutes	Continuous	Batjes, 2012
2	S-World	Global	30 arc sec	Continuous	Stoorvogel, 2014
3	AfSIS	Continental	30 arc sec	Continuous	ISRIC—World Soil Information, 2013
4	Local DSM	Regional	3 arc sec	Continuous	Mora-Vallejo et al., 2008
5	KenSOTER	National	1:1 million	Discrete	Batjes and Gicheru, 2004
6	FURP	National	1:1 million	Discrete	FURP, 1987 ; FURP, 1994

of the World (FAO, 1995) and the soil characteristic data from the ISRIC-WISE Harmonized Global Soil Profile dataset (Batjes, 2009). Machakos and Makueni counties were covered by ten mapping units of the ISRIC-WISE Derived Soil Properties dataset. Quantitative descriptions of the soil profile came from representative soil profiles.

2. S-World is a global digital soil map with a resolution of 30 arc sec (Stoorvogel, 2014). Data sources behind this dataset are the Harmonized World Soil Database (HWSD) (FAO et al., 2012), the ISRIC-WISE Harmonized Global Soil Profile dataset (Batjes, 2009), and various sources of auxiliary information. S-World disaggregates soil associations of the HWSD to obtain a map with single soil types. For Kenya, the HWSD is based on KenSOTER (Batjes and Gicheru, 2004). Subsequently, a model for soil formation is used to derive soil characteristics for each location based on ranges of soil characteristics per soil type derived from the ISRIC-WISE soil profile database.

3. The Africa Soil Information Service (AfSIS) produced a continental digital soil map with a resolution of 30 arc sec (ISRIC—World Soil Information, 2013). The digital soil map was produced from harmonized soil profile data (Africa Soil Profile Database) and auxiliary data. The Africa Soil Profile Database originates from more than 300 different soil data sources, including the ISRIC-WISE Harmonized Global Soil Profile dataset (Batjes, 2009). Twenty soil profiles of the Africa Soil Profile Database were located in our study area.

4. A Local DSM study for the counties Machakos and Makueni was performed by Mora-Vallejo et al. (2008). The study aimed to test digital soil mapping in an area with limited soil data and auxiliary data. The digital soil map is based on regression kriging of 95 composite soil samples of the topsoil (0–30 cm) and the map has a resolution of 3 arc sec. The composite samples were taken on terraced maize fields. The dataset provides soil characteristic maps of soil organic carbon and clay content. To get a description of the entire soil profile, the dataset was combined with subsoil data of KenSOTER (Batjes and Gicheru, 2004).

5. The Kenya Soils and Terrain Database (KenSOTER) (Batjes and Gicheru, 2004) is a 1:1 M polygon-based soil map based on the SOTER methodology (Van Engelen and Dijkshoorn, 2013). The discrete mapping units represent a unique combination of terrain and soil characteristics. The map is compiled from different soil data sources, e.g. Exploratory Soil Map of Kenya (Sombroek et al., 1982). Qualitative and quantitative soil profile descriptions were taken on representative locations and mapping units were defined from landform, lithology, surface form, slope, parent material and soils (Van Engelen and Wen, 1995). Our study area included 49 mapping units. Each mapping unit consists of one or more soil types and each soil type is described by at least one representative soil profile.

6. The Fertilizer Use Recommendation Project (FURP) yielded a point dataset. The project was established to provide fertilizer use recommendations for rain-fed maize areas in Kenya (FURP, 1987 ; FURP, 1994). Crop experiments were carried out in maize fields at representative locations and included chemical and physical analyses of the soil profile. The area was sub-divided in zones with similar agro-ecological

conditions based on Jaetzold and Schmidt (1982) and the Exploratory Soil Map of Kenya (Sombroek et al., 1982). In our study area, 10 soil profile descriptions and agro-ecological zones were located.

2.2.3 Comparison of the soil datasets

The various soil datasets are based on available soil data like the Exploratory Soil Map of Kenya (Sombroek et al., 1982) and in some cases additional field data collection. However, data processing differed per soil dataset which may result in differences between the datasets. The six soil datasets were compared to analyse the differences in soil characteristics. To overcome issues like scale differences between the datasets, 200 points were randomly selected within our study area. Through an overlay of these 200 points with the soil datasets, the average carbon content, texture and soil pH over 120 cm depth were determined.

2.2.4 Assumptions in deriving the soil datasets

Assumptions were made when soil datasets were established. Four assumptions were identified and verified by a field survey.

2.2.4.1 Assumption 1: Soil types are homogeneous

Soil types within a mapping unit were described by a limited number of representative soil profiles. Therefore, the internal variation of a soil type is often unknown. We tested the soil variability in two mapping units of the KenSOTER dataset Version 1.0 (Batjes and Gicheru, 2004). Both mapping units were classified by a single soil type. This does not mean that the mapping units were homogeneous, because mapping units were allowed to have a certain natural variability that is expected to occur at a scale of 1:1 M (Van Engelen and Dijkshoorn, 2013). The first mapping unit (781.5 km^2) was described by a Chromic Cambisol whereas the second mapping unit (47.2 km^2) was described by a Ferralic Arenosol. Chromic Cambisols are reddish coloured soils with little horizon differentiation evident from changes in colour, structure or carbon content (IUSS Working Group WRB, 2014). These soils are medium to fine-textured and originate from different parent materials. Red, sandy soils that lack any visible soil profile development are classified as Ferralic Arenosols

(IUSS Working Group WRB, 2014). The soil variability within the mapping units was tested by taking clustered random samples. First, 14 squares of 100 km^2 were randomly selected. Each square had the same chance of being reselected. Subsequently, for each square, samples were taken at 5 out of 10 randomly selected sampling locations, depending on accessibility of the locations. In the Chromic Cambisol 26 soil samples of the topsoil (0–20 cm) were taken and in the Ferralic Arenosol 16 soil samples of the topsoil were taken. To avoid effects of the within field variation, five samples were taken at each location and mixed thoroughly into a composite sample. In agricultural fields the composite samples were taken as one sample in the centre of the field and four samples 5 m towards each corner of the field. In natural areas the composite samples were taken on a distance of 5 m from each other. The soil variability in a soil type was tested by calculating the coefficient of variation (CV) on a chemical parameter (pH) and a physical parameter (texture). The CV is the ratio of the standard deviation to the mean.

2.2.4.2 Assumption 2: Soil mapping units can be delineated without considering land use and land management

Representative soil profiles describe and analyse soil types related to their representatives in nature (Soil Survey Division Staff, 1993). These soil profiles are dominantly at undisturbed locations. However, land use and land management have an effect on soil characteristics (e.g. Vågen et al., 2005). These effects of land use and land management on soil characteristics were tested by taking paired observations of soil conditions: 'agricultural land versus nature' (11 pairs), 'mono-cropping versus intercropping' (6 pairs) and 'terraced fields versus non-terraced fields' (8 pairs). Fields were selected with the help of District Agricultural Officers. At each field, composite samples of the topsoil were collected as described in Assumption 1. For each pair two samples were compared, except for the pairs 'terraced fields versus non-terraced fields'. For these pairs six soil samples were analysed. Samples at the top, in the middle and at the bottom of the terraced and non-terraced fields were compared. This avoided large deviations from the mean due to the influence of the slope (Herweg and Ludi, 1999). Soil samples were tested on pH and carbon content, because these soil characteristics were influenced by natural and human factors (e.g., mineral composition, fertilization) (Vågen et al., 2005). The samples were also tested on actual soil moisture content. Actual soil moisture contents in a pair were comparable, because we measured the actual soil moisture content in a short time span. Actual soil moisture content was variable within a field, therefore the average of 15 measurements around the sampling location was taken. Significant effects of land use and land management on soil characteristics were tested by a paired sample t-test ($p < 0.10$).

2.2.4.3 Assumption 3: Soil data sources can be combined

Harmonizing soil data is a standard procedure to develop soil datasets (Sulaeman et al., 2013). To make (harmonized) soil datasets applicable for RLUA, sometimes datasets need to be combined. Three-dimensional soil mapping techniques were explored to assure soil data requirements for RLUA (Kempen, 2011; Lacoste et al., 2014). However, applications of 3D soil mapping techniques in RLUA are still limited. With this assumption we tested the effect of combining two soil datasets. The datasets had different descriptions on the spatial variability. When topsoil data of the Local DSM and subsoil data of the KenSOTER dataset were combined, it is assumed that spatial variability in topsoil and subsoil could be described differently. Carbon content and pH of topsoil and subsoil were compared in seven KenSOTER mapping units. The seven mapping units were not homogeneous. The proportion of the dominant soil type in a mapping unit varied between 50% and 100%. KenSOTER delineated areas with distinctive patterns of landform, lithology, surface form, slope, parent material and soil (Van Engelen and Dijkshoorn, 2013). To test this assumption we assumed the KenSOTER mapping units as most representative polygons. This allowed for a comparison of the spatial variability in topsoil and subsoil. All sampling locations where it was possible to sample the subsoil were included to test this assumption. A composite sample of the topsoil (0–20 cm) (as described in 2.2.4.2) and one sample of the subsoil (50–60 cm) were taken at 65 locations.

2.2.4.4 Assumption 4: Soil characteristic maps can be established independently

In general, DSM results in single soil characteristic maps. Correlations between soil characteristics are considered indirectly in the underlying statistical models, i.e. by using covariates describing soil forming processes. The classification system of CSS keeps correlations between soil characteristics in the soil profile descriptions and soil analyses. Correlation coefficients between highly correlated soil characteristics (Yerima et al., 2009; Farrar and Coleman, 1967) were compared for the six soil datasets using linear regression. Following soil characteristics were correlated: 'carbon content and clay percentage' and 'carbon content and pH'. In addition, correlation coefficients were also estimated for the field data. This assumption was tested using the entire dataset, e.g. the Local DSM included data of Machakos and Makueni counties, while KenSOTER included data of Kenya. The field data consisted of 237 soil samples, including all samples used for testing the three assumptions, and 19 duplicate samples.

2.2.5 Measurement equipment and laboratorial analysis

To test the assumptions, a large number of soil samples were required. Therefore, the samples were tested by sensors. Following sensors were used: soil texture was measured using a turbidimeter AL250T-IR (for details see Appendix A) (Stoorvogel et al., in prep.), nitrogen (N) content was measured using the Nitracheck reflectometer (Eijkelkamp, 2004), pH-H_2O was measured using the Multimeter 18.50.01 and actual soil moisture content was measured using the Theta Probe ML2x (Eijkelkamp, 1999). To validate the turbidimeter AL250T-IR and the Nitracheck reflectometer, 19 samples were analysed in the laboratory on texture, nitrate content and carbon content. In the laboratory, the texture was analysed by the Hydrometer Method, nitrate by the Colorimetric Method and carbon by the Walkley and Black Method. Laboratory analysis resulted in a C:N ratio of 11.9. The C:N ratio had a correlation of 0.38. The low correlation was caused by the low contents (average carbon content was 1.6%). The actual soil moisture content was directly measured in the field and pH was, like in other researches (e.g. Adamchuk et al., 2004), successfully measured by the Multimeter 18.50.01. To test the quality of the laboratory results, 19 soil samples were analysed in duplicate.

2.2.6 Soil datasets effects on a regional land use analysis

2.2.6.1 Introduction

The GYGA project used different crop growth simulation models to estimate potential and water-limited yields for yield gap assessment. In this study, the WOrld FOod STudies (WOFOST) model (Boogaard et al., 2013) was chosen to simulate water-limited maize yields (Y_w). A sensitivity analysis showed the impact of soil characteristics on Y_w. To delineate maize cropping areas and to obtain soil input data for the WOFOST model, the GYGA project formulated two protocols. Both protocols were applied to the six soil datasets to analyse differences in the selection of maize cropping area and in simulated water-limited maize yield.

2.2.6.2 Crop growth simulation model

The WOFOST model requires crop phenology and genetic characteristics, weather and soil (Table 2.2). In addition, information on sowing and harvesting date and crop

Table 2.2. *Soil input data required for the crop growth simulation model and the availability of these required data per dataset.*

Dataset	FC[a]	Sand fraction	Clay fraction	OM content[b]	Infiltration rate	Maximum rooting depth
ISRIC-WISE	x	x	x	x	x	
S-World	x	x	x	x	x	x
AfSIS		x	x	x		
Local DSM			x	x		
KenSOTER	x	x	x	x	x	x
FURP	x	x	x	x	x	x

[a] Field capacity. Only FURP measured FC. FCs reported by other datasets were calculated using different pedotransfer functions.

[b] Organic matter content.

management is required. For the water balance, the model uses a simple water budget model including wilting point (WP in cm^3/cm^3), field capacity (FC in cm^3/cm^3), saturation point (SP in cm^3/cm^3), a runoff factor (fraction of rainfall lost through superficial runoff) and the maximum rooting depth (in cm). Although the six soil datasets differed in many characteristics, they contained the necessary soil input data for the WOFOST model (Table 2.2).

2.2.6.3 Sensitivity analysis

A sensitivity analysis evaluates the effect soil input parameters have on the results. With the partial sensitivity analysis we changed values of one soil variable, while other variables remained constant. For all soil types, the GYGA project used default values for WP (0.1 cm^3/cm^3) and SP (0.45 cm^3/cm^3). When FC was not available in the dataset, the parameter was calculated by the pedotransfer function of Saxton and Rawls (2006). The pedotransfer function required data on clay and sand percentage and organic matter content. The runoff factor required data on drainage. These data were derived from representative soil profiles. When maximum rooting depth was not available in the soil dataset, the value was estimated using the bottom soil layer (max. 100 cm) or the value of 100 cm was assumed. According to the GYGA project, the WOFOST model had three parameters that initially require soil data: FC, runoff factor and maximum rooting depth. For these three parameters a sensitivity analysis was performed. The values of the sensitivity analysis on FC differed from WP to SP in steps of 0.05 cm^3/cm^3. For our study area, the runoff factor can vary between 0% and 33%. For the sensitivity analysis of the rooting depth, the values differed between minimum rooting depth (60 cm), as defined by the GYGA project, and maximum rooting depth (100 cm) in steps of 10 cm.

2.2.6.4 Delineation of maize cropping areas

Countries with a national harvested crop area of more than 100,000 ha for a specific crop were included in the GYGA project for yield gap analysis. GYGA only included the most dominant cropping areas for the analysis. A protocol was developed to delineate these areas (Fig. 2.2). The delineation was based on harvested crop area maps, climate zonation maps, weather station data and soil datasets. The harvested

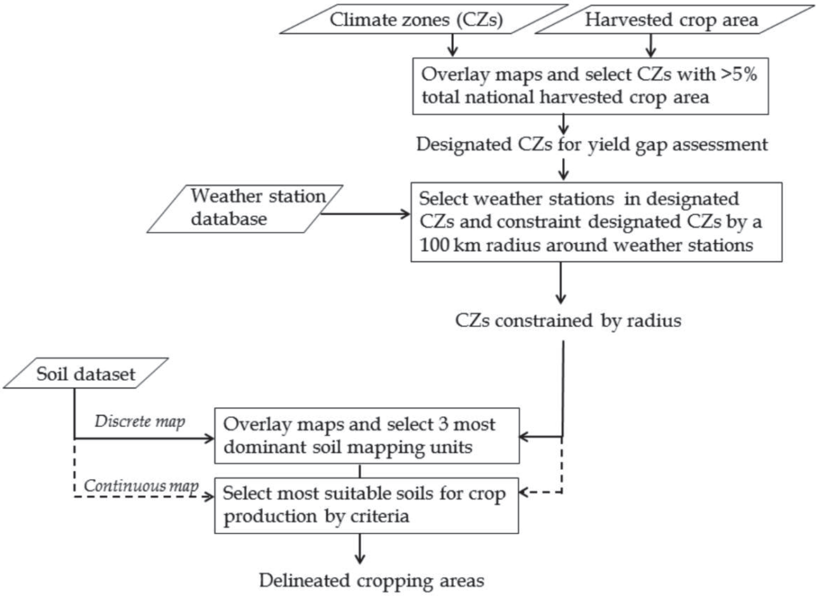


Figure. 2.2. *Protocol to select most dominant cropping areas. These areas are included for yield gap assessment in the Global Yield Gap Atlas (GYGA) project. The dashed lines are added to the original GYGA protocol.*

crop areas were selected from the HarvestChoice SPAM2000 database (You and Wood, 2006; You et al., 2009). The database contained harvested crop maps of 5 arc minutes for 20 major staple crops. The climate zonation scheme (GYGA Extrapolation Domain) of Van Wart et al. (2013) was created for and used by the GYGA project. The climate zones (CZs) were defined by differences in growing degree days, temperature seasonality, and aridity index. To select the most dominant harvested crop areas, an overlay of the climate zonation scheme and the HarvestChoice SPAM2000 database was made. CZs with more than 5% of the total national harvested crop area were selected (Van Wart et al., 2013). This results in a number of designated CZs for yield gap analysis. One weather station per designated CZ was selected from a weather station database. The selected weather stations were assumed to be representative for a radius of 100 km within the designated CZs. Finally, within these areas, the three dominant soil types were selected from discrete soil datasets. For spatially explicit soil datasets, most suitable soils for crop production were selected. As defined by the GYGA project, soils are suitable when the maximum rooting depth is more than 60 cm, average water holding capacity is above 7% and the average sand percentage is smaller than 75%. We applied the protocol to all six soil datasets to delineate the most dominant maize cropping areas. We compared the results with the most dominant maize cropping areas delineated from the ISRIC-WISE dataset, because this dataset was initially selected by GYGA for our study area.

2.2.6.5 Impact on simulated water-limited maize yields

The GYGA project formulated protocols to get model input data, as presented in Figure 2.3 for soil data. Only suitable soils were selected for the analysis. In datasets with discrete mapping units, each mapping unit consists of more than one soil type. In these datasets soils were selected until the proportion of soil types was 50%. For continuous datasets, we decided to restrict the selection only by discarding unsuitable soils. The impact parameters of the water budget model were WP, FC and SP. These parameters were often not measured, therefore the GYGA project assumed default values for WP (0.1 cm^3/cm^3) and SP (0.45 cm^3/cm^3). The water holding

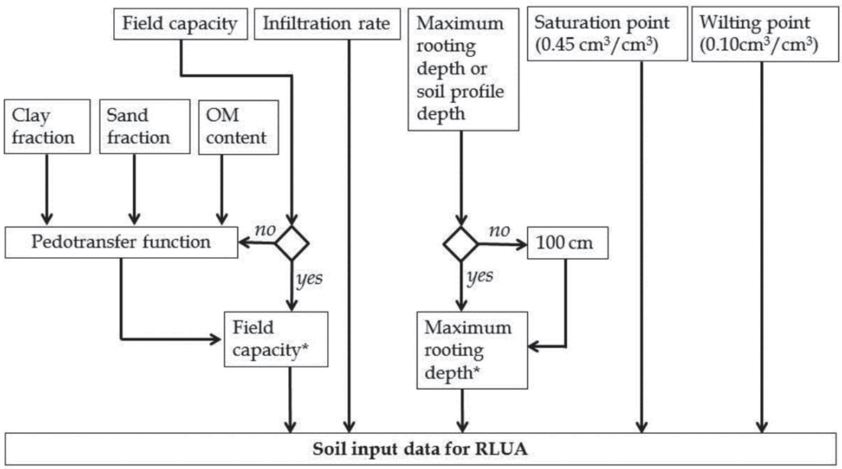


Figure 2.3. *Protocol to obtain soil input data for the crop growth simulation model WOFOST. Alternatives are given by a decision rhombus when the field capacity or the maximum rooting depth is not available in the dataset. Final field capacity and maximum rooting depth are indicated by *.*

capacity (WHC) is the difference between FC and WP. When WHC or FC were not given in the dataset, the FC was estimated by a pedotransfer function (Saxton and Rawls, 2006):

$$FC = \theta_{33} + (1.283(\theta_{33})^2 - 0.374\,\theta_{33} - 0.015$$

$$\theta_{33} = -0.251S + 0.195C + 0.011OM + 0.006(S * OM) - 0.027(C * OM) + 0.452(S * C) + 0.2993$$

where, θ_{33} is the moisture tension at 33 kPa, S is the sand fraction, C the clay fraction and OM is the organic matter content in %.

The GYGA project performed a literature search to estimate the runoff factor (Table 2.3). The runoff factor was based on drainage class and slope. The slope was

Table 2.3. *Fraction of rainfall lost through superficial runoff (%) based on slope and drainage class*[a].

Drainage class / Slope (%)	Very poor	Insufficient	Moderate	Well drained	Extremely well drained
0-2	20	13.3	6.7	0	0
2-6	26.7	20	13.3	6.7	0
6-10	33.3	26.7	20	13.3	6.7
>10	40	33.3	26.7	20	13.3

[a] According to: http://www.yieldgap.org/

estimated from a digital elevation model, e.g. SRTM DEM, or from topographical maps supplemented with the opinion of local agronomists. In our study, the slopes were estimated from SRTM DEM except for the FURP dataset. In FURP the slopes on the sampling location were given. When the drainage class was not available in the dataset, the data of KenSOTER were used. When the maximum rooting depth was not available in the soil dataset, the bottom soil layer (max. 100 cm) was used as maximum rooting depth. When the bottom soil layer was not available, the maximum rooting depth was assumed at 100 cm. Table 2.4 describes how model input data were derived from the six soil datasets.

Table 2.4. *Overview of how soil input data for crop growth simulation model WOFOST are derived from the soil datasets.*

Soil dataset	**Maximum rooting depth**	**Field capacity**	**Runoff factor**
ISRIC-WISE	Assumed at 100 cm.	WHC[1] available in dataset. FC[2] is WHC minus WP[3] (0.1cm^3/cm^3).	Infiltration rate available in dataset.
S-World	Soil profile depth is maximum rooting depth.	FC available in dataset.	Infiltration rate available in dataset.
AfSIS	Assumed at 100 cm.	FC is estimated by a PTF[4] of Saxton and Rawls (2008). Clay and sand fraction and OM[5] content available in dataset.	Infiltration rates of the KenSOTER dataset are used.
Local DSM	Assumed at 100 cm.	FC is estimated by a PTF of Saxton and Rawls (2008). Clay fraction and OM content available in the dataset. The sand fraction is assumed to be 100 - clay fraction, because the soils in the study area hardly contain silt.	Infiltration rates of the KenSOTER dataset are used.
KenSOTER	Bottom soil layer is maximum rooting depth (max. 100cm).	WHC is available in the dataset. FC is WHC minus WP (0.1cm^3/cm^3).	Infiltration rate available in dataset.
FURP	Bottom soil layer is maximum rooting depth (max. 100cm).	FC is available in the dataset. The FC in this dataset was measured by laboratory experiments.	Infiltration rate available in dataset.

[a] Water holding capacity, [b] Field capacity, [c] Wilting point, [d] Pedotransfer function, [e] Organic Matter.

We simulated 10 years (2004–2013) of water-limited maize yields with the WOFOST model. The yields were based on one cropping season. Sowing and harvesting date and crop management data were general information the WOFOST model required. Sowing dates vary per year in Kenya, because farmers shift their sowing date to the variable start of the rainy season (Müller et al., 2010). According to local agronomists, the average sowing date was around day 74 (day 1 is 1st of January). For each year, the optimum water-limited maize yield was estimated by adding and subtracting 10, 20 and 30 days from the average sowing date. The start of the water balance was initiated 90 days before. The initial available soil water is estimated at 5 cm and the maximum initial moisture content was estimated at 0.1 cm^3/cm^3, because the sowing date was at the start of the rainy season. The duration of crop growth was assumed until maturity, with a maximum of 120 days.

2.3 Results and discussion

2.3.1 Comparison of the soil datasets

Comparing the six soil datasets gave different results for carbon, sand and clay content and soil pH (Table 2.5), while some datasets were derived from the same soil data source. Sand content was 10.7% in AfSIS and 71.7% in the Local DSM. The other four datasets had comparable sand contents (36.2%–48.0%). The different values for AfSIS and the Local DSM might have been caused by the fact that both datasets are digital soil maps with a relatively low variance explained, 23.3% for AfSIS and 37% for the Local DSM. The clay content (11.7%) and pH (4.8) in the AfSIS dataset

Table 2.5. *Averages and standard deviations (in brackets) of four soil characteristics for six soil datasets.*

Dataset	Carbon (%)	Sand (%)	Clay (%)	pH
ISRIC-WISE	0.6 (0.1)	43.5 (6.5)	37.7 (4.1)	6.2 (0.7)
S-World	1.5 (1.2)	45.1 (16.5)	36.9 (13.4)	6.2 (0.4)
AfSIS	0.1 (0.1)	10.7 (4.9)	11.7 (4.0)	4.8 (0.7)
Local DSM	0.8 (0.2)	71.7 (17.6)	23.6 (8.8)	n.a.
KenSOTER	1.0 (0.6)	48.0 (21.0)	31.8 (16.7)	6.1 (1.1)
FURP	0.3 (0.0)	36.2 (5.0)	44.4 (7.2)	5.1 (0.7)

differed most from the other datasets. The AfSIS dataset showed an explained variance of 18.4% for clay and 30.7% for pH. Clay content in the remaining datasets ranged between 23.6% and 44.4% and a pH between 5.1 and 6.2.

Differences between soil datasets can be explained by different factors, because datasets were established using different data sources, assumptions and processing methods. The differences in soil characteristics make it difficult to decide which dataset to use for RLUA and soil characteristics differ too much to make a decision pragmatically. As the aim of the GYGA project was to estimate the yield gap for major staple crops, maize in our case, the Local DSM and the FURP dataset were established from soil samples taken under maize fields. Studies that need information on undisturbed soils (e.g., studies on nature conservation) are likely to prefer datasets that originally took soil samples in undisturbed soils.

2.3.2 Assumptions in deriving the soil datasets

2.3.2.1 Assumption 1: Soil types are homogeneous

The Soil Survey Manual (Soil Survey Division Staff, 1993) states that the value of the soil map is reduced when soil variability within a soil type is not described. Most KenSOTER mapping units consist of more than one soil type. However, this assumption was tested in two mapping units that gave proportions of 100% for the most dominant one. In the Chromic Cambisol the field data resulted in an average sand content of 78% with a CV of 7% (Table 2.6), a clay content of 21% with a CV of 41% and a pH of 5.9 with a CV of 8%. In the Ferralic Arenosol the field data resulted in an average sand content of 83% with a CV of 14%, a clay content of 15% with a CV of 73% and a pH of 6.3 with a CV of 9%. The field data showed largest variation in clay content. Sand content in the Chromic Cambisol and pH in the Chromic Cambisol and Ferralic Arenosol showed little variation.

The soil classification system is based on differentiating soil forming processes rather than soil characteristics. Nowadays, RLUA often use soil characteristics rather than soil classifications. This results in an increased need to quantify the spatial variation in soil properties within soil types. Derived soil characteristics indicate considerable soil variability, so we need to ensure that soil profile descriptions are indeed representative for a certain soil type.

Table 2.6. *The soil variability in two KenSOTER mapping units. The first mapping unit consist of Chromic Cambisols (CMx). The second mapping unit consists of Ferralic Arenosol (ARo). 'n' is the number of samples taken in the mapping unit. The averages and coefficients of variance (CV) resulting from field data.*

Mapping unit	Soil type	n	Area (km^2)	Average			CV		
				Sand (%)	Clay (%)	pH (-)	Sand (%)	Clay (%)	pH (%)
1	CMx	26	782	78	21	5.9	7	41	8
2	ARo	16	47	83	15	6.3	14	73	9

2.3.2.2 Assumption 2: Soil mapping units can be delineated without considering land use and land management

Some soil datasets do not take land use and land management into account. However, land use and land management have an effect on the soil characteristics (Table 2.7). The analysed soil samples showed a significant difference in carbon content in soils under nature compared to soils under agriculture ($p = 0.02$). This is unusual, but can be explained. In the study area, 'natural land' was former

Table 2.7. *Paired sample t-test to analyse the difference in soil characteristics between soil samples taken in nature and agriculture, terraced and non-terraced fields, and mono-cropping and intercropping fields. Significant ($p<0.10$) values are indicated by *.*

	pH	Actual soil moisture	Carbon
Nature vs agriculture	0.54	0.09*	0.1
Terraced vs non-terraced	0.5	0.03*	0.23
Mono-cropping vs intercropping	0.31	0.18	0.07*

agricultural land. To prevent increased land degradation, the areas where land degradation was significant were reforested (Tiffen et al., 1994). Low tree cover and active gully erosion showed the poor state of the 'natural areas'. The recovery of degraded land is slow in semi-arid environments.

The actual soil moisture content was significantly lower ($p = 0.09$) in soils under nature compared to soils under agriculture. The run-off factor in non-agricultural fields was high, causing lower actual moisture contents in natural soils. Nearly all agricultural fields were terraced. Normally, terraced fields improve the moisture content of the soil, but this could not be concluded from our analysis. The pairs 'mono-cropping versus intercropping' showed a significantly ($p = 0.07$) higher carbon content in areas where intercropping was applied. Intercropping with legumes enhances biological nitrogen fixation. The actual soil moisture content was lower in non-terraced fields. The paired sample t-test did not indicate any significant differences in pH.

This assumption showed that soil characteristics can be overestimated or underestimated in datasets that do not consider land use and land management. The AfSIS dataset, for example, was based on more than 12,000 soil samples using more than 300 different data sources without distinguishing soil samples taken under 'natural land' and 'agricultural land'.

2.3.2.3 Assumption 3: Soil data sources can be combined

The effect of combining soil datasets with different descriptions of soil variability was tested. There is a general assumption that subsoil is less variable in terms of soil characteristics than topsoil. In this study we also assumed less spatial variability in the subsoil, because composite samples of the topsoil were taken and only one sample of the subsoil was taken. KenSOTER mapping units were not homogeneous, but we assumed the mapping units to be most representative for the comparison of spatial variability in topsoil and subsoil. The mapping units showed indeed more variability in carbon content in the topsoil compared to the subsoil (Fig. 2.4). The pH did not show differences in topsoil and subsoil variability, indicating the low standard deviation of the pH. The soil variability in topsoil and subsoil differed per soil characteristic. Therefore, combining datasets with different descriptions of soil variability can affect any RLUA the data are used for.

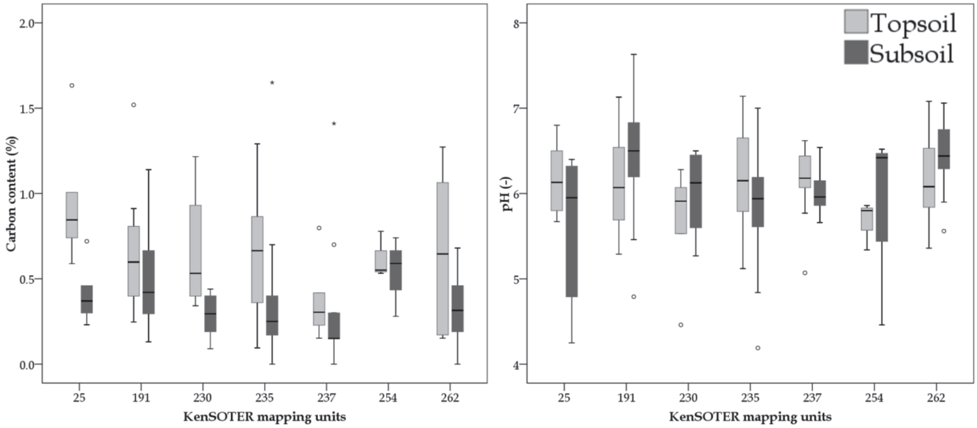


Figure 2.4. *The topsoil (0–20 cm) variability and subsoil (50–60 cm) variability in carbon content and pH are compared for seven KenSOTER mapping units. Note that each mapping unit includes more soil types.*

2.3.2.4 Assumption 4: Soil characteristic maps can be established independently

Linear regression showed very different correlation coefficients between all soil datasets (Table 2.8). The FURP dataset and the Local DSM focussed both on maize growing areas, but the correlation of clay and carbon content was 0.47 and − 0.13 respectively. The FURP dataset was based on soil profile descriptions and compared soil characteristics of the same soil profile, while the Local DSM was based on DSM and compared soil characteristic maps that were established independently. KenSOTER also described soil profiles, but this dataset showed no correlation (r^2 = 0.02). The AfSIS dataset is based on DSM, but the correlation coefficients were much closer to the correlation coefficients of the FURP dataset, r^2 = 0.47 for clay and carbon and r^2 = − 0.47 for pH and carbon. AfSIS is a digital soil map predicting soil characteristics (dependent variable) with limited auxiliary information (explanatory variable).

Table 2.8. *Correlation coefficients resulting from linear regression between clay and carbon (C) content and between pH and carbon content. Correlation coefficients are estimated for six soil datasets and field data.*

Soil dataset	Correlation coefficient	
	Clay-C	pH-C
ISRIC-WISE	0.13	-0.01
S-World	-0.01	-0.61
AfSIS	0.47	-0.47
Local DSM	-0.13	n.a.
KenSOTER	0.02	-0.24
FURP	0.47	-0.24
Field data	0.08	0.1

The dependent variable can only be explained by a limited number of explanatory variables, and therefore correlation coefficients can become higher. The field survey data showed a correlation coefficient of 0.08 for clay and carbon and 0.10 for pH and carbon. In literature, linear regression resulted in correlation coefficients of −0.44 for clay and carbon content and −0.51 for pH and carbon content (Yerima et al., 2009). Datasets were created using different methods to keep correlations between soil properties. This could have caused differences between datasets. Probably the extent of the datasets also affects the correlation coefficients.

2.3.3 Soil datasets effects on a regional land use analysis

2.3.3.1 Sensitivity analysis

The sensitivity of the WOFOST model was tested for three input parameters (FC, runoff factor and maximum rooting depth). To test the sensitivity of the model, one input parameter changed while others remained constant. We expressed the sensitivity of the model by analysing the effect different parameters have on simulated water-limited maize yields (Y_w) (Fig. 2.5).

The model simulation run was set for 10 years. Within these ten years, some years gave very low yields e.g., due to droughts. The first parameter (FC) showed a non-linear pattern when the FC changed from wilting point (0.10 cm^3/cm^3) to saturation point (0.45 cm^3/cm^3) (Fig. 2.5A). A small increase in FC from the wilting point caused a maize yield increase of 200 kg/ha/%. The field capacity reached an optimum between FC 0.18 cm^3/cm^3 and 0.30 cm^3/cm^3. On average, the Y_w decreased about 50 kg/ha/% after FC 0.30 cm^3/cm^3. The sensitivity of the runoff and maximum rooting depth parameters to water-limited maize yields was more linear (Fig. 2.5B). On average, an increase of the runoff factor with 10% resulted in a yield decrease of 290 kg/ha. For the maximum rooting depth the sensitivity analysis showed an average increase in yield of 234 kg/ha per 10 cm increase in maximum rooting depth (Fig. 2.5C). The model simulations were sensitive to all three parameters, but each parameter had different impact. When the maximum rooting depth was unknown, the bottom soil layer was assumed to be the maximum rooting depth (max. 100 cm) or the rooting depth is estimated at 100 cm. This relatively rough estimation could affect the results. A similar effect could happen to the FC. All soil types had the same default value for WP and SP, while these parameters differed per soil type. It is important to understand the impact soil characteristics and assumptions have on modelled results.

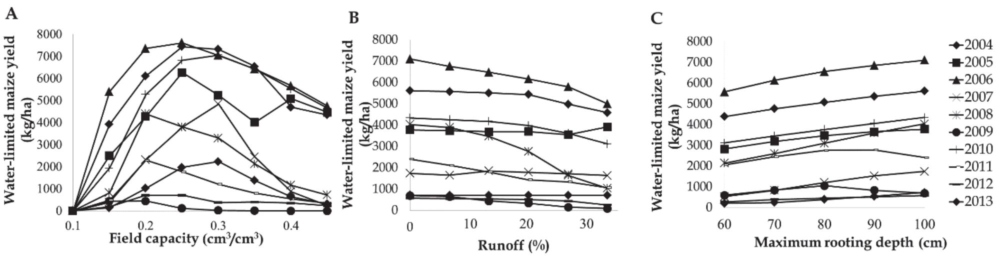


Figure 2.5. *Partial sensitivity analysis for crop growth simulation model WOFOST. The sensitivity of the field capacity (A), runoff factor (B) and maximum rooting depth (C) is tested by changing one parameter and analysing the effect this change has on water-limited maize yields.*

2.3.3.2 Delineation of maize cropping areas

The national harvested crop area of a target crop needed to be over 100,000 ha before Kenya can be selected for the GYGA project. According to the HarvestChoice SPAM2000 database, 2.16 million ha maize is cultivated in Kenya. The areas under maize cultivation are indicated in Fig. 2.6A. In Machakos and Makueni counties there was a climate zone (CZ) having more than 5% of the total national harvested crop area. Therefore, this CZ was selected for the yield gap analysis (Fig. 2.6B). The designated CZ had a weather station in Kambi Ya Mawe (1.554S, 37.322E). The designated CZ constrained by 100 km radius around the weather station (Fig. 2.6C) was further delineated by selecting the three most dominant soil mapping units from the ISRIC-WISE dataset (Fig. 2.6D). The final area for yield gap analysis was 972 km^2 and ranged in elevation between 913 m and 1400 m above sea level.

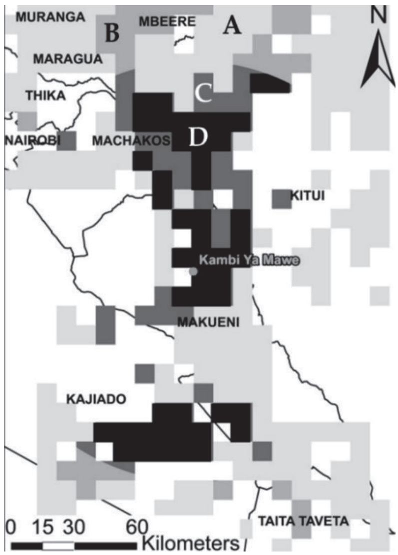


Figure 2.6. *The protocol of the Global Yield Gap Atlas (GYGA) project to select most dominant maize cropping areas in Machakos and Makueni counties. A) Select areas where maize is cultivated, B) designate climate zone with > 5% of total national harvested crop area, C) select one weather station in designated climate zone and delineate area by drawing a radius of 100 km around weather station Kambi Ya Mawe. Select the three most dominant soil mapping units in the area remaining from step A to C. Step D: final area for yield gap assessment in Machakos and Makueni counties.*

The protocol to delineate most dominant maize cropping areas in Machakos and Makueni counties was applied to all six soil datasets. This resulted in the selection of different areas (Fig. 2.7). The selected areas were especially different between discrete and continuous maps. However, the differences are also caused by the extent of the dataset. The Local DSM (Fig. 2.7D) and FURP (Fig. 2.7F) have a smaller extent than e.g. ISRIC-WISE. The AfSIS dataset had some areas with missing data. In S-World (Fig. 2.7B) some areas with a soil depth smaller than 60 cm were excluded. The overlap in delineation with the ISRIC-WISE dataset was for FURP 86% and for the Local DSM 41%. KenSOTER, AfSIS and S-World showed an overlap of 57%, 56% and

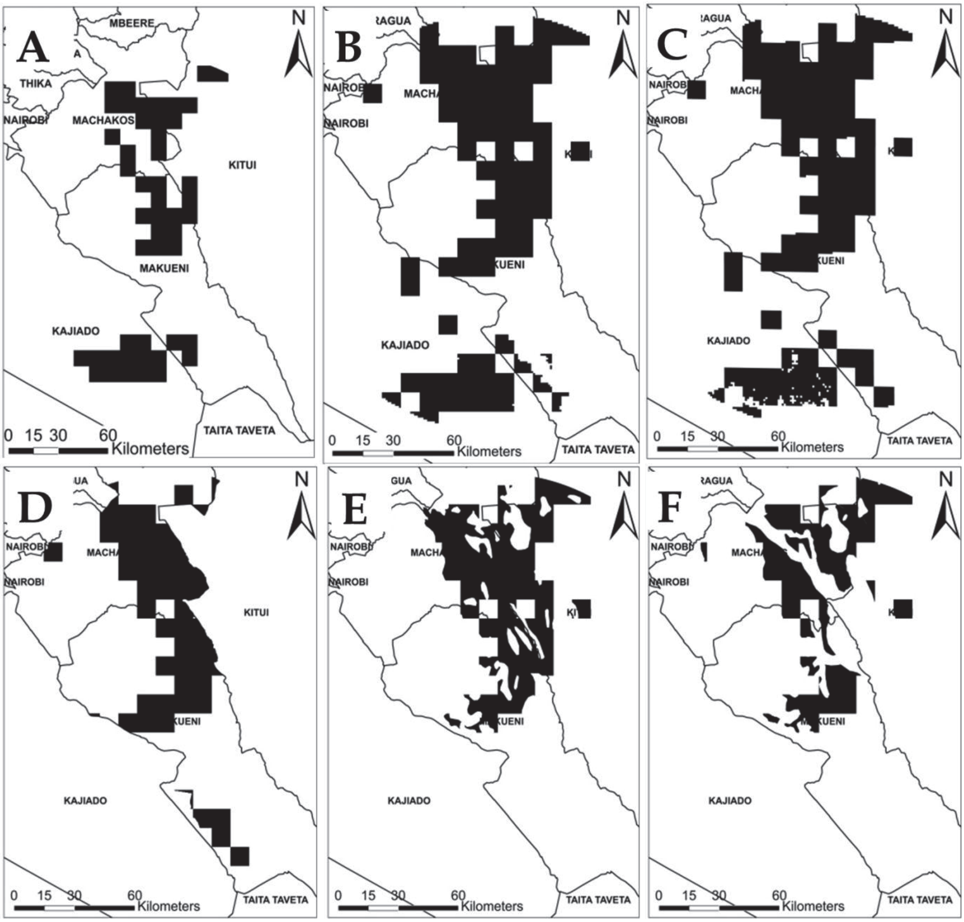


Figure 2.7. *Protocol to select most dominant maize cropping areas in Machakos and Makueni counties applied to the six soil datasets. A: ISRIC-WISE, B: S-World, C: AfSIS, D: Local DSM, E: KenSOTER, F: FURP.*

57% respectively. In this study, the choice of the soil dataset had influence on the delineation of maize cropping area for yield gap analysis.

2.3.3.3 Impact on simulated water-limited maize yields

Comparing the six soil datasets showed already differences in soil characteristics (Table 2.5). When the protocol for selecting model input data for the RLUA was applied to the six soil datasets, the input data also showed differences in input parameters (Table 2.9). Not only the input data differed, but also the procedure for selecting soil input data differed between discrete and continuous datasets. Datasets based on DSM had an average FC between 0.28 cm^3/cm^3 and 0.38 cm^3/cm^3, while datasets based on discrete mapping units showed less variability (0.18 cm^3/cm^3–0.23 cm^3/cm^3). The GYGA project assumed for all soil types the same default values for WP and SP, while these values differ per soil type. The FURP dataset was the only dataset that measured field capacity in maize fields (0.23 cm^3/cm^3) instead of estimating it from a pedotransfer function. In the datasets where the bottom soil layer was unknown, the maximum rooting depth was assumed to be 100 cm. The area is hilly, but the GYGA project discarded areas steeper than 10%. During the field survey, soil depth was not everywhere 100 cm and areas steeper than 10% were also cultivated. Restricting ourselves to the protocol of the GYGA project, the runoff factors varied between 0% and 26.7%.

Table 2.9. *The average, assumed or range of soil input parameters and the standard deviation (in brackets) for the crop growth simulation model WOFOST for six soil datasets.*

Dataset	**Max. rooting depth (cm)**	**Run-off factor (%)**	**FC[a] (cm^3/cm^3)**
ISRIC-WISE	100	0-6.7	0.19 (0.00)
S-World	109 (18)	0-6.7	0.38 (0.06)
AfSIS	100	0-26.7	0.28 (0.01)
Local DSM	100	0-6.7	0.29 (0.00)
KenSOTER	100	0	0.18 (0.01)
FURP	100	0-6.7	0.23 (0.05)

[a] Field capacity

From the input data, the FC varied most between the datasets. For the six different delineated cropping areas the average simulated water-limited maize yields and the standard deviations were estimated (Fig. 2.8). Crop failure took place in some of the years (2009 and 2012). In 2009, the rainfall in the cropping season was very low and in 2012 the rainfall came late in the cropping season. For some years, the different datasets showed yield differences of more than 4 tons/ha (2007), while in another year the yield difference was less than 2 tons/ha (2010). The effect of the chosen dataset on simulated water-limited maize yields mattered in some years more than in others. This did not depend on the rainfall amount, because years with equal

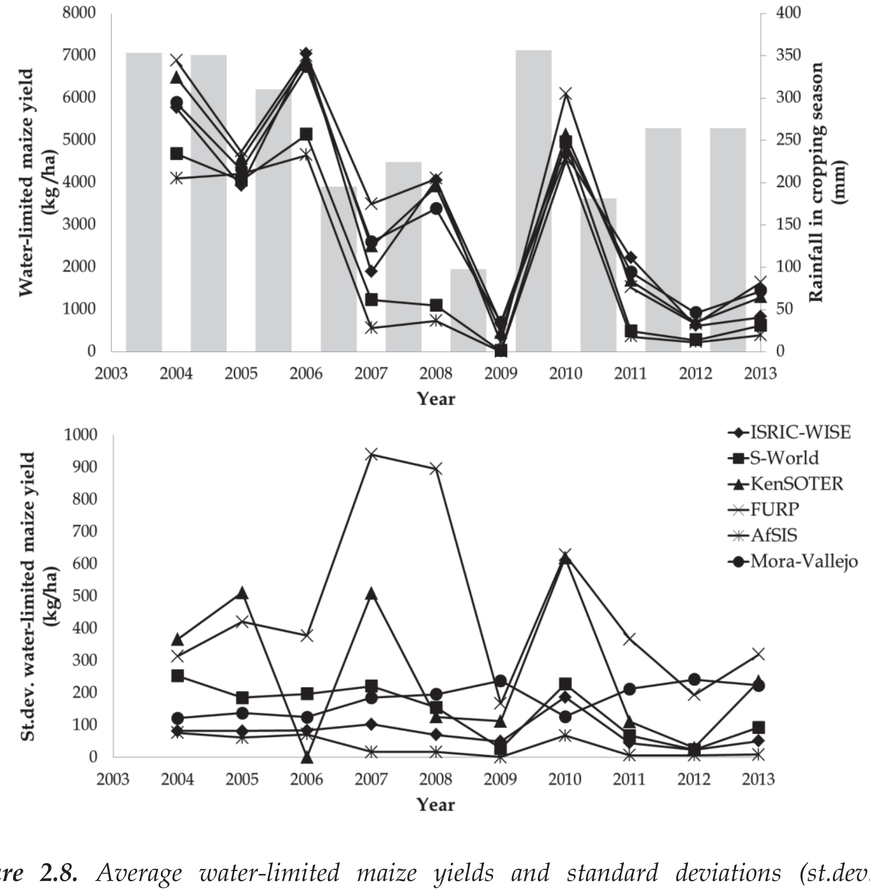


Figure 2.8. Average water-limited maize yields and standard deviations (st.dev.) for Machakos and Makueni counties. The crop growth simulation model WOFOST runs for six soil datasets from 2004 to 2013.

amounts of rainfall (e.g., 2004 and 2005, 2007 and 2012) also showed large yield differences between the datasets. The rainfall distribution differed per year and this explained the differences in yields between datasets. Water-limited maize yield was especially influenced by the rainfall at the start of the growing season. When there was rainfall throughout the growing season the difference in yields between the datasets was small. When there were days with and without water shortage in the growing season, the difference in yields between the datasets was largest.

2.4 General discussion

2.4.1 Challenges with soil data in RLUA

In this study, we presented some clear challenges with soil data in RLUA. The six soil datasets showed generalization and symbolization to highlight information and to suppress detail of lower priority (Monmonier, 1996). The information and level of detail different datasets have to provide, changed over the last decades (Hartemink and Sonneveld, 2013) and also depends on the type of RLUA.

The first challenge is the difference in soil characteristics between soil datasets. Except the Local DSM and FURP, all datasets used in this study were derivatives of other soil datasets (Fig. 2.9). The datasets are direct or indirect derivatives of the Soil Map of the World 1:5 M (FAO/Unesco, 1971-1981) and the Exploratory Soil Map and Agro-Climatic Zone Map of Kenya (Sombroek et al., 1982). This study showed that we should not only rely on available soil data, but also collect new soil data to validate the dataset and to test assumptions. While some datasets were derived from the same soil data, large differences occurred between soil characteristics in different soil datasets. In comparison to environmental models, soil datasets are hardly compared before a decision on which dataset to use for RLUA is made (e.g. Smith et al., 1997; Asseng et al., 2013). During the second phase of the GYGA project the ISRIC-WISE dataset was replaced by the AfSIS-GYGA dataset[1], without comparing the datasets beforehand.

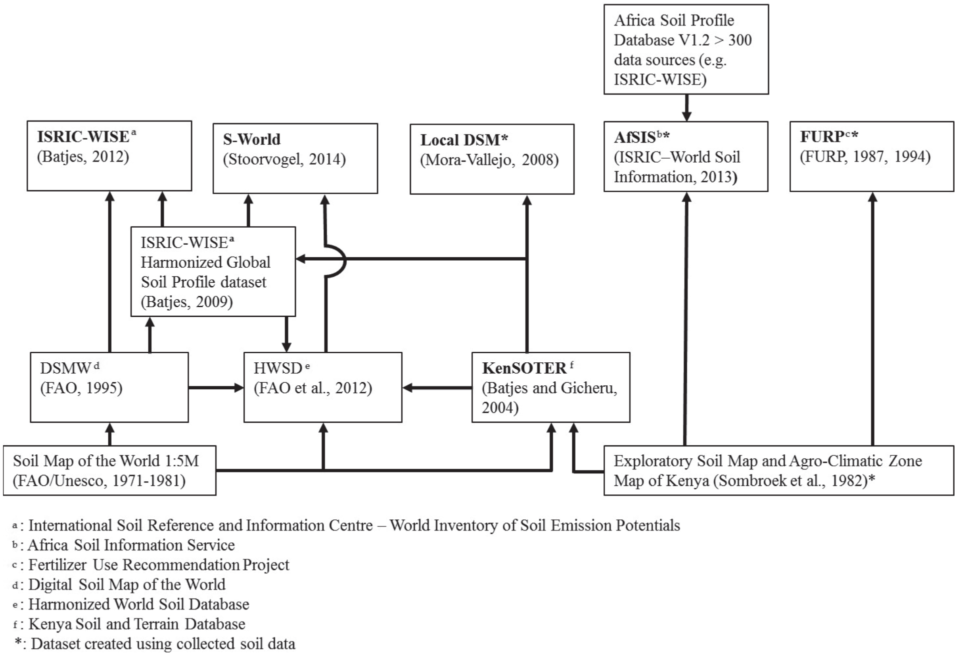


Figure 2.9. *The interrelations and origin of different soil datasets.*

The second challenge is related to the assumptions made to derive soil datasets. For example, some datasets do not consider land use and land management influencing soil characteristics, while we found significant differences in the field data. The third challenge with soil data in RLUA results from the combination of the first two challenges. As shown in this study, differences between datasets and the assumptions soil datasets underlie have consequences for the results of RLUA. For example, the model simulation exercise showed more than 4 tons/ha yield difference in some years by using different soil datasets. Results of studies on RLUA are often used for policy intervention (Bhatta and Aggarwal, 2015), but how reliable are policies derived from results that show such a difference? It is important to understand the differences and backgrounds of available datasets, the effect of assumptions and the sensitivity of the model parameters.

2.4.2 Challenges for users of soil datasets

The decision on which data to use for a RLUA is a difficult choice. In the land use analysis of Grassini et al. (2015) some challenges with soil, weather, crop management and actual yield data were noted. For example, soil data were often limited to the topsoil which caused problems in estimating rooting depth, a parameter required for their study (Grassini et al., 2015). Another problem Grassini et al. (2015) noted was the lack of actual measurements on soil water retention limits which forced them to use pedotransfer functions or default values. For Kenya, they selected the ISRIC-WISE dataset, because this dataset included nearly all required soil data on a global scale and the dataset is freely available.

Soil datasets are available and applicable, as this study showed. However, not all datasets are operational from an application point of view. Selecting soil input data for the RLUA is often based on pragmatic decisions. Soil datasets need to meet the scale and the data requirements of the RLUA. Some datasets that are not established from standardized protocols are difficult to understand for non-geoscientists. This makes it difficult to identify the best soil dataset. Another crucial point that hampers the decision is the low or unknown quality of soil datasets and the fact that datasets describe the quality differently. The quality of polygon-based soil maps is measured by the purity of mapping units in terms of equal classification and the variance of soil characteristics within mapping units (Bishop et al., 2001). The purity values are in general 70% to 80% (Bishop et al., 2001). For example, in Steur and Heijink (1991) the quality is indicated as the occurrence of different soil types that were not indicated in the mapping unit. A mapping unit should not contain more than 30% different soil types (Steur and Heijink, 1991). Most digital soil maps are not validated (Grunwald, 2009) and validation is essential to give an estimation of the quality of the map. The variance of prediction error is often used as quality indicator for digital soil maps (Bishop et al., 2001). However, the variance of prediction error depends on the distance between sampling locations. Higher sampling densities result in lower variance of prediction errors (Stein et al., 1989). The soil characteristic maps of AfSIS use, for example, the variance of prediction error and have a goodness-of-fit between 18% and 48% for different soil characteristics (ISRIC—World Soil Information, 2013). Different validation methods are available and should be used. The best method to obtain unbiased and valid estimates of the map quality is to obtain an independent dataset by probability sampling (Stehman, 1999). Less preferred are validations methods such as data-splitting and cross-validation, because these methods use biased datasets (Brus et al., 2011). The quality of the Local DSM was estimated using cross-validation. This validation method was required because of the limited number of soil samples. The cross-validation used by Mora-Vallejo et al. (2008) was based on clusters, which means that short distance variability was not included. The cross-validation of Mora-Vallejo et al. (2008) resulted in an explained variance of 18% for soil organic carbon and 37% for clay. In digital soil maps based on regression kriging the explained variance is often low, e.g. Balkovič et al. (2013); Hengl et al. (2004).

As postulated above, there is a strong need to validate soil datasets. If datasets lack validation, it is unknown which dataset is correct. In scenario (e.g., Goubanova and Li, 2007) or modelling studies (e.g., Rötter et al., 2011) the problem of lacking validation is solved by the approach of multi-data or multi-models using ensemble runs. Ensemble runs conduct multiple predictions using slightly different conditions. A global project that successfully applied multi-model ensembles is the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Elliott et al., 2015). As long as validations of soil datasets are absent and validation of RLUA and model input data are difficult, the use of multiple soil datasets (e.g. using ensemble runs) can increase the robustness of the soil data in RLUA.

2.4.3 Challenges for producers of soil datasets

For the case study six soil datasets were already available. The GYGA project selected initially the ISRIC-WISE dataset as it seemed to be the most operational dataset for the application. However, there are some challenges for producers of soil datasets to make available soil data or derivatives of available soil data more operational for studies on RLUA. One of these challenges is to validate soil datasets with an independent soil dataset using probability sampling. Another challenge is to bridge the gap between the available and required soil data. For example, soil datasets need to be combined to get information of the entire soil profile (e.g. Liu et al., 2007; Fischer et al., 2002). More demand-driven soil data supply would improve the functionality of the maps (Bacic, 2003). However, responding to the demand of RLUA is rather difficult, because there is large variety in RLUA methods and in the level of required detail. Nowadays, datasets that describe spatial variability in a continuous way (e.g. DSM) are more preferred for studies on RLUA than datasets that show spatial variability in a discrete way (e.g. CSS). Instead of pointing only to the producers of soil datasets, the producers of environmental models should keep in mind the availability and limitations of available soil data.

2.5. Conclusions

This study showed large differences between soil datasets in terms of soil characteristics (measured and derived) and in terms of assumptions that underpin the different datasets. These differences affect RLUA. Selecting a soil dataset that meets the data requirements for RLUA, often results in the selection of a soil dataset based on pragmatic decisions. Hence, the choice on which soil dataset to use for RLUA needs to be tailored to the aim of the RLUA. Main challenges with soil data in RLUA are: i) understand the assumptions in soil datasets, ii) create soil datasets that meet the requirements for regional land use analysis, iii) not only rely on available soil data but also collect new soil data and iv) validate soil datasets.

Appendix A

Samples for the field survey were tested on quantitative soil characteristics by proximal sensors. 19 samples were analysed in duplicate in the laboratory on soil texture and 19 samples were analysed on nitrate (N) and carbon (C) content.

The duplicate samples had correlation coefficient 0.82 for clay and sand percentage. Silt was hardly present, 1.9% (st.dev. = 1.7%), and had a low correlation coefficient ($r^2 = 0.07$). The textures measured by the turbidity metre had correlation coefficients of 0.43 for sand and 0.59 for clay content. The turbidity is converted to texture percentage by the statistical backward procedure. This resulted in the following equations: $Sand(\%) = 108.73 + 0.00024 * (\mathrm{ST} * \mathrm{LT}) - 1.91 * \mathrm{ST}^{0.5}$

$$Clay\ (\%) = -8.77 - 0.00022 * (\mathrm{ST} * \mathrm{LT}) + 1.75 * \mathrm{ST}^{0.5}$$

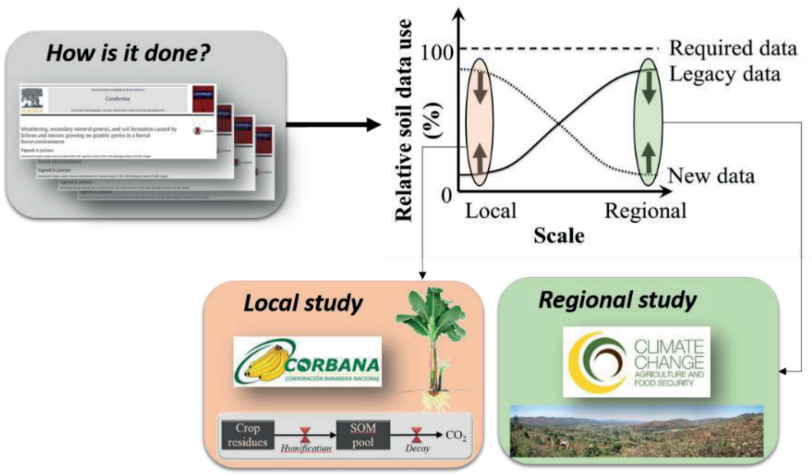
where, ST is the turbidity after stabilization of the soil-water solution (ratio 1:203) for 40 s and LT is the turbidity after stabilization of a soil-water solution for 1 hour.

Chapter 3

Combining new and available soil data to enrich local to regional land use analyses

Highlights

- Land use analyses hardly combine new and available soil data.
- In general, local land use studies tent to use new soil data, whereas regional studies tend to use available soil data.
- New soil data can contribute to regional land use analyses.
- Available soil data can contribute to local land use analyses.
- Clearinghouses on soil datasets can be used to make it easier to combine new and available soil data.



Slightly modified: Hendriks, C.M.J., Stoorvogel, J.J., Lutz, F. and Claessens, L.

Submitted to Geoderma

3.1 Introduction

Sustainable agricultural development gained increased attention, especially after the acceptance of the UN Sustainable Development of Goals in 2015. Land use analyses can be used to evaluate and quantify sustainable agricultural development (Hartemink et al., 2001). The type of questions that land use analyses need to answer changed rapidly over recent decades. Nowadays, many studies make use of e.g., quantitative simulation models. Available soil data were mainly collected for other applications and, as a result, these available soil data are not suitable for new land use analyses. For example, many soil surveys were carried out to support qualitative land evaluation (FAO, 1976). However, these surveys do not provide the necessary data required for crop growth simulation models or environmental impact models. In literature, available soil data are criticized as: (i) being dominantly qualitative, (ii) being outdated, (iii) being not spatially continuous, (iv) being only available at coarse scales, (v) being inconsistent, and (vi) lacking quality assessments (e.g., Sanchez et al., 2009; Heuvelink, 1998; Renschler and Harbor, 2002; Hengl et al., 2014). Alternatively, soil data can be obtained by new surveys. However, collecting new soil data also has its limitations: (i) it is expensive and time-consuming to collect and analyse new soil data, (ii) the soil data may not be collected spatially exhaustive, (iii) logistical issues are often faced during data collection (e.g., accessibility), (iv) the number of soil data that can be collected are often a trade-off between quality and quantity, and (v) new soil data are collected at one moment in time or over a relative short time span.

This study evaluates how land use studies use available data and/or new soil data. In addition, the potential of combining available soil data and new soil data to overcome some of the limitations of available soil data and new soil data is explored. On one hand, available soil data provide essential insights in dominant landscape units that can help to target soil data collection (e.g., Yang et al., 2011). On the other hand, new soil data can overcome some of the limitations that are hampering the use of available soil data (e.g., Kempen et al., 2009).

First, a literature review on 120 studies published in Geoderma is reviewed to analyse whether available soil data and new soil data collection are combined. Second, to analyse the potential for combining available soil data and new soil data, two case studies were carried out. These case studies are implemented to analyse the value of available data at the local scale (case 2) and the value of new soil data at the regional scale (case 3). Finally, we discuss our findings and draw conclusions how future land use studies can efficiently make use of a combination of available data and new soil data.

3.2 Literature review: how land use analyses obtain soil data

3.2.1 Introduction

We hypothesize that the relative use of new soil data and available soil data in land use analyses differ between scale levels as illustrated in the conceptual framework of Figure 3.1. Two scale levels are identified: i) local studies at, for example, fields, farms and villages, and ii) regional studies like for example, watersheds, landscapes and countries (FAO, 1993). On one hand, we expect local studies to obtain soil data by collecting new soil data through experiments, soil sampling, interviews and visual observations. On the other hand, we expect regional studies to use available soil data, including conventional soil surveys, soil profile data and digital soil maps. The decision of using either available soil data or new soil data depends on the soil data

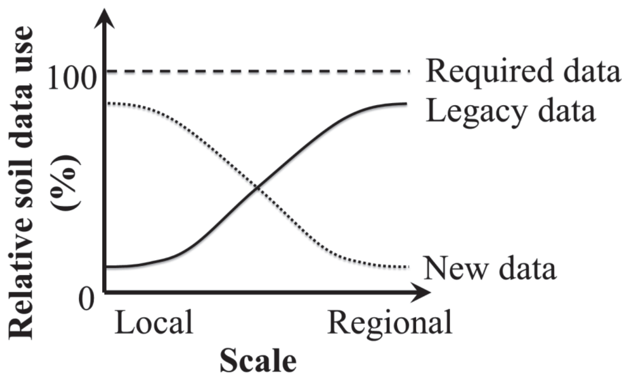


Figure 3.1. *Conceptual framework where we hypothesize that the relative use of new soil data and available soil data in land use analyses differ between scale levels.*

that are available, as well as on the type of land use analysis. The framework assumes that the relative use of new soil data and available soil data meet the soil data requirements for 100%.

There are several reasons why we expect regional studies to rely on available soil data. Peer-reviewed soil datasets have become readily available through the internet and various data portals now provide easy access like the European Digital Archive on Soil Maps of the World (Panagos et al., 2011) and the Web Soil Survey and the Soil Data Access Web Service of the National Resources Conservation Service of the US Department of Agriculture[2]. At the same time, collecting new soil datasets at the regional scale can be expensive. Local studies can, and sometimes have to, rely on the collection of new soil data. Available soil data often only include exploratory soil surveys at e.g., the national scale, which provide limited insight in local soil variability. For many local studies, these coarse scale levels do not provide enough detail.

3.2.2 Research implementation

To demonstrate the conceptual framework, 120 studies published in Geoderma were reviewed. The studies were analysed on: (i) the scale level, (ii) proportion of new soil data collected, (iii) proportion of available soil data used, and (iv) the use of auxiliary information. To analyse whether the use of soil data changed over time, 60 studies published between 1967 and 1971 were analysed and 60 studies in recent journal issues of 2015 and 2016 were analysed. Five different scale levels were identified ranging from plot to national. The scale of the study was derived from the research objective and conclusions. An overview of the results of the analysed literature is given in Figure 3.2. The studies of 1967-1971 collected new soil data and rarely used available soil data. However, in recent studies from 2015-2016 the use of available soil data increased and there is a clear trend that the relative importance of available data increases at higher scale levels. The results confirm the original hypothesis that local studies predominantly use new soil data, whereas regional studies predominantly used available soil data. This is particularly true for the recent studies. The studies between 1967 and 1971 only made limited use of available data,

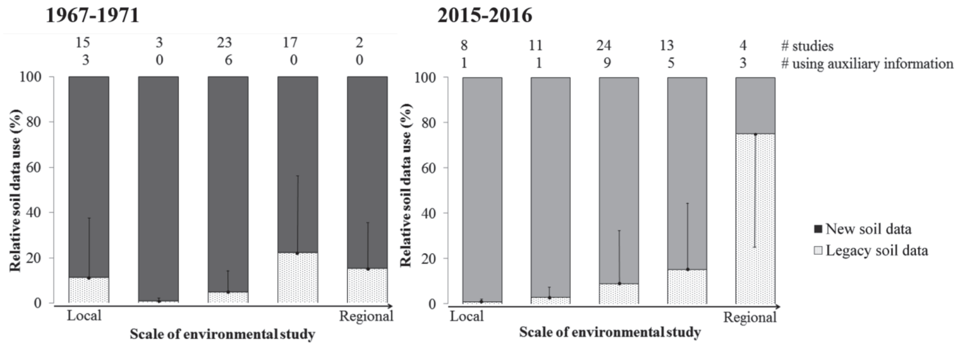


Figure 3.2. *The relative use, including the standard deviation (line), of field data and available soil data in old (1967-1971) and recent (2015-2016) studies published in Geoderma. Five scale levels were distinguished ranging from local to regional. The number of studies analysed at a particular scale level and using auxiliary information are mentioned above the bars.*

probably because of its limited availability. From the analysed literature, only 11 out of 120 studies combined new soil data and available soil data. In these studies, the relative use of available soil data and new soil data was over 30%. In general, more collected new soil data rather than using available soil data. The use of auxiliary information for land use analyses doubled in the 2015-2016 studies compared to the 1967-1971 studies. This is probably caused by the increase in availability of spatial exhaustive, high resolution and high quality auxiliary information.

The relative soil data use was estimated semi-quantitatively by the authors. To test the reliability of the analysis, twenty studies of 1967-1971 and twenty studies of 2015-2016 were analysed independently in duplo. The relative use of new soil data and available soil data were not significantly different ($p<0.05$) between the duplos. The estimation of the relative use of new soil data had correlation coefficients of 0.94 in the studies of 1967-1971 and 0.90 in the studies of 2015-2016. The estimation of the relative use of available soil data had correlation coefficients of 0.95 in the studies of 1967-1971 and 0.90 in the studies of 2015-2016. The scale of the study was most difficult to analyse, because the goal of the study frequently had a different scale compared to the analysis. Especially in studies of 2015-2016, the scale levels in the duplos differed one, or, in three cases, two scale levels.

3.2.3 Discussion

The case study confirms the initial hypothesis that regional studies rely more on available soil data compared to local studies. An example of a local study that collected new soil data, is a study that investigated and compared the weathering, secondary mineral-synthesising and soil-forming activities of different species of lichens and mosses (Jackson, 2015). The study required detailed and specific soil data that only could be obtained by collecting new soil data. An example at the regional scale, Wilford et al. (2015) used the National Geochemical Survey of Australia (NGSA, De Caritat and Cooper, 2011) to model the abundance of soil calcium carbonate across Australia using geochemical survey data and environmental predictors. The soil data requirements were available in the NGSA and therefore this study could use available soil data. The use of available soil data for land use analysis is a relatively new phenomenon and there are also some exceptions to the general trends. De Vos et al. (2015) studied at regional scale soil organic carbon stocks in forest floors and in mineral and peat forest soils. Despite the scale level of this study, soil data on the carbon concentration, bulk density, coarse fragments and effective soil depth were collected at almost 5000 locations in Europe. The review did not reveal any local studies that did not collect new soil data. Few of the older studies made use of available soil data and, as a result, there were minor differences between the scale levels for the older studies.

3.3 Case study 1: combining available and new soil data at local scale

3.3.1 Introduction

Costa Rica is one of the main banana exporters with one of the highest productions worldwide (≈ 50t bananas/ha/yr; FAO, 2016). However, because of intensive use of agro-chemicals and large monoculture plantations, the Costa Rican banana sector is under pressure to produce bananas in a more sustainable way. In a wide range of initiatives, the sector aims to make the production more environmentally friendly (Stoorvogel et al., 2004). The production of bananas coincides with the production of large quantities of crop residues. The crop residues are left on the field (stems and leaves) or returned to the field in a later stage (mainly bunch stalks from the packing plant). As such, the crop residues recycle large amounts of nutrients to the soils, and maintain soil organic matter stocks. However, with the increasing attention for biofuels and other secondary products, the crop residues of the banana plants are also seen as a valuable asset of raw material. Crop residues can be used in various ways like fibre for paper and biomass for biofuel. A recent development is the use of banana fibres for the production of ecologically friendly pallets by the Dutch Limited company Yellow Pallet B.V (www.yellow-pallet.com). For a proper business plan, it was important to know whether crop residues can be removed from banana plantations while sustaining soil fertility and crop productivity. The location specific repercussions for soil management had to be analysed and included in the business plan. The study was implemented on two banana plantations in the humid lowlands in the northeast of Costa Rica: the Banana Tica plantation (10°20′10″ N, 83°40′38″ W) with Eutropepts and Dystric Vitrudands the San Pablo plantation (10°6′45″ N, 83°22′53″ W) with Eutropepts and Humitropepts (soil classifications based on the Soil Survey Staff (1992) by Wielemaker and Vogel, 1993).

3.3.2 Research implementation

The long-term effects of management changes on soil organic matter stocks in a perennial crop can be analysed in different ways. One could do long-term experiments, but in the case of Yellow Pallet, the available resources were limited and commercial interests required answers within a year. Alternatively, various modelling approaches are available that one could make use of since soil organic matter dynamics have been studied intensively (Shibu et al., 2006). However, although one could rely on existing studies and data, it became apparent that most of these studies did not focus on the banana crop. As a result, it was decided to combine available data with a simple soil organic matter model and use field studies to collect very specific data for the banana crop. The organic matter model is described in

Figure 3.3. It deals with a single soil organic matter pool, organic matter inputs through crop residues, and a decay of soil organic matter through mineralization. Two conversion factors are important in the model, the humification rate of crop residues into the soil organic matter pool, and the decay rate from soil organic matter towards CO_2, i.e., the mineralization rate.

Figure 3.3. *The soil organic matter model simulates changes in the soil organic matter stock of a banana plantation.*

Data for model calibration were collected from different sources:

- Available soil data showed that soil organic matter contents in banana plantations under current management are stable (Fig. 3.4).
- Crop residue production in banana plantation was measured in the field.
- The soil organic matter pool was measured at different locations in the field.
- A field experiment was done in which soil organic matter contents were monitored during one year on plots that did not receive any crop residues and on plots that received normal crop residues.
- Literature data provided insight in specific elements of the system like crop residue production (Vargas and Flores, 1995) and decomposition (Geissen et al., 2009).

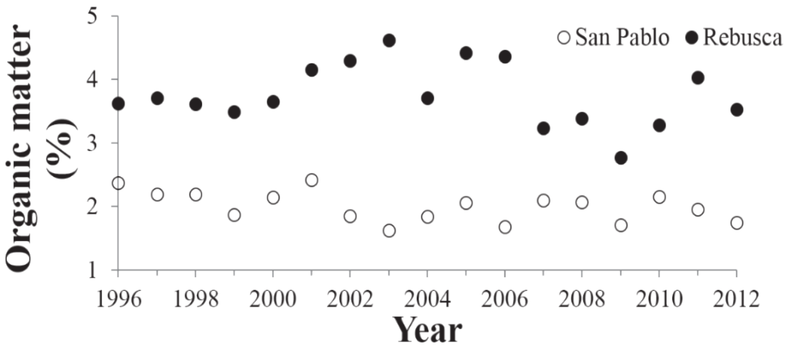


Figure 3.4. *Long-term soil organic matter contents in two Costa Rican banana plantations in San Pablo (sedimentary soils) and La Rebusca (volcanic soils), annual measurements at the end of the year.*

With the above information, the model was calibrated under normal management conditions in such a way that a steady state was obtained. Measurements showed that the banana plantation produced 26.2 t/ha of crop residues (dry weight). The soil organic matter pool was found to be 121 t/ha. The calibration resulted that only 11.2% of the crop residues ends up in the soil organic matter pool through humification and that annually 2.4% of the soil organic matter pool is mineralized and lost. Subsequently, the model was run for a 20 year period under a situation in which 75% of the crop residues are removed. It was assumed that production did not decline and that potentially reductions are compensated by proper fertilizer management. The results show that in a 20 year period, the soil SOM stock would be reduced by 10% to 110 t/ha.

3.3.3 Discussion

The results of the study provided a quick answer to the questions being asked by Yellow Pallet. The expected changes in soil organic matter stocks can be interpreted to assess the required changes in soil management and the repercussions in terms of costs for soil fertility maintenance. The case study is a good example of how available and new soil data complement each other. The analysis was facilitated by the soil organic matter model that integrated all the data. Currently, Costa Rican banana growers do not remove crop residues in banana plantations. Consequently, data on the impact of this management strategy simply were not available. Long-term trials were no option due to Yellow Pallet's urgent need for a business plan. Literature and models lacked basic knowledge on the banana production system. Therefore, it was impossible just to carry out the data analysis without data collection. The combination of literature, field data, and models proved to be an efficient procedure to provide the required answers for the company.

3.4 Case study 2: combining available and new soil data at regional scale

3.4.1 Introduction

Between 1981 and 2002, maize yields declined globally due to climate change and land degradation (Lobell and Field, 2007). The CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) promotes the adoption of Climate-Smart Agricultural (CSA) practices (e.g., terraces, stone bunds, dams, intercropping, integrated soil fertility management) which aims to sustainably increase productivity, adapt and build resilience to climate change, and reduce greenhouse gas emissions. Different studies demonstrated that CSA practices increased the productivity under current climate conditions (Paustian et al., 2016), but whether this increased productivity withstands under different climate conditions is often unknown. In this regional study, the effects of CSA practices on potential rain-fed maize yields under different climate scenarios were analysed. The study focussed on the semi-arid Machakos and Makueni counties (Kenya) covering approximately 8,000 km^2. Soils are classified as Rhodic Ferrasols, Chromic Cambisols, Eutric Vertisols, Haplic Lixisols and Chromic Luvisols. In the region, intercropping and terracing are already widely adopted as CSA practices.

3.4.2 Research implementation

The effect of CSA practices on potential water-limited maize yields require soil data that were taken at agricultural fields. The Fertilizer Use Recommendation Program (FURP, 1987; FURP, 1994) carried out intensive agronomic experiments. The data include i) a general description on the land use and land management in the counties, ii) management data on maize, and soil profile descriptions (including physical and chemical analyses of representative locations, and iii) crop response. Maize cultivation took already place for over 50 years in which organic fertilizer is applied resulting in actual average maize yields of approximately 2.7 ton/ha. The effect of CSA practices were evaluated using a crop-growth simulation model. Potential water-limited maize yields were simulated using the WOFOST (World Food Studies) crop-growth simulation model (Control Centre version 2.1; Boogaard et al., 2014). The model requires soil data on water holding capacity (WHC), the run-off factor and the maximum rooting depth. The WHC was derived from the pedotransfer functions of Saxton and Rawls (2006). This pedotransfer function requires sand, clay and organic matter content. The maximum rooting depth was fixed at 100 cm, except when in the field other limitations were observed. Additional new soil data on soil texture and soil organic matter were collected on fields with and without adoption of CSA practices. The new soil data were collected in pairs. In total, 11 pairs were sampled to compare the effect of terracing and 13 pairs were sampled to compare the effect of intercropping. A composite sample of the topsoil (0-20 cm) and a single sample of the subsoil (50-60 cm) were taken. Terracing resulted in a run-off factor of 0%. However, it did not directly result in an increase in soil organic matter content (on average -0.1%) and finer soil textures (on average -1.6%) as expected. This can be caused by the soil displacement from topsoil and subsoil when terraces were made. Intercropping resulted in a slight increase in carbon (on average 0.1%) content and finer soil textures (on average 1.3%). Differences between terraced and non-terraced fields and between intercropped and mono-cropped fields were analysed. Daily weather data were available from 2004 till 2012. From these data, a growing season (Oct-Dec) with an average amount of rainfall (205 mm), the wettest season (407 mm) and the driest season (131 mm) were selected. For these three seasons, four climate scenarios (RCP2.6, 4.5, 6.0, 8.5) were derived from WorldClim Version 1.4[3]. The methodology on how these scenarios were derived is described by Hijmans et al. (2005). The effects of terracing and intercropping on potential water-limited maize yields were simulated for the three seasons times four climate scenarios. The model was run with different climate scenarios to assess the effects of climate change. The results are presented in Table 3.1. Terracing seems to be a suitable CSA practice as potential maize yields increase significantly ($P<0.05$) under almost all future climate scenarios (except for the relatively wet season). This is caused by the soil and water conservation aspects of terracing in drier and average seasons. Intercropping, on the other hand, did not significantly increase the simulated water-limited maize yields. This is because the potential beneficial effects of intercropping on soil nutrients (N sequestration), is not reflected in the water-limited maize yield.

Table 3.1. *The effect of terracing and intercropping on potential water-limited maize yields For a relatively dry (DS), average (AS) and wet season (WS) and for four climate scenarios (CS).*

	Precipitation (mm/season)	**Mean min.T (°C)**	**Mean max. T (°C)**	**Effect of terracing (%)**	**Effect of intercropping (%)**
DS	131	20.5	28.0	50.4	2.0
CS 1	136	22.1	28.6	9.7	-1.6
CS 2	148	22.5	29.3	8.5	0.0
CS3	159	22.5	28.9	8.4	0.6
CS4	97	23.2	29.5	13.7	-3.0
AS	205	18.2	27.9	4.0	1.7
CS 1	220	19.6	28.6	4.0	0.5
CS 2	204	20.0	29.2	1.0	-0.2
CS3	240	20.0	28.9	2.1	1.3
CS4	179	20.6	29.4	3.0	-1.7
WS	407	18.5	28.2	3.8	0.9
CS 1	645	20.0	28.9	-0.6	-1.2
CS 2	464	20.3	29.6	0.0	0.1
CS3	484	20.3	29.2	0.0	0.2
CS4	392	20.9	29.7	0.0	0.0

Data for running the model were collected from different sources:

- The FURP dataset was used to obtain general information on the soil properties and agricultural practices in the study area.
- Pedotransfer function of Saxton and Rawls (2006) was used to derive the WHC.
- Soil data were collected in pairs to compare the effect of CSA strategies on water-limited maize yield.

3.4.3 Discussion

Not all current CSA practices will sustain increased productivity under future climate conditions. Alternative CSA practices need to be considered to increase the resilience towards a changing climate (e.g. water storage systems, agroforestry, conservation agriculture, etc.). By combining available soil data and new soil data we were able to analyse the effects of CSA practices for the Machakos and Makueni counties. Soil properties were affected by CSA practices. These effects could not have been analysed with available soil data alone. In the study area, studies on the effect of CSA practices were often focussing on individual farms. However, policies (e.g. governmental programs) are formulated based on the regional agricultural system instead of individual farms (Smit et al., 1996). It is therefore suggested to frame studies on CSA practices in a wider context by combining new soil data and available soil data. A good example is the study of Saiz et al. (2016), where new soil data were combined with the World Reference Base (IUSS Working Group WRB, 2015). When only available soil data were used for this study, soil data were generalized because of a lack in information on land use, land management and actual soil properties on agricultural fields. When only new soil data were used, the maximum rooting depth and the WHC had to be estimated in the field. The combination of pedotransfer functions, available soil data and field data proved to be efficient.

3.5 General discussion

3.5.1 Trends in obtaining soil data for land use analyses

The review showed that the number of studies that collect new soil data on a regional scale decreased. This trend was confirmed by Hartemink et al. (2001). They found that the number of land use analyses that collected new soil data decreased from 29% to 18% between 1970 and 1990. Two contradictions were found during the literature review: (i) regional studies increasingly use available soil data, while literature criticizes available soil data, and (ii) local studies still collect predominantly new soil data despite the increased availability of available soil data. Available soil data can contribute to local studies in different ways, whereas new soil data can contribute to regional studies (Fig. 3.5). Probably, there is also a relation between the scale, resolution or quality of the available soil data, and the number of new soil data collected. Soil maps at detailed scale and maps that are validated require less supplementary soil data.

Issues with available soil data could often be solved by collecting limited additional soil data. For example, outdated soil data can be updated by collecting some additional soil data on locations where most change is expected (Kempen et al., 2009; Yang et al., 2011) or the spatial variation could be described in more detail when available soil data and new soil data are combined (Song et al., 2016). New soil data could check assumptions that were made when soil datasets were established. For example, in a study of Pelegrino et al. (2016) the prediction of soil classes improved because the areas of uncertainty were identified using some additional new soil data. Additional new soil data could also contribute to the validation and verification of available soil data (Brus et al., 2011; Jones et al., 2005). Issues that are faced with new soil data can often be solved by using available soil data. Spatial patterns could for example be identified from available soil data to make new soil data collection more efficient. Many studies still use a random sampling design (Rodríguez Martín et al., 2016; Liu et al., 2006), while the number of observations could be reduced when the study area is divided in strata using the spatial patterns of available soil data. The

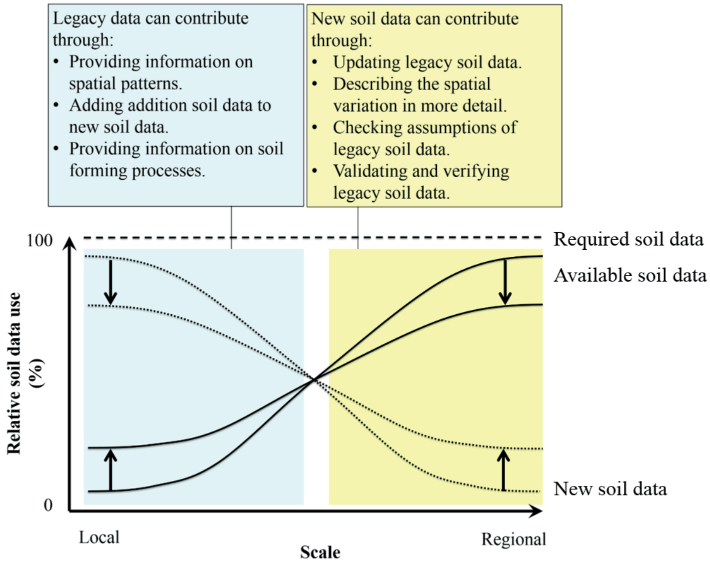


Figure 3.5. *Available soil data can contribute at local scale and new soil data can contribute at local scale.*

spatial prediction of soil properties could improve when available soil data are added to the new soil data (Song et al., 2016). Conventional soil surveys include information on soil forming processes which could be used for improving the predictions on soil properties or to obtain more functional soil properties.

3.5.2 How to combine available soil data and new soil data

It is not always feasible to collect soil data in the field exhaustively, but available soil data only do often not meet the data requirements. The local case study took advantage of available soil data by using available soil data on the mineralization rate. Collecting these data in the field would have been too expensive and time-consuming. The regional case study took advantage of new soil data by collecting additional data on land use and land management. These data were missing in available soil datasets. Some studies explicitly search for a combination of new soil data and available soil data. For example, Tarnocai et al. (2009) used available soil data to analyse soil organic carbon pools in the northern circumpolar permafrost region. However, because data on very deep carbon pools were missing, these data were collected in the field. Other studies combine new soil data and available soil data implicitly. In a study of Franco et al. (2016) new soil data were collected through a sampling design that was based on the spatial patterns of the available soil data.

As mentioned by Grassini et al. (2015), transparent, reproducible and robust guidelines are needed to obtain soil data for land use analysis studies. The inconsistency in terms of environmental input data, soil properties, quantitative methods, and evaluation, and validation strategies (Grunwald et al., 2011) makes it difficult to decide which soil data to use and which soil data to collect. To provide soil data in a transparent, reproducible and robust way, clearinghouses are essential. Through a clearinghouse, data can be searched, viewed, transferred, ordered, advertised, and disseminated. Clearinghouses can ease the identification of missing data and it can improve the access to the required data (Franco, 1992). Good examples of such a clearinghouse for soil data is the European Digital Archive on Soil Maps of the World (Panagos et al., 2011). In addition, standardized procedures for soil data collection could help users to interpret the datasets like the SOTER initiative (Igue et al., 2004 and Goyens et al., 2007), but also the older, still highly relevant, Soil Survey Manual (Soil Survey Staff, 1993). Finally, the development of larger harmonized datasets that brings together the existing data into new regional or global maps like the Harmonized World Soil Database (HWSD; FAO/IIASA/ISRIC/ISSCAS/JRC, 2012), SoilGrids (Hengl et al., 2014), S-World (Stoorvogel et al., 2017), and the global database with harmonised soil profile information for a World Inventory of Soil Emission Potentials (WISE; Batjes, 2009).

3.6 Conclusions

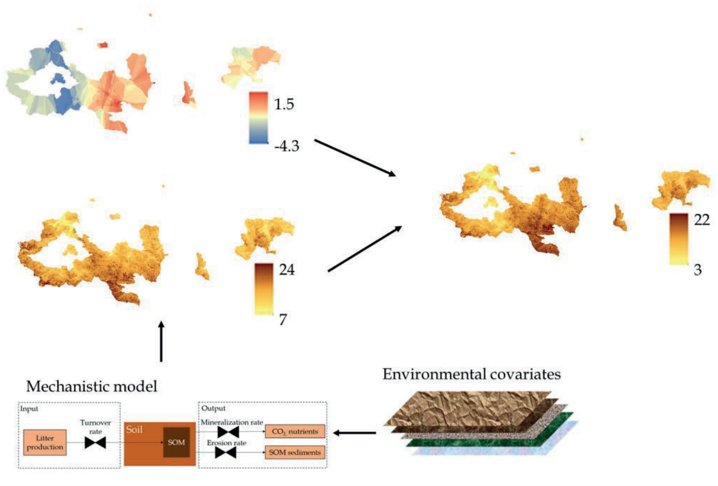
The number of land use analyses that combine available soil data and new soil data is low. Less than 10% of the studies in the literature review combined available soil data and new soil data. However, the case studies clearly showed the added value of combining available soil data and new soil data. Two contradictions can be concluded: i) regional studies rely dominantly on available soil data, while these data are criticised in literature and ii) local studies still rely dominantly on new soil data, while the number of available soil data increased. Awareness of these developments was raised in this study. Preconceptions on available soil data (e.g., the scale of the soil data is too coarse) and on new soil data (e.g., collecting new soil data is costly), should not determine the exclusion of one of both, because available soil data as well as new soil data provide complementary or supplementary information that can enrich the land use analysis. Studies need more often to consider a combination of available soil data and new soil data. To enrich the soil data for land use analyses, clearinghouses are recommended as an opportunity to obtain soil data and to identify missing data more easily.

Chapter 4

A mechanistic model for digital soil mapping to predict the soil organic matter content in nature areas

Highlights:

- Processes that influence the Soil Organic Matter (SOM) content are interrelated.
- The SOM content is difficult to predict using statistical models for Digital Soil Mapping (DSM).
- Mechanistic processes that influence a soil property are known and should be incorporated in DSM.
- SOM content predictions improve and the number of covariates reduces using a mechanistic model for DSM.



Based on: Hendriks, C.M.J., Stoorvogel, J.J., Álvarez-Martínez, J.M. and Claessens, L.

Submitted to European Journal of Soil Science

4.1. Introduction

Soil organic matter (SOM) is an important soil property that influences chemical and physical soil properties such as the nutrient and water availability, soil structure, aggregate stability and water holding capacity (WHC). SOM also determines different soil functions like the carbon sequestration capacity, the diversity and activity of soil organisms and the absorption and retention capacity of pollutants (Weil and Brady, 2016). These soil functions are essential for current land use analyses on e.g., climate change mitigation, agro-ecosystem functioning, soil health and habitat monitoring.

To meet the soil data requirements for current land use and land cover analyses, the need for spatially continuous soil data has increased over recent decades (Zimmermann et al., 2008). Digital Soil Mapping (DSM) is one of the mapping techniques that can be used to obtain these data (McBratney et al., 2003). Regression kriging is a commonly used DSM technique that consists of two steps. In the first step, a regression model is fitted between the observed soil property data and spatially exhaustive environmental covariates that represent the soil forming factors defined by Jenny (1941): climate, organisms, relief, parent material and time. The difference between the observed and predicted soil property is called a residual. In a second step, the spatial auto-correlation of the residuals is examined and interpolated when there is spatial auto-correlation. Different interpolation techniques can be used, e.g., Empirical Bayesian Kriging, ordinary kriging, simple kriging, universal kriging, anisotropic kriging, CoKriging. Combining the map that results from the regression model and the map of the interpolated residuals we can provide a digital soil map that improves from simple regression approaches.

Soil organic matter is the most complex and least understood component of soils (Magdoff and Weil, 2004). The SOM content is driven by the continuous admission of organic material and its transformation is caused by biological, chemical and physical factors (Kononova et al., 1966). The complex processes that influence the spatial variation in SOM content cannot always be explained by a statistical model. Especially at regional scale, the regression approach often results in poor predictions (e.g., Zhao et al., 2014; Dorji et al., 2014). Regression models only search for statistical relationships between the observed soil property and spatially exhaustive environmental covariates, while we have much knowledge on the mechanistic processes that influence the SOM content. The use of only statistical relationships can be the reason for the poor predictions of the SOM content. We assume that a mechanistic approach can improve its performance. This study analyses a mechanistic approach for DSM to predict the SOM content in nature areas.

The mechanistic model is developed using the knowledge we have on the mechanistic processes that influence the SOM content. This knowledge was collected from literature and from available dynamic soil models that describe associated C and N flows. The number of soil models has increased rapidly over recent decades (Campbell and Paustian, 2015; Shibu et al., 2015). These models can be static or dynamic and they are developed at different spatial scales (Manzoni and Porporato, 2009). Examples of regional mechanistic carbon cycle models are CENTURY (Parton et al., 1994), RothC (Coleman and Jenkinson, 2014) and Ecosys (Grant et al., 1995). These models simulate mechanistic processes of carbon between the Earth's spheres; biosphere, atmosphere, hydrosphere and lithosphere. Probably, the mechanistic processes cannot directly be applied in the model for DSM, because the processes often require variables that are not available spatially exhaustive. Spatially exhaustive environmental covariates can be used as proxy to describe a process. For example, the vegetation height can be used as proxy for the biomass production. Processes that cannot be explained by proxies or default values need to be excluded from the model. The study was carried out in the Natura 2000 areas of Cantabria region (Spain). Detailed soil data on the SOM content were required for these areas to estimate the status of habitats in the nature areas.

4.2 Materials and methods

4.2.1 Study area

This study focusses on the Natura 2000 areas of the Cantabrian region (43°20′N, 4°00′W NW) (Fig. 4.1). The Cantabria region in Spain has an Atlantic climate along the coast and an Alpine climate in the mountainous areas. At sea level, the mean

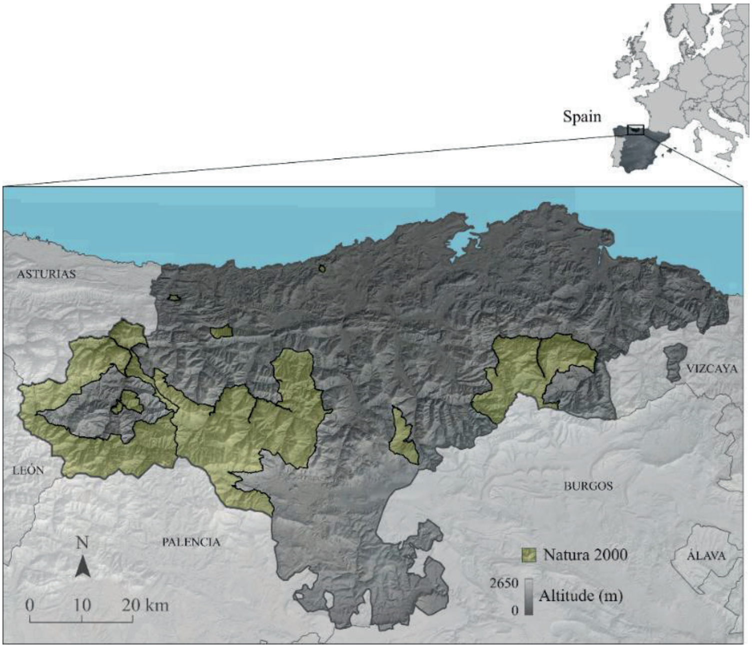


Figure 4.1. *The study was carried out in the Nature 2000 areas of Cantabria region (Spain).*

annual temperature is 15°C and at 2650m above sea level the mean annual temperature is 2°C. The mean annual precipitation has high pluviometry and ranges between 369 mm and 2369 mm. The landscape was formed by montane glaciation, periglacial phenomena, alluvial terraces and marine dynamics. This resulted in a hilly to mountainous landscape with steep slopes where erosion occurs. These steep slopes and the frequency of dry winds encourage anomalous wildfires in autumn and winter, mainly caused by local farmers for spring grazing. Geomorphological processes have formed a rich lithology in the area including shales, sandstone, limestone, conglomerates and slates. The most dominant soil types are the mollic, haplic, gleyic Solonetz, albic Luvisol, haplic Luvisol and the orthic Podzol (FAO/Unesco, 1981). The eutric Cambisol, dominates at sloping land (Gallardo et al., 2016).

The environmental heterogeneity of the area resulted in unique ecosystems. The area harbours a mix of temperate deciduous and sclerophyllous vegetation species, including beeches (*Fagus sylvatica*), oaks (*Quercus petraea, Q. robur*) and birches (*Betula spp*) in colder, wetter areas and other oak species (*Q. pyrenaica and Q. rotundifolia*) in warmer and dryer areas. Most of the mature forest was deforested over the past century for timber production and agriculture (Acton et al., 2013). Oppositely, the region is currently experiencing quick secondary succession because the area got a protected status and rural depopulation took place. The area is now recognized by mature forest, shrubs and abandoned pastures (Álvarez-Martínez et al. 2014). The abandoned pastures are now dominated by brambles (*Rubus* spp.), roses (*Rosa spp.*), hawthorn (*Crataegus monogyna*) and blackthorn (*Prunus spinosa*) (Álvarez-Martínez et al., 2017). In flatter alluvial terraces still agriculture takes place. At high altitudes, well-managed grasslands are used for extensive grazing.

4.2.2 Data collection

4.2.2.1 Environmental covariates

Many high resolution environmental covariates are available for the Cantabrian Mountains (Table 4.1). These covariates are used for defining the sampling scheme and as proxies in the mechanistic model. They can be categorized according to the five soil forming factors. Climatic data are obtained from Iberian Peninsula dataset of Ninyerola et al. (2007). The Continuous Geological Map of Spain is available from the Geological and Mining Institute of Spain (IGME) and can be used to obtain data on the parent material. The digital elevation model (DEM) and thetopographical layers were derived from LiDAR data at 5 m resolution obtained from the National Geographical Information Centre of Spain (CNIG, 2016). Data on land cover were obtained from the Landsat 8 Operational Land Imager (OLI) scene mosaic (Path 202, Row 30). Land cover data of 2013 to 2016 were analysed to composite cloud-free images (Álvarez-Martínez et al., 2017). From these data the Normalized Difference Vegetation Index (NDVI) (Rouse et al. 1973), the Normalized Difference Water Index (NDWI) and the Tasselled Cap (TC) transformation (Crist and Cicone 1984) were derived. Data on the vegetation height and canopy structure were derived from LiDAR data (Álvarez-Martínez et al., 2017). The environmental covariates obtained at 5m resolution were resampled to 30 meters using natural neighbour interpolation. The Mechanistic Digital Soil Map of the SOM content will be provided at 30m resolution.

Table 4.1. *Environmental covariates that are available for the Cantabria Region (Spain).*

Soil forming factor	Description	Variable	Unit	Code	Source	Scale/ Resolution
Climate	Mean and coefficient of variation 1981-2010	Precipitation	Mm	P	Spanish Climatic Map (Ninyerola et al., 2007)	30 m
		Max. temperature	ºC	T_{max}	Spanish Climatic Map (Ninyerola et al., 2007)	30 m
		Mean temperature	ºC	T_{mean}	Spanish Climatic Map (Ninyerola et al., 2007)	30 m
		Min. temperature	ºC	T_{min}	Spanish Climatic Map (Ninyerola et al., 2007)	30 m
	Mean solar radiation 2014	Solar radiation	$W/m^2/year$	S_{rad}	DEM (CNIG, 2016)	5 m
Organisms	Land cover 2014	Deciduous forest	% occupation	DF	Landsat 8 OLI (USG, 2016)	30 m
		Pine forest	% occupation	PF	Landsat 8 OLI (USG, 2016)	30 m
		Shrub land	% occupation	SL	Landsat 8 OLI (USG, 2016)	30 m
		Agricultural land	% occupation	AL	Landsat 8 OLI (USG, 2016)	30 m
		Grassland	% occupation	GL	Landsat 8 OLI (USG, 2016)	30 m
		Rock outcrops	% occupation	Rock	Landsat 8 OLI (USG, 2016)	30 m
		Urban	% occupation	Urban	Landsat 8 OLI (USG, 2016)	30 m
	Average Normalized Difference Vegetation Index (NDVI) 2014	NDVI	-	NDVI	Landsat 8 OLI (USG, 2016)	30 m
	Average Normalized Difference Water Index (NDWI) 2014	NDWI	-	NDWI	Landsat 8 OLI (USG, 2016)	30 m

Soil forming factor	Description	Variable	Unit	Code	Source	Scale/ Resolution
	Brightness vegetation	Tasseled Cap component 1, Brightness	-	Bn	Landsat 8 OLI (USG, 2016)	30 m
	Greenness vegetation	Tasseled Cap component 2, Greenness	-	Gn	Landsat 8 OLI (USG, 2016)	30 m
	Wetness vegetation	Tasseled Cap component 3, Wetness	-	Wn	Landsat 8 OLI (USG, 2016)	30 m
	Average vegetation height 2014	Vegetation height	M	Veg_{height}	LiDAR PNOA (CNIG, 2016)	5 m
Relief	Digital Elevation Model	Altitude	M	Alt	DEM (CNIG, 2016)	5 m
		Slope	degrees	Sl	DEM (CNIG, 2016)	5 m
		Southness	-	South	DEM (CNIG, 2016)	5 m
		Eastness	-	Easth	DEM (CNIG, 2016)	5 m
		Topographic Wetness Index	-	TWI	DEM (CNIG, 2016)	5 m
Parent material/ Time	Geological map	Geological unit	Categorical	GU	GEODE, CNIG, 2016	1:50,000
		Lithology class 1	Categorical	Lit1	GEODE, CNIG, 2016	1:50,000
		Lithology class 2	Categorical	Lit2	GEODE, CNIG, 2016	1:50,000
		Age 1	Categorical	Age1	GEODE, CNIG, 2016	1:50,000
		Age 2	Categorical	Age2	GEODE, CNIG, 2016	1:50,000

4.2.2.2 Sampling scheme and laboratory analysis

Soil observations were obtained by collecting soil data across the study area. The soil samples that were collected needed to cover most of the spatial variation. Therefore, the area was divided in clusters using a Principal Component Analysis (PCA) and a Cluster Analysis (CA) through the algorithm IsoData available in ArcGIS 10.4. The algorithm IsoDATA randomly places cluster centres. The standard deviation within each cluster and the distance between cluster centres are calculated. Clusters split if one or more standard deviations are greater than the user-defined threshold and clusters merge if the distance between the clusters is less than the user-defined threshold. The correlations between different environmental covariates were analysed in the PCA. The cluster analysis on the results of the PCA was carried out to divide the study area in 12 strata with common characteristics (Fig. 4.2A). Subsequently, the lithology classes of the Continuous Geological Map were reclassified to 12 major classes (Fig. 4.2B). The accessibility of the study area is limited and therefore the samples were taken within a 2km buffer around the paved roads (Fig.4.2C). Within a stratum, in each lithology class and under each vegetation type a composite sample was taken at a representative location. The composite samples were taken as a square of 5m; one sample in the centre of the square and one sample in each corner. Besides that, soil- and land characteristics, e.g., soil profile depth, were noted.

The soil samples were analysed in the laboratory on organic carbon (OC) and organic matter content. For these analyses the Walkley and Black method was used (Walkley 1947; Walkley and Black, 1934). The soil data and correlations between the organic matter content and the environmental covariates are analysed by exploratory data analysis. Besides organic carbon and organic matter, also the pH was measured using the Potentiometric Method and the textural class and the coarse fragments were measured using the Particle Size Analysis. The soil data and correlations between the organic matter content and the environmental covariates are analysed by exploratory data analysis.

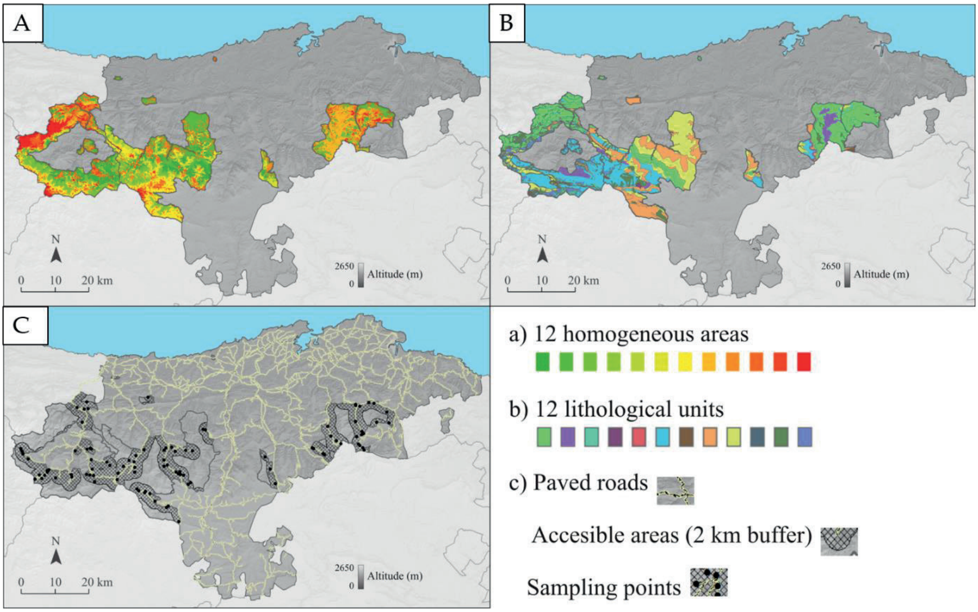


Figure 4.2. *Sampling scheme based on homogeneous strata that resulted from a Principal Component Analysis (PCA) and a Cluster Analysis (CA) (A). Major lithology classes were reclassified from the Continuous Geological Map (B). Within a stratum, in each lithology class and under each vegetation type a soil sample was taken within a 2km buffer around the roads (C).*

4.2.3 Mechanistic model

4.2.3.1 Conceptual framework

According to Jenkinson and Rayner (1977) and Pimm (1991), it can be assumed that systems that are under mature forest or that are abandoned for a long time, which is dominantly the case in our study area, have reached an equilibrium state. This means that carbon input equals carbon output. This assumption is used as starting point for the mechanistic model that comprises three main steps:

1. Selecting major processes that influence the SOM content. Processes that dominate at regional scale need to be selected. There are many dynamic soil models that describe associated C and N flows (Grace and Merz, 2001; Shibu et al., 2006). Comparing different soil models will help the selection of the most important

processes (Fig. 4.3.). The level of detail and complexity of these models differ, but major processes are quite similar and can be framed in a generic structure (Grace and Merz, 2001). First, part of the organic matter inputs, which are roots, wood and leaves, are decomposed. Second, the resistant plant material breaks down into CO_2, microbial biomass and humified organic matter. Third, humified organic matter contributes to the soil organic matter content as it enters the organic carbon pool. Some models even subdivide the carbon pool into three components: the active, slow and passive pool. Fourth, the carbon pool releases CO_2 and nutrients due to mineralization. The final process is that the release of nutrients contributes to plant growth, which again contributes to the organic matter inputs. The combination of mineralization and erosion can cause a severe depletion of the organic carbon pool (Lal, 2003). Although this factor is not included in the illustrated mechanistic carbon cycle models of Figure 4.3, erosion is an important process that needs to be considered.

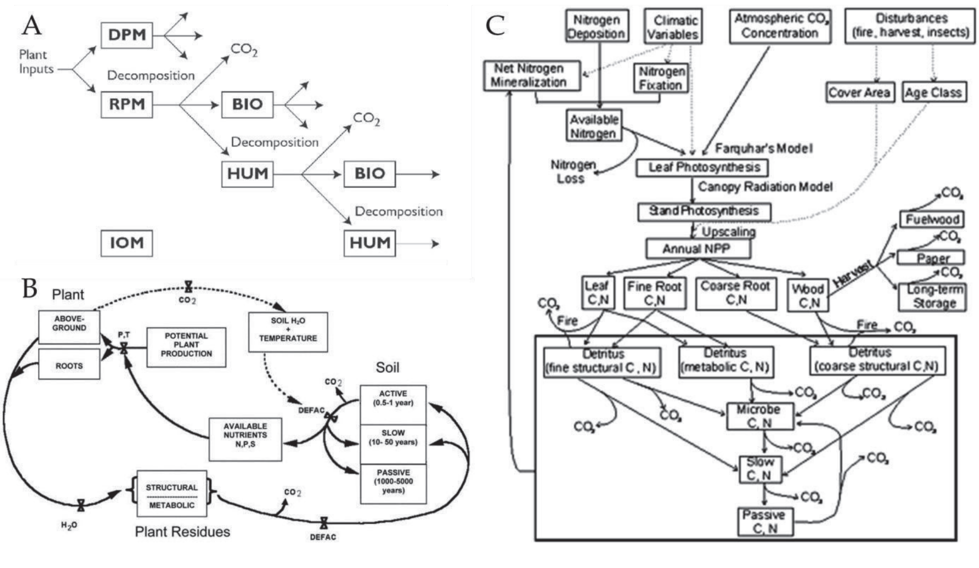


Figure 4.3. Roth-C (A), CENTURY (B), and Ecosys (C) are three carbon cycle models that include different levels of detail and complexity.

2. Defining the relationships. Much research has been done on the understanding of the carbon and nitrogen flows in the soil system (Shibu et al., 2006). The relationships

that are included in mechanistic carbon and nitrogen models are used in the mechanistic model for DSM. For example, consider a single carbon pool, the amount of SOM that turns into CO_2 and nutrients is exponentially related to the potential mineralizable carbon, the mineralization rate and time (Stanford and Smith, 1972).

3. Finding proxies or default values for the processes that influence the SOM content. The processes that are described in soil models often include variables that are not available spatially exhaustive. In this particular work, the environmental covariates of Table 4.1 were used as proxies to describe the process of interest. In addition, the decision on which environmental covariates to use for describing a process is in some cases easier than in others. For example, rainfall and slope are expected to be good proxies for the rainfall erosivity factor, which is required for estimating the erosion rate. A proxy for the clay content, which is required for the mineralization rate, is more difficult. Lithology class can be a proxy for clay. However, these data are nominal and only available for the representative soil type per mapping unit.

4.2.3.2 Calibration of the mechanistic model

The model is built based on the assumption that the system has reached an equilibrium state. The model included several constants and boundary conditions. These constants and boundary conditions ensure that each process is described by realistic values. The mechanistic model makes use of iterations to estimate the SOM content, because the amount of SOM that mineralizes and erodes depends on the observed SOM content. The constants were optimized using the 'Solver'-function in Excel. The objectives of the 'Solver'-function were: (i) the SOM balance (SOM_{bal}) has a value of 0 and (ii) the correlation between the observed and predicted SOM content was maximized.

There needs to be checked whether the model is pushed towards the boundary conditions. Samples where the difference between the observed and predicted SOM content are large need to be checked as well. The model sensitivity is tested by pushing the constants systematically out of balance.

4.2.4 Mechanistic digital soil map

The mechanistic model provides a spatially continuous map of the SOM content. The spatial auto-correlation of the residuals was analysed by plotting a semi-variogram. When the residuals show spatial auto-correlation, the residuals are interpolated using Empirical Bayesian Kriging. This interpolation technique accounts for the error in estimating the underlying semi-variogram through repeated simulations.

To indicate the quality of the mechanistic digital soil map, the mean error (ME) and the RMSD were estimated by:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(SOM_{obs,i} - SOM_{pred,i})$$

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(SOM_{obs,i} - SOM_{pred,i})^2}$$

In which, SOM_{obs} are the observed and SOM_{pred} are the predicted SOM contents.

To indicate the percent of variation that can be explained by the mechanistic digital soil map, the amount of variance explained (AVE), i.e. coefficient of determination, was calculated:

$$AVE = 1 - \frac{SSE}{SST}$$

In which, SSE is the sum of squares of residuals and SST is the total sum of squares.

4.2.5 Statistical model for DSM

The mechanistic digital soil map will be compared with the digital soil map that results from a standard DSM technique regression kriging. The spatially exhaustive covariates that are available for the mechanistic DSM are used as input data for the regression model. A backward linear regression is used to predict the SOM content. The spatial auto-correlation of the residuals is examined and interpolated using Empirical Bayesian Kriging when there is spatial auto-correlation. Combining the map that results from the regression kriging and the map that results from the Empirical Bayesian Kriging results in the digital soil map of the SOM content.

4.3 Results and discussion

4.3.1 Data collection

4.3.1.1 Environmental covariates

The average and standard deviations of available environmental covariates are given in Table 4.2. The data confirm that the area is dominated by deciduous forest (38%) and shrubland (43%). The data also confirm the large variation in temperature and precipitation. Some environmental covariates showed a strong linear relation. For example, NDVI and NDWI (r^2=0.93), mean annual temperature and altitude (r^2=0.99), temperature coefficient of variation (CV) and altitude (r^2=0.92) and radiation and southness (r^2=0.79). When two environmental covariates show a strong linear relation, one can replace the other including most functional relationship with the foreseen process.

Table 4.2. *Average and standard deviation (st. dev) of the environmental covariates and the coefficient of variation (CV) for the Natura 2000 areas in the Cantabria region (Spain).*

	Mean (st. dev)	**CV mean (st. dev)**
Precipitation (mm)	1260 (217)	2.9 (0.3)
Max. temperature (˚C)	14.6 (2.1)	3.9 (1.0)
Mean temperature (˚C)	9.2 (2.2)	5.6 (2.2)
Min. temperature (˚C)	3.9 (2.5)	11 (138)
Solar radiation (W*m²*year)	1107690 (218554)	
Deciduous forest (%)	38	
Pine forest (%)	0.4	
Shrub land (%)	43	
Agricultural land (%)	0.1	
Grassland (%)	12	
Rock outcrops (%)	6	
Urban (%)	2	
NDVI (-)	0.6 (0.2)	

	Mean (st. dev)	CV mean (st. dev)
NDWI (-)	0.5 (0.1)	
Brightness	0.9 (0.2)	
Greenness	0.3 (0.1)	
Wetness	0.2 (0.1)	
Vegetation height (m)	4.4 (6.6)	
Altitude (m)	1067 (428)	
Slope (˚)	24 (10)	
Southness	-0.16 (0.71)	
Eastness	0.08 (0.68)	
Topographic Wetness Index	9.8 (1.2)	

4.3.1.2 Laboratory and data analysis

At each sampling location one soil sample was taken, because hard rock or saprolite was reached within 10 cm to 27 cm soil depth. In total, 100 soil samples were taken. Among them, 48 were collected in (abandoned) pastoral grasslands, 43 in deciduous forest and 9 in meadow. The organic matter content of the soil samples was positively skewed (Fig. 4.4). The median of the organic matter content is 9%. Quartile 1 and 3 are 7% and 12% respectively. The soils in the study area are slightly acid. The soils haven an average pH is 5.2 and a standard deviation of 0.9. On average, the sand content is 44.3%, the silt content 33.8% and the clay content 21.9%. The

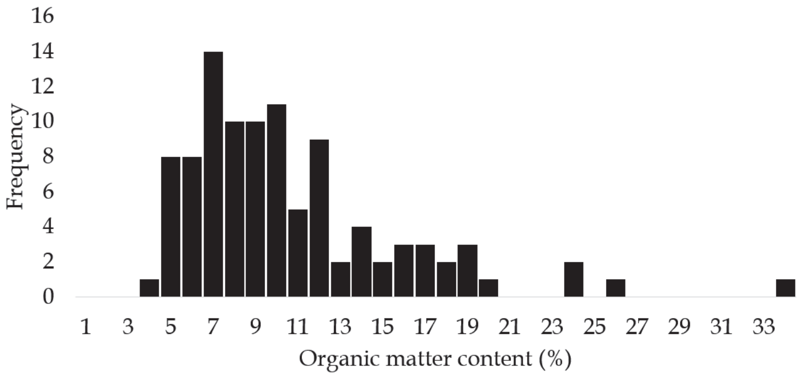


Figure 4.4. *The frequency distribution of the organic matter content resulting from 100 soil samples that were taken in the Natura2000 areas of the Cantabria region (Spain).*

percentage of coarse fragments ranged between 0% and 75%. The organic matter content is positively correlated to the vegetation height (0.11), but surprisingly negatively correlated to NDVI (-0.10). There is also a positive correlation between the organic matter content and the altitude, probably because the mineralization rate decreases with temperature. Besides that, the organic matter content is negatively correlated to slope (-0.26), which indicates that erosional processes play a role in the study area.

In a study of Rodríguez Martín et al. (2016), organic matter contents between 2.6% and 14.3% were measured in the Cantabria region, which corresponds to the majority of the samples that we took. Soil samples with an organic matter content of 17% or higher were tested twice in the laboratory and resulted in nearly the same values. Nine out of 13 samples with an organic carbon content above 17% were taken in pastoral grasslands. These grasslands are frequently burned by farmers to enrich the soil.

4.3.2 Mechanistic model

4.3.2.1 Conceptual framework

The starting point of the mechanistic model is that the study area has reached an equilibrium state. From this starting point, the mechanistic model was built following the three steps presented previously:

1. Selecting major processes. The mechanistic model for DSM required some simplifications compared to dynamic soil models, because the processes include often variables that are not spatially exhaustive available. Proxies or default values needed to be used to explain the processes. The SOM content is predicted including the following processes: turnover, mineralization and erosion. The conceptual framework of the mechanistic model is given in Figure 4.5. Because the system has reached an equilibrium state, the SOM content can be predicted by dividing the SOM input by the SOM output.

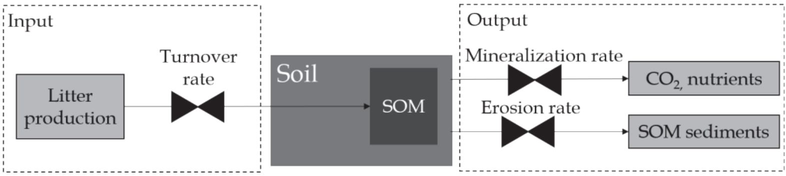


Figure 4.5. *Conceptual framework of the mechanistic model that will be used for predicting the soil organic matter (SOM) content using digital soil mapping.*

2 and 3. Defining the relationships and finding proxies or default values for the mechanistic processes. The SOM input was estimated by multiplying the litter production (LP) by a turnover rate. The LP depends on vegetation type and cover. Temperate deciduous forests produce between 8800 and 14100 kg/ha per year (Tateno et al., 2004) and the Carpathian grasslands produce about 1470 to 2870 kg/ha litter per year (Galvánek and Lepš, 2012). The litter production of Scots pine forests range between 651 kg/ha and 4912 kg/ha (Ukonmaanaho et al., 2008).

The vegetation height is taken as a proxy for the LP:

$$LP = c_1 + c_2 VH$$

In which, c_1 and c_2 are constants and VH is the vegetation height (m).

The turnover rate (TR) of humified organic matter to soil organic matter depends on the clay content. The lithology class can be used to represent the clay content. However, lithology is a categorical variable, while the model required a nominal variable. Because soil samples were taken in each lithology class, the average clay content per lithology class was taken for estimating TR. The poor correlation between the observed clay content and the average clay content per lithology class, made us decide to fix the TR at 0.32. This value is the average TR that results from the formula of the RothC model (Coleman and Jenkinson, 2014):

$$TR = \frac{1}{(3.09+2.7e^{(\text{CL})})}$$

In which, CL is the observed clay content (%).

The soil releases nutrients and CO_2 through mineralization. The mineralization rate (MR) is exponentially related to the temperature and soil moisture. Temperature and precipitation are correlated (0.28) and therefore one covariate replaces the other. The temperature showed strongest correlation with the observed organic matter content (-0.31) and therefore the MR is estimated as:

$$MR = c_3 e^{c_4 T}$$

In which c_3 and c_4 are constants and T is temperature (˚C).

The USLE equation is a commonly used equation to estimate the erosion rate (ER) (Wischmeier and Smith, 1965; Wischmeier and Smith, 1978). In this equation, the erosion rate depends on the slope, precipitation and the vegetation cover. The area is vegetated, which results in a constant value for vegetation cover. The slope is much stronger correlated to the organic matter content (-0.26) than the precipitation (0.07) and therefore the erosion is estimated as:

$$ER = c_5 + c_6 * S$$

In which c_5 and c_6 are constants and S is the slope (˚).

To predict the SOM content, the balance $SOM_{in} = SOM_{out}$ need to be optimized:

$$LP * TR = (MR * SOM_{stock} + ER * [SOM])$$

Which results in:

$$SOM_{stock} = \frac{LP * TR}{MR + \frac{ER}{BD * SD * 100{,}000}}$$

In which, BD is the bulk density (g/cm^3), which is fixed at 1.2 g/cm^3 and SD is the soil depth (cm), which is fixed at 20cm.

4.3.2.2 Calibration of the mechanistic model

The model is calibrated using 99 soil samples. One sample with an organic matter content of 34% was eliminated from the model, because there are no covariates that could clarify the extremely high organic matter content at this location. Optimizing the constants based on minimizing the RMSD and maximizing the correlation coefficient, resulted in the constants listed in Table 4.3. The model resulted in a correlation coefficient of 0.44 and a RMSD of 105,984 kg/ha, which is approximately 4%. The average balance SOM_{in} = SOM_{out} is slightly negative (-821kg/ha). It seemed that soil samples with low organic matter content were systematically overestimated and soil samples with high organic matter contents were systematically underestimated. The systematic over –and under estimation of SOM content by models used for DSM was also noticed by Angelini et al. (2016) and Yang et al. (2016). This systematically over- and underestimation of the SOM content by models used for DSM can have different causes, e.g., processes that influence extremely high and low organic matter contents were not included in the model or environmental covariates cannot explain the variation in SOM content.

Table 4.3. *Constant values used to fit the model.*

Constant	Value
c1	16033
c2	175
c3	0.007
c4	0.007
c5	0.000
c6	413

The litter production and the mineralization rate were constraint by boundary conditions. The litter production varied between 16033 and 20000 kg/ha, which means that the maximum boundary was reached. The mineralization rate varied between 0.01 and 0.02, which means that the minimum boundary condition was reached.

The constant values that are used to calibrate the model differed in sensitivity (Fig. 4.6). The constants are not responding symmetrically to lower or higher constant values. For example, decreasing the value of c_3 gives an exponential increase in the

RMSD, while increasing the value of c_3 changes the RMSD more gradually following a linear relationship. The same happens with the correlation coefficients. The model is most sensitive for c_1, c_3 and c_4, which means that LP and MR have a strong influence on the performance of the model. The constants that are used to estimate the erosion rate are hardly sensitive in the model.

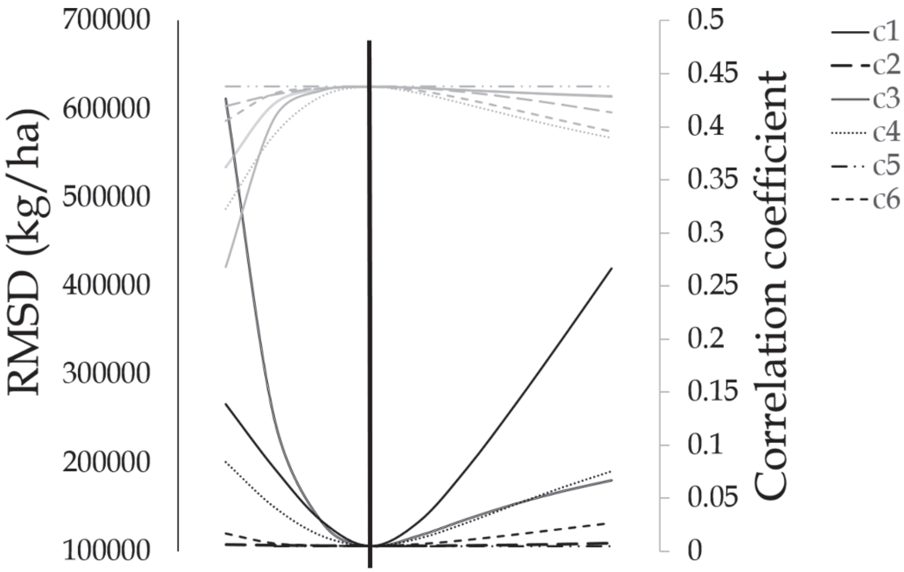


Figure 4.6. *Sensitivity analysis of the mechanistic model. The calibrated model is the central starting point (thick black line). From this point constant 1 to 6 were systematically taken out of equilibrium.*

4.3.3 Mechanistic digital soil map

The map that results from the mechanistic model is illustrated in Figure 4.7A. The residuals were calculated and the spatial dependency of the residuals was estimated by a semivariogram (Fig.4.7B), which presents the relationship between the distance and the semi-variance. The semi-variance is half the variance of the differences between all possible points spaced a constant distance apart. The semivariogram shows a spatial dependency of approximately 1 km. The residuals were interpolated using Empirical Bayesian Kriging. This resulted in a spatial exhaustive map of the areas where the SOM content is over- or underestimated by the mechanistic model (Fig.4.7C). The mechanistic model underestimated the SOM content in the central part of the study area and overestimated the SOM content in the eastern part of the study area. Combining the map that results from the mechanistic model and the map that results from the Empirical Bayesian Kriging, provides the mechanistic digital soil map of the SOM content (Fig.4.7D). In general, high organic matter contents were predicted in the southern part of the study area. This area is dominated by forest. The mechanistic DSM has a RMSD of 118,148 kg/ha, which is approximately 4.9%. The amount of variance explained is 57%.

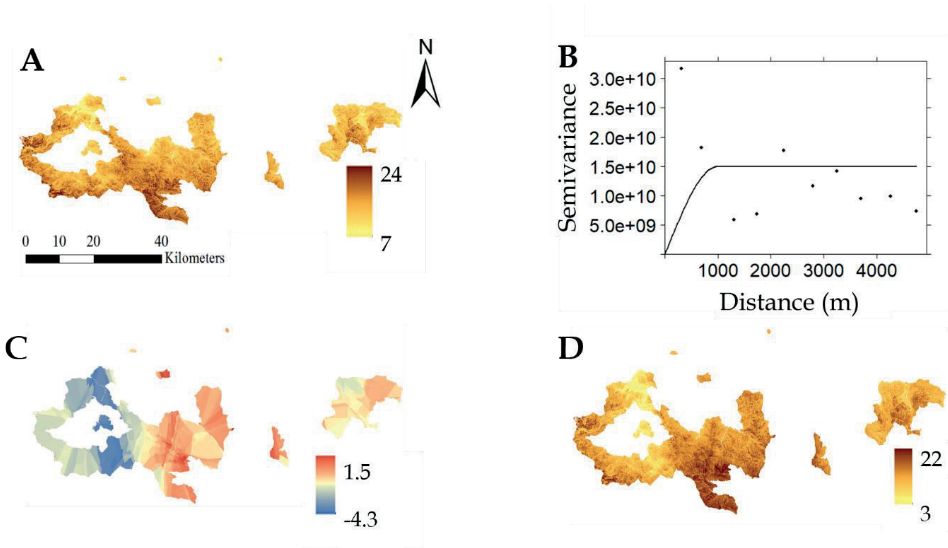


Figure 4.7. *Predicting the soil organic matter content (%) using a mechanistic model (A), plotting the residuals in a semivariogram (B), interpolating the residuals using Empirical Bayesian Kriging (C). The Mechanistic Digital Soil Map (D) results from map A and map C.*

4.3.4. Statistical model for DSM

The digital soil map that resulted from regression kriging had a RMSD of 929,025 kg/ha and a correlation coefficient of 0.46. The amount of variance explained is 29%. While the correlation coefficient performs slightly better than the mechanistic DSM, the RMSD is much higher and the amount of variance explained is much lower. Some interesting differences were noticed comparing the two maps. First, different environmental covariates were selected by the model. The mechanistic model only required three environmental covariates, where the map that resulted from regression kriging selected five environmental covariates: altitude, radiation, mean annual precipitation, mean annual temperature and southness. Only the environmental covariate mean annual temperature was selected by both models.

The differences between the mechanistic model for DSM and the statistical model for DSM become especially visible when zooming in to a smaller part of the study area (Fig.4.8). In places where no soil samples were taken, the mechanistic model showed sometimes opposite results compared to the statistical model (Fig.4.8).

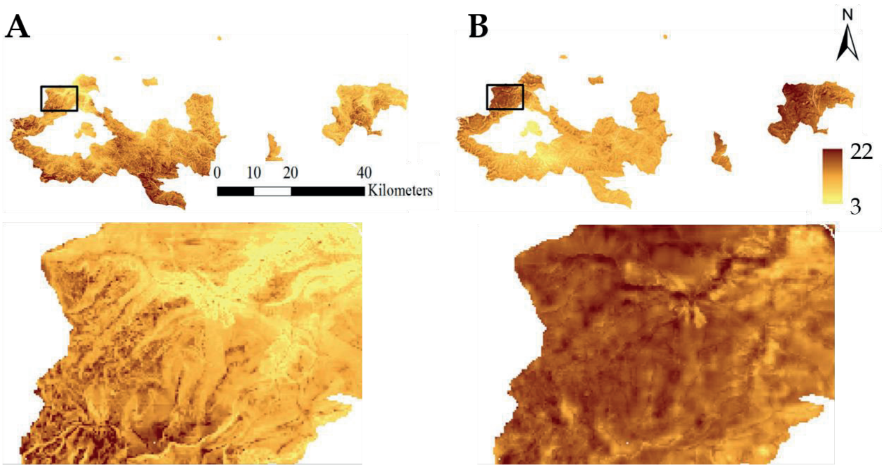


Figure 4.8. *Zooming in to the black rectangular, a clear difference between the use of a mechanistic model for DSM (A) and a statistic model for DSM (B) became visible.*

4.4 General discussion

4.4.1 Mapping the soil organic matter content

This study showed a different approach to DSM by using a mechanistic model instead of a statistical model for DSM. Nowadays, many studies on DSM search for complex, statistical models to improve the predictions on SOM content. However, many of these models still often result in poor predictions, especially at regional scale (Kempen et al., 2011; Mora-Vallejo et al., 2010). The mechanistic model used less environmental covariates and resulted in a lower RMSD compared to the statistical model. New statistical models are being developed for DSM to improve the understanding of the spatial patterns of interrelated soil properties. Two types of models are Boosted Regression Tree (BRT) and Random Forest (RF). Yang et al. (2016) concluded that variables that represent vegetation were most important in the BRT and RF models. These variables should be taken as main environmental indicator when mapping SOM content in Alpine environments. However, the importance of vegetation on SOM content is included in much detail in soil models as well (e.g., Guo and Gifford, 2002; Deng et al., 2016). The Ecosys model divided the vegetation even in four components (leaves, fine roots, coarse roots and wood) to get a better estimation of the amount of litter that enters the soil system.

There are few other studies that explore the use of mechanistic models for DSM. For example, Minasny et al. (2006) incorporated the within-profile transport of nutrients by using the same relationship as two mechanistic models (Elzein and Balesdent, 1995; Rosenbloom et al., 2001). Another example is the study of Angelini et al. (2016). They explored the use of structural equations modelling (SEM) in DSM. Hereby, the model equations were derived from known causal, often mechanistic relationships, while estimating the model parameters using available data (Angelini et al., 2016). While the studies of Minasny et al. (2006) and Angelini et al. (2016) use pedological knowledge, still statistical algorithms for predictive modelling were used.

Mechanistic models can have different advantages above statistic models for DSM. They describe the processes and predicted soil properties by values that typically stay within realistic boundaries. Due to this characteristic, it is likely that the mechanistic models have a greater potential for extrapolation. Statistical models can often exaggerate the error associated with the interpolation (Robinson and Metternicht, 2006). This can result in strong differences in accuracy between the interpolation and extrapolation area (e.g., Grinand et al., 2008). There is expected that mechanistic models can more easily be used for extrapolations.

4.4.2. A mechanistic approach for DSM

Using a mechanistic model for DSM is more complex than using a statistical model for DSM. However, the mechanistic approach for DSM should be considered more often as it is important to incorporate the knowledge we have on the processes that influence a soil property. The results of the mechanistic DSM were significantly different from the soil map that resulted from regression kriging.

Environmental covariates have some limitations (e.g., Da Costa et al., 2017; Rey et al., 2014). Therefore, the use of environmental covariates as proxies for driving certain processes has limitations as well. These limitations should be acknowledged. For example, environmental covariates that are derived from satellite imagery only observe the surface conditions. Processes that occur lateral (e.g., groundwater flow) or vertical (e.g., leaching) are not included. LiDAR data provides data on the vegetation height. However, other studies showed that LiDAR underestimated vegetation heights and with that it can underestimate the litter production (Streutker et al., 2006). The litter production does not only depend on litter production above the ground. Litter is also produced by the roots and lower vegetation, which is not included in the model. The model is highly simplified when environmental covariates are used as proxies. For example, the turnover rate was estimated by a default value, while the turnover rate depends on mechanistic feedbacks between biomass growth and production, tissues allocation, litter quality and nutrient availability (Potter and Klooster, 1997).

For decades, studies have tried to find proxies for describing processes. For example, Macduff and White (1985) measured and predicted the mineralization and nitrification rate in clay soils from soil temperature and moisture content. The increased availability of spatially exhaustive environmental data increased the potential of using these data to find relationships between soil processes and environmental covariates (Conant et al., 2011). How well a proxy describes the variable in a process differs and depends on the strength of the relationship between the variable and the environmental covariate. For example, the relationship between NDVI or LiDAR and litter production is more frequently studied and resulted in good estimates (e.g., Wang et al., 2004) compared to studies on the relationship between lithology class and mineralization rate (e.g., Hartmann and Moosdorf, 2011).

4.5 Conclusion

For soil mapping it is important to incorporate pedological knowledge. This knowledge can come from dynamic soil models that describe associated carbon and nitrogen flows. The environmental covariates that are selected by mechanistic DSM differ from the statistical approach. Therefore, we strongly advise the incorporation of pedological knowledge for DSM.

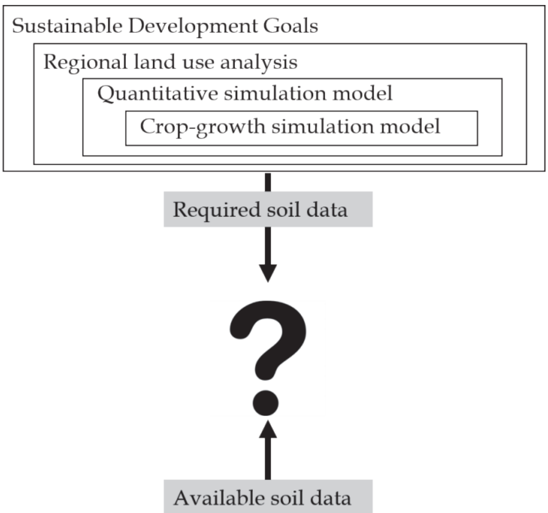
The three main advantages of a mechanistic approach above a statistical approach, that result from this study, are: (i) the RMSD can decrease, (ii) the number of environmental covariates that are selected can reduce, and (iii) the predictions become more realistically at locations where the error of statistical models exaggerated.

Chapter 5

How to obtain soil data for regional land use analyses?

Highlights

- Available soil data are not directly applicable for regional land use analyses (RLUA).
- Complex mapping techniques are used to obtain soil data for RLUA.
- Required soil data can be obtained more targeted using less complex mapping techniques.
- Soil data on the spatial variation need to be in line with the required spatial variation.



Based on: Hendriks, C.M.J., Stoorvogel, J.J., Claessens, L., Heuvelink, G.B.M.

Submitted to Agronomy for Sustainable Development

5.1 Introduction

The impact of population growth and climate change puts pressure on natural resources. Sustainable use of natural resources is needed to guarantee future human well-being. This need resulted in the development of global sustainable development programs like the United Nations Development Programme (UNDP) and the United Nations Environmental Programme (UNEP). In 2000, the development agenda of different sustainable development programs were converted into a Millennium Declaration (UN General Assembly, 2000). This declaration was signed by leaders of 189 countries and committed to achieving eight Millennium Development Goals (MDGs) by 2015. To build on these MDGs, a proposal for 17 Sustainable Development Goals (SDGs) and 169 associated targets was accepted in 2015. The SDGs focus on building a sustainable world wherein environmental sustainability, social inclusion, and economic development are equally valued (UN-SDSN, 2014).

The results of regional land use analyses (RLUA) are essential for achieving the SDGs (Keesstra et al., 2016). Over recent decades, the focus of RLUA changed. In the past, RLUA dominantly included qualitative land evaluation and land use planning. With the introduction of the SDGs, the focus of current RLUA changed towards quantitative, interdisciplinary impact assessment studies. These studies increasingly use quantitative simulation models, such as crop-growth models, erosion models and hydrological models. The use of quantitative simulation models for RLUA has changed the required soil input data. In general, RLUA require quantitative soil profile data that include detailed information on the spatial variation. These data often cannot be obtained from available soil data because: i) available soil data are often only available at coarse scale, while RLUA often require more detail on the spatial variation, ii) the location of a soil type within a mapping unit is unknown when a mapping unit consists of more than one soil type, while RLUA often require the soil profile data for a specific location and iii) the spatial variation within a soil type and the effect of different land use or land management within a soil type, while RLUA are often interested in these effects.

To meet the soil data requirements for RLUA, data on the spatial variation became increasingly important. The rapid increase in soil mapping tools and techniques and the increased availability of high resolution auxiliary data provided a range of new mapping techniques that can be used for mapping the spatial variation in more detail. However, the new mapping techniques often do not meet the data requirements for RLUA or the soil data were obtained using highly complex mapping techniques. For example, many digital soil maps only provide soil data of the topsoil (e.g., Mora-Vallejo et al., 2011; Schillaci et al., 2017; Jeong et al., 2017) and complex three-dimensional digital soil mapping techniques are used to obtain spatially exhaustive soil data that include variation over depth (e.g., Kempen et al., 2011; Kidd et al., 2015; Mulder et al., 2016). We assume that the required soil data can be obtained more targeted at RLUA and that the complexity of the techniques that are used to obtain soil data for RLUA can be reduced. To substantiate these assumptions, this study aims to analyse how different studies obtain the required soil data for RLUA.

This study carried out three case studies. The spatial variation at which the soil data are required differ per case studies. In the first case study, very general data on the spatial variation are required to assess the potential for crop intensification in climate zones. In the second case study, more detailed data on the spatial variation are required to analyse the potential of areas to adapt to the climate smart adaptation strategy agroforestry. In the third case study very detailed data on the spatial variation are required for an integrated assessment study. To obtain the required soil data for a RLUA, the studies should make efficient use of: available soil data, project resources (e.g., time, budget, capacity), mapping tools and techniques, auxiliary data and pedological knowledge (Fig. 5.1). In general, the required soil data can be obtained by transforming available soil data or processing collected soil data. For consistency among the three case studies, all studies require soil data for the crop-growth simulation model Decision Support System for Agrotechnology Transfer (DSSAT V4.0.2) (Hoogenboom et al., 2015; Jones et al., 2003). The case studies are carried out in different regions in Kenya, which are all characterized by rain-fed agriculture with maize as the major staple food crop. The DSSAT model was used to simulate water-limited maize yields over five years (2000-2004). In each case study, a Strength, Weaknesses, Opportunities and Threat (SWOT) analysis is carried out to provide an objective assessment of how the soil data were obtained.

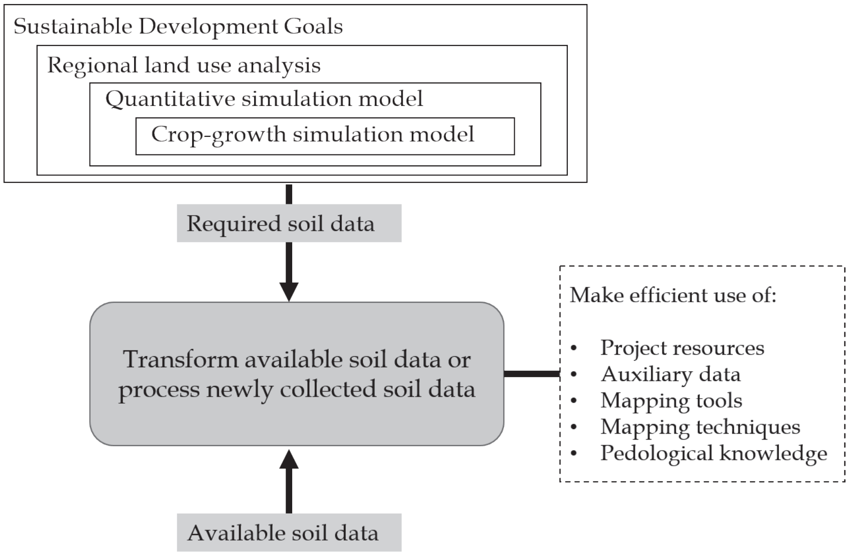


Figure 5.1. *To obtain the required soil data for regional land use analyses that use quantitative simulation models, it is required to make efficient use of available soil data, project resources, auxiliary data, mapping tools, mapping techniques and pedological knowledge. In general, the required soil data can be obtained by transforming available soil data or processing newly collected soil data.*

5.2 Crop-growth simulation model

There are many different crop-growth simulation models available. These models and the processes these models include differ in complexity. The complexity can range from qualitative to quantitative and from empirical to mechanistic. For example, to model the water balance a simple, so-called, tipping-bucket model or a complex mechanistic model can be used. The available soil data are limited for Kenya. Therefore, we decided to use a model that estimated the water balance based on a tipping-bucket model. The DSSAT model is applied for many crop-growth simulation studies in Africa (Jones et al., 2003). The model simulates crop growth, crop development and crop yield as a function of soil-plant-atmosphere dynamics.

The model comprises crop growth simulation models for more than 42 crops and includes database management, utility and application programs. Besides soil data, the database management program requires crop management and weather data. The model has widely been used in studies at different spatial and temporal scale to analyse, for example, different management strategies, management practices for optimum resource use and sustainable crop production, effects of economic return after decision making and alternative management practices (Hoogenboom et al., 2015; Jones et al., 2003).

For all three case studies, water-limited maize yields were simulated. Crop management was the same for all three case studies. In general, the smallholder farming systems do not use irrigation and they use limited amounts of fertilizer. Fixed amounts of nitrogen (25kg/ha) and organic fertilizer (incorporation percentage of 100%) were applied at each fertilization date. Planting takes place between day 70 and 102 depending on the soil moisture (between 40% and 100%) and the soil temperature (between 10˚C and 40˚C). The crops are harvested at maturity. For the weather files, data on solar radiation, maximum temperature, minimum temperature and rainfall were required. Weather data from 2000 to 2004 were selected from two weather stations in or closest to the study area. The data of the two weather stations were interpolated using a digital elevation model.

From origin, DSSAT is a one-dimensional model that requires soil profile data at point locations. When spatially exhaustive soil data are available, each mapping unit or pixel is taken as observation point. The model requires more than one soil horizon per soil profile. The soil profile data that can differ per case study are: soil depth, permanent wilting point (PWP), field capacity (FC), saturation point (SP), bulk density (BD), organic carbon (OC) content, clay and silt fraction, coarse fraction (CF), pH in water, pH in buffer and the cation exchange capacity (CEC). More information on the DSSAT model can be obtained from Hoogenboom et al. (2015) and Jones et al. (2003).

5.3 Case study 1: yield gap analysis that requires general data on spatial variation

5.3.1 Introduction

Population growth puts pressure on agricultural systems. Sustainable increase of agricultural production fits in the context of SDG 2 'Zero Hunger'. Yield gap analyses can identify regions with greatest potential for investment in agricultural development. The Global Yield Gap Atlas (GYGA) project assessed yield gaps, defined as the difference between potential or water-limited yield and actual yield, for countries and regions across the globe (Van Ittersum et al., 2013). The results are used to indicate areas with highest potential for crop intensification. This study aims to estimate the average water-limited maize yield for a climate zone (4106 km^2) that includes parts of the counties Machakos and Makueni (Kenya).

5.3.2 Process to obtain soil data

The GYGA project decided not to spend available project resources on collecting new soil data in the countries and areas they operate, but on agronomic assessments. Therefore, the required soil data were obtained from available soil data. Initially, it was decided to obtain the soil data from the most detailed soil dataset, which was for Kenya the Kenya Soils and Terrain Database Ver.2.0. (KenSOTER) at 1:1M scale (Dijkshoorn, 2007). However, for many regions in Sub-Saharan Africa no soil data were available. For these regions the Global Soil Profile dataset ISRIC-WISE ver.3.1. (Batjes, 2009) was used. This dataset is at coarse scale and lacks data on e.g., water retention parameters and rooting depth (Grassini et al., 2015). A more consistent method was required to improve the soil data for Sub-Saharan Africa.

The project decided to compile a spatially continuous soil dataset that covers Sub-Saharan Africa using different available soil datasets: African Soil Profiles database ver.1.2. (AfSP), AfSIS sentinel site soil point data, SoilGrids 1km layers and AfSoilGrids250m. Advanced digital soil mapping techniques and a wide array of auxiliary data were used to transform the available soil datasets into a map of the water holding capacity (WHC) and the rooting depth. The pedotransfer function of

Hodnett and Tomasella (2002) was used to calculate the water holding capacity. This pedotransfer function was calibrated for tropical soils. The rooting depth was estimated considering: (i) maximal rooting depth of maize, (ii) depth of the soil, (iii) depth of aerated soil and (iv) root restricting soil factors (Leenaars et al., 2015). The resulting dataset, denominated AfSIS-GYGA, provides spatially continuous data on the WHC and rooting depth and covered Sub-Saharan Africa. The GYGA project only used the WHC and rooting depth as soil input data for their quantitative simulation models. In this study, other soil properties were obtained from the KenSOTER dataset. Soil profile data were obtained by taking depth-weighted averages.

The spatial variation was described in more detail than the GYGA project required. Therefore, the WHC was categorized in seven classes and the rooting depth in four classes. These classes were combined in, so-called, soil classes and only the five most dominant classes were included for the analysis or the area coverage of crop harvested area had reached 50% (http://yieldgap.org). The area coverage of crop harvested area reached was estimated using the harvested crop area maps of the Spatial Production Allocation Model (SPAM) (You et al., 2009). Soils with a rooting depth less than 60cm, a WHC of $0.07 cm^3/cm^3$, sand content over 75% and soils on a slope steeper than 10% were discarded from the analysis. A maximum rooting depth of 1.5 m was assumed for maize in rain-fed agricultural systems. The area fractions per soil class was used to estimate the average yield gap for the climate zone. The GYGA project only used the WHC and rooting depth as soil input data for their quantitative simulation models. The GYGA project assumed that rooting depth equals soil depth. The PWP was fixed at $0.1 cm^3/cm^3$. For consistency, we used the DSSAT model for all three case studies. However, the GYGA project initially selected different crop-growth simulation models. The DSSAT model required more soil properties, e.g., pH, CEC, CF, BD, than the crop-growth simulation models that were selected by the GYGA project. The soil data that were not obtained by the soil data analysis of the GYGA project, were obtained from the KenSOTER dataset. For each soil class, most representative soil type was selected. Depth weighted averages of the representative soil profile were taken to obtain the properties.

5.3.3 Results

The averages and standard deviations of the five most dominant soil classes are provided in Table 5.1. These soil properties were used as input data for the DSSAT model. PWP and SP were fixed and the soil depth is one of the criteria for the soil classes. The area has clayey soils and the average organic carbon content is at maximum 1%. The standard deviations within the soil classes are high for most soil properties.

Table 5.1. *Averages and standard deviation (in brackets) of the soil properties that were used as input data for crop-growth simulation model DSSAT.*

Soil class	OC[a] (%)	Clay (%)	Silt (%)	BD[a] (g/cm³)	PWP[a] (cm³/cm³)	FC[a] (cm³/cm³)	SP[a] (cm³/cm³)	CF[a] (%)	pH_{H2O} (-)	pH_{KCl} (-)	CEC[a] (cmolc/kg)	Soil depth (cm)
1	0.74 (0.42)	35 (22)	18 (7)	1.26 (0.14)	0.10	0.17 (0.00)	0.40	12 (4)	6.9 (1.0)	5.6 (0.9)	15 (10)	115
2	0.88 (0.59)	38 (19)	21 (9)	1.23 (0.15)	0.10	0.16 (0.03)	0.40	11 (4)	7.3 (1.1)	6.1 (1.1)	20 (12)	115
3	1.00 (0.59)	36 (18)	24 (8)	1.23 (0.13)	0.10	0.17 (0.03)	0.40	11 (4)	7.5 (1.1)	6.3 (1.1)	23 (14)	115
4	0.68 (0.44)	31 (22)	19 (7)	1.25 (0.15)	0.10	0.17 (0.00)	0.40	13 (3)	6.7 (1.0)	5.4 (0.9)	13 (9)	150
5	0.85 (0.59)	29 (20)	20 (7)	1.23 (0.14)	0.10	0.17 (0.00)	0.40	13 (3)	6.8 (0.9)	5.6 (0.9)	15 (10)	150

These input data needed to be aggregated, because the GYGA project required a single average value of the water-limited maize yield per climate zone. The climate zone covers parts of Machakos and Makueni counties and resulted in a water-limited maize yield that differed between 1953kg/ha in 2002 and 4026kg/ha in 2004 (Fig. 5.2). The standard deviation differed between 332kg/ha in 2002 to 607kg/ha in 2001. The soil moisture content during the growing season was the main factor for stress during crop growth. The GYGA project simulated an average water-limited maize yield of 3100 kg/ha for this climate zone, which corresponds to our results.

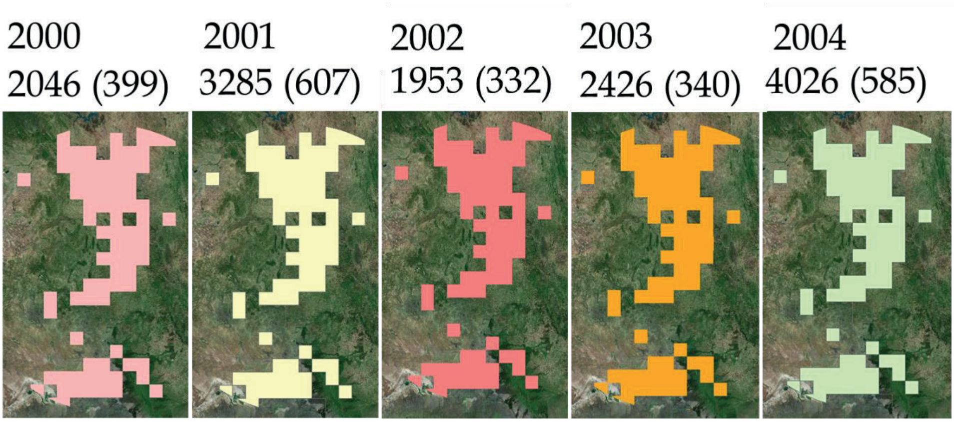


Figure 5.2. *The water-limited maize yield and standard deviation between 2000 and 2004 for a climate zone that covers part of Machakos and Makueni counties (Kenya).*

5.3.4 SWOT-analysis

A SWOT-analysis provides an objective assessment of how the soil data were obtained:

Strengths	**Weaknesses**
S1. Complex soil properties are derived from basic soil properties. S2. The five most dominant soil classes were selected from complex soil property maps instead of, for example, the most dominant soil type within a mapping unit. S3. The soil profile is considered because depth-weighted averages of the root-zone depth were taken. S4. The analysis results in consistent soil data that cover Sub-Saharan Africa.	W1. The soil data analysis does not include a quality assessment on the available soil data, the pedotransfer functions and the default values that were used. W2. For crop growth it is important to differentiate more than one soil horizon. Depth-weighted averages are therefore not optimal to represent variation over depth. W3. Different soil data sources were used for the analysis. Differences between the data sources are not considered.

Opportunities	Threats
O1. The resolution of auxiliary data increases. O2. New mapping tools and techniques become available and provide new analyses. O3. The quality of available soil data or the pedotransfer functions improve.	T1. The available soil data that are used for the analysis are of poor or unknown quality. T2. Areas that are discarded by the analysis are not chosen realistically. For example, in Kenya agricultural practices are still taking place in areas steeper than 10%.

5.4 Case study 2: study on climate-smart agriculture that requires moderate detail on the spatial variation

5.4.1 Introduction

The CGIAR research program on Climate Change, Agriculture and Food Security (CCAFS) aims to reduce the vulnerability of farmers to climate change. The measures taken to reduce the vulnerability differ per research site. This study was carried out in one of the 100 km^2 study sites of CCAFS; the Lower Nyando Basin (34.978E to 35.068E, 0.269S to 0.361S). Agroforestry is seen as one of the potential climate-smart adaptation strategies in the study area. Agroforestry is the intentional use of trees in cropping systems. Trees provide extra income by producing fuel wood and contribute to soil fertility by reducing soil erosion and increasing soil organic carbon content (Lorenz and Lal, 2014). Analyses on the effect of climate-smart agriculture fit in the context of SDG 2 ‘Zero Hunger’.

The altitude in the Lower Nyando Basin ranges from 1173 to 1746m ASL and the diversity in soil types is high according to the KenSOTER dataset. The area includes, for example, the soil types Dystric Regosol, Eutric Cambisol, Eutric Vertisol, Haplic Luvisol, Stagnic Solonetz, Luvic Phaeozem, Humic Nitisol and Eutric Planosol. Before implementing agroforestry on the large scale, the effect of this climate-smart adaptation strategy needs to be analysed. Agroforestry systems build-up about 0.07%/year more organic carbon than systems without agroforestry (Albrecht and Kandji, 2003). This study aims to analyse the effect of agroforestry on water-limited maize yields. This effect is analysed by comparing the water-limited maize yield that results from current agricultural system with the water-limited maize yield that results from an agricultural system that build-up 1.4% more soil organic carbon in 20 years of agroforestry. The positive effects of agroforestry on crop production were analysed by Schwab et al. (2015). In this study, the increase in organic carbon content influenced the water retention parameters as well. The project needed to identify areas with greatest potential for agroforestry.

5.4.2 Process to obtain soil data

To identify areas with greatest potential for agroforestry, a long term field experiment would be preferred. However, studies on agroforestry already confirmed the positive effect of agroforestry in the study area (Thorlakson, 2012). Therefore, the project needed to focus on how to expand the adaptation strategy in the area. This was done by identifying areas with highest potential for increased crop yields due to agroforestry. The study area was subdivided in strata. The strata were based on the mapping units of the KenSOTER dataset and the altitude (<1500 m and >1500 m ASL), assuming that these factors explain much of the spatial variation. In each stratum new soil data were collected on agricultural fields, because available soil data did not describe the spatial variation in enough detail. The number of soil samples that were collected depended on the area of the stratum. The fieldwork campaign took six weeks between October and November 2015 and aimed to obtain data on the spatial variation within the strata. The sampling locations were well distributed within a stratum. Within a buffer of 200m around the sampling locations, a maize-growing agricultural field was sampled. The samples were analysed in the laboratory on texture and organic carbon content, to obtain data on the current status of the soil fertility. Laboratory analyses on complex soil properties were more expensive and would therefore result in a reduced number of soil samples that could be collected. The consequence of this decision was that the PWP and FC were estimated from pedotranfer functions. The SP was fixed at 0.55cm^3/cm^3 and the BD was fixed at 1.2g/cm^3. At each sampling location, a sample of each soil horizon of the soil profile was taken. A composite soil sample of the first soil horizon was taken and one soil sample of the other soil horizons were taken. The composite soil sample included one sample in the centre of the field and four samples five meters towards each corner of the field. The soil profile data were used as input data for the DSSAT model. Soil properties that were not collected were obtained from the KenSOTER dataset. The representative soil profile of the soil type that was sampled in the field was taken from the KenSOTER dataset. The soil horizon depths of the KenSOTER dataset were adapted to the soil horizon depths that were distinguished in the field.

For each strata the average water-limited maize yield and the standard deviation was estimated. This analysis was repeated, but for a 1.4% higher organic carbon content. To create a map that identifies areas with greatest potential for agroforestry, the difference between the two maps was provided.

5.4.3 Results

In total, 73 sampling locations were visited. The average soil properties with and without the CSA strategy agroforestry are provided per strata in Table 5.2. The average carbon content was between 1.4% and 2.8%, which means that in some strata the organic carbon content doubled when increasing the OC content with 1.4%. The PWP and FC slightly increased due to the increase in OC content. In general, all soils in the study area had a high clay content and the pH was between 5.1 and 7.9. We expected a significant increase in water-limited maize yield when increasing the OC content with 1.4%, because this meant that at some locations the organic carbon content doubled. However, the effect of agroforestry was low or even negative (Fig. 5.3). It turned out that due to the higher organic matter content in the soil, the crop grows faster at the beginning of the cropping season. This fast growth results in more evapotranspiration, which finally leads to water stress during the cropping season. To confirm the modelling results, field experiments are preferred.

Table 5.2. *Averages of the soil properties that are used in crop-growth simulation model DSSAT to estimate the water-limited maize yield in different strata of the Lower-Nyando Basin (Kenya). Some soil properties were affected by the implementation of a climate-smart agricultural (CSA) practice.*

Stratum	**# samples**	**No CSA OC (%)**	**CSA OC (%)**	**Clay (%)**	**Silt (%)**	**BD (g/cm³)**	**No CSA PWP (cm³/cm³)**	**CSA PWP (cm³/cm³)**	**No CSA FC (cm³/cm³)**	**CSA FC (cm³/cm³)**	**SP (cm³/cm³)**	**CF (%)**	**pH_{H2O} (-)**	**pH_{KCl} (-)**	**CEC (cmolc/kg)**	**Soil depth (cm)**
Eutric Cambisol - high	15	2.1	2.8	50	16	1.2	0.18	0.19	0.3	0.31	0.55	10.4	6.3	5.1	14	46
Eutric Cambisol - low	8	1.4	2	52	24	1.2	0.19	0.19	0.32	0.33	0.55	2.5	7.1	5.3	15	54
Dystric Cambisol – low	4	1.6	2.3	48	17	1.2	0.18	0.18	0.29	0.3	0.55	1.1	7.7	4.8	13	56
Dystric Regosol – high	4	1.8	2.6	57	20	1.2	0.2	0.21	0.33	0.34	0.55	2	5.7	NA	NA	32
Humic Nitisol – high	4	2.4	3.1	69	13	1.2	0.23	0.24	0.35	0.36	0.55	8.6	5.1	4.8	28	48
Luvic Phaeozem - high	9	2.8	3.6	58	18	1.2	0.21	0.22	0.34	0.35	0.55	7.3	6.4	4.2	41	41
Haplic Luvisol – high	12	1.7	2.4	44	13	1.2	0.16	0.17	0.27	0.28	0.55	3.6	6.7	6.4	27	44
Haplic Luvisol – low	8	1.7	2.3	57	15	1.2	0.2	0.21	0.32	0.33	0.55	1.4	7.8	6.4	27	65
Stagnic Solonets – low	9	1.8	2.5	57	11	1.2	0.2	0.2	0.31	0.32	0.55	0	7.5	5.4	30	59

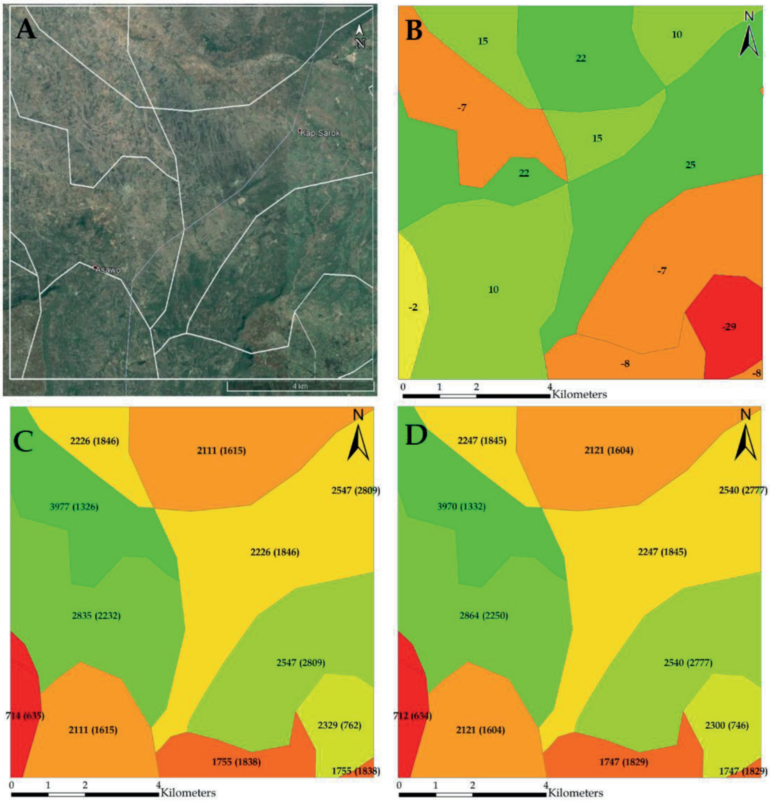


Figure 5.3. *The study area in the Lower Nyando Basin (Kenya) combined with the strata and the digital elevation model (A), the difference (B) in water-limited maize yield (kg/ha) between the current system (C) and a system with 1.4% higher organic carbon contents (D). The standard deviation of the water-limited maize yields within a stratum is given between brackets (C and D).*

5.4.4 SWOT-analysis

SWOT-analysis provides an objective assessment of how the soil data were obtained:

Strengths	**Weaknesses**
S1. The spatial variation between and within discrete units is provided. S2. Soil profile descriptions provide much detail on the variation over depth	W1. The spatial variation in water-limited maize yield within and between strata is not significant.
Opportunities	**Threats**
O1. The availability of proximal sensors can increase the number of soil observations. O2. The pedotransfer functions that are used to estimate permanent wilting point and field capacity improve.	T1. The climate smart agricultural practice influences, besides soil properties, other factors in the crop-growth simulation model as well (e.g., evapotranspiration).

5.5 Case study 3: integrated assessment study that requires detailed soil data on spatial variation

5.5.1 Introduction

Integrated assessment studies are frequently used for analysing policy impacts on economic, societal and environmental development (Valdivia et al., 2012). This study fits in the context of SDG 2 'Zero Hunger'. In Machakos and Makueni counties (13,500 km^2) in Kenya, climate change is high on the political agenda (Government of Kenya, 2016). The integrated assessment study aims to get more insight in the effect of climate change on the food security in the counties. Sufficient food production is one of the indicators for food security. The maize production in the study area is low and even decreasing due to nutrient depletion. In order to come up with proper recommendations on how to decrease the vulnerability of the people in the study area to climate change, the soil fertility needs to be mapped in much detail. These data will be used to study the inherent productivity of the soil. More detail on the study area was given in case study 1. The soil data are linked to other food security indicators (e.g., population density, distance to the market) and will help local and regional policy makers.

5.5.2 Process to obtain soil data

For integrated assessment studies detailed soil data are required to run e.g., nutrient balances. To obtain these detailed soil data on maize growing fields, new soil data needed to be collected. However, the project was limited by its available resources and the poor accessibility of the area. Digital soil mapping requires a limited number of soil observations to predict a soil property map. Compared to the topsoil, the spatial variation in the subsoil is lower (Vasenev et al., 2013) even under different land uses (Jaiyeoba, 2003). Therefore, we decided to collect new soil data on the topsoil (0-30cm) and to use available soil data on the subsoil. The study area is dominated by terraced maize fields and therefore new soil data were collected on these fields. Auxiliary data was used to exclude nature areas. The accessibility of the study area was limited and therefore the samples were collected in clusters of five sampling locations. The clusters were well-distributed in the study area to maximize the spatial coverage. To avoid the effect of within field variation, composite samples were taken by collecting five samples randomly in the terraced maize field and mix them thoroughly. The soil samples were analysed in the laboratory on OC and clay content, because these two properties are good indicators for soil fertility. The soil properties that were not analysed in the laboratory, but that were required for the DSSAT model, were obtained from the KenSOTER dataset.

The digital soil mapping technique regression kriging was used to create spatially continuous maps on the OC and clay content. This mapping technique makes intensive use of spatial exhaustive auxiliary data. Data on the subsoil were obtained from the KenSOTER dataset. Each pixel of the digital soil map is combined with subsoil data on the most dominant soil type within the mapping unit. In this case, each pixel provides data on the soil profile. These soil profile data were used as input for the DSSAT model.

5.5.3 Results

In total, 95 soil samples were collected in 15 clusters. The averages and standard deviations of the input data for the DSSAT model are provided in Table 5.3. The organic carbon content was between 0.35% and 1.18% and the clay content was between 5% and 57%. The auxiliary data that were used to predict the OC and clay content are described by Mora-Vallejo et al. (2008). To avoid overfitting due to the clustered sampling design, a leave-cluster-out cross validation was carried out to assess the quality of the digital soil maps. The high short-distance variation within the study area makes it difficult to predict soil properties at regional scale using only 95 soil samples. The digital soil maps had a variance explained of 13% for the OC content and 37% for clay the clay content. The average water-limited maize yield over five years (2000-2004) was between 40 kg/ha and 4895 kg/ha (Fig. 5.4). The spatial variation in water-limited maize yield was high, especially in the southern part of the study area.

Table 5.3. *Averages and standard deviations (in brackets) of soil properties that were used as input data for crop-growth simulation model DSSAT.*

OC (%)	**Clay (%)**	**Silt (%)**	**BD (g/ cm³)**	**PWP (cm³/ cm³)**	**FC (cm³/ cm³)**	**SP (cm³/ cm³)**	**CF (%)**	**pH_{H2O} (-)**	**pH_{KCl} (-)**	**CEC (cmolc/ kg)**	**Soil depth (cm)**
0.6 (0.7)	35 (18)	13 (8)	1.3 (0.1)	0.21 (0.10)	0.32 (0.12)	0.48 (0.01)	0	6.4 (1.2)	4.5 (2.3)	18 (13)	97 (48)

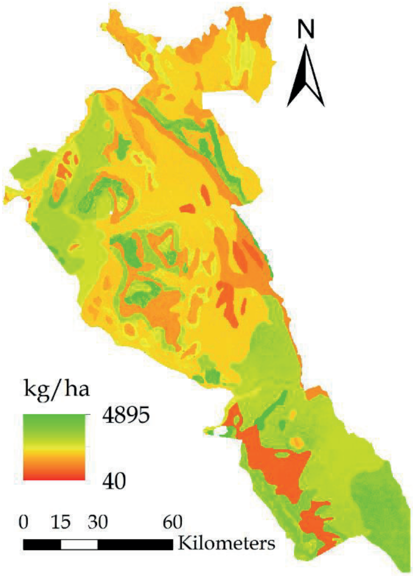


Figure 5.4. *Spatially continuous map of the water-limited maize yield that can be used for the integrated assessment study.*

5.5.4 Discussion and SWOT-analysis

A SWOT-analysis provides an objective assessment of how the soil data were obtained:

Strengths	Weaknesses
S1. The soil data analysis described the spatial variation in much detail, while spending limited project resources. S2. The analysis provides soil profile data.	W1. No correction was applied for the unlikely case that the subsoil data of the KenSOTER dataset showed higher OC contents than the topsoil data. W2. The correlation between clay and OC content is ignored, because basic soil properties (SOM, clay) were mapped individually without taking into account internal correlations.

Opportunities	Threats
O1. Proximal sensors can help collecting more soil data, which can improve the digital soil map. O2. The spatial resolution of auxiliary data increases. O3. Different auxiliary data become available that better explains the spatial soil variability.	T1. The assumption that spatial variation in the subsoil is less compared to the topsoil is not true for the study area. T2. The spatial variation in OC content cannot be described by spatially exhaustive auxiliary data. Different factors play a role (e.g., income, number of cattle).

5.6 General discussion

5.6.1 Transforming available soil data

Each RLUA required different soil data which can be obtained from unique approaches. This study illustrated that there are relatively simple techniques available to transform available soil data. Due to these techniques, missing data can be obtained and the spatial variation of soils can be described at a higher level of detail. Soil data on rooting depth or water retention parameters are, for example, often missing in available soil data sources. Case study 1 illustrated the potential of transforming basic soil properties into complex soil properties. Unless quantitative simulation models require dominantly basic soil properties, complex soil properties or other integrated soil data such as land qualities, soil functions, can contribute to achieving the SDGs. The AfSIS-GYGA dataset (Leenaars et al., 2015) and the RUSLE2015 map (Panagos et al., 2015) are examples of transformed soil data that aim to provide soil data for land use analyses.

At global scale there have been some initiatives on transforming available soil data. Most of these initiatives aim to provide basic soil property maps. Hengl et al. (2017) transformed, for example, many soil data sources into spatially exhaustive gridded soil property maps that represent the soil properties at fixed depth intervals. This global soil map, called SoilGrids, is available at 1km and 250m resolution. Advanced mapping tools and techniques, e.g., automated soil mapping, were used to fit statistical models between the available soil data and spatially exhaustive auxiliary data that represent the soil forming factors. Another initiative came from Stoorvogel et al. (2017). Stoorvogel et al. (2017) transformed the Harmonized World Soil Database (HWSD) by disaggregating the mapping units that consist of more than one soil type and resulted in the global soil map S-World. This study made use of pedological knowledge to define which environmental factors or which combination of environmental factors caused the occurrence of a soil type at a location. Hendriks et al. (2016) noticed that the soil properties of local to global soil datasets differ. Therefore, it is preferable that regional land use analyses use regional soil data.

For transforming available soil data, the three case studies made use of pedotransfer functions to estimate the water retention parameters. Pedotransfer functions are location specific. For example, Balland et al. (2008) focussed on a wide variety of soils in Canada, Khodaverdiloo et al. (2011) focussed on limestones in Iran and Hodnett and Tomasella (2002) focussed on tropical soils. Bouma (2016) emphasizes the need for validating pedotranfer functions in the area where the land use analysis is applied. Therein against, there are studies that argue that the performance of pedotranfer functions exceeds the performance of laboratory analyses on the water retention parameters, due to high within-field variability (Alaya et al., 2017)

5.6.2 Collecting new soil data

The change in focus of RLUA resulted in an increased need for new soil data. The availability of improved mapping tools and techniques and the increased availability of high resolution spatially exhaustive auxiliary data cannot replace the need for collecting new soil data. Project resources are often the limiting factor for collecting new soil data. Case study 2 and 3 illustrated the advantages of collecting new soil data. New soil data were obtained from different land uses, from the entire soil profile, and at locations where highest spatial soil variability was expected. Collecting new soil data can become efficient by making use of available soil data, auxiliary data, pedological knowledge and mapping tools and techniques. Often it is not necessary to collect completely new soil datasets, but often additional soil data or less intensive sampling schemes can meet the soil data requirements for RLUA.

5.6.3 The process to obtain soil data

The process on how to obtain soil data for RLUA differed between the three case studies. The input data of all three case studies was the same, but the aim and location of the case studies differed. To illustrate the process on how to obtain soil data for RLUA we illustrate the diagram of Hoosbeek and Bryant (1992), which was slightly adapted by Bouma and Hoosbeek (1996) (Fig.5.5). To obtain soil data for a RLUA, different models can be used. These models differ in degree of complexity and degree of computation. The degree of complexity ranges from empirical to mechanistic and the degree of computation ranges from qualitative to quantitative.

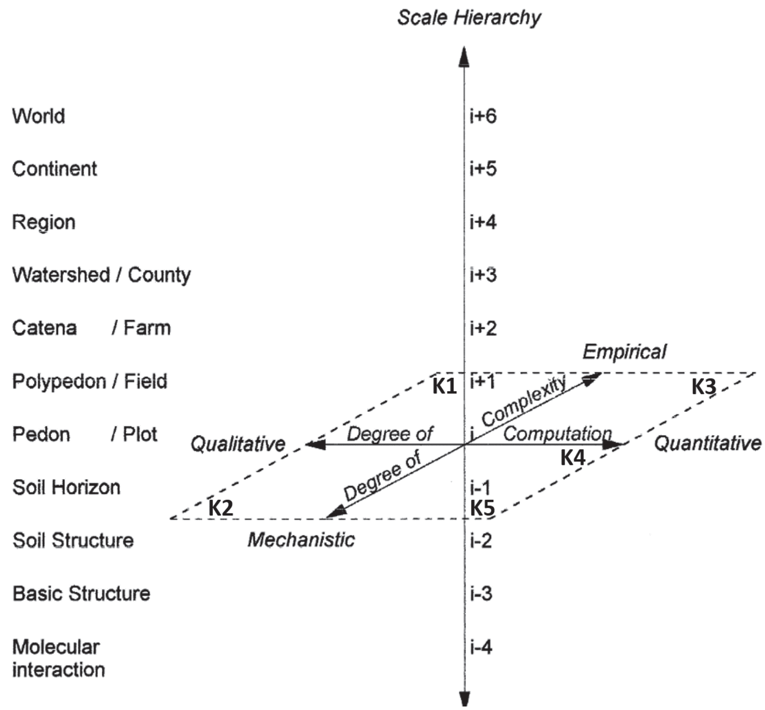


__Figure 5.5.__ To obtain the required soil data, different models can be used. These models can be classified based on hierarchical scale level, degree of computation and degree of complexity.

The spatial scale ranges between molecular and world. Different knowledge levels can be attributed to different models. K1 includes user expertise, K2 includes expert knowledge, K3 includes generalized holistic models, K4 includes complex holistic models and K5 includes complex models for parts of the system to be studied. To run the analyses with the DSSAT model, all three case studies required data at K5 level and at plot scale. However, the k-level and scale hierarchy at which the analyses were done can change from the k-level and scale hierarchy at which the study operates. For the analyses of case study 1, K5 level was required, but the study operates at K2 or K3 level. Case study 2 operates at K3 level and case study 3 operates at K5 level. This study illustrated that RLUA still operate at different k-levels, while the soil data that are collected are often processed to serve studies at K5 level. Available soil data that are not transformed only serve K1, K2 and in some cases K3 levels. The number of studies that operate at K4 or K5 level increased over recent decades, but this does not mean that soil data should only be transformed or processed to serve studies at K5 level.

Each method on how to obtain soil data for RLUA has its strengths, weaknesses, opportunities and threats. However, one soil data analysis can be more efficient than the other. For example, in a study of Rodríguez Martín et al. (2016) new soil data were collected using an intensive sampling scheme, while for the study area high quality auxiliary data were available and could be used for a more efficient sampling scheme. To obtain soil data for RLUA, it is important to define the modelling approach first (Hoosbeek and Bryant, 1992). Defining the modelling approach can help specifying the required soil data. Many RLUA do not define the modelling approach. When the modelling approach is not clearly defined, soil data analyses are often highly simplified, e.g., selecting only the most dominant soil type per mapping unit, or the soil data analyses are highly complex, e.g., providing three-dimensional soil properties that keep correlations between soil properties and variation over depth (e.g., Angelini et al., 2016). The highly simplified soil data analyses often require more detail on the spatial variation, while the highly complex soil data analyses often bring confusions about the use of the dataset for RLUA. The spatial variation of the input data should match with the spatial variation that is required by the RLUA (Fig. 5.5). The case studies required quantitative soil data for the quantitative, empirical DSSAT model (Bouma, 1997), but the spatial variation of the soil input data and required data differed, which made the soil data analysis differ. Nowadays, it is often assumed that spatially continuous data are required. However, many simulation models were not developed for two or three-dimensional purposes and therefore require not necessarily spatially continuous soil data (e.g., DSSAT, WOFOST, Nutmon) (Bouman et al., 1996). However, RLUA require increasingly spatially exhaustive results. Instead of providing spatially continuous soil data, the map on water-limited maize yields can be created after the model run for the point observations (Case study 2).

5.7 Conclusions

For RLUA that use crop-growth simulation models it is important to consider variation over depth, obtaining soil data at the spatial variation that is in line with the required spatial variation and obtaining functional soil data (e.g., complex soil properties). However, there is not a single solution to the question 'how to obtain soil data for regional land use analyses?'. Studies need to define the modelling approach before they start obtaining the required soil data. After defining the modelling approach, soil data can be obtained more targeted to the aim of the RLUA. The complexity and computation of the mapping technique need to be in line with the study. The complexity and computation of the study can differ from the quantitative simulation model. In the end, 'smart' analyses are required to obtain the required for RLUA. These analyses make efficient use of available soil data, project resources, auxiliary data and mapping tools and techniques and pedological knowledge.

Chapter 6

Synthesis

6.1 Introduction

The growing demand for quantitative soil profile data at detailed scale widens the gap between the required and available soil data for regional land use analyses (RLUA). For about 70% of the global surface there are no soil maps at a scale larger than 1:1million available (Nachtergaele and Van Ranst, 2003) and the soil data that are available, often do not meet the data requirements. Complementary soil data are required to narrow the gap between the required and available soil data for RLUA. However, which complementary soil data to obtain differs per RLUA. For example, in some RLUA complementary soil data consist of collecting more data on the spatial variation (Chapter 5), while in other RLUA complementary soil data consist of collecting soil data on different land use and management (Chapter 3). Different solutions are provided in this thesis to obtain complementary soil data. These solutions aim to bridge the gap between the available and required soil data for RLUA.

This synthesis assesses and discusses how the gap between the required and available soil data for RLUA can be bridged. In section 6.2 the research findings provide the lessons learnt and answers the sub-questions of this thesis: (i) does it matter which available soil data are used for RLUA (section 6.2.1), (ii) what complementary data are needed to meet the required soil data demand for RLUA (section 6.2.2), and (iii) how to obtain the required soil data for RLUA in an effective manner (section 6.2.3)? Implications of the research findings are discussed in section 6.3. In section 6.4 the aim of this study and the hypothesis are discussed. This section 6.4 provides a flowchart that helps obtaining the best soil data for RLUA and recommendations to the soil science community and the community that works with RLUA. Section 6.5 focusses on future perspectives of soil data for RLUA.

6.2 Research findings

6.2.1. Lessons learnt

The figure that illustrated the outline of the thesis in the introduction is used to illustrate the lessons learnt per chapter (Fig. 6.1). In Chapter 2 is learnt that available

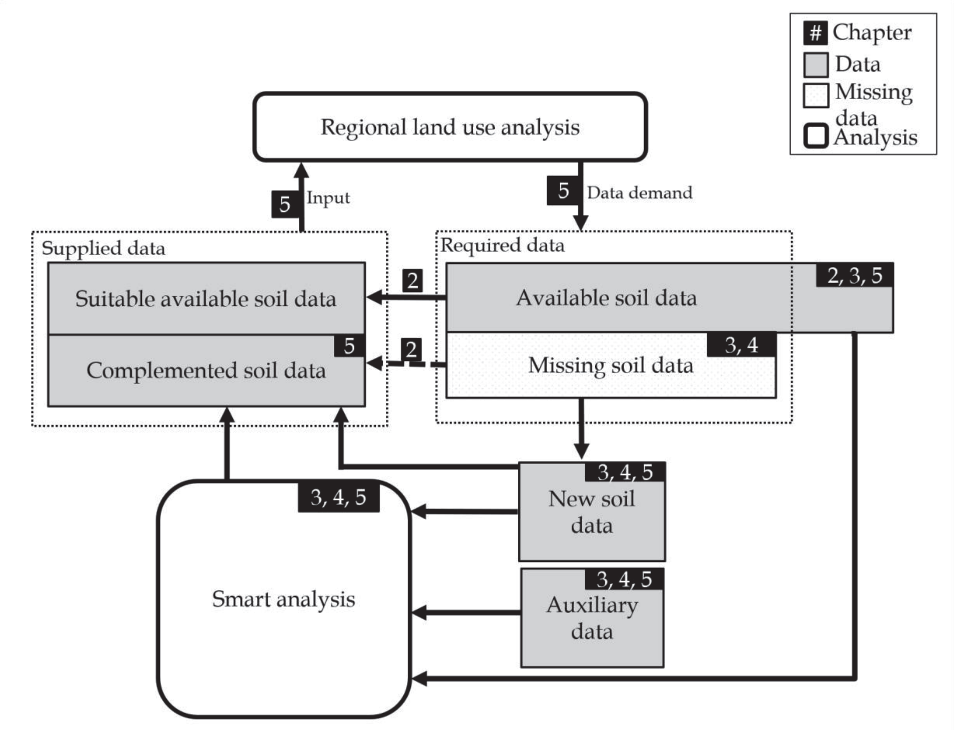


__Figure 6.1.__ The lessons learnt per chapter are illustrated in this flowchart. It illustrates how the gap between the required and available soil data for regional land use analyses (RLUA) can be addressed. The missing soil data, i.e., the gap was tried to be bridged by complementing missing data. Complemented soil data were obtained by collecting and/or processing new soil data or by transforming available soil data using 'smart analysis'. The soil data that resulted from the 'smart analysis', complement the suitable available soil data for RLUA. In this case, the supplied soil data that serve as input data for the RLUA meet the required soil data that RLUA demand.

soil data are not always most suitable for RLUA. The soil properties differ between soil datasets. This could have been caused by differences in the assumptions that were made in collecting soil datasets, differences in the scale at which the soil datasets were established, differences in the quality of soil datasets, and differences in the representation of the spatial variation. Missing soil data were, in Chapter 2, directly complemented by using default values or assumptions, while Chapter 3, 4 and 5 illustrated that default values or assumptions are not always necessary, because there are other methods to obtain missing soil data. Chapter 3 concludes that RLUA hardly combine available soil data and new soil data, while this combination can enhance the soil data for land use analyses. Especially at the regional scale available soil data are often used, while we learnt in Chapter 2 that these data not always result in most suitable soil data for RLUA. In Chapter 4 is learnt that new soil data should not always be processed using statistical models for DSM. Pedological knowledge can be incorporated in DSM by using a mechanistic model for DSM. The mechanistic model used less environmental covariates and predicted soil properties by values that typically stay within realistic boundaries. In Chapter 5 is learnt that input data for RLUA can be obtained by processing new soil data or transforming available soil data that were initially not suitable for RLUA. For example, the organic matter content and the clay content are used to derive the Water Holding Capacity from a pedotransfer function. The study should define the modelling approach and obtain the required soil data for this modelling approach, making efficient use of available soil data, auxiliary data, project resources, mapping tools and techniques and pedological knowledge. In this case, the required soil data are obtained using less complex and more targeted methods.

6.2.2 Effect of using different soil data for a regional land use analysis

Soil data that are used for RLUA are usually not critically evaluated (Renschler and Harbor, 2002). The literature study of Chapter 3 (Hendriks et al., in review) confirms this by noticing that many RLUA did not substantiate the decision on the soil data that were used. It is often unknown which soil data have the best quality for a particular application, and that makes the decision on which soil data to use for a RLUA difficult. However, the decision has significant effect on the results (Chapter 2) and therefore needs to be critically evaluated before using the soil data for RLUA. One of the reasons for not critically evaluating soil data is because soils often do not play a central role in RLUA and thus less attention is payed to the way soil data are obtained. Another reason is that assumptions are not well documented (Hendriks et al., 2016), which makes the boundary conditions for the application of a soil dataset unclear. For example, soil data that were collected in nature areas are used for agricultural studies (e.g., Van Ittersum et al., 2013), while the soil properties significantly differ between nature and agricultural areas (Chapter 2). To make available soil data more suitable for studies on agricultural land, additional soil data on agricultural land can be collected (Chapter 3). This was done in a study of Wu et al. (2003), where new soil data were collected on agricultural land and combined with the National Soil Survey of China. This resulted in the establishment of a new soil dataset that was corrected for the carbon loss that was faced from agricultural land and could be used for agronomic RLUA (e.g., Xiong et al. 2014).

6.2.3 Complementary soil data needed

Complementary soil data are needed to meet the soil data requirements for RLUA. The complementary soil data that are needed differs per RLUA. In general, the soil data for RLUA can be enriched by combining available soil data and newly collected soil data (Chapter 3). Surprisingly, many RLUA do not consider this option (Chapter 3). The change in RLUA resulted in the need for quantitative soil profile data but also data on the spatial variation. To meet these requirements, RLUA often require complementary soil data on: i) the spatial variation, ii) variation over depth, iii) quantitative land qualities or complex soil properties or iv) the quality of the soil data for a particular application.

Soil data are nowadays often provided as a package of individual soil properties, while soils are living bodies in the landscape. For analyses on crop growth, the availability of nutrients and water are of major importance. This availability is influenced by complex interactions between multiple soil properties, and not by an individual soil property or by a single soil layer. Functional soil data, such as complex soil properties, soil functions and land qualities, provide soil data that consider variation over depth and that keep correlations between soil properties. Functional soil data are increasingly requested for RLUA (Grassini et al., 2015). Chapter 5 illustrates that complex soil properties can be obtained by transforming available soil data. Although functional soil data are very useful for RLUA and these data improve the communication about soils among different disciplines, simulation models often cannot deal with soil functions and land qualities yet. Nowadays, emphasis is put on the spatial representation of soil properties, while for crop-growth simulation models, as well as for hydrological or climatological models, it is necessary to consider variation over depth (Chapter 5).

6.2.4 Obtaining the required soil data

Over recent decades, the use of quantitative simulation models for RLUA increased. Complex mapping techniques are used to meet the soil data requirements for these models. However, complex mapping techniques are not always required to meet the soil data requirements. Making smart use of available resources, mapping tools and techniques, auxiliary data and pedological knowledge, can result in simpler, more targeted methods to obtain the required soil data (Chapter 5). There is often thought that spatially continuous soil data are required. However, Chapter 5 illustrates that RLUA and the quantitative simulation models that are used for RLUA not always require spatially continuous soil data. To obtain the required soil data, it is essential to first define the modelling approach. Defining the modelling approach makes clear which soil data to obtain and how much detail on the spatial variation and on the variation over depth is required for the model and for the RLUA.

Obtaining soil data became more cost efficient over recent decades due to the development of new mapping techniques and the availability of high resolution auxiliary data. New mapping techniques make it possible to quantitatively estimate the degree of accuracy and uncertainty associated with a soil map. Moreover, soil maps can easily be updated when new soil data or auxiliary data become available. This makes it possible to refine the model until the accuracy and uncertainty standards are met (Stumpf et al., 2017). The power of new mapping techniques was confirmed by several studies (McBratney et al., 2003; Hengl et al., 2014; Stoorvogel et al., 2017). Sanchez et al. (2009) and Omuto et al. (2013) speculate that when the development of mapping techniques continues, soil maps with greater global coverage, greater accuracy for specific soil properties, and finer spatial resolution will become available. However, to obtain more detail on the spatial variation, new soil data are still required. The need for collecting new soil data cannot be replaced by the increased availability of auxiliary data or different mapping techniques (Chapter 3 and 5).

To obtain the required soil data, it is not only case to obtain the individual soil properties. Nowadays, soils or soil properties are dominantly predicted using statistical models for digital soil mapping (DSM) (e.g., Grimm et al., 2008; Sanchez et al., 2009; Hengl et al., 2017), while we have much knowledge on the processes that influence a soil property. Pedological knowledge is required to interpret the possibilities of soil data for RLUA. Therefore, a study on incorporating pedological knowledge in DSM is illustrated in Chapter 4. In this study a mechanistic model is used for DSM instead of a statistical model. Statistical models for DSM are often a 'black-box' where, first, all available spatially exhaustive auxiliary data are used as input data for the statistical model. Second, a type of algorithm for predictive modelling is chosen, e.g., regression trees, data mining, machine learning, to fit the model and minimizing the number of explanatory variables. The mechanistic model only selects the processes that influence the soil properties. Environmental covariates are used as proxy to explain these processes. In Chapter 4, this resulted in the selection of only three covariates as proxies to describe the processes that influence the SOM content, whereas a statistical model selected five environmental covariates.

6.3 Implications of research findings

6.3.1 Before using available soil data: evaluate and validate

The number of available soil datasets increased over recent decades. For example, six soil datasets were already available for Machakos and Makueni counties (Kenya). This makes it possible to select the most suitable soil dataset for the land use analysis or to use multiple soil datasets for ensemble runs (Chapter 2). A small literature study on 20 recent RLUA shows that conventional soil surveys (80%) and point observations (15%) are preferred above digital soil maps (5%) and remotely sensed soil data (0%) (Table 6.1). The decision on which type of soil data to choose is influenced by: the availability of different types of soil data, the suitability of a soil

Table 6.1. Types of soil data that are dominantly used for regional land use analyses.

	Source	Type	Soil data
1	Shrestha et al., 2017	Conventional	Soil map with mapping units
2	Pagani et al., 2017	Conventional	Soil Geographical Database of Europe (1:1,000,000)
3	Rinaldi et al., 2017	Conventional	Carta pedologica della Regione Basilicata
4	Sexton et al., 2017	Point	Point observations of Stewart et al. (2006) and Thorburn et al. (2010)
5	Mottaleb et al., 2017	Conventional	World Inventory of Soil Emission Potentials (WISE)
6	Pereira et al., 2017	Conventional	USDA Soil Conservation Service (USDA-SCS)
7	Ćosić et al., 2017	Point	Point observations of Cosic et al. (2015) and Djurovic et al. (2016).
8	Guilpart et al., 2017	Conventional	Soil map of the Soil Resources Development Institute (SRDI)
9	Chen et al., 2017	Point	Point database of Tan et al. (2014)
10	Cordeiro et al., 2017	Conventional	Manitoba Land Initiative (MLI) database (1:20,000 to 1:126,720)
11	Ouyang et al., 2017	Conventional	China Soil Scientific Database (1:1,000,000)
12	Yang et al., 2017	Conventional	China Soil Scientific Database (1:1,000,000)
13	Tukiainen et al., 2017	Conventional	Geological Survey of Finland (1:200,000)
14	Dumedah and Walker, 2017	Conventional	Digital Atlas of Australian Soils
15	Eddy et al., 2017	Conventional	USDA-NRCS (1986)
16	Rahman and Rosolem, 2017	Conventional	Harmonized World Soil Database (HWSD)
17	Himanshu et al., 2017	Conventional	Soil map from the National Atlas and Thematic Mapping Organization
18	Weissteiner et al., 2017	Conventional	Harmonized World Soil Database (HWSD)
19	Palazón and Navas, 2017	Digital soil map	Digital Soil Map of Aragón (Machín, 2000)
20	Malagò et al., 2016	Conventional	Harmonized World Soil Database (HWSD)

dataset for RLUA, and the visibility of new soil datasets. Nowadays, the decision on which soil data to use is often chosen pragmatically. Therefore, it needs to become easier to evaluate the boundary conditions of available soil data.

Evaluating the suitability of available soil data can be stimulated by sharing available soil data in Soil Data Warehouses (SDW) (Soil Science Division Staff, 2017). Examples of SDW are the European Soil DAta Centre (ESDAC) of the Joint Research Centre (JRC) and the recently renewed ISRIC-World Soil Information Soil Data Hub. The Natural Resources Conservation Service (NRCS) introduced the Geospatial Data Warehouse that aims as a source for environmental and natural data, at any time, from anywhere, to anyone. SDW allow choosing an area of interest, browse and select required soil data, customizing the format, reviewing the usage and quality of the soil data and downloading the data. Soil data that were collected for other purposes than soil data collection, e.g., agronomic experiments, should be included in SDW, because they often contain valuable information for RLUA. For example, the soil data that were collected for the Fertilizer Use Recommendation Project (FURP, 1987; FURP, 1994) could be used for RLUA focusing on maize growing areas (Chapter 2).

To select the most suitable soil dataset, it is necessary that soil datasets come along with a quality assessment and a document that describes the boundary conditions of a soil dataset. In Chapter 3 is noticed that land use analyses increasingly use available soil data, especially at regional scale. As these data are increasingly being used in regional land use analyses, more attention needs to be paid to the validation of available soil data (Heuvelink, 2014). Current RLUA dominantly use the representative soil profiles of conventional soil surveys. However, it was only suggested to give a purity of the mapping units and not a quality indication of the representativeness of the sampled soil profiles. Validating these representative soil profiles is difficult, because most soil samples were analysed decades ago and geographical coordinates of some representative locations are lacking. Data-splitting or cross-validation using in-situ measurements is the most common validation method for digital soil maps. Digital soil maps frequently come along with an accuracy map to indicate areas of uncertainty. Different soil maps or soil property maps can also be compared for validation (e.g., Dirmeyer et al., 2004). There is no soil map that can be considered as the truth and available soil data are often compiled from different soil datasets (Chapter 2). This can cause a certain level of spatial auto-correlation.

6.3.2 Continuous need for complementary soil data

The continuous need for complementary soil data is clearly described by Hartemink and Sonneveld (2013). Hartemink and Sonneveld (2013) studied the development of the soil maps of The Netherlands and showed that the development is very dynamic over time. In the beginning of soil mapping in The Netherlands, there was a need for soil data at finer scale. This trend continued until 1985 when the 1:50,000 map was established. After 1985, soils maps were aggregated to coarser scale and used for regional, national and continental planning. Since 1990, quantitative soil maps at different scales and resolution are established and the trend for soil maps with greater global coverage, greater accuracy for specific soil properties, and finer spatial resolution continues (Sanchez et al., 2009; Omuto et al., 2015).

To reach the current soil data requirements, new soil data have to be collected. For about 70% of the global surface still no soil maps at scales larger than 1:1million are available (Nachtergaele and Van Ranst, 2003). The advances in soil mapping tools and techniques and the finer resolution of auxiliary data do not mean that investments in new soil data collection can be reduced (Brevik et al., 2016). When the trend for soil maps at finer spatial resolution continuous, the soil data collection should even increase to obtain more detail on the spatial variation of the soil. Over recent decades, the soil data requirements changed because the use of quantitative simulation models for RLUA increased. The soil data requirements will continue to change (Soil Science Division Staff, 2017). For example, to increase the contribution of soil science to the SDGs it is important to interpret soil data besides providing soil data. Besides that, soils are a dynamic and integral part of the environmental system, which means that soil properties change. This change should be monitored. The enrichment in soil data when collecting new or additional soil data is highlighted in Chapter 3.

New soil data can be collected for specific studies (e.g., AfSIS-GYGA was established specifically for the GYGA project; Leenaars et al., 2015), or for general purpose interpretations (e.g., the DSM of the SOM content in the Cantabria region, Chapter 4). There is no "one soil data product meets all user needs" (Soil Science Division Staff, 2017). The need for soil data on-demand increases, but soil data for general purpose interpretations stay important as well. New soil data can, for example, be used to overcome the limitations of available soil data (Chapter 3). New soil data can collect missing soil properties (Wösten et al., 1985), update outdated maps (Kempen et al., 2009) or validate available soil data (Brus et al., 2011). For the '4 per 1000 Soils for Food Security and Climate Initiative, launched at the United Nations Climate Change 21st Conference of the Parties (COP21) conference in 2015, issues were faced with outdated soil maps, but also with lack in soil data on e.g., peat depth, water table and oxidation rate of peat (Minasny et al., 2017). This initiative illustrates the need for new soil data when more detail on the spatial variation is required. The change in soil property value is not for all soil properties as dynamic in space and time. Maps that identify areas where updating available soil data should be prioritized can help collecting new soil data in a more targeted way.

New soil data are required to achieve more accurate predictions, because it is not per definition that soil properties are more accurately predicted when auxiliary data of higher resolution are used (Ye et al., 2009; Samuel Rosa, 2015). Soil scientists are nowadays limited in their ability to collect data on the actual status of our soil resources (Omuto et al., 2013). To discuss the need for new soil data, we need to go back to the sampling schemes that were used for conventional soil surveys. In conventional soil surveys it was recommended to collect 0.5-1 soil observation per cm^2 on the map (Reid, 1988; Schoknecht et al., 2008). This means that 600 to 1200 observations are required to map the Natura 2000 areas in the Cantabria region (Spain) (about 120,000 ha) at 1:100,000 scale. To create the digital soil map for this study, only 100 soil samples were taken. With DSM we are able to predict soil properties using a limited number of soil observations, but predictions that are based on such a small number of soil observations result at regional scale often in soil maps of poor quality (e.g., Mora-Vallejo et al., 2008; Kempen et al., 2009). The use of proximal sensors brings possibilities to collect a large number of soil data. The neutron probe is, for example, a sensor that measures the soil moisture (Rossel et al., 2011). Other soil properties that can be measured by proximal sensors are: soil nutrients, heavy metals, CEC, pH and texture. Field visits can help understand land use systems, interpret results and estimate the reliability of default values. For example, most crop-growth simulation models estimate crop yields for single cropping systems (Bouman et al., 1996), while intercropping occurs in many parts of Sub-Saharan Africa.

6.3.3 Obtaining the required soil data

To obtain the required soil data, it is important to use pedological knowledge of processes that influence soil properties (Chapter 4) in combination with new mapping techniques. For example, the mapping units of conventional surveys are often quite similar to the patterns of digital soil maps (Bazaglia Filho et al., 2013), while most conventional soil surveys were established before spatially exhaustive environmental variables became available. This indicates that the knowledge that was used for creating the mapping units can be used in combination with current mapping techniques for advanced soil mapping. Another example is to communicate, share and combine soil data globally, through standardized methods for DSM. There is a pressing need for standard protocols for DSM. The publication of the 'Stages and Processes of DSM' in the Soil Survey Manual (Soil Science Division Staff, 2017) demonstrates the potential of standardized protocols for DSM. This publication can nowadays be seen as the most standard protocol for DSM. A third example is to use 'smart analyses' for the supply of functional soil data. Conventional soil surveys provided data about the suitability and limitations of each soil for multiple uses as well as their likely response to management systems (FAO, 2006). To predict functional soil data using DSM techniques, the data conventional soil surveys provide can be useful.

To improve the soil data for RLUA it is also important to check the sensitivity of the model parameters. It is important to obtain accurate soil data for most sensitive parameters. For example, in Chapter 2 the rooting depth turned out to be most sensitive for the crop-growth simulation model having large effect on the results. However, a default value of 100cm was used for the rooting depth, because no data on rooting depth were available. I suggest that most accurate soil data should be obtained for most sensitive parameters in the quantitative simulation model.

6.4 Bridging the gap between the available and required soil data

Every RLUA requires a different approach for bridging the gap between the available and required soil data. This section provides a flowchart that helps bridging this gap and aims to help studies that use RLUA obtain the required soil data more easily. The flowchart is applicable for a wide range of land use analyses. Additionally, specific recommendations can be made to the soil science community (section 6.4.2) and to the people that are involved in RLUA (section 6.4.3).

6.4.1 Flowchart to use for obtaining the required soil data

To obtain the required soil data, the so-called ARDAIG approach (Fig.6.2) can be used:

1. Define the modelling **A**pproach. The modelling approach can range from qualitative to quantitative and from empirical to mechanistic (Hoosbeek and Bryant, 1992). Also the scale needs to be indicated in the modelling approach. At global scale different soil data are required compared to field scale.

2. Define the soil data **R**equirements. This does not only include information on the required soil properties, but information on the variation over depth and spatial variation as well. RLUA need to check how sensitive the quantitative simulation model is for soil data.

3. Inventorize available soil **D**ata. The SDW that are currently available can be consulted in existing inventories (e.g., European Soil Data Centre (ESDAC) of the Joint Research Centre and the recently renewed ISRIC Soil Data Hub). However, also soil data that come along agronomic experiments can be useful for RLUA and should be inventorized. When no soil data are available, continue to step 5.

4. **A**ssumptions and quality of the available soil data need critical evaluation. Different assumptions were made when a soil dataset was established and the quality of a soil dataset for a particular application differs as well. The suitability of a soil dataset for a particular RLUA needs to be evaluated and unsuitable soil datasets should be eliminated.

5. **I**dentify the gap between the available and required soil data. Not all RLUA have a gap, but it is important to verify the absence of the gap. If there is a gap, identify the missing soil data. Based on the soil data requirements defined in step 2 and the

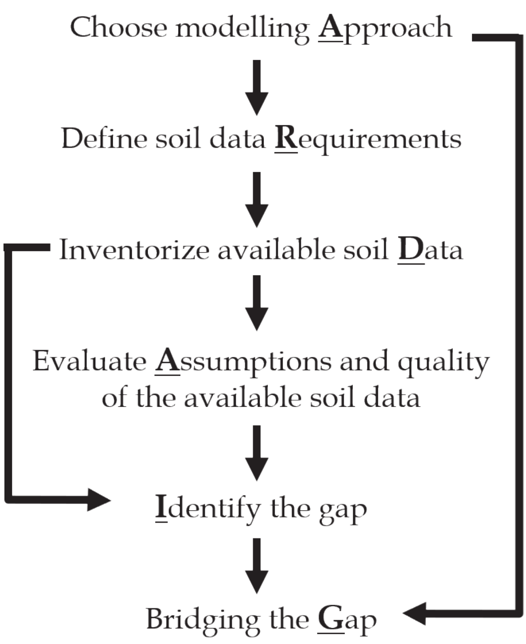


Figure 6.2. *The ARDAIG approach helps obtaining the required soil data for regional land use analyses.*

evaluation in step 4, the gap can be identified. Examples of gaps are: the required quality is not met, the spatial variation is not described in the required level of detail, the variation over depth is required but not provided by the soil dataset and different soil properties are required.

6. Bridging the **G**ap. Evaluating the available project resources, auxiliary data and mapping tools and techniques, to identify whether the gap can be bridged by transforming available soil data or collecting new or additional soil data. When it is impossible to bridge the gap, e.g., due to lack in resources, reconsider step 1.

6.4.2 Recommendations to the soil science community

The visibility of soils in RLUA that contribute to the SDGs can increase (Bouma, 2014). Therefore, specific recommendations are made for the soil science community to better meet the soil data requirements:

- Quantitative simulation models are increasingly being used for RLUA. Simulation models that emphasize the same category, e.g. crop-growth, require often similar input data (Cornelissen et al., 2013). While RLUA have an interdisciplinary character, the quantitative simulation models that are used for RLUA can often be categorized in agronomic, climatological, hydrological and ecological models (Keesstra et al., 2016). I recommend to publish more often soil data along these categories.
- The soil science community needs to acknowledge the need for collecting new data. It is a case that not only study-specific soil data are being collected, but collecting soil data for general purpose interpretation is important as well. Investments in new soil data collection cannot be replaced by advances in soil mapping tools and techniques and the finer resolution of auxiliary data. There is a continuous need for new soil data. The need for new soil data collection should be framed in an interdisciplinary context (e.g., Bonfante et al., 2017). For example, carbon sequestration is an international and interdisciplinary topic that requires detailed soil data on the carbon content. These data are nowadays simply lacking for many regions across the globe. I recommend to

illustrate the need for investing in new soil data collection in an interdisciplinary context.

- The added value of pedological knowledge in RLUA should be stated more prominently. The increased use of statistical models for digital soil mapping, made it possible that even non-soil scientists can create spatially exhaustive soil property maps. Soils are a complex and dynamic system and should be interpreted like that. Nowadays, soil scientists are often only involved in RLUA for the soil mapping. I recommend that soil scientists should stay involved in RLUA for interpreting the results.
- The soil science community provides different types of products; e.g., conventional soil surveys, many different types of digital soil maps, point observations, remotely sensed soil data. This diversity is good on one hand, but it is confusing for the user on the other hand. To reduce the confusion, I recommend to communicate clearly about the assumptions that were made when a soil dataset was established, to use a standard protocol for DSM and to share the data through soil data warehouses.
- Incorporating variation over depth or keeping correlations between soil properties while using digital soil mapping are nowadays addressed by complex statistical models. There is much knowledge on the processes that influence a soil property. Chapter 4 illustrated the importance of incorporating this knowledge in DSM. I recommend to use the knowledge we have on the soil system and move towards mechanistic digital soil mapping.
- Nowadays, much soil datasets have a poor or unknown quality or the data became outdated. Available soil data need to be validated. I recommend to explore new validation techniques, because global soil datasets do not allow for proper validation, due to the scale and the unequally distributed data density (Stoorvogel et al., 2017).
- For crop-growth simulation, soil data on the variation over depth were extremely important. Different case studies (e.g., Chapter 5) show that RLUA not always require spatially continuous soil data. I recommend to provide soil

data on the spatial variation and variation over depth at the level of detail that is required by the RLUA.

6.4.3. Recommendations to people involved in RLUA

The importance of soils for RLUA can get more attention. Therefore, specific recommendations are made for the people involved in RLUA:

- Nowadays, quantitative simulation models dominantly require soil properties as input data, while land qualities or complex soil properties would fit better in the interdisciplinary context of RLUA. I recommend to adapt quantitative simulation models in such a way they require land qualities or complex soil properties rather than individual soil properties.
- Soil scientists often do not stay involved in the RLUA after providing the required soil data. I recommend to involve soil science community actively in the search for the required soil data, but also in the interpretation of the results, because soils are more than just a series of individual soil properties.
- Every land use analysis requires different soil data. These data can often not be obtained from available soil datasets. The increased need for soil data that includes more detail on the spatial variation, forces RLUA to collect new soil data as well. I recommend to make use of proximal sensors and auxiliary data to collect new soil data in a cost efficient way.
- Many studies on RLUA do not communicate openly about the limitations of the used soil data. I recommend to communicate more openly about the limitations of the soil data, because working together towards a solution can help bridging the gap between available and required soil data.
- Many studies on RLUA do not substantiate the decision on which soil dataset was used for the analysis. I recommend studies on RLUA to use the ARDAIG approach to obtain the required soil data, because it forces the user to consider all inventoried soil datasets. In this way the decision on which soil dataset to use for RLUA can be substantiated.

6.4.3 Hypothesis

This thesis hypothesises that the need for new soil data can be minimized by making 'smart' use of available soil data. Many case studies included in this thesis, demonstrated that the required soil data could be obtained by making efficiently use of available soil data, project resources, auxiliary data, mapping tools and techniques and pedological knowledge. Approximately 30% of the global surface is covered by soil maps at scales larger than 1:1million (Nachtergaele and Van Ranst, 2003), which indicates a tremendous need for new soil data. In contrast to that, the effort that is needed to collect the vast amount of new soil data decreased due to the availability of high resolution auxiliary data and the development of new mapping techniques. In conclusion, the need for new soil data can indeed be minimized by making 'smart' use of available soil data, but the need for new soil data is still extremely high. The hypothesis is confirmed.

6.5 Future perspective

There are opportunities for soil science to contribute more efficiently to RLUA. If the soil science community and the people involved in RLUA will implement the presented recommendations, I foresee that the soil data requirements for many RLUA will be met. New investments in soil data collection are required to meet the rapidly increasing need for soil profile data that includes more detail on the spatial variation. Cost efficient soil data collection is stimulated by the development of proximal sensors, which help collecting a large number of soil data in a short time.

Following the ARDAIG approach should help RLUA find the most suitable soil data. In the future, soil input data will receive more attention in RLUA, because of the effect soil datasets have on the results of a RLUA is significant. The need for new soil data collection is acknowledged and the involvement of soil scientists is not only limited to the provision of soil property maps, but soil scientists stay involved in the RLUA for interpreting the results. Through functional soil data that is interpretable for non-soil scientists, it is important to communicate about soils in interdisciplinary land use analyses.

References

Adamchuk, V.I., Hummel, J.W., Morgan, M.T., Upadhyaya, S.K. (2004). On-the-go soil sensors for precision agriculture. Computers and Electronics in Agriculture 44: 71-91. DOI: 10.1016/j.compag.2004.03.002.

Alaya, I., Masmoudi, M.M., Lagacherie, P., Coulouma, G., Jacob, F., Ben Mechlia, N. (2017). Performance of Saxton and Rawls Pedotransfer Functions for Estimating Soil Water Properties in the Cap Bon Region-Northern Tunisia. In: Ouessar M., Gabriels D., Tsunekawa A., Evett S. (eds.) Water and Land Security in Drylands. Springer, Cham.

Albrecht, A., Kandji, S.T. (2003). Carbon sequestration in tropical agroforestry systems. Agriculture, Ecosystems & Environment 99: 15-27. DOI: 10.1016/S0167-8809(03)00138-5.

Álvarez-Martínez J.M., Jiménez-Alfaro B., Barquín J., Ondiviela B., Recio, M., Silió-Calzada, A., Juanes, J.A. (2017). Modelling the area of occupancy of habitat types with remote sensing. Methods Ecol Evol, n/a. DOI: 10.1111/2041-210X.12925.

Álvarez-Martínez, J.M., Suárez-Seoane, S., Stoorvogel, J.J., de Luis Calabuig, E. (2014). Influence of land use and climate on recent forest expansion: a case study in the Eurosiberian–Mediterranean limit of north-west Spain. J Ecol 102: 905-919. DOI: 10.1111/1365-2745.12257.

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., Morrás, H.J.M. (2016). Mapping the soils of an Argentine Pampas region using structural equation modelling. Geoderma 281: 102-118. DOI: 10.1016/j.geoderma.2016.06.031.

Arnold, R.W., Sparks, D.L. (2016). Perspectives About the National Cooperative Soil Survey, Advances in Agronomy. Academic Press, pp. 1-26. DOI: 10.1016/bs.agron.2015.11.003.

Asseng S., Ewert F., Rosenzweig C., Jones J.W., Hatfield J.L., Rueane A.C., Boote K.J., Thorburn P.J., Rötter R.P., Cammarano D., Brisson N., Basso B., Martre P., Aggarwal P.K., Angulo C., Bertuzzi P., Biernath C., Challinor A.J., Doltra J., Gayler S., Goldberg R., Grant R., Heng L., Hooker J., Hunt L.A. et al. (2013).

Uncertainty in simulating wheat yields under climate change. Nature Climate Change 3: 827-831. DOI:10.1038/nclimate1916.

Bacic, I.L.Z. (2003). Demand-driven land evaluation: with case studies in Santa Catarina, Brazil. PhD thesis. Wageningen, The Netherlands.

Balkovič, J., Rampašeková, Z., Hutár, V., Sobocká, J., Skalský, R. (2013). Digital soil mapping from conventional field soil observations. Soil and water resources 8: 13-25. ISSN 1805-9384. Retrieved from: http://www.i-scholar.in/index.php/SWRCAAS/article/view/59768 (accessed: November 2015).

Balland, V., Pollacco, J.A.P., Arp, P.A. (2008). Modeling soil hydraulic properties for a wide range of soil conditions. Ecological Modelling 219: 300-316. DOI: 10.1016/j.ecolmodel.2008.07.009.

Batjes, N.H. (1997). A world dataset of derived soil properties by FAO–UNESCO soil unit for global modelling. Soil Use and Management 13(1), 9-16. DOI: 10.1111/j.1475-2743.1997.tb00550.x.

Batjes, N.H. (2009). Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use and Management 25: 124-127. DOI: 10.1111/j.1475-2743.2009.00202.x.

Batjes, N.H. (2012). ISRIC-WISE derived soil characteristics on a 5 by 5 arc-minutes global grid (ver.1.2). Report 2012/01: 52, Wageningen, The Netherlands.

Batjes, N.H., Gicheru, P. (2004). Soil data derived from SOTER for studies of carbon stocks and change in Kenya (ver. 1.0; GEFSOC Project). Report 2004/01, ISRIC - World Soil Information, Wageningen, The Netherlands.

Bazaglia Filho, O., Rizzo, R., Lepsch, I.F., Prado, H.D., Gomes, F.H., Mazza, J.A., Demattê, J.A.M. (2013). Comparison between detailed digital and conventional soil maps of an area with complex geology. Revista Brasileira de Ciência do Solo 37: 1136-1148. DOI: 10.1590/S0100-06832013000500003.

Bhatta, G.D., Aggarwal, P.K. (2015). Coping with weather adversity and adaptation to climatic variability: a cross-country study of smallholder farmers in South Asia. Climate and Development: 1-13. DOI: 10.1080/17565529.2015.1016883.

Bindraban, P.S., Stoorvogel, J.J., Jansen, D.M., Vlaming, J., Groot, J.J.R. (2000). Land quality indicators for sustainable land management: proposed method for yield gap and soil nutrient balance. Agriculture, Ecosystems & Environment 81(2), 103-112. DOI: 10.1016/S0167-8809(00)00184-5.

Bishop, T.F.A., McBratney, A.B., Whelan, B.M. (2001). Measuring the quality of digital soil maps using information criteria. Geoderma 103: 95-111. DOI: 10.1016/S0016-7061(01)00071-4.

Bockheim, J.G., Gennadiyev, A.N., Hammer, R.D., Tandarich, J.P. (2005). Historical development of key concepts in pedology. Geoderma 124(1), 23-36. DOI: 10.1016/j.geoderma.2004.03.004.

Bond, R., Curran, J., Kirkpatrick, C., Lee, N., Francis, P. (2001). Integrated Impact Assessment for Sustainable Development: A Case Study Approach. World Development 29(6), 1011-1024. DOI: 10.1016/S0305-750X(01)00023-7.

Bonfante, A., Sellami, M.H., Abi Saab, M.T., Albrizio, R., Basile, A., Fahed, S., Giorio, P., Langella, G., Monaco, E., Bouma, J. (2017). The role of soils in the analysis of potential agricultural production: A case study in Lebanon. Agricultural Systems 156: 67-75. DOI: 10.1016/j.agsy.2017.05.018.

Boogaard, H., Wolf, J., Supit, I., Niemeyer, S., Van Ittersum, M. (2013). A regional implementation of WOFOST for calculating yield gaps of autumn-sown wheat across the European Union. Field Crops Res. 143: 130–142. DOI: 10.1016/j.fcr.2012.11.005.

Bouma, J. (1997). The role of quantitative approaches in soil science when interacting with stakeholders. Geoderma 78(1), 1-12. DOI: 10.1016/S0016-7061(97)00014-1.

Bouma, J. (2014). Soil science contributions towards Sustainable Development Goals and their implementation: linking soil functions with ecosystem services. Z. Pflanzenernähr. Bodenk. 177(2), 111-120. DOI: 10.1002/jpln.201300646.

Bouma, J. (2016). Hydropedology and the Societal Challenge of Realizing the 2015 United Nations Sustainable Development Goals. Vadose Zone Journal 15: 1-7. DOI: 10.2136/vzj2016.09.0080.

Bouma, J., Hoosbeek, M. R. (1996). The Contribution and Importance of Soil Scientists in Interdisciplinary Studies Dealing with Land. In: R. J. Wagenet, J. Bouma, editors, The Role of Soil Science in Interdisciplinary Research, SSSA Spec. Publ. 45. SSSA and ASA, Madison, WI. p. 1-15. doi:10.2136/sssaspecpub45.c1.

Bouman, B.A.M., van Keulen, H., van Laar, H.H., Rabbinge, R. (1996). The 'School of de Wit' crop growth simulation models: A pedigree and historical overview. Agricultural Systems 52: 171-198. DOI: 10.1016/0308-521X(96)00011-X.

Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A. (2016). Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma 264: 256-274. DOI: 10.1016/j.geoderma.2015.05.017.

Brus, D.J., Kempen, B., Heuvelink, G.B.M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science 62: 394-407. DOI: 10.1111/j.1365-2389.2011.01364.x.

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J. (2013). A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192: 341-353. DOI: 10.1016/j.geoderma.2012.08.020.

Canada Department of Agriculture (1970). The system of soil classification for Canada. Queen's Printer for Canada, Ottawa.

Chen, S., Shao, D., Gu, W., Xu, B., Li, H., Fang, L. (2017). An interval multistage water allocation model for crop different growth stages under inputs uncertainty. Agricultural Water Management 186: 86-97. DOI: 10.1016/j.agwat.2017.03.001.

Claessens, L., Antle, J.M., Stoorvogel, J.J., Valdivia, R.O., Thornton, P.K., Herrero, M. (2012). A method for evaluating climate change adaptation strategies for small-scale farmers using survey, experimental and modeled data. Agricultural Systems 111: 85-95. DOI: http://dx.doi.org/10.1016/j.agsy.2012.05.003.

Commission on Revenue Allocation (2011). Kenya county fact sheets. Retrieved from: http://siteresources.worldbank.org/INTAFRICA/Resources/257994-1335471959878/Kenya_County_Fact_Sheets_Dec2011.pdf (accessed: June 2015).

Conant, R.T., Ogle, S.M., Paul, E.A., Paustian, K. (2011). Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. Frontiers in Ecology and the Environment 9: 169-173. DOI: 10.1890/090153.

Cordeiro, M.R.C., Wilson, H.F., Vanrobaeys, J., Pomeroy, J.W., Fang, X. (2013). Simulating cold-region hydrology in an intensively drained agricultural watershed in Manitoba, Canada, using the Cold Regions Hydrological Model. Hydrol. Earth Syst. Sci. 21: 3483-3506. DOI: 10.5194/hess-21-3483-2017.

Cornelissen, T., Diekkrüger, B., Giertz, S. (2013). A comparison of hydrological models for assessing the impact of land use and climate change on discharge in a tropical catchment. Journal of Hydrology 498, 221-236. DOI: 10.1016/j.jhydrol.2013.06.016.

Ćosić, M., Stričević, R., Djurović, N., Moravčević, D., Pavlović, M., Todorović, M. (2017). Predicting biomass and yield of sweet pepper grown with and without plastic film mulching under different water supply and weather conditions. Agricultural Water Management 188: 91-100. DOI: 10.1016/j.agwat.2017.04.006.

CSSD (2015). China Soil Scientific Database, 2015. Map of Chinese Soils (1:12 Million), http://www.soil.csdb.cn/page/showEntity.vpage?uri=cn.csdb.soil.taxonomy1.cstTugang (accessed: September, 2015.

De Bakker, H., Schelling, J. (1966). A System of Soil Classification for The Netherlands, the Higher Levels. Pudoc, Wageningen.

De Caritat, P., Cooper, M. (2011). National Geochemical Survey of Australia: The Geochemical Atlas of Australia. Geoscience Australia Record, 2011/20 (2 Volume + Dataset), Canberra, Australian Capital Territory, Australia (available at: http://www.ga.gov.au/metadata-gateway/metadata/record/gcat_71973).

De Jager, A., Van Keulen, H., Mainah, F., Gachimbi, L.N., Itabari, J.K., Thuranira, E.G., Karuku, A.M. (2005). Attaining Sustainable Farm Management Systems in Semi-Arid Areas in Kenya: Few Technical Options, Many Policy Challenges. International Journal of Agricultural Sustainability 3: 189-205. DOI: 10.1080/14735903.2005.9684756.

De Vos, B., Cools, N., Ilvesniemi, H., Vesterdal, L., Vanguelova, E., Carnicelli, S. (2015). Benchmark values for forest soil carbon stocks in Europe: Results from a large scale forest soil survey. Geoderma 251-252: 33-46. DOI: 10.1016/j.geoderma.2015.03.008.

Deng, L., Zhu, G., Tang, Z., Shangguan, Z. (2016). Global patterns of the effects of land-use changes on soil carbon stocks. Global Ecology and Conservation 5: 127-138. DOI: 10.1016/j.gecco.2015.12.004.

Dijkshoorn, J.A. (2007). Soil and terrain database for Kenya (ver. 2.0) (KENSOTER). ISRIC- World Soil Information, Wageningen.

Dirmeyer, P.A., Guo, Z., Gao, X. (2004). Comparison, Validation, and Transferability of Eight Multiyear Global Soil Wetness Products. Journal of Hydrometeorology 5: 1011-1033. DOI: 10.1175/JHM-388.1.

Dorji, T., Odeh, I.O.A., Field, D.J., Baillie, I.C. (2014). Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. Forest Ecology and Management 318: 91-102. DOI: 10.1016/j.foreco.2014.01.003.

Dumedah, G., Walker, J.P. (2017). Assessment of model behavior and acceptable forcing data uncertainty in the context of land surface soil moisture estimation. Advances in Water Resources 101: 23-36. DOI: 10.1016/j.advwatres.2017.01.001.

Eddy, M.C., Moreda, F.G., Dykes, R.M., Bergenroth, B., Parks, A., Rineer, J. (2017). The Watershed Flow and Allocation Model: An NHDPlus-Based Watershed Modeling Approach for Multiple Scales and Conditions. J Am Water Resour Assoc 53: 6-29. DOI: 10.1111/1752-1688.12496.

Eijkelkamp (1999). ThetaProbe Soil Moisture Sensor, type ML2x User Manual. Retrieved from http://upgmbh.com/fileadmin/produkte/support/ML2x_Theta_ProbeUserManual_v1.21.pdf (accessed: June 2015).

Eijkelkamp (2004). Operating instructions, 18.40 Nitrachek Reflectometer. Retrieved from: https://en.eijkelkamp.com/products/field-measurement-equipment/functional-specs-nitrachek-reflectometer.html (accessed: June 2015).

Elliott J., Müller C., Deryng D., Chryssanthacopoulos J., Boote K. J., Büchner M., Foster I., Glotter M., Heinke J., Iizumi T., Izaurralde R. C., Mueller N. D., Ray D. K., Rosenzweig C., Ruane, A. C., Sheffield J. (2015). The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0). Geoscientific Model Development 8: 261-277. DOI: 10.5194/gmd-8-261-2015.

FAO/Unesco (1971-1981). The FAO-Unesco Soil Map of the World. Legend and 9 volumes. UNESCO, Paris, France.

FAO (1976). A framework for land evaluation. Soils Bulletin 32. FAO Rome, Italy.

FAO (1985). Guidelines: Land evaluation for irrigated agriculture - FAO soils bulletin 55. FAO Land and Water Development Division, Rome.

FAO (1987). Land evaluation for forestry – FAO forestry paper 48. Food and Agriculture Organization of the United Nations, Rome.

FAO (1993). Guidelines for land use planning. FAO Development Series 1. FAO Rome, Italy.

FAO (1995). Digital Soil Map of the World and derived properties (ver. 3.5). FAO Land and Water Digital Media Series # 1. Food and Agriculture Organization of the United Nations, Rome.

FAO (2006). Guidelines for soil description (4th ed). Food and Agriculture Organization of the United Nations, Rome.

FAO (2016). FAOSTAT database collections. FAO, Rome. Access date: 2016-05-05. URL: http://faostat.fao.org.

FAO Statistics Devision (2015). Faostat.fao.org. Retrieved from: http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor (accessed: 23 June 2015).

FAO/IIASA/ISRIC/ISSCAS/JRC (2012). Harmonized World Soil Database (version 1.2). Food and Agriculture Organization of the United Nations, Rome and IIASA, Laxenburg.

FAO-UNESCO (1988). Soil map of the world: Revised legend (with corrections and updates): World Soil Resources Report 60. Food and Agriculture Organization of the United Nations, Rome.

Farrar, D.M., Coleman, J.D. (1967). The correlation of surface area with other properties of nineteen british clay soils. Journal of Soil Science 18: 118-124.

Fischer, G., Van Velthuizen, H., Shah, M., Nachtergaele, F.O. (2002). Global agro-ecological assessment for agriculture in the 21st century: methodology and results. Research Report RR-02-02. IIASA, Laxenburg, Austria.

Fischer, G., F. Nachtergaele, S. Prieler, H.T. van Velthuizen, L. Verelst, D. Wiberg (2008). Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008). IIASA, Laxenburg, Austria and FAO, Rome.

Fraga, H., García de Cortázar Atauri, I., Malheiro, A.C., Santos, J.A. (2016). Modelling climate change impacts on viticultural yield, phenology and stress conditions in Europe. Glob Change Biol 22(11), 3774-3788. DOI: 10.1111/gcb.13382.

Franco, A.L.C., Bartz, M.L.C., Cherubin, M.R., Baretta, D., Cerri, C.E.P., Feigl, B.J., Wall, D.H., Davies, C.A., Cerri, C.C. (2016). Loss of soil (macro)fauna due to the expansion of Brazilian sugarcane acreage. Science of the Total Environment 563-564: 160-168. DOI: 10.1016/j.scitotenv.2016.04.116.

Franco, J. (1992). Nitrate management program - fertilizer research and education program: Progress report 1991. Communications in soil science and plant analysis 23: 2111-2134. DOI: 10.1080/00103629209368729.

Fresco, L., Huizing, H., Van Keulen, H., Luning, H., Schipper, R. (1990). Land evaluation and farming system analysis for land use planning. FAO guidelines: Working Document.

FURP (1987). The Fertilizer Use Recommendation Project, Final Report. Annex 1: Fertilizer trial documentation (FERDOC). Minestry of Agriculture, Nairobi, Kenya.

FURP (1994). Fertilizer use recommendations, volumes 1 to 24. Fertilizer Use Recommendation Project, Kenya Agricultural Research Institute, National Agricultural Research Laboratories, Nairobi, Kenya.

Galvánek, D., Lepš, J. (2012). The effect of management on productivity, litter accumulation and seedling recruitment in a Carpathian mountain grassland. Plant Ecology 213: 523-533. DOI: 10.1007/s11258-011-9999-7.

Gasch, C.K., Hengl, T., Gräler, B., Meyer, H., Magney, T.S., Brown, D.J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The Cook Agronomy Farm data set. Spatial Statistics 14(Part A), 70-90. DOI: 10.1016/j.spasta.2015.04.001.

Ge, Y., Thomasson, J.A., Sui, R. (2011). Remote sensing of soil properties in precision agriculture: A review. Frontiers of Earth Science 5(3), 229-238. DOI: 10.1007/s11707-011-0175-0.

Geissen, V., Peña-Peña, K., Huerta, E. (2009). Effects of different land use on soil chemical properties, decomposition rate and earthworm communities in tropical Mexico. Pedobiologia 53: 75-86. DOI: j.pedobi.2009.03.004.

Gelens, H.F., Kinyan1ui, H.C.K., Van De Weg, R.F. (1976). Soils of the Kapenguna area. Reconnaissance soil survey report R2, Kenya Soil Survey. Nairobi.

Geological Service NRW (2004). Bodenkarte 1: 50.000 (BK50). Karte der schutzwürdigen Böden. Bear-beitungsmaßstab 1:50.000. CD-ROM. Geologischer Dienst Nordrhein-Westfalen. Krefeld. http://www.gd.nrw.de/g_bk50d.htm (accessed: January, 2015).

Geza, M., McCray, J.E. (2008). Effects of soil data resolution on SWAT model stream flow and water quality predictions. Journal of Environmental Management 88(3), 393-406. DOI: 10.1016/j.jenvman.2007.03.016.

Goubanova, K., Li, L. (2007). Extremes in temperature and precipitation around the Mediterranean basin in an ensemble of future climate scenario simulations. Global and Planetary Change 57: 27-42. DOI: 10.1016/j.gloplacha.2006.11.012.

Government of Kenya (2016). Sessional Paper No. 3 of 2016 on National Climate Change Framework Policy. Ministry of Environment and Natural Resources, State Department of Environment.

Goyens, C., Verdoodt, A., Van de Wauw, J., Baert, G., van Engelen, V.W.P., Dijkshoorn, J.A., Van Ranst, E. (2007). Base de données numériques sur les sols et le terrain (SOTER) de l'Afrique Centrale (RD Congo, Rwanda et Burundi). Etude et Gestion des Sols 14: 207-218.

Grant, R.F. (1995). Dynamics of energy, water, carbon and nitrogen in agricultural ecosystems: simulation and experimental validation. Ecological Modelling 81: 169-181. DOI: 10.1016/0304-3800(94)00169-I.

Grassini, P., Van Bussel, L.G.J., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., De Groot, H., Van Ittersum, M.K., Cassman, K.G. (2015). How good is good enough? Data requirements for reliable crop simulations and yield-gap analysis. Fied Crops Research 177: 49-63. DOI:10.1016/j.fcr.2015.03.004.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146: 102-113. DOI: 10.1016/j.geoderma.2008.05.008.

Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152: 195-207. DOI:10.1016/j.geoderma.2009.06.003.

Grunwald, S., Thompson, J.A., Boettinger, J.L. (2011). Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. Soil Sci. Soc. Am. J. 75:1201-1213. DOI: 10.2136/sssaj2011.0025.

Guilpart, N., Grassini, P., Van Wart, J., Yang, H., Van Ittersum, M.K., Van Bussel, L.G.J., Wolf, J., Claessens, L., Leenaars, J.G.B., Cassman, K.G. (2017). Rooting for food security in Sub-Saharan Africa. Environmental Research Letters 12: 114036. DOI: 10.1088/1748-9326/aa9003.

Guo, L.B., Gifford, R.M. (2002). Soil carbon stocks and land use change: a meta analysis. Global Change Biology 8: 345-360. DOI: 10.1046/j.1354-1013.2002.00486.x.

Hartemink, A.E., McBratney, A.B., Cattle, J.A. (2001). Developments and trends in soil science: 100 volumes of Geoderma (1967-2001). Geoderma 100: 217-268. DOI: 10.1016/S0016-7061(01)00024-6.

Hartemink, A.E., Sonneveld, M.P.W. (2013). Soil maps of the Netherlands. Geoderma 204-205: 1-9. DOI: 10.1016/j.geoderma.2013.03.022.

Hartmann, J., Moosdorf, N. (2011). Chemical weathering rates of silicate-dominated lithological classes and associated liberation rates of phosphorus on the Japanese Archipelago—Implications for global scale analysis. Chemical Geology 287: 125-157. DOI: 10.1016/j.chemgeo.2010.12.004.

Hendriks, C.M.J., Stoorvogel, J.J., Claessens, L. (2016). Exploring the challenges with soil data in regional land use analysis. Agricultural Systems 144: 9-21. DOI: 10.1016/j.agsy.2016.01.007.

Hendriks, C.M.J., Stoorvogel, J.J., Lutz, F., Claessens, L. (in review). Combining new and legacy soil data to enrich local to regional land use analyses. Geoderma.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R. (2014). SoilGrids1km — Global Soil Information Based on Automated Mapping. PLOS ONE 9: e105992. DOI: 10.1371/journal.pone.0105992.

Hengl, T., Heuvelink, G.B.M., Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120: 75-93. DOI: 10.1016/j.geoderma.2003.08.018.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE 12: e0169748. DOI: 10.1371/journal.pone.0169748.

Herweg, K., Ludi, E. (1999). The performace of selected soil and water conservation measures - case studies from Ethiopia and Eritrea. Catena 36: 99-114. DOI: 10.1016/S0341-8162(99)00004-1.

Heuvelink, G.B.M. (1998). Uncertainty analysis in environmental modelling under a change of spatial scale. Nutrient Cycling in Agroecosystems 50: 255-264. DOI: 10.1023/A:1009700614041.

Heuvelink, G.B.M. (2014). Uncertainty quantification of GlobalSoilMap products. D. Arrouays, N.J. McKenzie, J.W. Hempel, A. Richer de Forges, A.B. McBratney (Eds.), GlobalSoilMap. Basis of the Global Spatial Soil Information System, Proceedings of the First GlobalSoilMap Conference, Orleans, France, 7–9th October 2013, Taylor & Francis, Balkema, The Netherlands (2014)

Heuvelink, G.B.M., Webster, R. (2001). Modelling soil variation past, present, and future. Geoderma 100: 269-301. DOI: 10.1016/S0016-7061(01)00025-8.

Hiepe, C. (2008). Soil Degradation by Water Erosion in a Sub-humid West-African Catchment: A Modelling Approach Considering Land Use and Climate Change in Benin. University of Bonn, Bonn.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. Int. J. Clim., 25, 1965–1978. DOI: 10.1002/joc.1276.

Himanshu, S.K., Pandey, A., Shrestha, P. (2016). Application of SWAT in an Indian river basin for modeling runoff, sediment and water balance. Environmental Earth Sciences 76: 3. DOI: 10.1007/s12665-016-6316-8.

Hodnett, M.G., Tomasella, J. (2002). Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: a new water-retention pedo-transfer functions developed for tropical soils. Geoderma 108: 155-180. DOI: 10.1016/S0016-7061(02)00105-2.

Hoffmann, H., Zhao, G., Asseng, S., Bindi, M., Biernath, C., Constantin, J., Coucheney, E., Dechow, R., Doro, L., Eckersten, H., Gaiser, T., Grosz, B., Heinlein, F., Kassie, B.T., Kersebaum, K.-C., Klein, C., Kuhnert, M., Lewan, E., Moriondo, M., Nendel, C., Priesack, E., Raynal, H., Roggero, P.P., Rötter, R.P., Siebert, S., Specka, X., Tao, F., Teixeira, E., Trombi, G., Wallach, D., Weihermüller, L., Yeluripati, J., Ewert, F. (2016). Impact of Spatial Soil and Climate Input Data Aggregation on Regional Yield Simulations. PLOS ONE 11(4), e0151782.

Hoogenboom, G., Jones, J.W., Wilkens, P.W., Porter, C.H., Boote, K.J., Hunt, L.A., Singh, U., Lizaso, J.I., White, J.W., Uryasev, O., Ogoshi, R., Koo, J., Shelia, V., Tsuji, G.Y. (2015). Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.6. DSSAT Foundation, Prosser, Washington.

Hoosbeek, M.R., Bryant, R.B. (1992). Towards the quantitative modeling of pedogenesis — a review. Geoderma 55(3), 183-210. DOI: 10.1016/0016-7061(92)90083-J.

Igue, A.M., Gaiser, T., Stahr, K. (2004). A soil and terrain digital database (SOTER) for improved land use planning in Central Benin. European Journal of Agronomy 21, 41-52. DOI: 10.1016/S1161-0301(03)00062-5.

Isbell, R., National Committee on Soil and Terrain (2016). The Australian soil classification, 2nd edition. CSIRO Publishing, Clayton.

ISRIC–World Soil Information (2013). Soil property maps of Africa at 1 km. Retrieved from: http://www.isric.org/data/soil-property-maps-africa-1-km (accessed: June 2015).

IUSS Working Group WRB (2015).World Reference Base for Soil Resources 2014, Update 2015 International Soil Classification System For Naming Soils And

Creating Legends For Soil Maps. World Soil Resources Reports No. 106. FAO, Rome.

Jackson, T.A. (2015). Weathering, secondary mineral genesis, and soil formation caused by lichens and mosses growing on granitic gneiss in a boreal forest environment. Geoderma 251-252:78-91. DOI: 10.1016/j.geoderma.2015.03.012.

Jaetzold, R., Schmidt, H. (1982). Farm Management Handbook of Kenya. Volume II: natural conditions and farm management information. Part B: Central Kenya. Ministry of Agriculture in Cooperation with German Agency for Technical Cooperation, Nairobi, Kenya.

Jaiyeoba, I.A. (2003). Changes in soil properties due to continuous cultivation in Nigerian semiarid Savannah. Soil and Tillage Research 70: 91-98. DOI: 10.1016/S0167-1987(02)00138-1.

Jamagne, M., Le Bas, C., Berland, M., Eckelmann, W. (1995). Extension of the EU database for the soils of central and eastern Europe. In: King, D., Jones, R.J.A., Thomasson, A.J. (Eds.). EUR 16232 EN, Office for Official Publications of the European Communities, Luxembourg.

Jenny (1941). Factors of soil formation: A system of quantitative pedology. Dover Publications, Inc., New York.

Jeong, G., Choi, K., Spohn, M., Park, S.J., Huwe, B., Ließ, M. (2017). Environmental drivers of spatial patterns of topsoil nitrogen and phosphorus under monsoon conditions in a complex terrain of South Korea. PLOS ONE 12: e0183205.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T. (2003). The DSSAT cropping system model. European Journal of Agronomy 18: 235-265. DOI: 10.1016/S1161-0301(02)00107-7.

Jones, R.J.A., Hiederer, R., Rusco, E., Montanarella, L. (2005). Estimating organic carbon in the soils of Europe for policy support. European Journal of Soil Science 56: 655-671. DOI: 10.1111/j.1365-2389.2005.00728.x.

Keeble, B.R. (1988). The Brundtland report: 'Our common future'. Medicine and War 4(1), 17-25.

Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett, R.D., Moolenaar, S., Mol, G., Jansen, B., Fresco, L.O. (2016). The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. SOIL 2(2), 111-128. DOI: 10.5194/soil-2-111-2016.

Kempen, B. (2011). Updating soil information with digital soil mapping. PhD thesis. Wageningen.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J. (2009). Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151: 311-326. DOI: 10.1016/j.geoderma.2009.04.023.

Kempen, B., Brus, D.J., Stoorvogel, J.J. (2011). Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. Geoderma 162: 107-123. DOI: 10.1016/j.geoderma.2011.01.010.

Khodaverdiloo, H., Homaee, M., van Genuchten, M.T., Dashtaki, S.G. (2011). Deriving and validating pedotransfer functions for some calcareous soils. Journal of Hydrology 399: 93-99. DOI: 10.1016/j.jhydrol.2010.12.040.

Kidd, D., Webb, M., Malone, B., Minasny, B., McBratney, A. (2015). Eighty-metre resolution 3D soil-attribute maps for Tasmania, Australia. Soil Research 53: 932-955. DOI: 10.1071/SR14268.

Lacoste, M., Minasny, B., McBratney, A.B., Michot, D., Viaud, V., Walter, C. (2014). High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213: 296-311. DOI: 10.1016/j.geoderma.2013.07.002.

Lagacherie, P., McBratney, A.B., Lagacherie, P., McBratney, A.B., Voltz, M. (2006). Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping, Developments in Soil Science. Elsevier, pp. 3-22. DOI: 10.1016/S0166-2481(06)31001-X.

Lal, R. (2003). Soil erosion and the global carbon budget. Environment International 29: 437-450.

Leenaars, J.G.B., Hengl, T., Ruiperez González, M., Mendes de Jesus, J.S., Heuvelink, G.B.M., Wolf, J., Van Bussel, L.G.J., Claessens, L., Yang, H., Cassman, K.G. (2015). Root Zone Plant-Available Water Holding Capacity of the Sub-Saharan Africa Soil: Gridded functional soil information (dataset RZ-PAWHC SSA version 1.0). ISRIC Report 2015/02. Collaboration project of Africa Soil Information Service (AfSIS) and Global Yield Gap and Water Productivity Atlas (GYGA). ISRIC - World Soil Information, Wageningen.

Li, Q., Yue, T.., Wang, C.., Zhang, W.., Yu, Y., Li, B., Yang, J., Bai, G. (2013). Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach. CATENA 104: 210-218. DOI: 10.1016/j.catena.2012.11.012.

Liu J., Williams J.R., Zehnder A.J.B., Yang H. (2007). GEPIC - modelling wheat yield and crop water productivity with high resolution on global scale. Agricultural Systems 94: 478-493. DOI: 10.1016/j.agsy.2006.11.019.

Liu, X., Wu, J., Xu, J. (2006). Characterizing the risk assessment of heavy metals and sampling uncertainty analysis in paddy field by geostatistics and GIS. Environmental pollution 141: 257-264. DOI: 10.1016/j.envpol.2005.08.048.

Liu, Y., Yang, S., Li, S., Chen, F. (2012). Application of the Hybrid-Maize model for limits to maize productivity analysis in a semiarid environment. Scientia Agricola 69, 300-307. DOI: 10.1590/S0103-90162012000500003.

Lobell, D.B., Field, C.B. (2007). Global scale climate–crop yield relationships and the impacts of recent warming. Environ.Res.Lett. 2: 014002. DOI: 10.1088/1748-9326/2/1/014002.

Lorenz, K., Lal, R. (2014). Soil organic carbon sequestration in agroforestry systems. A review. Agronomy for Sustainable Development 34: 443-454. DOI: 10.1007/s13593-014-0212-y.

Malagò, A., Efstathiou, D., Bouraoui, F., Nikolaidis, N.P., Franchini, M., Bidoglio, G., Kritsotakis, M. (2016). Regional scale hydrologic modeling of a karst-dominant geomorphology: The case study of the Island of Crete. Journal of Hydrology 540: 64-81. DOI: 10.1016/j.jhydrol.2016.05.061.

Manzoni, S., Porporato, A. (2009). Soil carbon and nitrogen mineralization: Theory and models across scales. Soil Biology and Biochemistry 41: 1355-1379. DOI: 10.1016/S0160-4120(02)00192-7.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M. (2000). An overview of pedometric techniques for use in soil survey. Geoderma 97: 293-327. DOI: http://dx.doi.org/10.1016/S0016-7061(00)00043-4.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B. (2003). On digital soil mapping. Geoderma 117: 3-52. DOI: http://dx.doi.org/10.1016/S0016-7061(03)00223-4.

Milne, G. (1936). A provisional soil map of East Africa. East Africa Agricultural Research Station, Amani, Memoir, 34.

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L. (2017). Soil carbon 4 per mille. Geoderma 292: 59-86. DOI: 10.1016/j.geoderma.2017.01.002.

Minasny, B., McBratney, A.B. (2006). Mechanistic soil–landscape modelling as an approach to developing pedogenetic classifications. Geoderma 133: 138-149. DOI: 10.1016/j.geoderma.2006.03.042.

Ministry of Agriculture and Ministry of Livestock and Fisheries Development (2004). Strategy for Revitalizing Agriculture, 2004-2014. Ministry of Agriculture, Kenya.

Monmonier, M. (1996). How to lie with maps (2nd edition). University of Chicago Press, USA.

Mora-Vallejo, A., Claessens, L., Stoorvogel, J.J., Heuvelink G.B.M. (2008). Small scale digital soil mapping in Southeastern Kenya. CATENA 76: 44-53. DOI: http://dx.doi.org/10.1016/j.catena.2008.09.008.

Mottaleb, K.A., Rejesus, R.M., Murty, M., Mohanty, S., Li, T. (2017). Benefits of the development and dissemination of climate-smart rice: ex ante impact assessment of drought-tolerant rice in South Asia. Mitigation and Adaptation Strategies for Global Change 22: 879-901. DOI: 10.1007/s11027-016-9705-0.

Mulder, V.L., De Bruin, S., Schaepman, M.E., Mayr, T.R. (2011). The use of remote sensing in soil and terrain mapping — A review. Geoderma 162: 1-19. DOI: http://dx.doi.org/10.1016/j.geoderma.2010.12.018.

Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D. (2016). National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263: 16-34. DOI: 10.1016/j.geoderma.2015.08.035.

Müller, C., Bondeau, A., Popp, A., Waha, K., Fader, M. (2010). Climate change impacts on agricultural yields. Development and Climate Change Background Note to the World Development Report 2010. Potsdam Institute for Climate Impact Research, Germany.

Nachtergaele, F., Van Ranst, E. (2003). Qualitative and Quantitative Aspects of Soil Databases in Tropical Countries. In: Stoops (ed.). Evolution of Tropical Soil Science. Past and Future, 107–126. Royal Academy of Overseas Sciences, Brussels.

Ninyerola, M., Pons, X., Roure, J.M. (2007). Objective air temperature mapping for the Iberian Peninsula using spatial interpolation and GIS. Int. J. Climatol. 27: 1231-1242. DOI: 10.1002/joc.1462.

Norwegian Forest and Landscape Institute (2014).Kilden–til Arealinformasjon. Ås, Norway: Norwegian Forest and Landscape Institute. http://www.skogo-glandskap.no/kart/kilden (accessed: February 2016).

Omuto, C., Nachtergaele, F., Vargas Rojas, R. (2013). State of the Art Report on Global and Regional Soil Information: Where are we? Where to go? Global soil partnership, technical report. Food and Agricultre Organization of the United Nations, Rome.

Omuto, C.T., Paron, P., Olago, D.O., Omuto, C.T. (2013). Chapter 11 - Major Soil and Data Types in Kenya, Developments in Earth Surface Processes. Elsevier, pp. 123-132.

Ouyang, W., Cai, G., Tysklind, M., Yang, W., Hao, F., Liu, H. (2017). Temporal-spatial patterns of three types of pesticide loadings in a middle-high latitude agricultural watershed. Water Research 122: 377-386. DOI: 10.1016/j.watres.2017.06.023.

Pagani, V., Guarneri, T., Fumagalli, D., Movedi, E., Testi, L., Klein, T., Calanca, P., Villalobos, F., Lopez-Bernal, A., Niemeyer, S., Bellocchi, G., Confalonieri, R. (2017). Improving cereal yield forecasts in Europe – The impact of weather extremes. European Journal of Agronomy 89: 97-106. DOI: 10.1016/j.eja.2017.06.010.

Palacios-Orueta, A., Ustin, S.L. (1998). Remote sensing of soil properties in the Santa Monica Mountains I. Spectral Analysis. Remote sensing of environment 65: 170-183. DOI: 10.1016/S0034-4257(98)00024-8.

Palazón, L., Navas, A. (2016). Case Study: Effect of Climatic Characterization on River Discharge in an Alpine-Prealpine Catchment of the Spanish Pyrenees Using the SWAT Model. Water 8: 471. DOI 10.3390/w8100471.

Panagos, P., Borrelli, P., Poesen, J., Ballabio, C., Lugato, E., Meusburger, K., Montanarella, L., Alewell, C. (2015). The new assessment of soil loss by water erosion in Europe. Environmental Science & Policy 54: 438-447. DOI: 10.1016/j.envsci.2015.08.012.

Panagos, P., Jones, A., Bosco, C., Senthil Kumar, P.S. (2011). European digital archive on soil maps (EuDASM): preserving important soil data for public free access. International Journal of Digital Earth 4: 434-443. DOI: 10.1080/17538947.2011.596580.

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G.P., Smith, P. (2016). Climate-smart soil. Nature 532: 49-57. DOI: 10.1038/nature17174.

Pelegrino, M.H.P., Silva, S.H.G., de Menezes, M.D., da Silva, E., Owens, P.R., Curi, N. (2016). Mapping soils in two watersheds using legacy data and extrapolation for similar surrounding areas. Ciencia e Agrotechnologia 40: 534-546. DOI: 10.1590/1413-70542016405011416.

Pereira, F.F.S., Sánchez-Román, R.M., Orellana González, A.M.G. (2017). Simulation model of the growth of sweet orange (Citrus sinensis L. Osbeck) cv. Natal in response to climate change. Climatic Change 143: 101-113. DOI: 10.1007/s10584-017-1986-0.

Persson, T., Kværnø, S. (2017). Impact of projected mid-21st century climate and soil extrapolation on simulated spring wheat grain yield in Southeastern Norway. The Journal of Agricultural Science 155(3), 361-377. DOI: 10.1017/S0021859616000241.

Potter, C.S. (1997). An ecosystem simulation model for methane production and emission from wetlands. Global Biogeochem. Cycles 11: 495-506. DOI:

10.1029/97GB02302.

Rahman, M., Rosolem, R. (2017). Towards a simple representation of chalk hydrology in land surface modelling. Hydrol. Earth Syst. Sci. 21: 459-471. DOI: 10.5194/hess-21-459-2017.

Reid, R.E. (1988). Soil survey specifications. In: Guidelines for Conducting Surveys: Australian Land and Survey Handbook. Gunn, R.H., Beattie, J.A., Reid, R.E., Van der Graaf, R.H.M. (eds.). Inkata Press, Melbourne.

Renschler, C.S., Harbor, J. (2002). Soil erosion assessment tools from point to regional scales—the role of geomorphologists in land management research and implementation. Geomorphology 47: 189-209. DOI: 10.1016/S0169-555X(02)00082-X.

Rey, A., Belelli-Marchesini, L., Etiope, G., Papale, D., Canfora, E., Valentini, R., Pegoraro, E. (2014). Partitioning the net ecosystem carbon balance of a semiarid steppe into biological and geological components. Biogeochemistry 118: 83-101. DOI: 10.1007/s10533-013-9907-4.

Rinaldi, M., Scalcione, E., Perniola, M., Maddaluno, C., Garofalo, P. (2017). Soil tillage and residues management in wheat continuous cropping in Southern Italy: A model application for agronomic and soil fertility assessment. Computers and Electronics in Agriculture 140: 77-87. DOI: 10.1016/j.compag.2017.05.037.

Ritung, S., Wahyunto, K.N., Sukarman, H., Suparto, C.T. (2011). Petalahan gambut Indonesia. Skala 1:250.000 (Maps of Peat-land Distribution in Indonesia). Balai Besar Sumber DayaLahan Pertanian (BBSDLP) Ministry of Agriculture, Republic of Indonesia.

Robinson, T.P., Metternicht, G. (2006). Testing the performance of spatial interpolation techniques for mapping soil properties. Computers and Electronics in Agriculture 50: 97-108. DOI: 10.1016/j.compag.2005.07.003.

Rodríguez Martín, J.A., Álvaro-Fuentes, J., Gonzalo, J., Gil, C., Ramos-Miras, J.J., Grau Corbí, J.M., Boluda, R. (2016). Assessment of the soil organic carbon stock in Spain. Geoderma 264: 117-125. DOI: 10.1016/j.geoderma.2015.10.010.

Rötter R.P., Carter T.R., Olesen J.E., Porter J.R. (2011). Crop-climate models need an overhaul. Nature Climate Change 1: 175-177. DOI: 10.1038/nclimate1152.

Saiz, G., Wandera, F.M., Pelster, D.E., Ngetich, W., Okalebo, J.R., Rufino, M.C., Butterbach-Bahl, K. (2016). Long-term assessment of soil and water conservation measures (Fanya-juu terraces) on soil organic matter in South Eastern Kenya. Geoderma 274: 1-9. DOI: 10.1016/j.geoderma.2016.03.022.

Samuel Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C. (2015). Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243-244: 214-227. DOI: 10.1016/j.geoderma.2014.12.017.

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.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L. (2009). Digital Soil Map of the World. Science 325: 680-681. DOI: 10.1126/science.1175084.

Saxton, K.E., Rawls, W.J. (2006). Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions. Soil Science Society of America Journal 70: 1569-1578. DOI: 10.2136/sssaj2005.0117.

Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., Acutis, M. (2017). Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. Geoderma 286: 35-45. DOI: 10.1016/j.geoderma.2016.10.019.

Schoknecht, N., Wilson, P.R., Heiner, I. (2008). Survey specification and planning. In: Guidelines for Surveying Soil and Land Resources. 2nd ed. McKenzie, N.J., Grundy, M.J., Webster, R. Ringroase-Voase, A.J. (eds.). CSIRO Publishing, Melbourne.

Schwab, N., Schickhoff, U., Fischer, E. (2015). Transition to agroforestry significantly improves soil quality: A case study in the central mid-hills of Nepal. Agriculture, Ecosystems & Environment 205: 57-69. DOI: 10.1016/j.agee.2015.03.004.

Sexton, J., Everingham, Y.L., Inman-Bamber, G. (2017). A global sensitivity analysis of cultivar trait parameters in a sugarcane growth model for contrasting production environments in Queensland, Australia. European Journal of Agronomy 88: 96-105. DOI: 10.1016/j.eja.2015.11.009.

Shibu, M.E., Leffelaar, P.A., Keulen, H. van, Aggarwal, P.K. (2006). Quantitative description of soil organic matter dynamics - A review of approaches with reference to rice-based cropping systems. Geoderma 137: 1-18. DOI: 10.1016/j.geoderma.2006.08.008.

Shrestha, S., Chapagain, R., Babel, M.S. (2017). Quantifying the impact of climate change on crop yield and water footprint of rice in the Nam On Irrigation Project, Thailand. Science of The Total Environment 599-600: 689-699. DOI: 10.1016/j.scitotenv.2017.05.028.

Sintondji, L. (2005). Modelling the Rainfall-Runoff Process in the Upper Ouémé Catchment (Térou in Benin Republic) in a Context of Global Change: Extrapolation from the Local to the Regional Scale. Shaker, Aachen.

Smit, B., McNabb, D., Smithers, J. (1996). Agricultural adaptation to climatic variation. Climatic Change 33: 7-29. DOI:10.1007/BF00140511.

Smith, P., Smith, J.U., Powlson, D.S., McGill, W.B., Arah, J.R.M., Chertov, O.G., Coleman, K., Franko, U., Frolking, S., Jenkinson, D.S., Jensen, L.S., Kelly, R.H., Klein-Gunnewiek, H., Komarov, A.S., Li, C., Molina, J.A.E., Mueller, T., Parton, W.J., Thornley, J.H.M., Whitemore, A.P. (1997). A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. Geoderma 81: 153-225. DOI:10.1016/S0016-7061(97)00087-6.

Soil Science Division Staff (2017). Soil survey manual. C. Ditzler, K. Scheffe, and H.C. Monger (eds.). USDA Handbook 18. Government Printing Office, Washington, D.C.

Soil Survey Division Staff (1993). Soil survey manual. Soil Conservation Service. U.S. Department of Agriculture Handbook 18. Washington DC.

Soil Survey Staff (1992). Keys to soil taxonomy. 5th ed. SMSS Tech. Monogr. 19. Pocahontas Press, Blacksburg, VA.

Soil Survey Staff (2014). Keys to soil taxonomy, 12th edition. USDA-Natural Resources Conservation Service, Washington, DC.

Sombroek, W.G., Braun, H.M.H., Van Der Pouw, B.J.A. (1982). Exploratory Soil Map and Agro-Climatic Zone Map of Kenya, 1980 (scale 1:1,000,000). Exploratory Soil Survey Report No. E1, Kenya Soil Survey, National Agricultural Laboratories, Ministry of Agriculture, Nairobi, Kenya.

Song, X., Brus, D., Liu, F., Li, D., Zhao, Yang, J.L., Zhang, G.L. (2016). Mapping soil organic carbon content by geographically weighted regression: A case study in the Heihe River Basin, China. Geoderma 261: 11-22. DOI: 10.1016/j.geoderma.2015.06.024.

Stanford, G., Smith, S.J. (1972). Nitrogen Mineralization Potentials of Soils1. Soil Science Society of America Journal 36: 465-472. DOI:10.2136/sssaj1972.03615995003600030029x.

Stehman, S.V. (1999). Basic probability sampling designs for thematic map accuracy assessment. International Journal of Remote Sensing 20: 2423-2441. DOI: 10.1080/014311699212100.

Stein, A., Bouma, J., Mulders, M.A. (1989). Using spatial variability studies to predict physical land qualities of a level river terrace. Soil Technology 2: 385-403. DOI: 10.1016/0933-3630(89)90005-6.

Steur, G.G.L., Heijink, W. (1991). Bodemkaart van Nederland Schaal 1:50000, Algemene begrippen en indelingen. Van der Wiel en Smit B.V., Nederland.

Stoorvogel, J.J. (2014). S-world: A global map of soil properties for modelling. In: Proceedings of the 1st GlobalSoilMap Conference on GlobalSoilMap: Basis of the global spatial soil information system. Orleans, France, 7 - 9 October 2013. Taylor & Francis Group UK: 227-231.

Stoorvogel, J.J., Antle, J. (2001). Regional land use analysis: the development of operational tools. Agricultural Systems 70(2-3), 623-640. DOI: 10.1016/S0308-521X(01)00062-2.

Stoorvogel, J.J., Bakkenes, M., Temme, A.J.A.M., Batjes, N.H., Ten Brink, B.J.E. (2017). S-World: A Global Soil Map for Environmental Modelling. Land Degradation and Development. 28: 22-33. DOI: 10.1002/ldr.2656.

Stoorvogel, J.J., Bouma, J., Orlich, R.A. (2004). Participatory research for systems analysis: Prototyping for a Costa Rican banana plantation. Agronomy Journal 96: 323-336.

Stoorvogel, J.J., Hendriks, C.M.J., Claessens, L. In prep. Field procedure for measuring the soil texture.

Streutker, D.R., Glenn, N.F. (2006). LiDAR measurement of sagebrush steppe vegetation heights. Remote Sensing of Environment 102: 135-145. DOI: 10.1016/j.rse.2006.02.011.

Stumpf, F., Schmidt, K., Goebes, P., Behrens, T., Schönbrodt-Stitt, S., Wadoux, A., Xiang, W., Scholten, T. (2017). Uncertainty-guided sampling to improve digital soil maps. CATENA 153: 30-38. DOI: 10.1016/j.catena.2017.01.033.

Sulaeman, Y., Minasny, B., McBratney, A.B., Sarwani, M., Sutandi A. (2013). Harmonizing legacy soil data for digital soil mapping in Indonesia. Geoderma 192: 77-85. DOI: 10.1016/j.geoderma.2012.08.005.

TAC (1988). Sustainable agricultural production: implications for international agricultural research. TAC, Rome.

Tarnocai, C., Canadell, J.G., Schuur, E.A.G., Kuhry, P., Mazhitova, G., Zimov, S. (2009). Soil organic carbon pools in the northern circumpolar permafrost region. Global biogeochemical cycles 23: GB2023. DOI: 10.1029/2008GB003327.

Tateno, R., Hishi, T., Takeda, H. (2004). Above- and belowground biomass and net primary production in a cool-temperate deciduous forest in relation to topographical changes in soil nitrogen. Forest Ecology and Management 193: 297-306. DOI: 10.1016/j.foreco.2003.11.011.

Thorlakson, T., Neufeldt, H. (2012). Reducing subsistence farmers’ vulnerability to climate change: evaluating the potential contributions of agroforestry in western Kenya. Agriculture & Food Security 1: 1-13. DOI: 10.1186/2048-7010-1-15.

Tiffen, M., Mortimore, M., Gichuki, F. (1994). More people, less erosion: environmental recovery in Kenya. John Wiley and Sons Ltd, UK.

Tiktak, A., de Nie, D.S., Piñeros Garcet, J.D., Jones, A., Vanclooster, M. (2004). Assessment of the pesticide leaching risk at the Pan-European level. The

EuroPEARL approach. Journal of Hydrology 289(1), 222-238. DOI: 10.1016/j.jhydrol.2003.11.030.

Tomasella, J., Hodnett, M. G. (1998). Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Science, 163(3), 190-202. DOI: 10.1097/00010694-199803000-00003.

Tukiainen, H., Bailey, J.J., Field, R., Kangas, K., Hjort, J. (2017). Combining geodiversity with climate and topography to account for threatened species richness. Conservation Biology 31: 364-375.

Uda, S.K., Hein, L., Sumarga, E. (2017). Towards sustainable management of Indonesian tropical peatlands. Wetlands Ecology and Management 25(6), 683-701. DOI: 10.1007/s11273-017-9544-0.

UN General Assembly (2000). United Nations Millennium Declaration, Resolution Adopted by the General Assembly. A/RES/55/2, available at: http://www.refworld.org/docid/3b00f4ea3.html (accessed: December 2017).

UN-SDSN (2014). United Nations-Sustainable Development Solutions Network. An Action Agenda for Sustainable Development. Report for the UN Secretary-General. Prepared by the Leadership Council of the Sustainable Development Solutions Network.

Vågen T.G., Lal R., Singh B.R. (2005). Soil carbon sequestration in sub-Saharan Africa: a review. Land Degradation & Development 16: 53-71. DOI: 10.1002/ldr.644.

Valdivia, C., Barbieri, C., Gold, M.A. (2012). Between Forestry and Farming: Policy and Environmental Implications of the Barriers to Agroforestry Adoption. Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie 60: 155-175. DOI: 10.1111/j.1744-7976.2012.01248.x.

Van Engelen, V.W.P., Wen, T.T. (1995). Global and national soils and terrain databases (SOTER). Procedures manual (revised edition). UNEP-ISSS-ISRIC-FAO, Wageningen.

Van Engelen, V.W.P, Dijkshoorn, J.A. (2013). Global and national Soils and Terrain Databases (SOTER). Procedures Manual, version 2.0, ISRIC Report 2013/04, ISRIC - World Soil Information, Wageningen.

Van Ittersum, M.K., Cassman, K.G., Grassini, P., Wolf, J., Tittonell, P., Hochman, Z. (2013). Yield gap analysis with local to global relevance—A review. Field Crops Research 143: 4-17. DOI: http://dx.doi.org/10.1016/j.fcr.2012.09.009.

Van Ittersum, M.K., Leffelaar, P.A., van Keulen, H., Kropff, M.J., Bastiaans, L., Goudriaan, J. (2003). On approaches and applications of the Wageningen crop models. European Journal of Agronomy 18(3), 201-234. DOI: 10.1016/S1161-0301(02)00106-5

Van Wart, J., Van Bussel, L.G.J., Wolf, J., Licker, R., Grassini, P., Nelson, A., Boogaard, H., Gerber, J., Mueller, N.D., Claessens, L., Van Ittersum, M.K., Cassman, K.G. (2013). Use of agro-climatic zones to upscale simulated crop yield potential. Field Crops Research 143: 44-55. DOI: http://dx.doi.org/10.1016/j.fcr.2012.11.023.

Vargas, R., Flores, C.L. (1995). Retribución nutricional de los residuos de hojas, venas de hojas, pseudotallo y pinzote de banano (Musa AAA) en fincas de diferentes edades de cultivo. CORBANA (C.R.) 20: 33-47.

Vasenev, V.I., Stoorvogel, J.J., Vasenev, I.I. (2013). Urban soil organic carbon and its spatial heterogeneity in comparison with natural and agricultural areas in the Moscow region. CATENA 107: 96-102. DOI: 10.1016/j.catena.2013.02.009.

Verdoodt, A., Van Ranst, E. (2003). Land Evaluation for Agricultural Production in the Tropics. A Large-Scale Land Suitability Classification for Rwanda. Ghent University. Laboratory of Soil Science.

Viscarra Rossel, R. A. and McKenzie, N. J. and Adamchuk, V. I. and Sudduth, K. A. and Lobsey, C. (2011) Proximal soil sensing. An effective approach for soil measurements in space and time. Advances in Agronomy, 113. pp. 237-282. DOI: 10.1016/B978-0-12-386473-4.00010-5.

Wanders, N., Wada, Y. (2015). Human and climate impacts on the 21st century hydrological drought. Journal of Hydrology 526(Supplement C), 208-220. DOI: 10.1016/j.jhydrol.2014.10.047.

Wang, J., Rich, P.M., Price, K.P., Kettle, W.D. (2004). Relations between NDVI and tree productivity in the central Great Plains. International Journal of Remote Sensing 25: 3127-3138. DOI: 10.1080/0143116032000160499.

Weissteiner, C.J.; Ickerott, M.; Ott, H.; Probeck, M.; Ramminger, G.; Clerici, N.; Dufourmont, H.; de Sousa, A.M.R. Europe’s Green Arteries—A Continental Dataset of Riparian Zones. Remote Sens. 2016, *8*: 925. DOI: 10.3390/rs8110925.

Wielemaker, W.G., Vogel, A.W. (1993). Un Sistema de Información de Suelos y Tierras para la Zona Atlántica de Costa Rica. Atlantic Zone Program Report 22. Wageningen Agricultural University, Wageningen, The Netherlands

Wilford, J., De Caritat, P., Bui, E. (2015). Modelling the abundance of soil calcium carbonate across Australia using geochemical survey data and environmental predictors. Geoderma 259-260: 81–92. DOI: 10.1016/j.geoderma.2015.05.003.

World Reference Base (2015). World reference base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports 106. Food and Agriculture Organization of the United Nations, Rome.

Wösten, J.H.M., Bouma, J., Stoffelsen, G.H. (1985). Use of Soil Survey Data for Regional Soil Water Simulation Models. Soil Science Society of America

Journal 49: 1238-1244. DOI: 10.2136/sssaj1985.03615995004900050033x.

Wu, H., Guo, Z., Peng, C. (2003). Land use induced changes of organic carbon storage in soils of China. Global Change Biology 9: 305-315. DOI: 10.1046/j.1365-2486.2003.00590.x.

Xiong, X., Grunwald, S., Myers, D.B., Ross, C.W., Harris, W.G., Comerford, N.B. (2014). Interaction effects of climate and land use/land cover change on soil organic carbon sequestration. Science of The Total Environment 493: 974-982. DOI: 10.1016/j.scitotenv.2014.06.088.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A., Hann, S., Burt, J.E., Qi, F. (2011). Updating Conventional Soil Maps through Digital Soil Mapping. Soil Sci. Soc. Am. J. 75:1044-1053. DOI: 10.2136/sssaj2010.0002.

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., Li, D.-C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators 60: 870-878. DOI: 10.1016/j.ecolind.2015.08.036.

Yang, Z., Gao, C., Zang, S., Yang, X. (2017). Applicability evaluation of the SWIM at river basins of the black soil region in Northeast China: A case study of the upper and middle Wuyuer River basin. Journal of Geographical Sciences 27: 817-834. DOI: 10.1007/s11442-017-1408-2.

Ye, X., Zhang, Q., Liu, J., Li, L., Zuo, H. (2009). Effects of Spatial Resolution of Soil Data on Hydrological Processes Modeling. Progress in Geography 28: 575-583. DOI: 10.1002/hyp.7823.

Yerima, B., Van Ranst, E., Verdoodt, A. (2009). Use of correlation relationships to enhance understanding of pedogenic processes and use potential of vertisols and vertic inceptisols of the Bale Mountain area of Ethiopia. Tropicultura 27: 223–232. ISSN: 0771-3312. Retrieved from: http://lib.ugent.be/catalog/pug01:909199 (accessed: November 2015).

You, L., Wood, S. (2006). An entropy approach to spatial disaggregation of agricultural production. Agricultural Systems 90: 329-347. DOI: http://dx.doi.org/10.1016/j.agsy.2006.01.008.

You, L., Wood, S., Wood-Sichra, U. (2009). Generating plausible crop distribution maps for Sub-Saharan Africa using a spatially disaggregated data fusion and optimization approach. Agricultural Systems 99: 126-140. DOI: 10.1016/j.agsy.2008.11.003.

Zhao, M.-S., Rossiter, D.G., Li, D.-C., Zhao, Y.-G., Liu, F., Zhang, G.-L. (2014). Mapping soil organic matter in low-relief areas based on land surface diurnal temperature difference and a vegetation index. Ecological Indicators 39: 120-133. DOI: 10.1016/j.ecolind.2013.12.015.

Zimmermann, B., Zehe, E., Hartmann, N.K., Elsenbeer, H.C.W. (2008). Analyzing spatial data: An assessment of assumptions, new methods, and uncertainty using soil hydraulic data. Water Resour. Res. 44: n/a. DOI: 10.1029/2007WR006604.

Zinck, J.A. (1995). Soil survey: perspectives and strategies for the 21st century: an international workshop for heads of national soil survey organization November 1992, Enschede. Food and Agriculture Organization of the United Nations, Rome.

Summary

The United Nations pledged to achieve the Sustainable Development Goals by 2030. Regional land use analyses (RLUA) have an essential contribution to achieving these goals. To better meet the needs for achieving sustainable development, RLUA became more quantitative and more interdisciplinary over recent decades. This change resulted in an increased use of quantitative simulation models, which changed the type and nature of input data as well. Soil data are one of the input data RLUA require. Available soil data often do not meet the soil data requirements anymore, due to the change in RLUA. Therefore, a gap exists between the available and required soil data. This thesis aims to find possible solutions to bridge this gap. The gap differs per RLUA and therefore there is no straightforward solution on how to bridge the gap. This thesis identifies the gap and provides potential solutions on how to bridge the gap by trying to answer the following research questions: i) does it matter which available soil data are used for a regional land use analysis, ii) what complementary data are needed to meet the required soil data demand for regional land use analysis and, iii) how to obtain the required soil data for regional land use analyses in an efficient manner? The thesis hypothesises that the need for new soil data can be minimized by making 'smart' use of available soil data.

In Chapter 2, different soil datasets are compared to identify the gap and to analyse the effect of using different soil datasets as input for a regional land use analysis. As resources to collect new data are limited, RLUA often rely on available soil data. Six soil datasets were available for the study area that partly covered Machakos and Makueni counties (Kenya). The soil datasets showed large differences in reported soil properties. For example, average clay percentages varied between 11.7% and 44.4%

for the same location. The soil datasets were developed under different assumptions on, for example, soil variability. Four assumptions were verified using a field survey. An ongoing RLUA, the Global Yield Gap Atlas (GYGA) project, was taken as case study to analyse the effect of using different soil datasets. The GYGA project aims to assess yield gaps defined as the difference between potential or water-limited yields and actual yields. Rain-fed maize is the dominating cropping system in Machakos and Makueni counties. The GYGA project uses soil data for the selection of the most dominant maize growing areas and to simulate water-limited maize yields. The protocols developed by the GYGA project were applied to the six soil datasets. This resulted in the selection of six different maize-growing areas and different water-limited maize yields. Our study demonstrates the large differences between soil datasets. Main challenges with soil data in RLUA are: i) understanding the assumptions in soil datasets, ii) creating soil datasets that meet the requirements for regional land use analysis, iii) not only rely on available soil data but also collect new soil data and iv) validate soil datasets.

In general, two sources of soil data are at hand: i) available soil data and ii) newly collected soil data, i.e., new soil data. Chapter 3 analyses what complementary data are required to bridge the gap between the available and required soil data by combining available soil data and new soil data. Often a choice is made between the two data sources, while a combination of both might be more efficient. This study discusses and illustrates the possibility to combine available soil data and newly collected soil data for land use analyses. This study first looks back into the literature and describes the sources of soil data used in 120 Geoderma publications. Second, two case studies look forward by implementing a local (case 1) and a regional (case 2) study that combine available soil data and new soil data. The literature study indicated that less than 10% of the studies combined available soil data and new soil data. In regional studies, the relative use of available soil data increased despite some of the limitations mentioned in the literature. The two case studies showed that the combination of available soil data and new soil data opens new opportunities for RLUA at different scale levels. We suggest this option to be considered more often. The use of available data should be enhanced by, for example, the introduction of clearing houses for soil datasets. In Chapter 4, a mechanistic model for digital soil mapping (DSM) is developed and the potential of mechanistic digital soil mapping is explored. Soil organic matter (SOM) is an important soil property that is difficult to predict using DSM. DSM uses statistical models to predict relationship between the SOM content observed and the spatially exhaustive environmental covariates. This study analyses a mechanistic approach for DSM to predict the SOM content in nature areas. Mechanistic processes may be useful for a better prediction of the complex processes driving the SOM content. The approach makes use of dynamic soil models that include carbon and nitrogen fluxes. The mechanistic model was developed in three steps: i) select major processes that influence the SOM content, ii) study the relationship of processes that influence the SOM content, and iii) find proxies for variables that are not spatially exhaustive available. The mechanistic model resulted in a spatially continuous map of the SOM content. To improve the mechanistic map, the residuals are interpolated to estimate areas where the SOM content was systematically over- or underestimated. These two maps together resulted in a mechanistic digital soil map on the SOM content. The mechanistic digital soil map was compared to the digital soil map that resulted from a statistical model. The RMSD and the number of variables selected for the model were lower in the mechanistic model. Mechanistic models incorporate pedological knowledge and predict soil properties soil properties by values that typically stay within realistic boundaries.

The use of quantitative simulation models in RLUA increased over recent decades. As a result, complex mapping techniques are increasingly being used to better meet the data requirements. These complex mapping techniques transform available soil data or process new soil data. Chapter 5 analyses whether the required soil data can be obtained more targeted to RLUA using less complex mapping techniques. It is often thought that soil data for RLUA need to be spatially continuous. However, it depends on the RLUA at which level of detail the spatial variation needs to be described. Chapter 5 includes three case studies that require the spatial variation at different levels of detail. The studies require soil data for the same crop-growth simulation model, but the method that is used to obtain the required soil data differ.

How to obtain the required soil data depends on the objective of the study and the available soil data, but also on the efficient use of project resources, mapping tools and techniques, auxiliary data and pedological knowledge. To obtain the required soil data for RLUA it is important to first define the modelling approach, which can range from qualitative to quantitative and from empirical to mechanistic. The complexity of quantitative simulation models can differ from the complexity required by the RLUA. Therefore, the spatial variation at which the soil properties are provided need to be in line with the spatial variation at which the RLUA operate.

In the synthesis (Chapter 6), the research findings, implementation of the research findings, the hypothesis and the future perspective are discussed. Most important research findings are: i) the decision on which soil data to use for RLUA significantly influences the results, ii) complementary soil data include dominantly data on the spatial variation, the variation over depth, quantitative land qualities or complex soil properties and the quality of available soil data for a particular application, iii) soil data can be obtained more targeted to RLUA and using less complex mapping techniques.

The decision on which soil data to use for RLUA is often taken pragmatically, while the decision significantly influences the results. Soil data should be evaluated and validated before using them for RLUA. This can be stimulated by providing available soil data in Soil Data Warehouses and describing the boundary conditions and limitations of each soil dataset. The advances in soil mapping tools and techniques and the finer resolution of auxiliary data do not imply that investments in new soil data collection can be reduced. For about 70% of the global surface still no soil maps at scales larger than 1:1million are available. Soil data can be collected more efficiently making use of the advances in soil mapping tools and techniques and the finer resolution of auxiliary data. Maps that identify areas where updating available soil data should be prioritized can help the collection of new soil data to be more targeted. The required soil data can be obtained using the ARDAIG approach. This approach first defines the modelling approach that is used for the RLUA. When the modelling approach is defined, the required soil data can be formulated and available soil data can be inventorized. If there are no soil data available, the gap can be identified. Otherwise, the assumptions and quality of available soil data need to be evaluated first before identifying the gap. When the gap is identified, solutions on bridging the gap can be applied. These solutions consist in general of transforming available soil data or processing new soil data. When the gap cannot be bridged, the modelling approach needs to be reconsidered. There are opportunities for soil science to contribute more efficiently to RLUA. If the soil science community and the people involved in RLUA will implement the presented recommendations, I foresee that soil data requirements for many RLUA will be met.

Samenvatting

De Verenigde Naties streven er naar om de Duurzame Ontwikkelings Doelen (SDGs) in 2030 bereikt te hebben. Regionale landgebruiksanalyses (RLA) spelen een essentiële rol bij het bereiken van de SDGs. Om beter aan de vraag naar duurzame ontwikkeling te voldoen, werden RLA kwantitatiever en meer interdisciplinair over de afgelopen decennia. Deze verandering resulteerde in een toenemend gebruik van kwantitatieve simulatiemodellen. Dit veranderde de vraag naar gegevens. Bodemgegevens zijn gegevens nodig voor RLA. Vanwege de verandering in RLA, voldoen beschikbare bodemgegevens vaak niet meer aan de gevraagde invoergegevens. Hierdoor is er een kloof ontstaan tussen de beschikbare en gevraagde bodemgegevens. Dit proefschrift heeft als doel om mogelijke oplossingen te vinden die deze kloof overbruggen. De kloof verschilt per regionale landgebruiksanalyse en daarom is er niet één voor de hand liggende oplossing om de kloof te overbruggen. Dit proefschrift identificeert de kloof en brengt mogelijke oplossingen voor het overbruggen van de kloof door de volgende onderzoeksvragen proberen te beantwoorden: i) maakt het uit welke bodemgegevens er gebruikt worden voor een regionale landgebruiksanalyse, ii) welke aanvullende gegevens zijn er nodig om aan de gevraagde bodemgegevens van RLU te voldoen, en iii) hoe verkrijg je de gevraagde bodemgegevens voor RLA op een efficiënte manier? De hypothese van dit proefschrift luidt: de vraag aan nieuwe bodemgegevens kan geminimaliseerd worden door 'slim' gebruik te maken van beschikbare bodemgegevens.

In Hoofdstuk 2 worden verschillende bodemgegevenssets vergeleken om de kloof te identificeren en om te analyseren wat het effect is als verschillende bodemgegevenssets worden gebruikt als invoergegevens voor een regionale landgebruiksanalyse. De middelen om nieuwe gegevens te verzamelen zijn beperkt, en daarom zijn veel RLA afhankelijk van beschikbare bodemgegevens. Zes bronnen voor bodeminformatie zijn beschikbaar voor het studiegebied dat delen van de provincies Machakos en Makueni (Kenya) omvat. De bronnen toonden echter grote verschillen in bodemeigenschappen. Bijvoorbeeld, het gemiddelde kleipercentage verschilde tussen de 11.7% en 44.4% voor exact dezelfde locaties. De bodemgegevenszijn verzameld aan de hand van verschillende aannames, bijvoorbeeld aannames voor het beschrijven van de bodemvariabiliteit. Vier aannames werden geverifieerd aan de hand van een veldonderzoek. Een lopende landgebruiksanalyse, het Global Yield Gap Atlas (GYGA) project, werd gebruikt als casus om te analyseren wat het effect is als verschillende bodemkaarten worden gebruikt voor een landgebruiksanalyse. Het GYGA project heeft als doel om de opbrengstkloof van gewassen vast te stellen. De opbrengstkloof is het verschil tussen potentiële of water-gelimiteerde gewasopbrengsten en de actuele gewasopbrengsten. Door regenwater gevoede maïs is het meest dominante gewassysteem in de provincies Machakos en Makueni. Het GYGA project gebruikt bodemkaarten om de gebieden waar voornamelijk maïs verbouwd wordt te selecteren en om water-gelimiteerde maïsopbrengsten te simuleren. De protocollen van het GYGA project werden toegepast op de zes bodemkaarten. Dit resulteerde in de selectie van zes verschillende gebieden waar voornamelijk maïs verbouwd wordt en in verschillende maïsopbrengsten. De grootste uitdagingen met bodemgegevens in RLA zijn: i) het begrijpen van de aannames onderliggend aan bodemkaarten, ii) het ontwikkelen van bodemkaarten die aan de vraag van RLA voldoen, iii) niet alleen vertrouwen op bestaande bodemkaarten, maar ook nieuwe bodemkaarten verzamelen, en iv) het valideren van bodemkaarten.

In het algemeen zijn er twee mogelijke bronnen om aan bodemgegevens te komen: i) beschikbare bodemgegevens gebruiken en ii) nieuwe bodemgegevens verzamelen, i.e. nieuwe bodemgegevens. Hoofdstuk 3 analyseert welke aanvullende bodemgegevens er nodig zijn om de kloof tussen de beschikbare en gevraagde bodemgegevens te overbruggen door beschikbare en nieuwe bodemgegevens te combineren. Vaak kiezen studies één bron om aan bodemgegevens te komen, terwijl een combinatie van beiden misschien wel efficiënter is. Deze studie bediscussieerd en illustreert de mogelijkheden om beschikbare en nieuwe bodemgegevens te combineren voor landgebruiksanalyses. Deze studie kijkt eerst terug in de literatuur om te beschrijven welke bronnen 120 publicaties gebruikten om aan bodemgegevens te komen. Daarna blikken we, aan de hand van twee casussen, vooruit door een lokale (casus 1) en een regionale (casus 2) studie uit te voeren die beschikbare en nieuwe bodemgegevens combineert. Uit het literatuuronderzoek kwam naar voren dat minder dan 10% van de studies beschikbare en nieuwe bodemgegevens combineert. Het aantal regionale studies dat gebruik maakt van beschikbare bodemgegevens is gestegen, ondanks dat de literatuur sommige beperkingen van beschikbare bodemgegevens benadrukt. De twee casussen lieten zien dat de combinatie van beschikbare en nieuwe bodemgegevens nieuwe mogelijkheden biedt voor landgebruiksanalyses op verschillende schaalniveaus. Wij stellen voor om deze optie vaker te overwegen. Het gebruik van beschikbare bodemgegevens zou bijvoorbeeld verbeterd kunnen worden door het introduceren van 'clearing houses' voor bodemgegevenssets.

In Hoofdstuk 4 wordt een mechanistisch model voor digitale bodemkartering (DSM) ontwikkeld en de potentie van deze methoden van karteren wordt onderzocht. Bodemorganische stof (OS) is een belangrijke bodemeigenschap die moeilijk te karteren is aan de hand van huidige DSM technieken. De huidige DSM technieken maken gebruik van statistische modellen om een relatie tussen bodemeigenschappen en ruimtelijk expliciete omgevingsfactoren te voorspellen. Deze studie analyseert een mechanistische benadering voor DSM om het OS-gehalte in natuurgebieden te voorspellen. De achterliggende gedachte van deze benadering is dat complexe processen die ten grondslag liggen aan het OS-gehalte wellicht beter voorspeld kunnen worden aan de hand van een mechanistische benadering. De methode maakt gebruik van dynamische bodemmodellen die koolstof- en nitraatfluxen beschrijven. Het mechanistische model voor DSM is ontwikkeld in drie stappen: i) selecteer de hoofdprocessen die van invloed zijn op het OS-gehalte, ii) bestudeer de relatie van processen die van invloed zijn op het OS-gehalte, en iii) zoek ruimtelijk expliciete vervangers voor variabelen die niet ruimtelijk expliciet beschikbaar zijn. Het mechanistische model resulteerde in een ruimtelijk doorlopende kaart van het OS-gehalte. Om de kaart te verbeteren worden de residuen geïnterpoleerd. Op deze manier worden de gebieden waar het OS-gehalte systematisch over -en onderschat wordt in kaart gebracht. De twee kaarten samen resulteren in een mechanistische digitale bodemkaart van het OS-gehalte. De mechanistische digitale bodemkaart werd vergeleken met de bodemkaart die resulteert uit een statistisch model voor DSM. De gemiddelde kwadratische afwijking en het aantal geselecteerde variabelen die geselecteerd zijn voor het model waren lager in het mechanistische model. Mechanistische modellen verwerken pedologische kennis en de waarden van de bodemeigenschappen vallen binnen realistische grenzen.

Het gebruik van kwantitatieve simulatiemodellen in RLA is toegenomen over de afgelopen decennia. Als gevolg hiervan worden complexe karteertechnieken in toenemende mate gebruikt om beter aan de vraag naar bodemgegevens te voldoen. Deze complexe karteertechnieken manipuleren beschikbare bodemgegevens of verwerken nieuwe bodemgegevens. Hoofdstuk 5 analyseert of de gevraagde bodemgegevens doelgerichter voor RLA verkregen kunnen worden door minder complexe karteertechnieken te gebruiken. Er wordt vaak gedacht dat bodemgegevens voor RLA ruimterijk doorlopend moeten zijn. Heet hangt echter van de RLA af op welk detailniveau de ruimtelijke variatie beschreven dient te worden. Hoofdstuk 5 omvat drie casussen die elk op een ander detailniveau bodemgegevens nodig hebben. De casussen hebben allemaal bodemgegevens voor hetzelfde gewasgroei model nodig, maar de methoden om aan deze bodemgegevens te voldoen verschilt. Hoe de benodigde bodemgegevens wordt verkregen hangt af van het doel van de casus en de beschikbare bodemgegevens, maar ook van het efficiënt gebruik maken van beschikbare middelen binnen het project, karteer mogelijkheden en technieken, hulpgegevens en pedologische kennis. Om aan de gevraagde bodemgegevens voor RLA te voldoen, is het belangrijk om eerst de modelaanpak te beschrijven. Deze varieert van kwalitatief tot kwantitatief en van empirisch tot mechanistisch. De complexiteit van kwantitatieve simulatiemodellen kan verschillen van de complexiteit die RLA nodig hebben. Daarom is het belangrijk dat het detailniveau waarop de bodemgegevens de ruimtelijke variatie beschrijft in overeenstemming is met het detailniveau waarom de regionale landgebruiksanalyse wordt uitgevoerd.

In de synthese (Hoofdstuk 6) worden de onderzoeksresultaten, de implementaties van de onderzoeksresultaten, de hypothese en het toekomstbeeld bediscussieerd. De belangrijkste onderzoeksresultaten zijn: i) het besluit over welke bodemgegevens te gebruiken voor een RLA beïnvloedt de resultaten significant, ii) aanvullende bodemgegevens die RLA vaak nodig hebben zijn gegevens die de ruimtelijke variatie, de variatie over de diepte, kwantitatieve landkwaliteiten of complexe bodemeigenschappen en de kwaliteit van beschikbare bodemgegevens voor een bepaalde toepassing beschrijven, iii) bodemgegevens kunnen doelgerichter voor RLA verkregen worden en gebruik makend van minder complexe karteertechnieken.

Vaak wordt er pragmatische besloten welke bodemgegevens te gebruiken voor RLA, terwijl deze beslissing de resultaten significant beïnvloed. Bodemgegevens zouden geëvalueerd en gevalideerd moeten worden voordat ze gebruikt worden voor RLA. Dit kan worden gestimuleerd door bodemgegevens beschikbaar te maken in zogeheten 'Soil Data Warehouses' en door het potentiële gebruik van de bodemgegevens te beschrijven. De vooruitgang in bodemkarteermethodes -en technieken en de hoge resolutie waarop gegevens tegenwoordig beschikbaar zijn, nemen niet weg dat investeringen in het verzamelen van nieuwe gegevens gereduceerd kunnen worden. Voor ongeveer 70% van het aardoppervlak zijn nog steeds geen bodemkaarten op schaal 1:1 miljoen of hoger beschikbaar. Bodemgegevens kunnen efficiënt worden verzameld door gebruik te maken van de vooruitgang in de bodemkarteermethoden -en technieken en de hoge resolutie waarop andere gegevens beschikbaar zijn. Kaarten die gebieden identificeren waar het vernieuwen van beschikbare bodemgegevens prioriteit moet krijgen, kunnen het doelgerichter verzamelen van nieuwe bodemgegevens vergemakkelijken. De benodigde bodemgegevens kunnen verkregen worden aan de hand van de ARDAIG methode. Met deze methode moet de gebruiker eerst de modelaanpak van de RLA beschrijven. Als deze modelaanpak beschreven is, kan de nodige bodemgegevens geformuleerd worden en kan de beschikbare bodemgegevens geïnventariseerd worden. Als er geen bodemgegevens beschikbaar zijn, dan kan de kloof geïdentificeerd worden. Om aan de gevraagde bodemgegevens te voldoen, bestaan de oplossingen in het algemeen uit het manipuleren van beschikbare bodemgegevens of het verwerken van nieuwe bodemgegevens. Als de kloof niet overbrugd kan worden, dan moet de modelaanpak worden heroverwogen. Er zijn mogelijkheden om de bodemkundige expertise effectiever bij RLA te betrekken. Als de bodemkundige gemeenschap en de mensen die betrokken zijn bij RLA de genoemde aanbevelingen implementeert, dan voorzie ik dat er in veel RLA voldaan wordt aan de gevraagde bodemgegevens.

About the author

Chantal Hendriks was born at 15 July 1989 in Weert, The Netherlands. In 2007, Chantal graduated from secondary school where she obtained her VWO diploma. She started her study Soil, Water and Atmosphere at Wageningen University. Chantal did a minor in ‘Policies on, management of and modeling of land’, and her thesis was about developing a carbon cycle model that evaluate the sustainability of mixed smallholder farming systems. During her studies, Chantal was Activity Commissioner of the lifesaving association and of the student swimming associations. Chantal got her Bachelor degree in 2010. She decided to continue in Soil Science for her master degree at Wageningen University. During the master, Chantal decided to do two internships, one at Grontmij Houten (The Netherlands) and one at the International Livestock Research Institute in Addis Ababa (Ethiopia). Her thesis was about the dynamics of inland drift sand areas. During the study, Chantal had different jobs as student assistant. For the Education and Competence Studies group she gave Modular Skills Training, for the Systemwide Livestock Programme she worked on the carbon cycle model which she had developed during her bachelor thesis, for the Physical Chemistry and Colloid Science group she assisted practicals and for the AcKlas project she developed and educative program for high school students. Chantal started her PhD in March 2013. During the PhD, Chantal published papers, educated in several courses, supervised bachelor and master students, and attended different conferences. During the European Geoscience Union conference of 2017, Chantal’s PICO presentation was awarded the best Student Poster and Pico Award, which she will receive in April 2018. In 2017, Chantal was appointed as secretary and treasurer of the Dutch Soil Association. Since January 2018, she works at the University of Oxford as Landscape Modeller for the Malaria Atlas Project.

Peer-reviewed publication:

Peer-reviewed publications C.M.J. Hendriks, J.J. Stoorvogel, L. Claessens. 2016. Exploring the challenges with soil data in regional land use analysis. Agricultural systems 144: 9-21.

Publications in review:

C.M.J. Hendriks, J.J. Stoorvogel, F. Lutz, L. Claessens. In review. Combining field and legacy soil data to enrich local to regional environmental studies. Geoderma.

C.M.J. Hendriks, J.J. Stoorvogel, L. Claessens, G.B.M. Heuvelink, J. Bouma. In review. How to obtain soil data for regional land use analyses? Agronomy for Sustainable Development.

C.M.J. Hendriks, J.M.A. Martinez, L. Claessens, J.A. Rodríguez, J.J. Stoorvogel. In review. A mechanistic model for digital soil mapping to predict the soil organic matter content in nature areas. European Journal of Soil Science.

Main conference presentations

C.M.J. Hendriks, J.J. Stoorvogel, L. Claessens. 2015. Land qualities or soil properties? A novel approach for soil data to address food security. Wageningen Soil Conference. The Netherlands.

C.M.J. Hendriks, J.J Stoorvogel, L. Claessens. 2015. A novel approach to map water holding capacity and nitrogen availability. Tropilakes conference 2015. Ethiopia.

C.M.J. Hendriks, J.J. Stoorvogel, J.M. Álvarez-Martínez, L. Claessens, S.A. Aparicio, I.P. Soils, J.B. Ortiz. 2016. Mechanistic soil modelling as an approach to improve habitat monitoring. Austin International Conference on Soil Modeling. USA.

C.M.J. Hendriks, J.J. Stoorvogel, L. Claessens. 2017. How to feed environmental studies with soil information to address SDG ‘Zero hunger’. European Geosciences Union General Assembly 2017. Austria.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature

- Exploring the gaps: soil data in regional land use analysis

Writing of project proposal

- The supply and demand of soil data for regional land use analysis

Post-graduate courses

- Hands-on global soil information facilities; ISRIC (2013)
- International on-site course: land dynamics - getting to the bottom of Mount Kenya; PE&RC (2015)

Laboratory training and working visits

- Working visit to IH Cantabria (2015)

Invited review of (unpublished) journal manuscript

- Nutrient Cycling in Agro-ecosystems: effects of management on soil fertility in south Ecuadorian smallholder farming systems by modelling soil nutrient balances (2014)

Deficiency, refresh, brush-up courses

- Basic statistics; PE&RC (2013)
- Data management; PE&RC (2014)

Competence strengthening / skills courses

- Scientific writing; PE&RC (2013)
- Competence assessment; PE&RC (2013)
- Techniques for writing and presenting a scientific paper; PE&RC (2014)
- Personal leadership and self-direction while performing; PE&RC (2014)

PE&RC Annual meetings, seminars and the PE&RC weekend

- PE&RC First years weekend (2014)
- PE&RC Last years weekend (2016)

Discussion groups / local seminars / other scientific meetings

- Modelling and Statistics Network (MSN) (2013-2015)
- Two times a year a theme day of the Dutch Soil Association (Nederlandse Bodemkundige Vereniging, NBV) (2013-2017)
- Organising theme days with the topic 'soil' for the Junior Science Lab (2014-2015)
- Land dynamics (2016-2017)

International symposia, workshops and conferences

- Global yield gap atlas workshops (2014, 2016)
- EGU (2015, 2015)
- Wageningen soil conference (2015)
- Austin international conference on soil modeling (2016)
- TropiLake conference (2016)

Lecturing / Supervision of practicals / tutorials

- Environmental data collection and analysis (2013, 2014)
- Landscape geography (2013, 2015, 2016)
- Integration course (2014, 2015)
- Inventory techniques (2016)

Supervision of MSc students

- Exploring the potential of soil and water conservation as an adaptation strategy to climate change
- Modelling soil redistribution and gully head retreat; analysing the effect of climate and land management
- A mechanistic approach to soil variability at different scale levels

The research described in this thesis was implemented as part of the CGIAR Research Programme on Climate Change, Agriculture and Food Security (CCAFS), which is carried out with support from CGIAR Fund Donors and through bilateral funding agreements. The Crop Research Institute for Semi-Arid Tropics (ICRISAT) and Wageningen University & Research also financially supported this research.