A SPATIAL STATISTICAL STUDY ON UPSCALING IN THE SDI FRAMEWORK: THE CASE OF YIELD AND POVERTY IN BURKINA FASO

Muhammad Imran

PhD dissertation committee

Chair		
	To be decided	University of Twente
Promoter		
	prof. dr. ir. A. Stein	University of Twente, ITC
Assistant promoter		
	dr. R. Zurita-Milla	University of Twente, ITC
Members		
	prof. dr. M.J. Kraak	University of Twente, ITC
	prof. dr. P.J.M. Havinga	University of Twente, EWI
	prof. dr. ir. M.K. van Ittersum	Wageningen University
	prof. C. Brunsdon	University of Liverpool, UK

ITC dissertation number 234
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN: 978-90-6164-362-3
Printed by: ITC Printing Department, Enschede, The Netherlands

UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

A SPATIAL STATISTICAL STUDY ON UPSCALING IN THE SDI FRAMEWORK: THE CASE OF YIELD AND POVERTY IN BURKINA FASO

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, October 23, 2013 at 16:45

by

Muhammad Imran
born on August 01, 1974
in Khushab, Pakistan

This dissertation is approved by:

prof. dr. ir. A. Stein (promotor)
dr. R. Zurita-Milla (assistant promotor)

Abstract

Cropping conditions in West-Africa are highly spatially and temporally variable. Because of this, a variety of computational models have been developed based on understanding agricultural processes at different spatial scales. Farmers and extension workers need scientific tools that allow accessing, combining, and assessing data and models to obtain sustainable solutions at a farm location. Ideally such tools should be part of an agricultural spatial data infrastructure (SDI) so that wall to wall services are possible. In this work, we carried out four studies that support the creation of such an agricultural SDI in Burkina Faso.

The first study proposes and deploys a flexible framework system for upscaling datasets and for linking such datasets with regional simulation models. The proposed framework is based on SDI technology. The service-oriented architecture of SDI allows datasets and models to be deployed as re-usable web services. The study investigates how to use an open and interoperable SDI environment to integrate data and models for deploying location-based wall to wall services. It also studies how this environment can allow models to be adapted for variables upscaled from ground-based surveys. It provides access to datasets and models as re-usable web services by means of standard wrapper implementations. The proposed framework is deployed for on-farm decision-making in Burkina Faso. To do so, the wrapper implementation deploys a farm simulation model following the "Model-as-a-Service" paradigm and the datasets as spatial data services. Orchestrating these services enables community participation by integrating the several farming resources. The study found that the model benefits from various spatial data services in state-of-the-art SDI-based implementations. It concluded that adaptation of the variables from the country's agricultural surveys in the application of SDI services required the application of spatial statistical models and the use of remote sensing to upscale the survey data to the national scale.

The second study uses data on biophysical, socioeconomic and human resources of terroirs in Burkina Faso to estimate crop yields and to upscale the yield estimates to the national scale. The study explores the application of remote sensing (RS) data to investigate yield spatial variability. A time series of SPOT-VEGETATION (NDVI) data 1 km 10-day composites for the period covering the crop growing season was used. Field observations for crop yields were obtained from ground

surveys published in the national statistical database and sub-Saharan auxiliary datasets, originally developed using RS, were obtained from online repositories. Geographically weighted regression was applied to interpolate crop data from the field scale towards the national scale. Estimates thus obtained were stored in the geodatabase. The spatial data services deployed on top of the geodatabase can adequately initialize a farm simulation model for a terroir location. Uncertainty due to limited data availability, likely prohibits the stability of statistical models to fully capture the high spatial variability of yields in a highly heterogeneous landscape. This required to model uncertainty associated with crop yield models at regional scales. The study concludes that statistical methods and RS technology can be used for upscaling crop yield estimates for the entire country.

The third study quantifies the uncertainty in crop yield modeling at a national scale, using the crop yield observations obtained from countrywide georeferenced surveys and the spatial statistical upscaling. It presents a hybrid approach integrating ordinary kriging and geographically weighted regression. This geographically weighted regression-kriging approach was applied to crop yields in Burkina Faso. The study shows that quantifying uncertainties in large-area crop models can help to improve the sources of uncertainty given by the sampling design and the model structure. Moreover, the uncertainty maps obtained in this way can increase the confidence of end-users by taking into account the accurately estimated prediction uncertainty of crop yields.

The fourth study investigates regional and global datasets, including RS products, for modeling marginality status of terroir communities as upscaled from the targeted household surveys in Burkina Faso. It also upscales marginality estimates to the national scale. To do so, it assumes that the socioeconomic status of the terroir communities largely depends upon the agroclimatic potential of the farming systems, This can be identified from regional and global datasets. Data on biophysical factors that affect the agroclimatic potential of terroirs were obtained from SPOT-VEGETATION NDVI values and from rainfall estimates extracted from TAMSAT data. An indicator was developed that quantifies human, social and financial capital assets. A statistical analysis was performed to spatially relate the agroclimatic potential of terroirs to the asset indicator of farming communities. This relation was upscaled to estimate marginality at the national scale. Geographically weighted regression could delineate the farming systems to obtain a better understanding on the marginality status of farmers within the terroirs, thereby allowing integrative models to initialize at the national and regional scale. Such initialization requires approximating of the marginality status in order to be able to assess the capability of terroirs to apply fertilizers, pest control, and crop varieties.

To summarize, spatial statistics is applied for upscaling crop yields and communal marginality status at the terroir level to the scale of Burkina Faso. Quantification and propagation of uncertainties should from now on be an integral part of research on spatial modeling and upscaling.

To do so, the statistical methodology was adequate as it quantifies jointly and systematically the uncertainties present when sampling representative terroirs in Burkina Faso and in upscaling spatial model output. The use of an SDI framework may thus provide a robust environment for integrating datasets and spatial upscaling to farm simulations models for developing wall to wall agricultural services.

Acknowledgments

While writing these acknowledgments, many people and things flashed into my mind right from first day in this lovely country and in the faculty of Geo-information Science and Earth Observation (ITC). Today, all my words are getting smaller to express my gratitude and appreciation to those people, who stood with me through the hard times, gave me courage to overcome all kinds of troubles, and supported me to complete this task successfully. God bless you all.

The first and most one I would like to thank my promotor Prof.dr. ir. A. (Alfred) Stein and assistant promotor dr. R. (Raul) Zurita-Milla for their dedicative and enthusiastic efforts to my research. I also wish to gratefully acknowledge dr. ir. R.A. (Rolf) de By for providing valuable contributions to my research proposal and during fieldwork studies. I always enjoyed your guidance in the field of geo-information processing and earth observation to solve its complex application problems. Your gentle working style has greatly influenced my growth and research skills unambiguously. Thanks for apprising me with the scientific knowledge in geo-information processing and earth observation, inspiring me to challenge the new research topic, always quickly responding to my queries, and contributing constructive comments to every scientific output. I appreciate your encouragements during my PhD studies whenever I feel losing my temperament. I always remember your warm and comfort words when I came across hard times in my PhD studies. It was indeed a great pleasure to meet you and good fortune to work with you.

I wish to extend my gratitude to Prof. dr. Mujahid Kamran, vice-chancellor university of the Punjab, Pakistan. He is the person who motivated me for PhD study and so helped me in referring to Western institutions. His moral support and kind cooperation always encouraged me to continue my studies abroad. I always appreciate his friendly behavior, courageous, energetic and cheerful personality, and his attitudes towards personal and professional life. During my MSc study at the department of physics, he encouraged me for further study, and thereon, supported in all my problems that I faced during my student life. I would not have had a chance to study abroad without his support.

I would also like to thank all staff in the department of geo-information processing and earth observation for giving me useful feedback during research meetings. Particularly, I would like mention Ir. V. (Bas) Retsios

for helping me solve tool-oriented problems. I also wish my other PhD colleagues and country fellows who gave me full moral support during my studying period in ITC and I enjoyed every brainstorming sessions with them about problems in scientific research. They were always very patient to me in person, but critical to my research and giving me useful suggestions. I wish you much success in your life and bright future.

I would like to thank the Higher Education Commission (HEC) of Pakistan, the Netherlands Organization for International Cooperation in Higher Education (NUFFIC) and ITC, University of Twente, Netherlands for providing me this research opportunity and funding assistance for carrying out this research work. Many thanks go to the people in ITC and in Statistiques Agricoles du Burkina Faso (aka AGRISTAT) who provided valuable contributions to my research during my fieldwork. Here I would like to mention Ms. Ir. L.M. (Louise) van Leeuwen (ITC) and Moussa Kabore (Director, AGRISTAT) for their kind co-operation in my fieldwork arrangements as well as interviewing the agricultural professionals and farmer communities in Burkina Faso. I also acknowledge Guissou S. Richard, Bazongo Baguinebie Marcellin, Adama Koursangama and Nakelse Victor interviewed from the AGRISTAT, Burkina Faso for their guidance and assistance in the survey data collection.

Finally, I wish to express my deepest gratitude to my parents and my family for their moral and mental supports. I would like to express my indebted appreciation to my best friends Ch. M. (Adnan), Sh. (Abid) Mansoor, (Elsbeth) Meijer (Hengelo), Rao. (Ihsan)-ul-haq, (Liang) Zhou, dr. ir. P.R. (Pieter) van Oel (Enschede), Ch. Sanaullah (Sana), Ch. (Tanveer) for their help and everlasting friendship. I wish you all a healthy and happy life into a bright and prosperous future.

Contents

List of Figures

List of Tables

Introduction 1

[1]This chapter is based on: Imran, M. (2010) SDI - based architecture for integrated agricultural assessments and decision - making by farmer communities in sub - Saharan Africa. In: Proceedings of the GIScience 2010 doctoral colloquium, Zurich, Zwitserland, September 2010 / J.O. Wallgrün, A.-K. Lautenschütz. - Heidelberg: Akademische Verlagsgesellschaft, 2010. - 86 p. ; 24 cm. ISBN 978-3-89838-640-1. pp.45–50

1.1 Motivation and outlook

1.1.1 Spatial statistics in agricultural systems research

Spatial statistics concerns the quantitative analysis of spatial variables, including their spatial variability, their spatially varying relations, and their spatial inference and uncertainty quantification. It has been widely applied in the past to the biophysical and socioeconomic domains in agricultural systems research at different spatial scales. These applications mainly focus on analyzing and quantifying the spatial variability of several variables related to crops/farms and to their environment (Peeters *et al.*, 2012), characterization of farms and farming systems (Chomitz & Thomas, 2003; Baltenweck *et al.*, 2004; la Rosa *et al.*, 2004) and farmers adoptions of agricultural production technologies (Holloway *et al.*, 2002; Staal *et al.*, 2002).

For example, (Peeters *et al.*, 2012) used K-means clustering technique to identify significant clusters of correlations between plant-related variables and environmental variables, and then they used geographically weighted regression to determine the driving mechanisms behind the recognized clusters and to develop management zones. (Chomitz & Thomas, 2003) used spatial analysis to explain spatial variation in Amazon farming systems, which used the census tract-level data to relate forest conversion and pasture productivity to precipitation, soil quality, infrastructure and market access, proximity to past conversion and protection status. (Baltenweck *et al.*, 2004) used geographical information system (GIS) to link the household and location characteristics (i.e. biophysical and socioeconomic conditions), applied a logit model to relate the relative probability for the different farming systems at a certain location, and, thus, identified the spatial distribution of farming systems in Kenya without the need to extensively map all the farming systems across a large region. (la Rosa *et al.*, 2004) related land vulnerability (response variable) to the selected land characteristics (explanatory variables) and characterized the Mediterranean farms (i.e. land units) based on the predicted soil productivity. (Holloway *et al.*, 2002) applied Bayesian statistics to estimate spatial neighborhood effects in technology adoption models. (Staal *et al.*, 2002) used logistic regression to incorporate GIS-derived measures of external rural environments into a farmers adoption model, which is based on georeferenced data on household characteristics, and it can potentially differentiate the multiple impacts of location on choices of farming practices.

1.1.2 Spatial statistics for upscaling biophysical and socioeconomic variables to regional scale

Spatial statistics is the core methodology applied in this thesis for upscaling biophysical and socioeconomic variables from field/household level to farm and to regional scale. In many developing countries, ground-based field/farm surveys are the primary means to collect data on a range

of variables of agricultural performance such as crop yields (AGRISTAT, 2010; FAO, 2005). These surveys are often conducted for selected sites in various administrative units. Agricultural responses in large areas however are often highly spatially variable because of varying agroecological conditions like soil types, weather, and management factors. The ground surveys cannot properly reflect such spatial variability to characterize the agricultural land units (i.e. individual farms or groups of farms) in large areas (Lambin *et al.*, 1993; Prasad *et al.*, 2006; Sharma *et al.*, 2011). They are thus insufficient to present spatial heterogeneity at regional scale, which requires the collection of spatially-explicit data that represents spatial variations in biophysical and socioeconomic variables (Therond *et al.*, 2011). This issue demands ensuring methods and techniques are available to upscale estimates at the field/farm level to large heterogeneous areas and even to whole countries.

Variables of agricultural performance exhibit a high degree of spatial variation across land units, because of their underlying factors often related to several natural resources (e.g. soil, climate, topography), socioeconomic condition (e.g. ability to apply modern inputs), and resource base (e.g. water reservoirs, available labor) (Shaner *et al.*, 1982; Dixon & Gulliver, 2001). Spatial statistics can be applied to upscale the variable estimates to large areas through modeling their high spatial variability. Models of spatial structure can be applied to weigh sampled values of the variable estimates based on their spatial neighborhood. Moreover, the variable estimates can be statistically related to their collocated factors, causing significant spatial variability, mainly belonging to the biophysical and socioeconomic contexts (Faivre *et al.*, 2004; Challinor *et al.*, 2009). The spatial weights and/or relationships obtained in this way can be used in the quantitative upscaling of spatial variables. The spatially-explicit results from upscaling can be obtained on densely gridded maps, which can then be used as the basis for developing a range of location-based applications at regional scale.

Uncertainty is inherent in the quantitative analysis of spatial variables. Since, spatial data have different representations with different levels of inherited inaccuracies. Moreover, linking data of different spatial and temporal scales/resolutions to upscaling procedures may introduce uncertainty. Uncertainties may also be associated with measurements, sampling and experimental design, flaws in the upscaling procedures or statistical analysis itself. Spatial statistics can be applied to quantify uncertainty in the process of spatial upscaling.

1.1.3 Spatial data infrastructures (SDIs) for integrating data and models

Despite the availability of crop data and of explanatory variables, it is not straightforward to link this data and to make it accessible to agricultural simulation models. The problem of integrating data is threefold (Groot & McLaughlin, 2000; Beare *et al.*, 2010; Foerster *et al.*, 2010):

1. **Technical**: are the syntactical problems related to differences in the formats in which models expect input data to be provided and the formats in which the chosen candidate datasets are originally found. Spatial data are characterized by spatial scales and levels of detail, implicit errors and uncertainties, missing data, and other fitness-for-use information that are often not communicated.
2. **Conceptual**: are the structural and semantic problems related to differences in the conceptual schemas (e.g. concepts, terminologies, and meaning) of datasets and the concepts conceived in models. Conceptual barriers are often posed by different interpretations of data, which need to adhere to common standards and semantics.
3. **Institutional**: are the barriers in data sharing posed by organizations, researchers, and surveyors through legal regulations of privacy, ownership and copyrights.

Geographic information systems are often deployed to store, describe and to analyze the domain-specific geospatial datasets. A modeling framework system however is a set of component sub-systems that can be assembled into a model application under a common architecture which is regarded as core of the framework (Rizoli *et al.*, 2008). Interoperability is the capability of cross-domain GI databases and modeling frameworks to interact through overcoming the technical, conceptual and institutional barriers (Janssen *et al.*, 2009; Reichardt, 2010; Granell *et al.*, 2010). To achieve geospatial data interoperability, the creation of spatial data infrastructure (SDI) methodology is being increasingly initiated recently (Kiehle, 2006; INSPIRE, 2008). SDI denotes the relevant base collection of technologies and standards to provide an ideal environment to couple datasets and applications (Nebert, 2004). This environment encompasses not only the datasets, but the metadata (i.e. documentation for fitness-for-use), and the technology to discover, visualize and to integrate datasets in application (i.e. catalogues and Web mapping).

The SDI technology has the potential to design interoperable frameworks in which different working groups can effectively participate in sharing their unique knowledge/resources for solving a given problem. Currently, much research in GIS and environmental modeling is focused on the use of enhanced interoperability offered by SDIs (Maué *et al.*, 2010). Following this, research communities have started to expose datasets and models as SDI services (Geller & Melton, 2008), for example, for sharing agro-geospatial data in the CropScape application (Han *et al.*, 2012), and for the hydrological models re-use (Granell *et al.*, 2010; Castronova *et al.*, 2013). Little research has been done so far in developing SDI-based frameworks to link data and models in agricultural systems research.

1.1.4 Upscaling and integration in an SDI-based framework

A particular focus of this study is to investigate SDI technology to propose a flexible framework for coupling spatial statistical upscaling to agricultural simulations models. The proposed framework essentially

allows to model and to upscale variables of agricultural systems, and to apply the upscaled outcomes to farm simulation models at regional scale. Farm simulation models, e.g., bio-economic farm models (Janssen *et al.*, 2009; Louhichi *et al.*, 2010) are usually deployed for integratively accessing information from several agricultural domains, and, thus, can be used for on-farm decision-making. Attention therefore focuses on: (i) upscaling biophysical and socioeconomic variables from field/household level to farming community and even to national scale, and (ii) adapting the simulation models to be coupled with the upscaling so that the wall to wall services for integrated agricultural assessment are possible.

Integrated assessment is an interdisciplinary process of combining and interpreting knowledge from diverse scientific disciplines in order to provide useful information to decision-makers (Rothman & Robinson, 1997). Technically, it is conducted in a framework system that integrates several datasets and models representing different spatial processes at various spatio-temporal scales and in different domains (Parker *et al.*, 2002; van Ittersum *et al.*, 2008). To do so, the spatial and temporal scales of data and the modeling scale are two factors in the context of problems associated with data quality and interoperability (Jakeman & Letcher, 2003; Janssen *et al.*, 2009). In this research, these two problems are emphasized in upscaling datasets, integrating the upscaling results to models, and also in quantifying and communicating the uncertainty associated with upscaling procedures. We will look into these issues more in-depth in the case of crop and marginality upscaling in next Section; here these are generally described in the context of the framework design in this thesis:

- The common case for applying models is a site-specific application on which all input datasets have already been prepared for a particular farm site. It contrasts to a spatially-explicit and wall to wall application of model, for which no specific site in a large area has yet been identified, and consequently no specific, targeted data sets are recognized. The later application demands upscaling variables to regional scale, discussed in Section 1.1.2. The spatially-explicit outcomes can then be applied for 'spatialising' a model over large area (Faivre *et al.*, 2004), for instance, a farm simulation model for its location-based initialization in country. In this context, this research contributes in developing spatial statistical models to upscale the biophysical and socioeconomic variables to regional scale and to quantify the associated uncertainty. The research output can be employed in deploying wall to wall agricultural services.
- It is challenging to integrate datasets and upscaling and agricultural simulation models in a wall to wall setting. Because, the models as used in this setting will have to be somewhat different from the original site-specific models in monolithic modeling frameworks. It therefore requires deploying an open framework system which is more explicit in handling data quality and interoperability in order to adapt models. Ideally such framework should be part of an SDI

so that wall to wall services are possible. In this context, this study investigates the SDI technology and how it can be used to design interoperable framework for integrating datasets and models.

1.2 Challenges for upscaling and integration in the SDI framework: the case of yield and poverty in Burkina Faso

Agriculture in Sahelian and West-African countries can in many cases be characterized as marginal, with subsistence farming being an important activity. In Burkina Faso, for instance, two-thirds of the population works in agriculture (USAID, 2009). Those mainly poor farmers make their livelihood in local communal systems, called *terroirs* (AGRISTAT, 2010). Each terroir is traditionally led by a chief. Individual households in a terroir contribute their small-holdings for cultivation and adopt common interventions. Cropping conditions in terroirs are highly spatially and temporally variable (see for example Figure 1.1). In particular, the variability of vegetation, soils, and topography has a serious impact on crop yields (Graef & Haigis, 2001). Moreover, marginality and poverty status of the terroir communities has an impact on their capability of applying modern inputs like fertilizers, pest control, and crop varieties. Therefore, an important reason for upscaling crop yields and marginality status is that farmers respond to indicators both from the biophysical environment and from their socioeconomic contexts. Data on various driving factors of crop yields and marginality, however, are usually not available in Burkina Faso (Roncoli *et al.*, 2001; Roncoli *et al.*, 2009).

Upscaling biophysical, social and economic variables can provide spatial-explicit data at the scale of Burkina Faso. Farmers and extension workers however search information channels that allow combining and assessing multidisciplinary data to develop location-specific sustainable strategies for farm interventions. To do so, they may benefit from integratively assessing up-to-date information on several driving forces of their terroir production (Lambin, 2003). More precisely, the farm simulation models can be used to simulate farms for several farm resources to their optimal allocation (Janssen & van Ittersum, 2007). These models typically combine data from biophysical and socioeconomic domains of land units for collectively assessing those variables, and, thus, can be important tools for on-farm decision-making in the country. However, a big challenge is to adopt such models to be coupled with the upscaling procedures so that wall to wall services are possible.

This research applies spatial statistics as a tool to upscale crop yields and marginality estimates to the country scale. This upscaling is based on modeling the spatial variability of their various driving factors in Burkina Faso terroirs. It further designs an application framework based on SDI in which datasets and upscaling and farm simulation models can be integrated to deploy location-based wall to wall services. In this

Figure 1.1 A heterogeneous cropping system related to a terroir in Burkina Faso.

spatially-explicit application, challenges are related to apply models at any terroir location in the country, thus obtaining results that are qualitatively comparable to those of site-specific model applications. It may help to solve issues in the following fields:

Limited data availability In a site-specific model application, more detailed datasets are usually obtained by intensively applied expensive technology, allowing highly standardized and dense data acquisition techniques, often leading to possibilities of 'precision agriculture' (Lee *et al.*, 2010). Such acquisition technology, however, is often not available in West-African subsistence farming. Data sources are scarce or even non-existent. A different strategy is therefore needed to meet the requirements of model applications for large and heterogeneous areas. In most cases, this provision demands securing third-party data sources that can reliably serve the data needs of models, even though these sources have not been specifically designed for that purpose. *In situ* data therefore need to be replaced with data obtained through remote sensing or upscaled from national statistical datasets (Faivre *et al.*, 2004).

Remote sensing is widely used in modeling various quantities from finer to regional scales, e.g., crop yields, mapping diseases, estimating biomass and moisture stress in plants (van Ittersum *et al.*, 2004; Dutta *et al.*, 2011). Utilizing high spatial and temporal RS

coverage is attractive in a regional scale modeling, as applications can use up-to-date information over large areas. RS primarily delivers images of land cover, which may be stored into a GIS to derive patterns of land cover based on RS reflectance characteristics. To some degree inferences about variables of agricultural systems can be made from patterns of land cover obtained from RS, but fully capturing spatial distributions of variables requires ground information. In West-Africa, national statistical organizations conduct geo-referenced surveys at selected sites to collect observations on a range of variables on farming systems. These observations may be upscaled to national and regional scales; however, it may not capture the landscape heterogeneity particularly in West-Africa. Alternatively, with an understanding of the reflectance characteristics and some ground observations it is possible to use remotely sensed data to obtain estimates of various model inputs, which are statistically upscaled for large areas using relatively small sets of field observations. There are two common ways in which this may be done in a farming system research: vegetative indices and land cover clustering and classification techniques (Dorigo *et al.*, 2007). The former approach has been mainly applied in this thesis.

The upscaled quantities may provide location-specific estimates of various existing potentials of the farming and cropping systems over large areas. In this thesis, yield and marginality estimates upscaled by spatial statistical models to the national scale of Burkina Faso will allow to initialize farm simulation models everywhere in the country (i.e. in a wall to wall setting). Such an upscaling in West Africa is challenging as it requires capturing biophysical and socioeconomic quantities with high spatial variability, caused by heterogeneous farming conditions.

Spatial data quality This thesis considers different spatial scales. The highest relevant scale is the national scale of Burkina Faso that encompasses 351 districts and almost 7000 terroirs (see Figure 1.2). A terroir is comprised of a few dozen to many hundreds of households. National statistical departments collect data for households in a single representative terroir in a district, thus producing 351 data points for observations (AGRISTAT, 2010). Upscaling as applied in this research, on one hand, quantifies variables from targeted household surveys to the lowest administrative level (i.e. terroir). On the other hand, it uses regional and global datasets including RS products to upscale those variables at the terroir level to the national-scale of Burkina Faso, and in doing so, it tends to recognize the heterogeneity that inevitably exists within such terroirs on the national scale. There are two issues of data quality here: (i) a choice for a combination of data has to be motivated by the questions to be addressed (i.e. objective of the study, level of analysis, and data availability); (ii) incorporating spatial dependency into the process of upscaling both in terms of conventional inference on

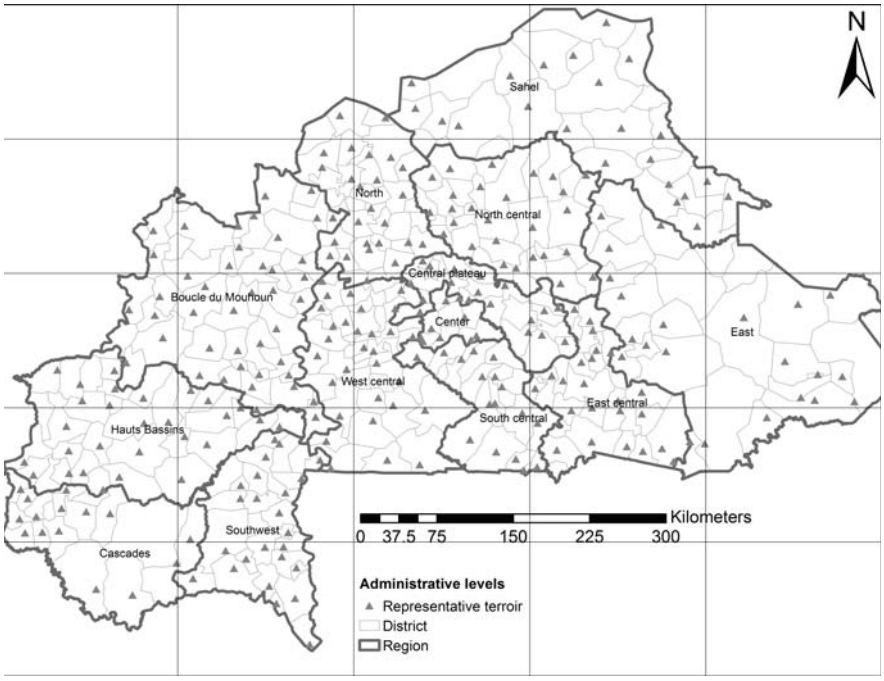


Figure 1.2 Different administrative levels in Burkina Faso.

variable coefficients and goodness of fit (Lambin, 2003). These two issues are highly challenging to tackle particularly for upscaling in heterogeneous West-African conditions.

Spatial data have the tendency to be spatially dependent (aka spatial autocorrelation). Ignoring spatial dependencies in data may lead to biased inference of variables, estimation of error variance, and testing of statistical significance (Cressie & Wikle, 2011). To deal with it statistically, the spatial statistical methods are mainly used in this thesis, such as spatial auto-regression (Anselin, 1995), geographically weighted regression (Fotheringham *et al.*, 2002), and geostatistical methods such as regression kriging (Diggle, 2003). They use different methods to incorporate spatial dependency in data. Uncertainties in the original datasets and models are required to be addressed. It is however too ambitious to materialize it in all its aspects, i.e., developing methods to quantify and characterize uncertainty associated with various datasets and model types and propagating the characterized uncertainty into agricultural assessments. This research focuses on modeling uncertainty in the spatial statistical models to upscale crop yield estimates to the scale of Burkina Faso. Ideally, the quantified uncertainty should be communicated when these estimates are linked to initialize simulation models.

Model adaptation and end-user communication A particular objective of this study is to propose a flexible framework to link upscaling models to farm simulation models at regional scale. The proposed framework should achieve an adequate level of technical and conceptual interoperability so that the regional models can be adapted in wall to wall services.

Overcoming technical barriers means that the proposed framework essentially uses same structure of datasets and input format of models. Two approaches are common in achieving so: (i) tool coupling, in which models are linked together in a framework with a common graphical user interface and data storage (Janssen *et al.*, 2009), and (ii) loose coupling, in which the model interaction is established during run-time and the data and models do not know each other in advance (Granell *et al.*, 2010). Recently, the SEAMLESS integrated framework (van Ittersum *et al.*, 2008) opted for the former approach for integrating legacy models (Janssen *et al.*, 2009). This approach however may cause dependencies on framework-specific libraries that may be difficult to resolve when using the models elsewhere. This kind of dependency can be overcome by loosely coupling data and models, i.e., they are exposed with XML-based standard interfaces and they can be linked run-time (Jakeman & Letcher, 2003). This research opts for the loose coupling approach.

The SDI technology can be seen as a realization of the service-oriented architecture (SOA) to disseminate data and services (Kiehle, 2006). In SOA, the geospatial datasets and processes (e.g. GIS algorithms and procedures, computational models) can be deployed as loosely-coupled and distributed web services, following the publish-find-bind paradigm. To overcome the technical barriers, web services implement standard interfaces that expose their quality through describing metadata and adopt common data formats for message exchange (Di, 2005). This can allow communities the introspection of interoperable data and models as web services and participate in communal problem solving. The SDI implementations generally adopt the open geospatial consortium (OGC) standard interfaces (OGC, 2008c). This research implements the OGC standards to deploy data and model web services.

Overcoming the conceptual differences means to achieve a shared understanding between datasets and models (Janssen *et al.*, 2009). More specifically, it requires to explicitly describing concepts in the datasets and in the parametric space of models. Concepts are formally described in a conceptual schema, by using conceptual formalism of a formal language, e.g., unified modeling language (UML), ontology web language (OWL) (ISO/IEC, 1996). OWL being in-line with the semantic web is powerful to declare the semantics explicitly. Whereas, UML has been widely used due to its strong expressiveness, web compatibility, technology independence, intu-

itiveness, and tool support (Francois *et al.*, 2009). This research uses the later approach to describe concepts in agricultural surveys and in the parametric space of models. However, for a wider use, the concepts are harmonized with the SEAMLESS ontology (Janssen *et al.*, 2009), which is developed through the collaboration of stakeholders in the agricultural domain.

The SDI-based framework in this thesis makes two types of adaptations to models: model-related and data-related. Model-related adaptation refers to the provision of a farm simulation model to farmers or extension workers as a web service for on-terroir decision-making. To run the model service on a terroir location, data-related adaptation refers to the robustness of the framework system against the choice of data services that provide input data, which are either upscaled from spatial statistical models in this thesis or provided by a third-party dataset. The farmers or extension workers may determine which data services are going to be used, and the system should allow binding them into the farm simulation models.

1.3 Research objectives

This study considers a terroir in Burkina Faso as the unit for analysis and uses three types of datasets: statistical surveys, GIS layers, and remote sensing products. The country's agricultural surveys were collected from representative terroirs countrywide for the year 2009. In the survey data, household records are maintained for the related farming and cropping systems details. The surveys contain data for total 4850 households that cover 351 representative terroirs of the country. Data were processed to obtain values of several variables of the terroir-level farming and cropping systems. GIS layers such as soil properties and market access were obtained from various regional and international organizations. Remote sensing products with the spatial coverage of the whole country were mainly obtained from SPOT. The main objective of this research is to use these datasets to upscale crop yields and marginality status from field/household level to terroir and to the national-scale of Burkina Faso and to design an interoperable framework system to apply the upscaled outcomes to farm simulation models deployed as wall to wall services. The identified means to achieve this objective use spatial statistics for upscaling variables and SDI technology for designing the framework. Based on this main objective and the available datasets, we formulated the following sub-objectives to achieve in four studies:

- To investigate SDI technology to propose a flexible framework to link spatial upscaling to simulation models at regional scale for deploying wall to wall services.
- To model the relationship between the observed crop yields and their collocated explanatory variables at the terroir level and to

upscale the yield estimates to the national-scale of Burkina Faso.

- To model uncertainty in the regional modeling and upscaling of crop yields in Burkina Faso.
- To model the welfare and marginality status at the terroir level using targeted household surveys, and to investigate regional and global datasets including RS products for upscaling the terroir-level marginality estimates to the national-scale of Burkina Faso.

To carry out this research, we set out a framework and accomplished the research objectives in providing its various components, explained in the next Section.

1.4 Research framework

Figure 1.3 shows schematic diagram of various components within this research framework:

Knowledge base This component was designed:

1. to store all spatial and thematic datasets which originate at inputs and outputs of various models in the research framework.
2. to store metadata about all datasets and models to enforce their consistency and integrity and to establish a meaningful exchange of datasets among the framework components.
3. to provide data transformations, e.g., aggregation, manipulation and formatting in the thematic and spatial attribute spaces.
4. to be served as a harmonized database for linking datasets and models.

National statistical datasets obtained from AGRISTAT were made spatially-aware. To do so, a spatial schema was designed to realize a GIS-based database in the PostGIS/PostgreSQL. Using the spatial schema, the datasets were extracted into the database. This database was served to provide field observations to a variety of models in the framework. The experimental data were transformed according to the input requirements of the models. The outputs from the statistical models were stored in the databases that were accessed by various agricultural web services to initialize a farm simulation model for a given terroir location.

Agricultural models – This component contains computational models mainly belonging to the following two categories:

- Spatial statistical models were developed to upscale biophysical and socioeconomic inputs from ground-based surveys to the national scale of Burkina Faso. Upscaled estimates from statistical models were used to initialize the farm simulation model for optimization at a terroir location, specifically for the

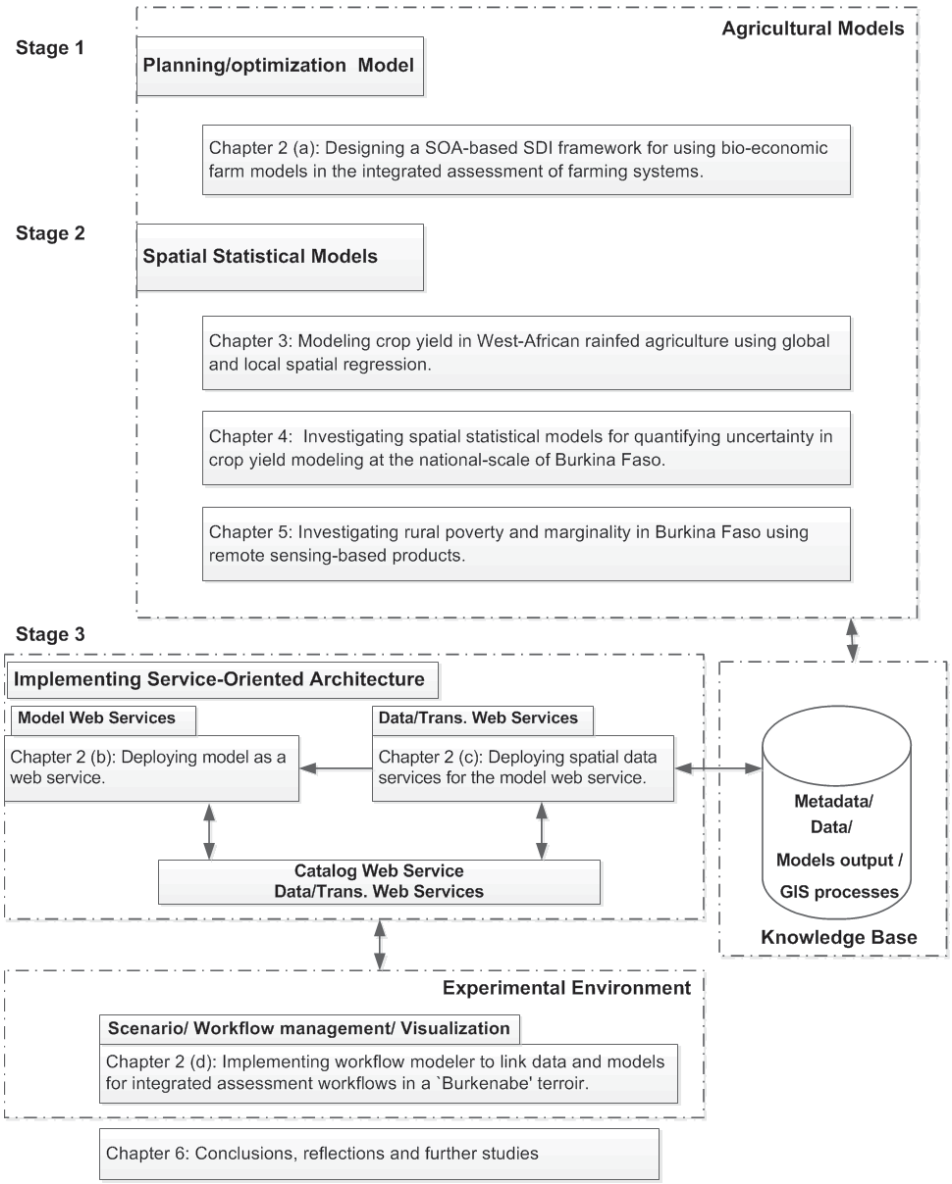


Figure 1.3 Research framework to upscaling in the Spatial Data Infrastructure (SDI) framework: the case of yield and poverty in Burkina Faso.

biophysical potential, i.e., crop yields (see objective 2) and the socioeconomic constraints, i.e., marginality status of terroir communities (see objectives 4).

- A bio-economic farm model (Janssen *et al.*, 2009; Louhichi *et al.*, 2010) was used as a farm simulation model for optimization/planning. The basic purpose of using this model is to illustrate: (i) its adaptation in developing wall to wall agricultural services for on-terroir decision-making , (ii) its adaptation as a geospatial standard web service, and (iii) its adaptation for regional modeling and upscaling in an SDI-based framework. The model was tested in a case study to adapt various input upscaled from the spatial statistical models. The model we deployed as an geospatial web service operates at the terroir level and it is comparatively static, i.e., it has no interdependence of outcomes across years, and model results represent the equilibrium situation for a single cropping system in a year.

Experimentation Environment – This component was designed mainly to construct a flexible interface for: (i) runtime linking the geospatial web services for data and models into workflows, (ii) executing the workflows, and (ii) visualizing and communicating results to end-users, i.e., farmer communities. This component delivers results by easy-to-understand means via reports and visualization tools.

These framework components were implemented in the service-oriented architecture of SDI. This implementation provides a web-based tool for decision-making allowing Burkinabé farmers and extension workers to obtain sustainable farming solutions on their terroir locations.

1.5 Thesis outlines

The thesis is carried out in three stages. In the first stage, the overall SDI technology is investigated to design a flexible and interoperable framework system for data and model integration. Based on this design in Chapter 2, an application is deployed that implements the proposed framework design to devise farmers the optimal plans for on-farm decision-making in Burkina Faso. It deploys data and models as standard web services which take benefit from current Web mapping technologies. Based on this framework deployment in Chapter 2, various issues are identified related to input data availability and spatial data quality in the wall to wall model application. To overcome these issues, the second stage deals with the spatial statistical methodology to upscale the biophysical and socioeconomic variables using ground-based surveys and other regional and global datasets including RS products. Chapters 3, 4, and 5 of this thesis report work done during this stage which comprises the major part of the thesis. The final Chapter provides conclusions, reflections and recommendations for further studies.

An SDI-based framework for the integrated assessment of agricultural information

2

[1]This chapter is based on: Imran, M., Zurita-Milla, R. and de By, R.A. Integrated environmental modeling: an SDI - based framework for integrated assessment of agricultural information. Presented at AGILE 2011: the 14th AGILE International Conference on Geographic Information Science, 18-21 April 2011, Utrecht, Netherlands. 9 p.

2. An SDI-based framework for the integrated assessment of agricultural information

Abstract

Monolithic framework systems pose obstacles to apply agricultural models at regional scales, and, thus, to develop location based wall to wall services. In particular in an integrated assessment, this requires linking a range of datasets and models. This Chapter proposes and deploys a flexible framework system for linking quantitative models for spatial upscaling with farming system simulation models at regional scale. The proposed framework is based on spatial data infrastructure (SDI) technology. The service-oriented architecture of SDI allows datasets and models to be deployed as re-usable web services. This study investigates how to use an open and interoperable SDI environment to integrate data and models for deploying location-based wall to wall services, and how this environment can allow models to be adapted for variables upscaled from ground-based surveys. Here, we provide access to datasets and models as re-usable web services through standard wrapper implementations. These services are loosely-coupled and the framework is robust against coupling the appropriate data services for location-based initializations of model. The proposed framework is deployed for on-farm decision-making in Burkina Faso. To do so, the wrapper implementations in the framework deploys a farm simulation model following the "Model-as-a-Service" paradigm and the datasets as spatial data services. Orchestrating these services allows enabling community participation in a common problem through assessing integratively the several farming resources. Testing the services for the study area the study found that the model benefits from various spatial data services in state-of-the-art SDI-based implementations. Moreover, it found that, to adapt variables from the country's agricultural surveys in the application of SDI services in Burkina Faso required applying spatial statistical models and use of remote sensing to upscale the survey data to the national scale. In this context, the next three studies were carried out to upscale the biophysical and socioeconomic variables measured at terroir locations in the country.

2.1 Motivation and outlook

Agriculture in Sahelian countries can in many cases be characterized as marginal, with subsistence farming being an important activity. Farmers often find themselves deprived of important inputs, whether they are good soils, seeds or fertilizers, or availability of water, workforce, or good-practice information (Roncoli *et al.*, 2001; Roncoli *et al.*, 2009).

Farm simulation models, i.e., models that simulate a farming system can be used to assess what-if scenarios of farm's production over a season. They can be important tools for on-farm decision-making (Janssen & van Ittersum, 2007). The Model-as-a-Service (MaaS) paradigm aims to bring the results of often complex and data-intensive computations towards a large user community (Reichardt, 2010; Granell *et al.*, 2010). MaaS may tackle the pre-processing of large amounts of data, the curation of datasets for future use, or simulation and forecasting. The use of the model service is typically offered through a robust computational mechanism that will handle the input data appropriately, i.e., through identifying data fitness-for-use and transforming data (discussed in Section 2.2.1) in the case when fitness-for-use is only partial. This is especially true for simulations and forecasts. The results are normally received on client applications like a web browser (Zhanng & Tsou, 2009).

The application of farm models in sub-Saharan Africa is a highly challenging domain because these models require large amounts of data as all farming resources (soil, crops, farming activities, etc) need to be properly characterized and because these models are sensitive to the quality of the input data. In the more developed economies, data availability and quality are achieved by intensively applied and expensive technology, allowing highly standardized and dense data acquisition techniques, often in situ, and leading even to possibilities of precision agriculture (Lee *et al.*, 2010). However, such acquisition technology is often not available in sub-Saharan settings. This is even stronger the case for location-specific data. Furthermore, running a model usually requires training that is generally not available or not feasible.

In this study, we address these challenges by diminishing the data exchange interaction between model and end-user. This would relieve the latter from issues of data generation, standardization and quality control. To do so, we attempt to replace the user-generated inputs to the model as much as possible by system-generated inputs obtained from reliable third-party sources. This may lead to a light-weight service consumption scenario, and more standardized inputs. It also may lead to technical challenges. More precisely, our focus is on bio-economic farm models (BEFMs), which allow one to simulate farm responses at a location (Janssen *et al.*, 2009). These models require a range of inputs covering the farm's biophysical, social and economic environments of associated farming system. The common case of using a BEFM is as a desktop application on which all input datasets have been prepared for a particular study site, and are under control of a highly skilled

end-user. We call this a site-specific application. It contrasts to a wall to wall application of the model, for which no specific site has yet been identified, and no specific targeted input datasets are available.

The challenge is to make wall to wall applications work everywhere, obtaining results qualitatively comparable to those of site-specific applications, or reaching at least a level of quality that is of use to its stakeholders, e.g. the farmers or farm extension workers. We recognize five important problem domains to be addressed in reaching these goals:

1. **availability**: securing third-party datasets that can reliably serve the data needs of the model, even though these sources have not been specifically designed for that purpose;
2. **scaling**: ensuring that methods and techniques are available to detect and resolve differences in spatial and temporal resolution between the available data and the inputs needed by the model;
3. **model adaptation**: the model as used in a wall to wall application has to be somewhat different from the original site-specific application model, a.o. in a more explicit handling of uncertainty;
4. **uncertainty**: creating the system/model capability of computing uncertainty into the model to allow it inclusion of either uncertainties innate in the original data, or caused by scaling processes, or innate in the model used;
5. **end-user communication**: guaranteeing that the results of the model runs are communicated with the stakeholders in optimal ways, reducing risks of misinterpretation, and subsequent issues of trust, as much as possible.

In this chapter, we address especially the matters of availability, scaling, and model adaptation, as approaches to them are highly interdependent. They are also considered to be the key factors to an overall success. The issues of uncertainty and end-user communication are only briefly addressed here, and will be followed-up in Chapter 4.

In this study, the wall to wall application needs to make adaptations to the underlying farm model. Those adaptations come in two kinds: spatial parameterization and spatial scaling. By spatial parameterization, we mean that the system that encapsulates the model proper is sensitive to, but also robust against choice of a specific farm location. This choice may determine which third-party datasets are going to be used, and the system should effectively accommodate this. External datasets may differ substantially from what the model naturally requires, but yet be the best available source. Therefore, fitness-for-use tests need to be available, and these should address scale, resolution and thematic content. By spatial scaling, we mean the system's internal functionality required for the spatial parameterization is flexible for taking alternative input data. Upon deciding to use such input data, in a data curation phase, data should be properly mapped, i.e., spatially scaled using spatial statistical (quantitative) models and thematically generalized using some GIS functionality. For this the computational framework needs to be defined during an input requirements analysis phase.

The main objective in this chapter is to propose a flexible framework to link spatial scaling to farm simulation models at regional scale for deploying wall to wall services. This will adapt a BEFM following the MaaS paradigm. To achieve this, the proposed framework will provide the model as a web service. This MaaS can be parameterized spatially by means of combining web services that provide the required data. The proposed MaaS will be used by farmers and extension workers to find optimal and feasible farming solutions at their farm locations.

In Section 2.2, we discuss this MaaS approach to three important challenges, and relate it to work reported in the literature. The architecture of our proposed framework is discussed in Section 2.3. Section 2.4 reports implementation of our proposed computational framework for analyzing farming systems in Burkina Faso. It further illustrates the implementation by the output of BEFM simulation for Burkinabé terroirs. Finally, Section 2.5 highlights conclusion and future work.

2.2 Challenges for providing model as a wall to wall service

A commonly applied methodology to explore a farm is to collectively assess a range of farm-related information in simulating the farm responses. To do so, (de Wit *et al.*, 1988) showed the potential of multiple-goal linear programming to develop a bio-economic farm model (BEFM). Based on this, Figure 2.1 illustrates the schematic diagram of the multiple-goal modeling adapted for this research. Such modeling demands a thematically wide range of data inputs. Site-specific applications can deal with data appropriate for the single location for which a model is used, but may not suited for wall to wall applications aiming at a large geographic area. The data sources used in site-specific cases then need to become functions of location.

The range of data themes typically needed by any BEFM covers the following three (Janssen *et al.*, 2010; Louhichi *et al.*, 2010):

1. **production technology**: aims to quantify the human influence on the farming system potential. The most common factors are related to farm inputs, e.g., irrigation, fertilization, crop protection such as pest/disease/weed, crop types and varieties, sowing date and density, household labour make-up and soil tillage.
2. **biophysics**: covers the data quantifying the agricultural response on a farm as a physical process of (plant) growth. The most common factors are either biotic or abiotic, and include terrain/soil data, weather/climate data, as well as land use characteristics such as extent, agricultural intensity and zonation.
3. **economic**: quantifies the functioning of economic agents, e.g. farms or market parties. Common factors include prices of farm inputs and outputs, e.g. crops and fringe products, but also purchase capacity of farmers and their participation levels in public markets.

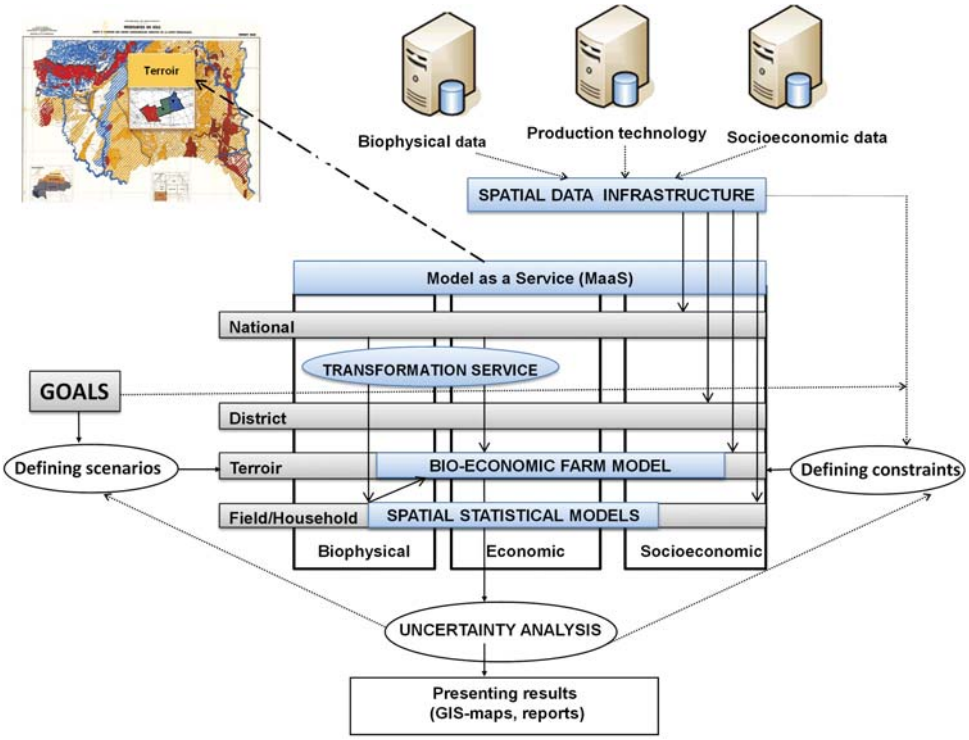


Figure 2.1 Various components of the multiple-goal modeling in this research; spatial data infrastructures (SDIs) provide a range of datasets of different scales in different domains; a bio-economic farm model (BEFM) is provided following the Model-as-a-Service (MaaS) paradigm; a transformation service transforms data for fitness-for-use for the model service; a spatial statistical (quantitative) model accomplishes scale-related data transformations; using these mapping outcomes at a farm location, several environmental and (socio-) economic constraints on the farm (or group of farms such as terroir in Burkina Faso) may be identified to evaluate farming activities for the goals of farmers.

Below, we look at issues to provide the fitness-for-use of datasets for BEFM inputs.

2.2.1 Availability

Location-based initialization of a range of model inputs concerns data availability related to farming and cropping conditions at a location. This is typically challenging for spatial parameterization of BEFM in a wall to wall setting. The field of Spatial Data Infrastructure (SDI) aims to share spatial datasets of different scales across the social, economic and environmental domains of sustainable agricultural development (Kiehle, 2006; Granell *et al.*, 2010). SDI deploys standard metadata procedures to provide information on data fitness-for-use and accuracies. Today, several national and global SDIs exist that provide spatial datasets for applications at different levels of scale (INSPIRE, 2008). For initializing a BEFM the direct question is whether any dataset can be identified to serve the model as input for one of the required themes. This means that an agent or service needs to understand thematic information content of the model input(s) and of the candidate datasets. It can then determine the quality of data use in the model and perform required content transformation when fitness-for-use is only partial. Thus, the role of a versatile GI transformation service becomes vital to detect and resolve the quality issues. Often this service is needed to perform scale-related and schema-related transformations, as explained below,

2.2.1.1 Availability and scale-related transformations

A BEFM simulates a farm, either a real farm or an average (representative) farm for a group of farms (Janssen & van Ittersum, 2007), which is referred to as the model simulation/spatial analysis unit. In this context, at the low-quality end, the data may be coarse-grain, at the national or even continental scale. The data source characteristics need to be known, as they serve to compare one data source candidate against the next. At the high-quality end, the data are fine-grain, at the district, village or even at the parcel/household level. Datasets may also be inferred from large and dense sample sets, often acquired at the parcel/field level. This often involves using spatial statistics for quantitative upscaling of field observations.

2.2.1.2 Availability and schema-related transformations

Here the question on availability is selecting the most appropriate dataset to serve as model input. This requires appropriate techniques to evaluate metadata, and compare with those of alternative data sources. Such an evaluation and comparison are typically multivariate optimization problems on metadata attributes. Moreover, the required optimization functions are poorly understood, as they should express fitness-for-use for one data input, for one specific BEFM. It is safe to state that today's

technology for metadata analysis is too immature to expect automated solutions to solve the above problems.

Agricultural datasets with global or regional coverage are, for instance, produced by FAO, JRC, and the US Geological Survey (USGS). These international organizations provide datasets for a broad range of applications, ignoring the specifics of certain end-user applications. These datasets are highly standardized between nations. National organizations, like ministries and census bureaus provide comprehensive data obtained from bottom-up approaches, typically working up the administrative area hierarchy by sensible aggregation. The first group of data has often undergone strong curation, and sometimes has lost some detail.

In a wall to wall BEFM application, the run-time discovery of candidate inputs, their selection, and subsequent preparation for use by the model, is further hindered by: data preparation (Granell *et al.*, 2010). The final question in this section is how to prepare the chosen dataset for its use as model input? Three important issues are identified to tackle this question in full, and they have been recognized for some time (Groot & McLaughlin, 2000; Beare *et al.*, 2010; Foerster *et al.*, 2010):

1. **Syntax**: are the problems related to the format in which the model expects the input data to be provided, and the difference with the formats in which the chosen candidate datasets are originally found. This might has the problems of different representations or different encoding. These may involve file formats (e.g., `shp` versus `gml` or `GeoJSON`).

2. **Structure**: Geospatial objects refer to features on the earth surface. In a system environment, they are interpreted with thematic and geometric descriptions (Molenaar, 1998). Geospatial objects are often grouped into several distinct classes with a list of attributes associated with each class. This grouping of classes and attributes is called a conceptual schema. Structural problems originate from differences in underlying conceptual schemas of the chosen datasets and of the model input, especially in the recognition of classes, their thematic attributes, and/or association types between the classes. In the thematic attribute space, for instance, the schema transformations may be required due to: (i) different class-attribute names, (ii) different class-attribute associations, (iii) attributes spread between multiple classes, and (iv) different methods of handling missing data (Steinman *et al.*, 2009).

3. **Semantics**: problems are related with different meaning of the information format, content and structure of the chosen candidate dataset and the model input. Concepts and terminologies conceived in the datasets may not match with those of the model inputs (Granell *et al.*, 2010). The problem usually arises where the natural language terms used are identical or seemingly synonymous, but their (human) interpretations are not. A simple example may well be that what one conceptual schema calls a farm presents

only a subset of the group of farms recognized by the other conceptual schema.

Interoperability is the capability of individual datasets and models to interact through overcoming the syntax, structure and semantic barriers. Syntactical problems are usually overcome by specifying standard interfaces for datasets and models and a standard message format for communication between them (e.g. gml for spatial data) (Foerster *et al.*, 2010). Various standards used for deploying BEFM as a MaaS are discussed in Section 2.2.2.

Structural and semantic transformations (so called schema mapping or schema transformation) usually involve mapping of the conceptual schema of datasets and the model inputs. To do so, functional maps are defined to allow precise data alignments between the conceptual schemas of candidate datasets and the model inputs/parameters, referred to in this thesis as source and integrated conceptual schemas, respectively. Schema mapping affects the structure or the content of datasets, but does not affect the level of detail (Foerster *et al.*, 2010). Although changing the data content may also involve changing level of detail of data, we restrict the schema mapping to the structural and semantic translations of the data in the thematic attribute space.

Structural problems can often be resolved but typically need human interaction (Bian, 2007; Brimicombe, 2009). Approaches have been proposed for automated, often ontology-based, structural repairs, but the level of success is variable (Kaza & Chen, 2008; Huang & Diao, 2008). Solutions to the semantic problems generally focus on developing a common and shared view of a particular domain through developing vocabularies and shared ontologies. These semantic models are used to annotate low level model formalism captured in the integrated schema as well as in defining transformations.

In this study, the spatial parameterization of BEFM demands the scale and schema transformations to be executed at the model run-time. This provision largely depends on the underlying framework for deploying the BEFM as a MaaS.

2.2.2 Model adaptation

An integrated assessment of farm resources involves linking several datasets and sub-models (e.g. a crop model, soil erosion model) to a BEFM. A framework provides an environment to integrate data and model components. Re-usability is the capability of data and model components to be linked outside their frameworks. A framework can either be closed or open. In a closed framework, components are tightly-coupled with the framework, i.e., several framework requirements needs to be accomplished for integrating a component. Consequently, the interoperability of components among closed frameworks is limited. For instance, OpenMI standard defines interfaces that need to be implemented at design-time which results a relatively closed framework. Because

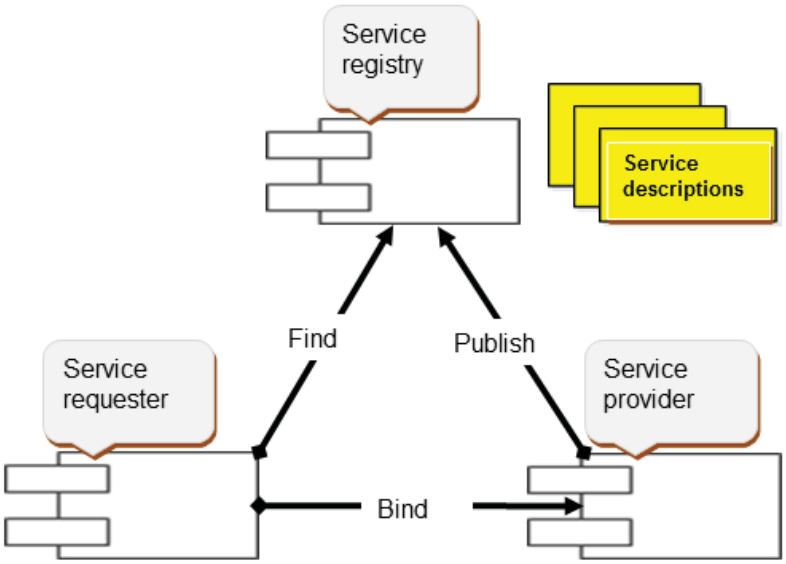


Figure 2.2 The publish-find-bind paradigm.

interfaces require language-specific implementation to integrate a component (Knapen *et al.*, 2009). Consequently, the dependencies may be difficult to resolve when using the models elsewhere. Open frameworks provide run-time integration of components, i.e., components at various organizational nodes may not know each other in advance. Components can be discovered and linked for use by even relatively non-technical users such as farmers and extension workers.

The BEFM used for illustrative purpose in this study was developed in the General Algebraic Modeling System (GAMS) framework (Louhichi *et al.*, 2010). The problems of (re-)using BEFM components (Figure 2.1) alternatively and of interoperability however remain unsolved. Problems are related to the monolithic nature of the framework that offers integration of data and models in a closed and stand-alone environment. It restricts a shared and collaborative environment in which a number of stakeholders can participate remotely to a common problem and share their interoperable resources, services and solutions.

An open framework is usually based on the Service-Oriented Architecture (SOA). SOA supports the *publish-find-bind* paradigm, which specifies three roles: service provider, service registry and service requester. The interaction of the three roles is shown in Figure 2.2:

- the service provider publishes datasets and models as web services with registries (catalogs) using standard metadata descriptions (entries).
- the service requester (i.e., an agent or automated mechanism), by

querying the registries, finds the appropriate web services according to the descriptions of an intended scenario.

- the service requester binds to the service and retrieves the desired functionality.

Based on the SOA, recently, model web started offering distributed data and models as interoperable web services (Geller & Melton, 2008; Reichardt, 2010; Granell *et al.*, 2010). A web service is defined as a software component that provides functionality to invoke a component (e.g. data, model) via a web-accessible interface in a programming language- and platform-independent manner (Foerster *et al.*, 2010). Web services use standard interfaces and information models. Web service interfaces are described in a machine-understandable way, which is a requirement for syntactical interoperability. As a result, web services can interact during run-time and they do not need knowing each other in advance. Web services communicate based on platform-independent data exchange formats (e.g. xml, gml). Web services and SOA provide the basis of the open service framework for data and model integration and re-usability. Conceptually, providing MaaS in this study requires adopting the BEFM components in the open service framework.

International organizations like the Open Geospatial Consortium (OGC) (OGC, 2008c) and ISO (ISO, 2005) specify standards for web service interfaces, and also exchange formats to allow consistent data flows between interfaces. OGC web services are also referred to geospatial services (Foerster *et al.*, 2010). A geospatial data service invokes a computer application for spatial data or related metadata (Di, 2005). Spatial datasets can be disseminated from a database, a physical sensor, an environmental model, or a geocomputation. Metadata describes spatial datasets and spatial data services to make them discoverable. The geospatial web services are described in Section 2.3 in the context of the architectural design of BEFM following the MaaS paradigm.

Using the open service framework, several national and global SDIs have been developed to disseminate spatial datasets through geospatial data services. SDIs generally provide a single access point to the geospatial services (Alameh, 2003). Technically, standards for datasets and metadata are adapted in a cooperative and multi-stakeholder approach (Bernard *et al.*, 2005), for example in the INSPIRE European SDI initiative (INSPIRE, 2008). Our implementation strategy to provide BEFM as a MaaS is centered on the concept of INSPIRE-based open service platform, in which data and model components are offered as loosely-coupled geospatial web services.

2.3 Proposed framework

This section provides the architecture of our proposed framework. Based on the open service platform of SDI (INSPIRE, 2008), the components shown in Figure 2.1 are essentially deployed as web services. Our pro-

posed framework has three tiers: physical tier, services tier and presentation tier (see Figure 2.3), as described below:

1. **Physical tier**: contains hardware and software components of MaaS, including a BEFM underlying modeling system and a database management system to store datasets and metadata.
2. **Services tier**: deploys geospatial web services to invoke the components at physical tier. To initialize BEFM for a location, they provide data, perform GI transformations (when required) and invoke BEFM functionality. Five service roles are distinguished:

   Discovery service plays a catalog role. For this, OGC provides specifications of the Web Catalog Service (CSW) (OGC, 2007a) for implementing a discovery service. BEFM components deployed as web services can be published and discovered in the catalog using metadata descriptions. End-users or other services can query these descriptions, and, thus, they can perform a fine-grained search to discover required services.

   View service supports visualization of datasets. Instead of the geographic and thematic content of geospatial objects, they render maps in a pictorial format such as GIF. OGC provides specifications of web map service (WMS) (OGC, 2008b) for implementing a view service.

   Invoke service refers to as the encoding of service input parameters for execution, but also to describe format for transferring the results. For this, XML-based standards are used in SOA, including the web service definition language (WSDL) to describe interfaces and the Simple Object Access Protocol (SOAP) to transfer results.

   Download service supports access to the geographic and thematic content of geospatial objects. OGC provides the implementation specifications of download services, including (i) the Web Feature Service (WFS) (OGC, 2005) to query the vector-based geographic data encoded in the Geographic Markup Language (GML), and (ii) the web coverage service (WCS) (OGC, 2008a) to access raster-based geographic data, e.g., digital elevation models, multispectral images encoded in a specified binary image format (Granell *et al.*, 2010).

   Geocomputation and transformation services the geocomputation service invokes binding to the published processes that operate on spatially referenced data. OGC provides specifications of the Web Processing Service (WPS) (OGC, 2007b) for implementing a geocomputation service. WPS is a generic web service interface which means that it can be configured for a geocomputation process (e.g. to perform a transformation or a spatial analysis) as well as for a model (i.e. MaaS). Because there are no restrictions on format, network location, platform and number of data inputs/outputs to/from a geospatial

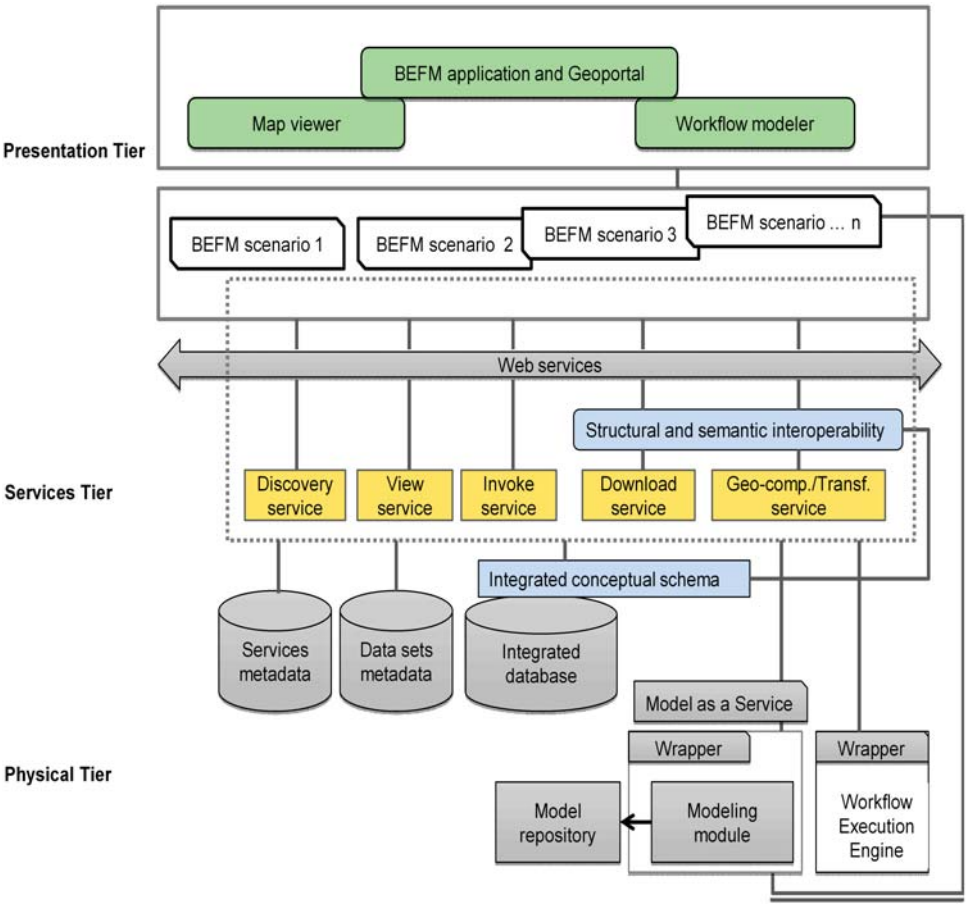


Figure 2.3 Physical, services, and presentation tiers of the proposed framework to deploy a bio-economic farm model (BEFM) as a service based on the open service platform of spatial data infrastructures; web services on the services tier interact with BEFM components on the physical tier; structural and semantic interoperability is obtained through integrated conceptual schema in the database; spatial statistical (quantitative) models are provided with Geocomputation and transformation services.

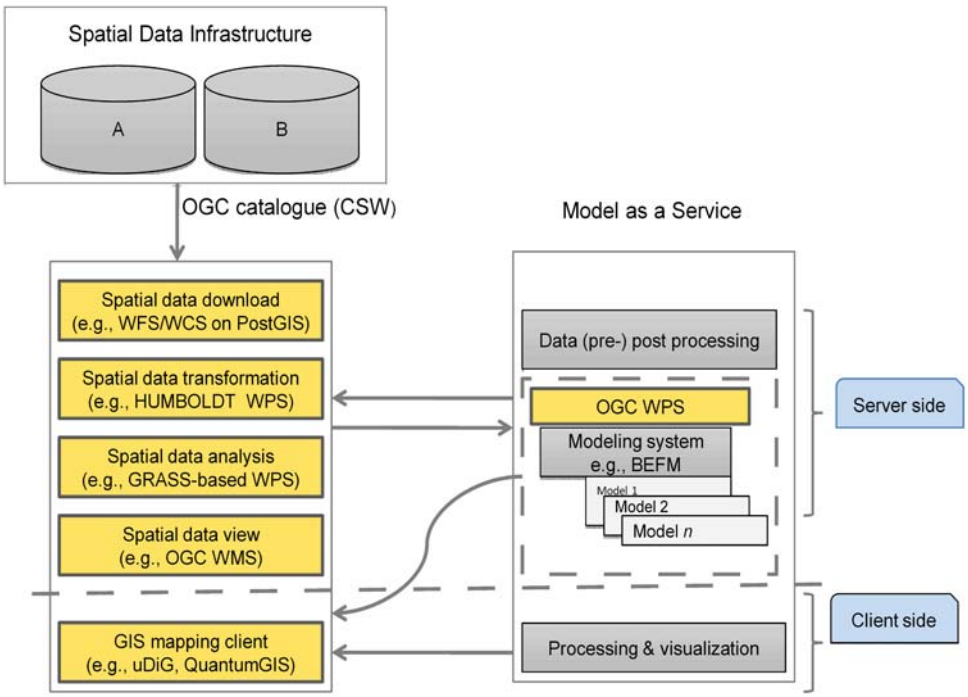


Figure 2.4 The proposed framework is shown in the traditional client-server view; the Open Geospatial Consortium (OGC) Web Processing Service (WPS) interface can be implemented to accomplish a geocomputation that may be: (i) a bio-economic farm model (BEFM) as a location-based service following the MaaS paradigm, (ii) a spatial data analysis, or (iii) a spatial data transformation; For location-based parameterization, these geocomputation services interact with geospatial data services (e.g. spatial data discovery, download and view services) offered by spatial data infrastructures.

process (OGC, 2007b), complex computational models can be wrapped with the OGC WPS, e.g., a BEFM as a service (see Figure 2.4).

The transformation services are the core of this framework. They can be deployed for accomplishing scale and schema transformations (discussed in Section 2.2.1) of the spatial datasets offered by third-party agricultural services to initialize a number of BEFM inputs/parameters. The structural and semantic difference in datasets and model inputs is a major difficulty in moving towards a more rigorous approach to their integrated use. Consequently, it requires parsing, transforming and spatially analyzing of the download services to match exactly the initialization requirements of a computational model. Current SDI initiatives, like INSPIRE (INSPIRE,

2008), do not enforce organizations to change and store existing datasets. Instead, the datasets may be transformed or mapped on-the-fly via transformation services. To do so, the following methods may be deployed in this framework:

- To overcome structural differences between diverse datasets and BEFM, the download and transformation services may comply with an integrated conceptual schema, in which datasets and models stand in relation to each other. The integrated conceptual schema makes explicit the BEFM formalisms (e.g. model inputs/parameters and their association with spatial objects) as well as provides the basis for designing transformation operators.
- A shared ontology is developed collaboratively by a group of researchers in a domain to achieve agreed-upon meaning of various concepts (e.g., classes, attributes and associations between classes) in the domain. The integrated conceptual schema can be annotated or harmonized with a shared ontology in the agricultural domain to align semantics of the internal formalisms of diverse datasets and BEFM inputs.
- Functional maps (referred to in this thesis as transformation operators) may be defined to accomplish transformations in a workflow for a farming system assessment. Defining functional maps is a semi-automated way to translate datasets (i.e. source conceptual schema) to the integrated conceptual schema. Transformations can also be implemented in a download service via filter encoding or with geocomputation services. The download services can be implemented through stored procedures in GIS database.

3. **Presentation tier**: For a location-based parameterization, the BEFM web service links the geospatial data services. For this, the presentation tier facilitates viewing metadata, visualizing and selecting appropriate geospatial services at the services tier, but also displaying the FSSIM results. To achieve this, a Geoportal is often implemented as a single access point to all geospatial web services (Zhanng & Tsou, 2009). Alternatively, workflow modeler can be used to link different BEFM components as pluggable services.

2.3.1 Spatial statistical (quantitative) models

Implementing transformation operators may be relatively straight forward for the production technology data (e.g. labour, fertilizers). These variables may be aggregated (average, sum, mean, or single values) from sampled data, for instance, for all households in a representative village. The aggregated values can then be served as representative values for the village-level wall to wall application of BEFM. In the same way,

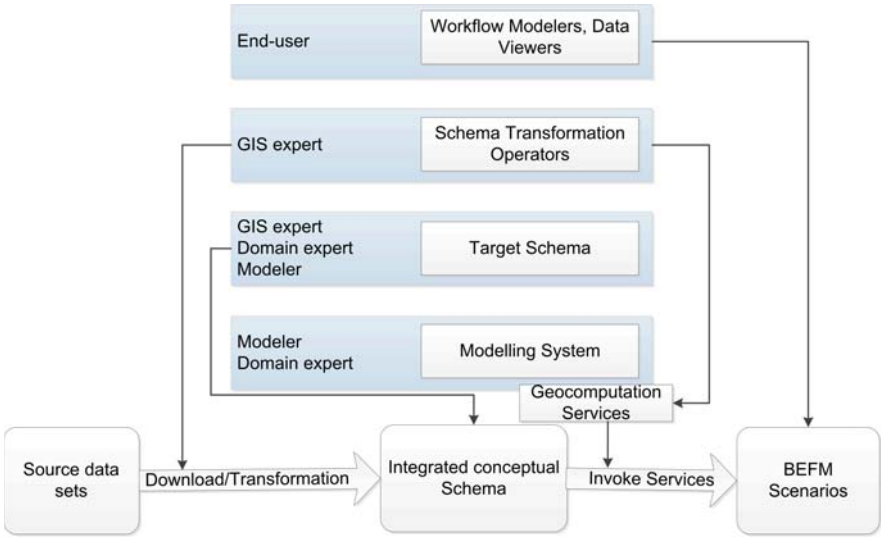


Figure 2.5 The course of interaction of various users to the components of proposed framework

however, aggregating biophysical data may not be simple. Spatial (dis-)aggregations, in this case, often need to change data resolution using scientifically rigorous (up-)down-scaling procedures. These scaling procedures are usually concerned with the variability of spatial processes, their states, and throughput of quantities (Bian, 2007). For example, the wall to wall application of BEFM at the farm/village level requires estimates of crop yields for all farms/villages in a large region (e.g. country). To obtain this, the crop yield observations in representative villages may be spatially interpolated towards non-sampled villages. Such a scale-related transformation is relatively more cumbersome to perform, because crop yields are highly spatially variable due to heterogeneous soil, weather, and topographic conditions. Constructing spatial variability of crop yields at a finer resolution requires explicit treatments, either using ancillary data or models to describe the unknown variability. Spatial statistical models may be applied in combination with remote sensing to model spatial structure/neighborhood for weighing data at one scale to generate data at another scale. The output of these quantitative models can be stored in GIS database. These data may be complied with the integrated conceptual schema to run a wall to wall BEFM application.

2.3.2 Interaction of system users to components of the proposed framework system

Figure 2.5 shows the course of interaction of various users to the components of proposed framework. Four types of user roles may be identified including, the modeler, the domain expert, the GIS expert and the end-user. A modeler often together with a domain expert interacts with the underlying modeling system at the physical tier to develop a BEFM. Modelers and GI experts may independently interact with the system. However, all the four user roles need to collaborate to develop an integrated conceptual schema. Using the integrated conceptual schema, the modelers and GIS experts determine required transformations. Based on the integrated conceptual scheme, GIS experts develop and advertise appropriate download, transformation and geocomputation web services at the services tier. Using thin clients at the presentation tier, end-users may discover and invoke data services for linking to a BEFM web service.

2.4 Implementing the proposed framework – the case of Burkina Faso

The proposed framework (in Section 2.3) is implemented to analyze terroirs at the national-scale of Burkina Faso (i.e. in a wall to wall setting). A terroir in Burkina Faso is a community-based land management approach that not only constructs a physical area, but also a social construct and the notion of natural resources and biophysical conditions. Thus, it constitutes a farming system in which farmers contribute their individual parcels and adopt common policies for agricultural production. In Burkina Faso approximately 92% of the country workforce is actively associated with the agricultural sector, of which 80% are small holder farmers living in terroirs (USAID, 2009). Farming and cropping conditions are highly variable in the country due to spatially varying biophysical and socioeconomic conditions and use of production technology, while the level of spatial variation might differ between regions in the country. The Statistiques Agricoles du Burkina Faso (aka AGRISTAT) collects various data for a representative terroir in a district (AGRISTAT, 2010). Like many West-African countries, the biophysical parameters are highly variable in the country. National policies, for instance, are spatially more constant than most biophysical parameters, but even they are district-specific.

Based on local farming and cropping conditions, agricultural production largely depends on formulating effective farming activities (e.g. allocating land, labour, and inputs to various crops) at terroir locations. A farming activity is characterized by a set of inputs and outputs that expresses the contribution of activity to the realization of defined goals or objectives (Janssen & van Ittersum, 2007). The inputs are the current (and alternative) biophysical, (socio-)economic and human resources, which may vary from one terroir location to another. The outputs (aka

Table 2.1 Inputs of farming activities related to the farm, labour, parcel, price and production at a terroir location in Burkina Faso.

Service	Model inputs	Description	Source
Farm	*f*	terroir *f* representing group of farms, i.e., a rural community in Burkina Faso	AGRISTAT
	i	farming activities *i*	–
	crops(i)	crop activities *i*	–
	food(i)	food crop activities *i*	–
	cash(i)	cash crop activities *i*	–
	past(i)	animal activities *i*	–
	annual(i)	annual crops activities *i*	–
	per(i)	perennial crops activities *i*	–
	luse(i)	other land use activities	–
	size(f)	size of a terroir *f*	–
	capstart(f)	starting capital of a terroir *f*	–
	cons(i)	minimum required consumption of crop *i* (per terroir)	–
Labour	*aez*	agri-ecological zone *aez* represents land quality based on soil conditions (percentage of soil calcareous, loam, sand, water logging capacity in top-soil), topography (slope and elevation) and weather (rainfall) at a terroir *f* location	HarvestChoice
	labreq(f,i,aez)	labour required for crop *i* on land quality *aez* of the terroir *f* (hours)	AGRISTAT
	famlab(f)	available family labour in a terroir *f*	AGRISTAT
Price	*price(f,i)*	price of crop *i* at the terroir *f* location	AGRISTAT
Parcel	*cropland(f,aez)*	cropland quantity of land quality *aez* available to a terroir *f* (ha)	AGRISTAT
	totland(f,aez)	total land of land quality *aez* available to a terroir *f* (ha)	–
	initland(f,i,aez)	initial land of land quality *aez* allocated per crop *i* by the terroir *f* (ha)	–
Production	*yield(f,i,aez)*	yield (kg ha^{-1}) per crop *i* in the terroir *f* depending on land quality *aez* obtained from crop yield modeling and upscaling	Chapter 3
	inputs(f,i)	non-labour costs for producing crop *i* in a terroir *f*. Proxy data are obtained from farmers marginality modeling and upscaling	Chapter 5

Table 2.2 Output decision variables for integrated assessments.

Output decision variables	Description
TLABOUR(f,i,aez)	hired temporary labour (hours) for crop *i* on land of quality *aez* by the terroir *f*
FLABOUR(f,i,aez)	family labour (hours) allocated to crop activity *i* on land of quality *aez* by the terroir *f*
A(f,i,aez)	land allocated (ha) to crop *i* on land of quality *aez* of the terroir *f*
C(f,i,aez)	capital assigned (CHF) to crop *i* on land of quality *aez* of the terroir *f*
OUTPUT(f,i,aez)	output of crop *i* on land of quality *aez* of the terroir *f*
FOODCOST(f)	terroir *f* costs of buying food from market
PRODCOST(f)	terroir *f* total costs for production
LANDVAL(f)	land value of terroir *f*
PURCH(f,i)	terroir *f* amount of food crop *i* purchased from market
REV(f)	terroir *f* gross revenue
OBJ(f)	terroir *f* objective function

decision variables) describe elements of farming activity for which a decision must be made by a terroir community. A BEFM can optimize resource allocations to a terroir, given the farming activities and constraints with respect to an objective function that reflects goal of the terroir community (Hazell & Norton, 1986). In this context, we want to implement a BEFM as a tool for on-terroir decision making in Burkina Faso. To do so, we focus on integrating location-based datasets and a bio-economic farm model adapted from FSSIM (Farm System Simulator, SEAMLESS (Louhichi *et al.*, 2010)). Here, the basic purpose of using this BEFM is to illustrate: (i) its use in developing wall to wall agricultural services for on-terroir decision-making in Burkina Faso using third-party datasets, including national agricultural surveys and regional datasets, (ii) its adaptation as a geospatial web service in the proposed SDI-based framework, and (ii) its adaptation for regional (quantitative) modeling and upscaling in the proposed SDI-based framework, as explained below,

2.4.1 Adapting the BEFM for on-terroir decision-making in Burkina Faso

This section investigates how to adapt the FSSIM formalisms (e.g. inputs/outputs, objective function and constraints) for subsistence farming situations in Burkinabé terroirs. The adapted FSSIM is essentially a linear programming model in the GAMS framework. The model formalisms are briefly described as,

- Table 2.1 presents inputs of farming activities related to the farm, labour, parcel, price and production inputs at a terroir location f.
- Table 2.2 describes different output decision variables.

- The objective function was set out to maximize the revenue of a terroir community, after meeting all production costs and securing enough food for all its household members. The mathematical formulations of adapted objective function and constraints are provided in Appendix A.

2.4.2 Adapting the BEFM as a geospatial web service in the proposed SDI-based framework

This section investigates how to adapt the FSSIM in implementing the proposed SDI-based framework in Section 2.3. Following the MaaS paradigm, the model was deployed as a web service that can be initialized for a terroir location through orchestrating various spatial data services. Various tiers outlined in the proposed framework (Figure 2.3) were systematically deployed (Figure 2.6), as explained below:

1. **Physical tier**: the General Algebraic Modeling System (GAMS) is the underlying framework of the FSSIM. GAMS native interfaces (i.e. JNI and GDX API) were wrapped with OGC WPS standard interface to allow the model to interact with other geospatial services. Its deployment at the physical tier (called inner wrapper in Figure 2.6) is responsible to change the model states at simulation run-time. Moreover, it provides a communication stack between the physical tier and the web services tier. For a terroir location, it analyzes and decomposes the geospatial objects into the model parameters. Similarly, PostgreSQL/PostGIS was deployed as an underlying database system to realize source and integrated conceptual schemas. Datasets obtained from AGRISTAT were stored in the database (see the source conceptual schema in Appendix C). This GIS database was served to provide field observations to the FSSIM and to several quantitative models in this thesis. The target conceptual schema is explained below.

2. **Services tier**:

   Discovery service The deployment of OGC CSW was based on GeoNetwork (GeoNetwork, 2010). Following the standard style operation, the web services published in the catalog can be discovered by other services and client applications (Geoportal, Workflow Modeler, Udig and so forth).

   Download service The deployment of OGC WFSs and WCSs was based on GeoServer (GeoServer, 2010). The GeoServer was deployed top of the PostgreSQL/PostGIS database to provide datasets as WFSs for parameterization of the FSSIM web service.

   Invoke service Following the current SDI implementations for geospatial web services, this study used open and non-proprietary internet standards like URL.

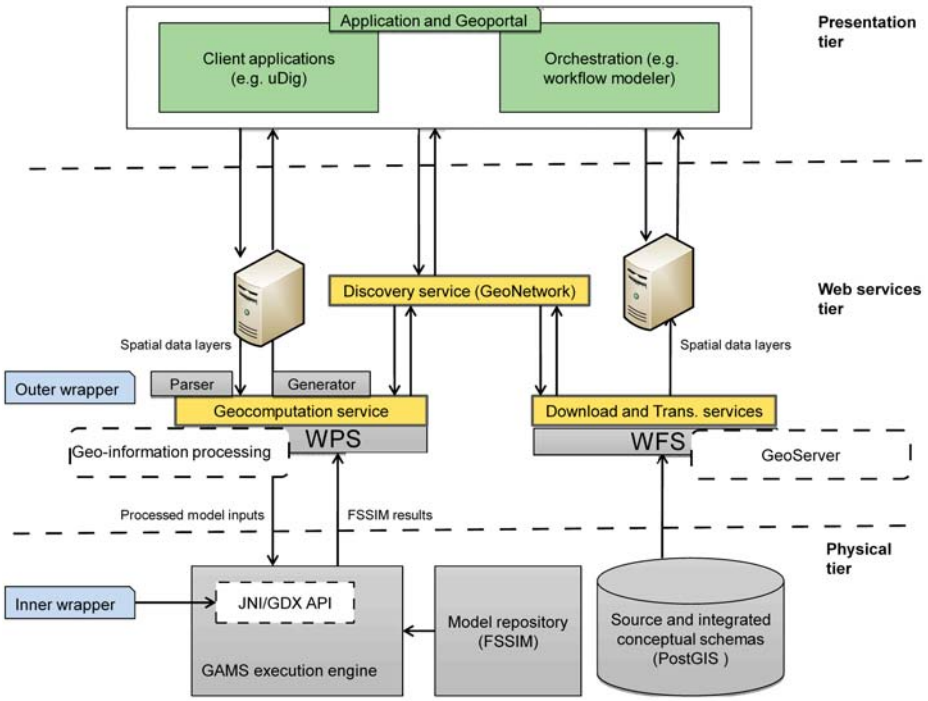


Figure 2.6 Implementing the Web Processing Service (WPS) interface for the Farming System Simulator Model (FSSIM) and the Web Feature Services (WFSs) for spatial data; Inner wrapper implements GDX API and JNI interfaces of the FSSIM modeling system, i.e., the General Algebraic Modeling System (GAMS) to develop a communication stack with 52North WPS process; Outer wrapper handles requests and responses of the FSSIM provided as WPS, i.e., the model as a service (Maas); Spatial data download and transformation services implement WFS interface based on GeoServer; Discovery service implements the Web Catalog Service (CSW) interface based on GeoNetwork.

Geocomputation service To interact with the FSSIM at the physical tier, the OGC WPS (OGC, 2007b) was deployed at the services tier. The deployment of this geocomputation service was based on the 52North WPS platform (Schäffer, 2009). The deployed geocomputation services encapsulated the FSSIM as well as existing GIS computational functions such as provided in GRASS, SEXTANTE, HUMBOLDT (Figure 2.4). Following the OGC WPS style operation, a list of all available models and geocomputational processes were included in the service capabilities document while metadata descriptions of their inputs were accessed through the standard WPS style operation (i.e. DescribeProcess).

Transformation service The syntactical and semantics problems of defining, associating and aggregating spatial objects were tackled through designing the integrated conceptual schema of the FSSIM model inputs (see Figure 2.7). This integrated schema was annotated with the SEAMLESS agricultural ontology (Janssen *et al.*, 2009). The SEAMLESS ontology has been developed in the SEAMLESS project (van Ittersum *et al.*, 2008) to integrate various datasets and models in the agricultural domain. The annotation of the integrated schema with the SEAMLESS ontology is explained as:

a) **Production technology/socioeconomic data**: AGRISTAT collects data for all households in a representative terroir. A household has one or many land parcels for growing one or many crops. In the integrated schema, a household is represented with the 'Household' class, which was derived from the 'Farm' class of the SEAMLESS ontology. The 'Household' class defines model inputs related to the production technology and socioeconomic data in the source schema. This data were aggregated from all households in a representative terroir to obtain the model inputs at the terroir-level, for instance, total family labour available in a terroir. The 'Household' class is associated with the 'Representative Terroir' class, which was derived from the 'Representative Farm in Agri-ecological Zone' class in the SEAMLESS ontology. The 'Representative Terroir' class is associated with the 'District' class, which was derived from the 'Administrative region' class in the SEAMLESS ontology. The thematic attributes associated with the 'Representative Terroir' class in the target schema were served as representative data, to run the model for all other terroirs in the associated 'District' class. For instance, the 'Available Family Labour' from the 'Household' class was mapped from aggregating the number of adult members in the 'Household Member' class of the source schema.

b) **Biophysical data**: The 'Representative Terroir' class in

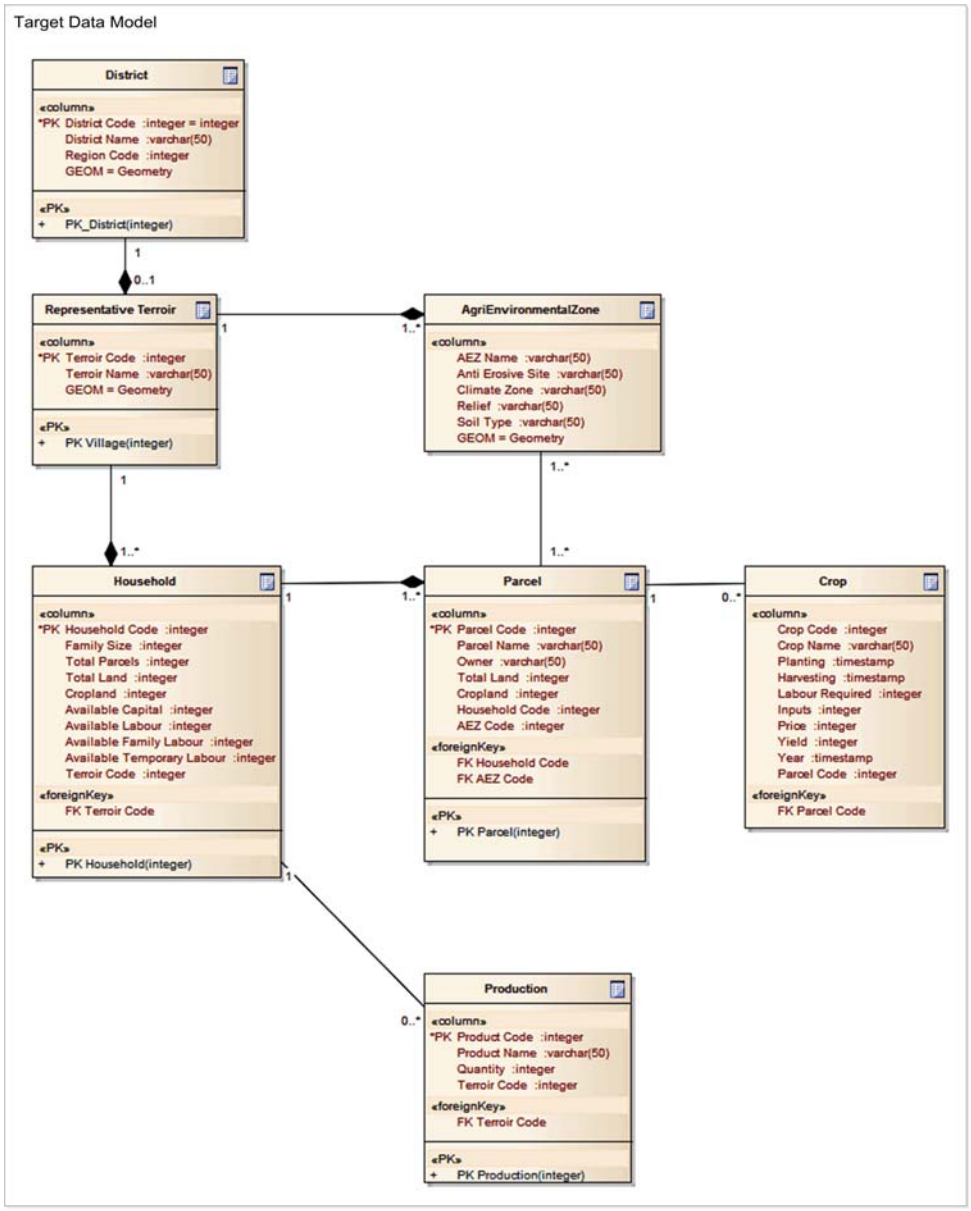


Figure 2.7 Integrated conceptual schema for datasets and models integration.

the target data model is further associated with the 'AgriEnvironmentalZone' class, which was derived from the 'Agri-ecological Zone' class in the SEAMLESS ontology. The 'AgriEnvironmentalZone' class characterizes the biophysical and environmental content relevant to the soil, topography, and climate data. For example, Chapter 3 provides a model input of crop yield, which is estimated as 'yield (kg ha^{-1}) per crop in a terroir depending on agri-ecological zone'.

The integrated schema was used to define a range of transformations (from source to integrated schema) to initialize the FSSIM inputs related to various variables of farming systems. They are described through a formal language for several transformation operators (Appendix D). These operators were implemented through database stored procedures. Transformed data in the GIS database were provided with WFSs to parameterize the FSSIM model (see Figure 2.6). In this way, WFSs served the transformation services in this study.

3. **Presentation tier**: Workflow modeler was deployed to achieve tasks of the presentation tier. Using this, the end-users can easily visualize the environment to compose and to deploy the FSSIM scenarios. Interactions among various components (in Figure 2.6) are explained in the next paragraph.

   Following the WPS style operation, the FSSIM web service (i.e. OGC WPS) can be discovered from the SDI catalog (i.e. discovery service). It can then be invoked by workflow modeler at the presentation tier. For this, the services tier handles requests and responses of the FSSIM service. The services tier parses (using parsers in the 52North WPS platform) geospatial data offered by WFSs. The parsers manipulate feature objects to prepare input data which are then passed to the physical tier. The physical tier then changes the FSSIM states at simulation run-time. In this way, physical tier provides a communication stack with the services tier. Its main tasks are to decompose the input data into model parameters and to handle the FSSIM simulation. Finally, the simulation results are delivered to a generator at the services tier to generate a response to the client application at the presentation tier. Like parsers, the 52North WPS platform provides various output data transformers for turning the result into appropriate format (i.e. for presenting the FSSIM results).

To find out the optimized resource allocations given the objective function and decision variables in Section 2.4.1, the implemented framework is tested to run the FSSIM web service in Burkinabé terroirs. To initialize the FSSIM web service for a terroir location f, the geospatial download services (i.e. WFSs) are deployed on GeoServer. These agricultural services provide terroir-based data from diverse disciplines to the FSSIM model inputs. The data are related to the farm, labour, parcel,

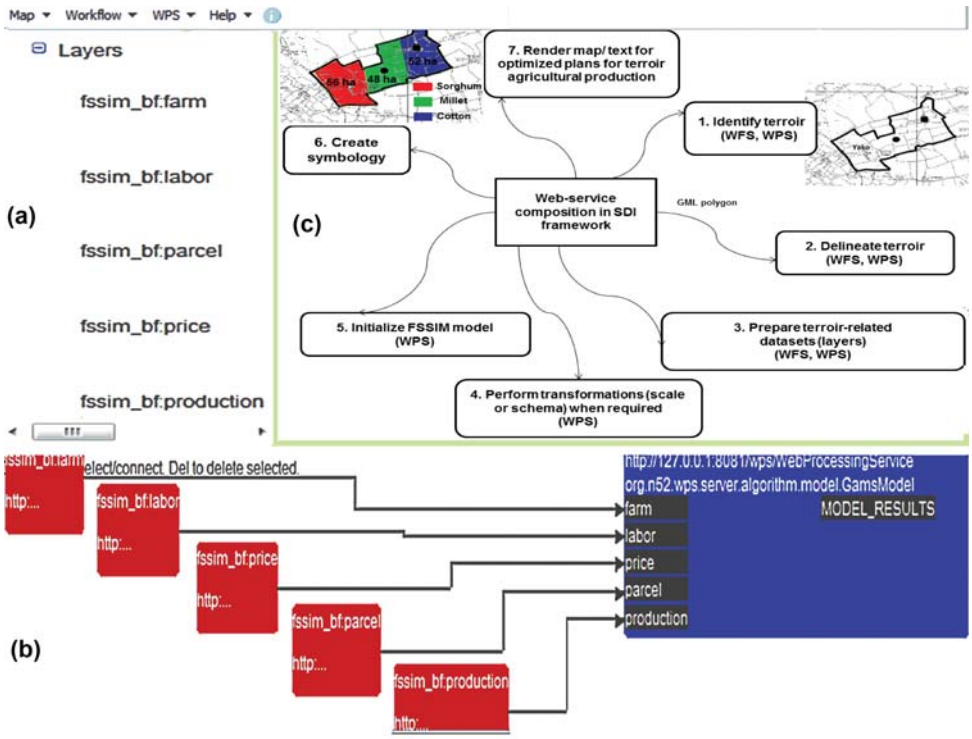


Figure 2.8 Web services providing data (layers) related to the farm, labour, parcel, price and production inputs of the Farming System Simulator Model (FSSIM) model (a) – Web services for data (WFSs, Web Feature Services) are linked to the web service (WPS, Web Processing Service) for the FSSIM model (b) – Various steps performed by the web service chain composed in the SDI (spatial data infrastructure) framework for rendering optimal cropland allocation (area) plans for 'Yako' terroir in Burkina Faso (c).

price and production inputs (Table 2.1). For instance, FSSIM model inputs pertaining to labour and production, e.g., *labreq(i,aez), yield(f,i,aez), inputs(f,i,aez)* depend on agri-ecological zone *aez*, crop type *i* and terroir *f*. These geospatial data services are linked to the FSSIM web service (i.e. WPS on 52North) for assessing farming resources in terroirs. The service results values of output decision variables which depict optimal allocations of resources in a terroir.

An extension worker or terroir community can easily link various agricultural services to the FSSIM web service and can visualize the optimal solutions for the decision variables. For example, Figure 2.8 shows WFSs of farm, labour, parcel, price and production are linked to the FSSIM WPS to find an optimal solution of allocating cropland at the location of 'Yako' terroir. The service recommends allocation of 56 ha, 48 ha, and 52 ha for sorghum, millet, and cotton crops respectively.

2.4.3 Adapting the BEFM for regional modeling and upscaling in the proposed SDI-based framework

In the proposed framework (in Section 2.3), two types of transformations are identified: schema-related and scale-related. In this implementation, the transformation operators are mostly used to transform spatial objects in the thematic attribute space, e.g., spatially aggregating all household members in a terroir to compute available family labour (i.e. management data). The scale-related transformations however can be accomplished through applying spatial statistical models, such as the models for upscaling crop yield and marginality to regional scale. This Chapter discusses the prospect for these transformation services to adapt the FSSIM as a wall to wall service in SDI-based framework. However, scale-related transformations and consequently, the data fitness-for-use at the model simulation unit, its effect on the model parameterization, and resulting uncertainty are the issues that need to be further investigated. The issues of spatial data quality, upscaling, and uncertainty will be investigated in Chapters 3, 4, and 5 of this thesis, in the context of adapting BEFM for regional quantitative models in the SDI-based framework.

2.5 Conclusions and future work

This study explores how models in agriculture, offered as geospatial web services, can take benefits of SDI-based frameworks. To do so, it set out a standard wrapper over a bio-economic farm model to allow it to be exposed as a web service, following the Model-as-a-Service paradigm. The study achieves this through the OGC-compliant implementations for datasets and for the model, and it discusses the prospect for more mature transformation services, in the context of our services framework. As a test case for a terroir location in Burkina Faso, the study initializes the

model inputs/parameters on-the-fly with the orchestration of geospatial data services.

The study addresses the syntactical, structural and semantic issues for data and models integration. The syntactical issues can be handled through deploying standard interfaces and data encoding. The structural issues can be overcome through defining an integrated conceptual schema for data and models integration. The integrated schema can be used to explicate the parametric space of models. The study derives concepts and classes in the integrated schema from the SEAMLESS ontology. The SEAMLESS ontology is useful to align the semantics between data and model formalisms in the agricultural domain. The structural translations (i.e. from a source schema to the integrated schema) can be provided in the development of transformation operators. Presently, these operators are deployed at design time in a semi-automated transformation mode. Providing dynamic transformations from source to target models is challenging.

We conclude that the geospatial web services provide a scalable way of discovering and linking data and models for integrated assessments. They support community-wide participation in understanding, developing and using those resources. Moreover, the geospatial web services overcome technical and conceptual barriers to support sharing of existing spatial datasets. Benefiting from SDI technology, the proposed framework has potential for linking spatial scaling to simulation models at regional scale to deploy wall to wall services. Provision of these services will enable wider exploitation of existing SDIs to facilitate the integrated assessments of various farm resources in developing countries.

Presently, it is possible to show results for a single decision variable (i.e. land allocated (ha) to crop *i*). Further investigation can provide efficient visualization tools for presenting results from the BEFM web service. Moreover, the deployed service requires a usability test and proper feedback from the extension workers in Burkina Faso. Nevertheless, this prototype implementation is first attempt to make a BEFM spatially-aware and to integrate it with various SDI services.

3

Modeling crop yield in West-African rainfed agriculture using global and local spatial regression

[1]This chapter is published as: Imran, M, Zurita-Milla, R, Stein, A. (2013). Modeling crop yield in West-African rainfed agriculture using global and local spatial regression. *Agronomy Journal*:105(4).

3. Modeling crop yield in West-African rainfed agriculture using global and local spatial regression

Abstract

Performance of crop yield models is generally evaluated without testing their ability to capture yield spatial variability over a large area. Local soil and environmental conditions or management factors usually cause significant crop yield variability. In West Africa, landscape heterogeneity and data scarcity pose yet additional challenges to crop yield modeling. In this study, conditional autoregression (CAR) and geographical weighted regression (GWR) were used to better understand spatial patterns in sorghum, millet, and cotton yields in Burkina Faso. A series of SPOT NDVI 1 km^2 10-day composite imageries spanning the crop growing season and observations of rainfall, topography, soil properties, and labor availability were used as explanatory variables in the CAR and GWR models. Regression analyses revealed that crop yield was significantly related to rainfall and topography in the semiarid and subhumid agroecological zones of Burkina Faso. Soil properties and labor availability mainly affected sorghum and millet yields in its semiarid zone. By addressing spatial dependency between crop yield observations in the two zones, GWR outperformed the CAR models. For CAR models, R_a^2 values for the sorghum, millet, and cotton yields were 0.76, 0.70, and 0.50, respectively, for the semiarid zone, and 0.54, 0.32, and 0.30, respectively, for the subhumid zone. For GWR models, R_a^2 values were 0.85, 0.70, 0.78, respectively, for the semiarid zone, and 0.76, 0.67, 0.65, respectively, for the subhumid zone. Thus, despite limited data availability, GWR can be used to model the spatial variability of crop yields over large areas in West Africa.

3.1 Motivation and outlook

West-African agricultural systems are characterized by large spatial variability of soil and weather conditions. Farmers cultivate crops in small fields, often growing multiple crops in the same fields due to the prevalence of rainfed subsistence farming. This induces large spatial variation in crop yields and makes these agricultural systems relatively complex and highly location-specific (Therond *et al.*, 2011). As a result, farmers adopt various strategies to increase their crop yields, to secure food supply and their economic profitability. Research aimed at designing sustainable farming strategies often makes use of bio-economic farm models (BEFMs), which typically combine methods and data from biophysical and economic disciplines (Janssen & van Ittersum, 2007). In particular they use crop models that are usually calibrated at a site-specific level from more detailed data sets (e.g. point-based or fine resolution). Application of BEFMs at a regional scale, however, requires adopting approaches that accommodate sparse amounts data while meeting their calibration requirements for large heterogeneous areas (Faivre *et al.*, 2004). In this study we focus on precisely this type of BEFM application to model spatially disaggregated estimates of crop yields in Burkina Faso. Such an undertaking for West Africa is challenging as it requires modeling crop yields with high spatial variability, caused by heterogeneous soil, climate, and management practices (Lambin *et al.*, 1993).

To estimate crop yields at a regional scale, crop-growth simulation and statistical models (Therond *et al.*, 2011) can be used. Models simulating crop growth use deterministic and semi-deterministic representations of biophysical processes (e.g. photosynthesis and leaf area development) to simulate rates of crop growth at a point-sampling area within a field or at field scale (Ncube *et al.*, 2009). Such models, for example used in combination with Geographical Information Systems, require daily estimation of input values or observations for a range of parameters pertaining to the vegetation soil energy system. A dense network of field stations is required to provide the necessary input data. However, the use of crop simulation models in West Africa has been criticized because of data scarcity and because only default or average parameters are available as input values (Therond *et al.*, 2011).

As in many developing countries, statistical crop data have been used in West Africa to estimate yield of major crops. The statistical data on crop area, yield, and production are collected during ground-based field surveys and are typically reported for various administrative units. Because of high operational costs, ground-based surveys are conducted for selected representative sites and then only once every several years. To improve yield estimates, observations are extrapolated for larger areas (Lambin *et al.*, 1993). Different extrapolation methods may be used to estimate the impact of specific factors on the observed crop yield, and to predict crop yields at unsampled locations using the estimated relations (Prasad *et al.*, 2006; Sharma *et al.*, 2011).

Derived from remote sensing (RS), various biophysical indicators have been widely used to monitor crop performance and other vegetation properties (Dorigo *et al.*, 2007), ultimately resulting in maps that indicate the presence, acreage or yield of a crop. For West-African countries, the combination of RS data and statistical methods has produced promising results in modeling crop yields at the regional scale (Challinor *et al.*, 2009). Estimated relationships between crop yields and their external covariates may, however, suffer from spatial and temporal instability (Ozdogan, 2010; de Beurs & Henebry, 2010). Relationships among environmental factors, crop management practices and crop yields typically vary with spatial location (Faivre *et al.*, 2004). Standard regression approaches for modeling these relationships assume stationary conditions, with mean and variance constant and independent of location (Griffith, 1988). Geostatistical methods allow spatial variation of crop yield to be modeled. Performance will be poor, however, if variables exhibit discontinuous spatial variation (Faivre *et al.*, 2004).

Conditional autoregressive (CAR) models estimate regression coefficients by incorporating a spatial neighborhood structure when dealing with first-order spatial dependence (Overmars *et al.*, 2003). Nonetheless, local dependence between crop yield and explanatory variables may violate the basic assumption of CAR models, namely that the variance is stationary in the study area. Geographically weighted regression (GWR) methods overcome this problem by using a series of distance-related weights to deal with spatial variation in the relationships between crop yield and local conditions (Fotheringham *et al.*, 2002). In doing so, GWR may identify more precisely those local factors that contribute to the spatial variability of crop yields in a highly heterogeneous landscape like that of West Africa (Bevan & Conolly, 2009). In this study we set out to investigate and map the relationship between crop yields and biophysical properties in Burkina Faso. Vegetation indices derived from remote sensing images and biophysical properties such as local weather data, soil characteristics, and labor availability were used as explanatory variables.

3.2 Study area

The study was conducted using data from farming systems throughout Burkina Faso, where more than 80% of the country's population makes their living from agriculture. Croplands represent 23% of the country's total area. Sorghum and millet, the top two staple food crops, are cultivated on 34% and 29%, respectively, of the total area cropped, whereas cotton, Burkina Faso's primary cash crop, is cultivated on 13% of cropland (FAO, 2012a). Administratively, the country is divided into 351 districts (Figure 3.1), which together contain some 7000 terroirs, or areas of land representing a specific combination of biophysical, social, and economic conditions. Within a terroir, typically farmers adopt common management practices and agricultural policies on their individual land

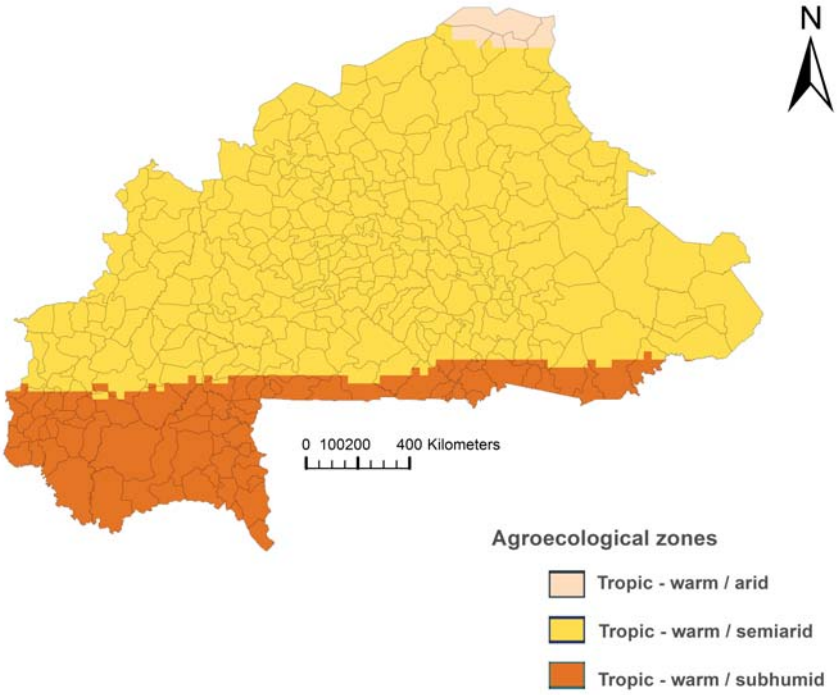


Figure 3.1 The three agroecological zones (AEZs) of Burkina Faso: arid, semi-arid, and subhumid and the boundaries of the 351 districts of Burkina Faso.

parcels. In our study, we used the land parcel within a terroir as the unit for spatial analysis.

The central and northern parts of Burkina Faso are usually subject to low but varying rainfall (300-950 mm), while higher and more homogeneous rainfall (950-1300 mm) occurs in the South (AQUASTAT, 2005). Elevation varies from 300 to 900 m in the central and northern parts of the country, as compared to elevations of 900-1300 m in the South (FAO, 2005). Soil characteristics such as texture and structure also vary across the country. Different rainfall, topography, and soils cause variation in average crop yields from one location to another. On a sub-Saharan scale, the HarvestChoice/International Food Policy Research Institute has defined a number of agroecological zones (AEZs) (HarvestChoice, 2012), three of which apply to Burkina Faso: the tropic-warm / arid and tropic-warm / semiarid zones found in the central and the northern parts of the country; and the tropic-warm / subhumid zone occurring in the southern part of the country (Figure 3.1).

3.3 Materials and methods

3.3.1 Data

In West Africa, estimations of crop yields and cropped acreages are based on ground surveys. Field surveyors supply Statistiques Agricoles (AGRISTAT, Burkina Faso) with seasonal crop-specific data for the predefined terroirs (AGRISTAT, 2010). AGRISTAT survey data for the year 2009 were used as a comprehensive database for crop yield modeling in this study.

For vegetation monitoring and crop yield assessment and forecasting data, the normalized difference vegetation index (NDVI), a common indicator for these purposes (Budde *et al.*, 2004), was used. For 2009, a time-series of SPOT VEGETATION NDVI composite images (S10 product) was obtained from the VGT4AFRICA project through GEONETCast (JRC, 2006). The quality of SPOT NDVI data is sufficient for small- to large-scale crop yield mapping in West Africa (Ramankutty, 2004). A total of 18 images were obtained for the 2009 crop growing season (June to November).

Data on soil properties, topography, weather conditions, and labor availability (Figure 3.2) were used as input for biophysical factors. Raster maps (1 km^2) showing soil properties of the topsoil were obtained from the HarvestChoice database (HarvestChoice, 2012). Elevation and slope were selected as topographical variables; these data were obtained from Hydro 1 km Africa datasets (USGS, 2012).

Climatic conditions are important determining factors for sorghum and millet yields (Graef & Haigis, 2001). The climatic research unit time-series datasets (CRU TS3.0) for long-term average annual rainfall (mm) (1901-2006) were obtained from the University of East Anglia (CRUTS, 2006). Rural population density was used as a proxy for labor availability

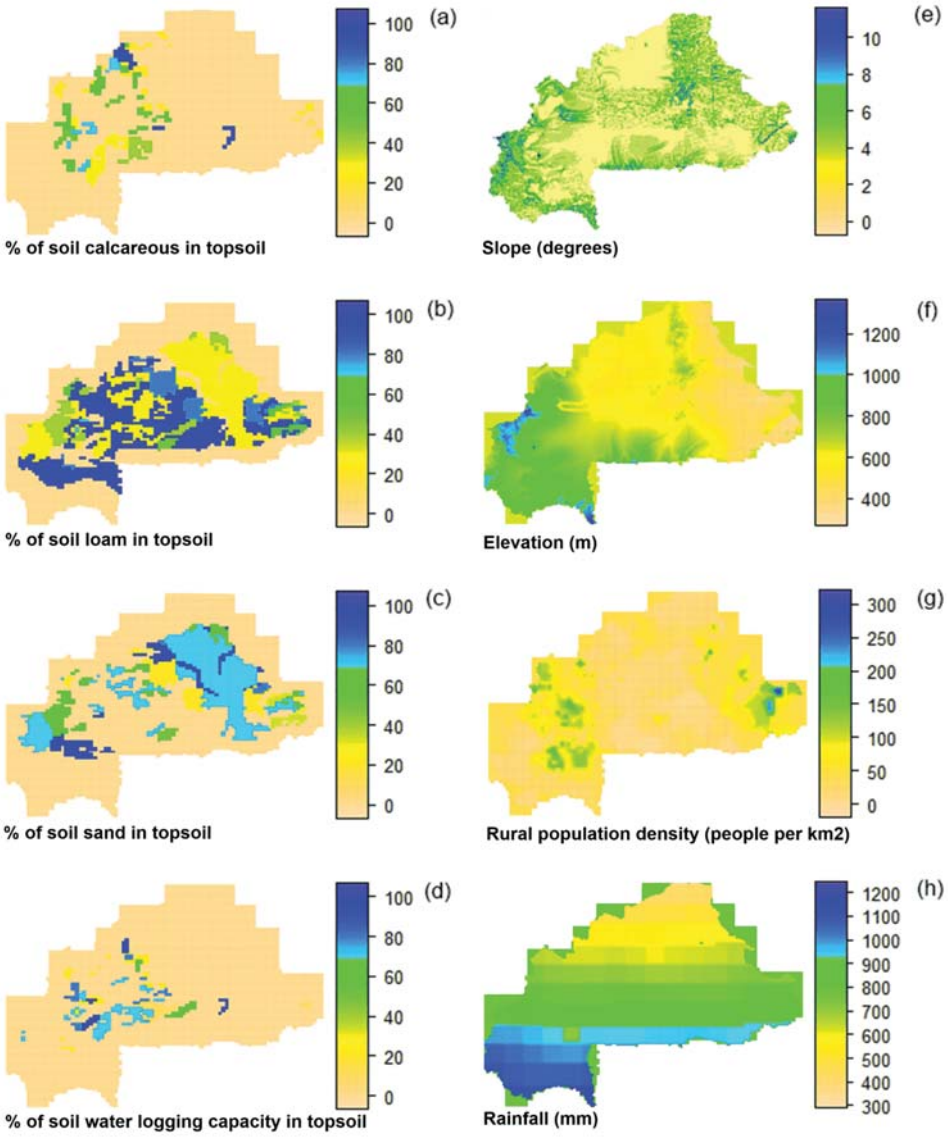


Figure 3.2 Explanatory variables: SoilCalc - percentage of area with carbonate in the topsoil (a); SoilLoam - percentage of area with loam in the topsoil (b); SoilSand - percentage of area with sand in the topsoil (c); SoilWL - percentage of area with soil-water holding capacity in the topsoil (d); Slope (degrees) (e); Elevation (m) (f); RURPD - rural population density (number of people per km^2) (g); Rainfall (mm) (h).

in Burkinabé terroirs. A raster map of rural population density (i.e. number of rural inhabitants per square kilometer) based on the 2005 census was obtained from the HarvestChoice database (HarvestChoice, 2012).

3.3.2 Methods

The modeling process comprised three distinct steps: (i) data pre-processing, (ii) global and local spatial regression to model spatial relationships for estimating crop yield, and (iii) spatial prediction and validation of crop yields.

3.3.2.1 Data pre-processing

AGRISTAT measures the area (ha) and geographical location of a household parcel being cropped in a representative terroir and registers the weight of crop (kg) harvested from the parcel. The total cultivated area of a crop on a household parcel and the weight of crop were used to calculate the observed crop yield (kg ha^{-1}) of a representative terroir for a district. AGRISTAT processes this survey data so that each database record contains the geographical location of the specific parcel being observed for its crop yield and the name of the associated representative terroir. For the georeferenced parcels of 351 terroirs that represent the entire study area, the observed crop yields for sorghum, millet, and cotton were used as the response variable.

The observed crop yields were linked to the two major AEZs in the study area: the tropic-warm / semiarid zone and the tropic-warm / subhumid zone (Figure 3.1). Observations for the third AEZ, tropic-warm / arid, contained less than 1% of the total number of observations and was therefore not considered in our study. AEZs are typically defined in terms of relatively homogeneous regions with a specific range of potentials and constraints for land use. The agricultural field survey data, however, are typically obtained for the administrative districts. By relating the agricultural survey data to AEZs, crop responses were linked to the areas of homogeneous cropping conditions, thus dividing up the heterogeneous landscape of the study area.

We first applied a Principal Component Analysis (PCA) to the 18 NDVI images for the 2009 crop growing season to reduce the temporal instability of regression coefficients during the growing season. PCA tends to obtain the best observation available for each pixel over the duration of crop growing period (de Beurs & Henebry, 2010). It captures the maximum variance within an image time-series to represent the multi-temporal content of the dataset (Hirosawan *et al.*, 1996), which allowed aligning the image data with maximum crop-canopy cover during the crop growing period.

Using PCA, we transformed the correlated spectral and temporal image bands into principal components of the NDVI data, referred to from here onwards as NDVI PCs. The principal components were used

Table 3.1 Explanatory variables used to model sorghum, millet, and cotton yields in Burkina Faso.

Variables	Description
NDVI.PC1, NDVI.PC2, NDVI.PC3	first three PCs obtained from 18 SPOT NDVI images covering the crop growing period.
SoilCalc, % of area[a]	area with carbonate in surface soil
SoilLoam, % of area	area with loam surface soil
SoilSand, % of area	area with sand surface soil
SoilWL, % of area	area with water holding capacity in the topsoil
Elevation, m	topographic elevation
Slope, degree	topographic slope
Rainfall, mm	long-term average annual rainfall
RURPD, persons/km^2	rural population density

[a]Percentage of area with soil component in the topsoil (plow layer or upper 20 cm, whichever was shallower).

to quantify photosynthetic activity of vegetation over the crop growing season. Observed crop yields per representative terroir were related to the co-located cell values of NDVI PCs and the explanatory variables described in Table 3.1.

3.3.2.2 Spatial modeling of crop yields

We carried out two different analyses to model crop yields, each taking into account the spatial processes represented by the variables in Table 3.1: (i) spatial regression producing a single global model, and (ii) geographically weighted regression to produce local models.

Global regression For each AEZ in Burkina Faso, a conditional autoregressive model (CAR) was specified and calibrated as a global spatial model of crop yield (Cliff & Ord, 1981; Wall, 2004). It explicitly includes recursive, higher-order neighborhood effects, or global spatial autocorrelation. Its properties are based on the ordinary linear regression model, expressed as

$$Y(s) = \beta_\circ + \sum_k \beta_k X_k(s) + \epsilon(s), \tag{3.1}$$

where $Y(s)$ is the estimated crop yield at a location s, $\beta_\circ$ is an intercept, β_k are coefficients for the kth explanatory variables $X_k(s)$ (i.e. values of NDVI, rainfall, soil, topography, and labor at location s), and $\epsilon(s)$ denotes the random error for location s. The expected value of $Y(s)$ is $E(Y(s)) = \mu$, where μ is the vector of means $\mu(s) = X\beta$. CAR models, however, assume that $Y(s)$ depends both on a set of explanatory variables at a location s and the response

variable at all other locations. This changes the $E(Y(s))$ to the conditional probability density of the response variable at location s given all other variables(Wall, 2004)

$$E(Y(s)|Y(-s)) \sim \left(\mu(s) + \sum_{t=1}^{n} C_{st}(Y(t) - \mu_t), \sigma_s^2\right), \tag{3.2}$$

where $Y(-s) = \{Y(t) : t \neq s\}$, σ_s^2 is the conditional variance, and C_{st} are weights that determine the relative influence of location s on location t. In this study, the weights were chosen by a proximity criterion, which was specified using a neighborhood weight matrix W, which is subject to a relative graph-based neighborhood. The graph was composed by connecting the geographical locations of the observed parcels in the 351 representative terroirs of Burkina Faso. Two locations s and t were connected by an edge (i.e. $W_{st} = 1$) if their districts shared common borders, and $W_{st} = 0$ otherwise. For fixed σ_s^2, we form matrices $C = (C_{st})$ and $T = diag\{\sigma_s^2\}$ (Cliff & Ord (1981), p. 170), so

$$Y \sim N(\mu, (I_n - C)^{-1}T), \tag{3.3}$$

Assuming a fixed σ_s^2, Eq 3.2 represents a global model of crop yield that is spatially stationary, i.e. the relationship between crop yield and the factors affecting it does not vary in the study area. At a regional scale in West Africa, however, it may not always hold true and we need to model local dependence, or the non-stationarity, of the spatial relationship.

Local regression - Geographically Weighted Regression -
Geographically Weighted Regression (GWR) deals with the non-stationarity of the spatial relationship. Separate models were established for each sampled location and local coefficients were estimated (Fotheringham et al., 2002). This changes the model in Eq 3.1 to

$$Y(s) = \beta_\circ(s) + \sum_k \beta_k(s)X_k(s) + \epsilon(s), \tag{3.4}$$

where $\beta_\circ(s)$ and $\beta_k(s)$ represent the estimated intercepts and coefficients at a location s, and $X_k(s)$ are explanatory variables. The model estimates local coefficients from

$$\hat{\beta}(s) = (X^T W(s) X)^{-1} X^T W(s) Y(s), \tag{3.5}$$

where $W(s)$ is a $n \times n$ diagonal matrix of spatial weights specified by a spatial kernel function. The kernel centers on a sampled

location s and weights the neighboring sampled locations t subject to a distance-decay. In our study, Gaussian weighting was used as the kernel function to specify the spatially weighted matrix:

$$W_{st} = \exp[-0.5(d_{st}/b)^2], \tag{3.6}$$

where d_{st} is the distance between the sth and tth sampled locations and b is the kernel bandwidth that measures the distance-decay in the kernel function. Minimizing of the Akaike's Information Criterion (AIC) (Hurvich *et al.*, 1998) was used for GWR as a criterion to determine the effective bandwidth for the kernel as well as providing a trade-off between goodness-of-fit and degrees of freedom.

Multicollinearity among the explanatory variables was analyzed before applying regression models. For this, we first applied the global regression using all explanatory variables, and crop yields as response variables. Only those variables that contributed to output at the minimum probability value ($p = 0.1$) were retained in the analysis. Next, variables with a variance inflation factor of 5 or more were removed from the analysis (Kutner (2005), p. 408). Finally, the condition indices were computed for the matrix of explanatory variables (Belsley *et al.*, 1980). The process of excluding variables was continued until all condition indices were below 30 and all variables contributed to the output.

Model comparison The properties of global and local crop-yield models were compared. Moran's I was used to analyze spatial autocorrelation in the model's residuals. Use of a regression model for spatial prediction is only appropriate when the residual errors are approximately independent. Monte Carlo testing was done to check the significance of the spatial non-stationarity of each GWR parameter estimate (Fotheringham *et al.* (2002), p.93). The null hypothesis of the test was that the surface representing an estimate of a local parameter arises from a stationary process and that the observed variation in the local estimate results solely from sampling variation. Rejection of the null hypothesis would indicate an improvement of GWR over CAR.

In this study, the R package spdep (Bivand, 2012) was used for the CAR analysis. It produced an adopted likelihood ratio for the best model, based on the Nagelkerke pseudo R^2 (Nagelkerke (1991), p.691). For the GWR analysis we used the R package spgwr (Bivand & Yu, 2012). It produced a global R^2, the calculation of which was based on the local R^2 statistic (Fotheringham *et al.* (2002), p.215). Consequently, the approximated R^2 in CAR and GWR may not alone be sufficient to identify the best model. We therefore interpreted the R_a^2 jointly with the AIC statistic (Fotheringham *et al.* (2002), p.61) and the ANOVA F-test. The AIC takes into account

the different number of degree of freedom in different models, so that their relative performance can be compared more precisely. ANOVA tests the improvement of GWR over a global model.

Crop yield spatial prediction and validation Geographically Weighted Regression (GWR) was applied as a multivariate interpolation technique to predict crop yields at unsampled locations. For this, GWR uses the estimated local relationship and cell values of explanatory grids as

$$\hat{Y} = x_{\circ}(s)^T \hat{\beta}(s), \tag{3.7}$$

where $x_{\circ}(s)$ is a vector of cell values of explanatory grids at grid cell. The crop yields were interpolated for a grid cell of 1 km^2.

The best models were validated using 10-fold cross-validation. The AGRISTAT data were randomly divided into 10 groups of approximately equal size. We fitted the CAR and GWR models using 90% of the data and generated predicted values of the remaining 10% (10-fold cross-validation). This approach enabled us to generate predicted values for the test data, independent of training data, for all crop yield observations. We then calculated the sum of squared errors (SSE) of prediction (mean over 10-fold) to evaluate the prediction accuracy of the models.

3.4 Results

3.4.1 Data pre-processing

Figure 3.3 shows the spatial distribution of crop yields (kg ha^{-1}) of sorghum, millet and cotton aggregated from the AGRISTAT data. These results represent 153, 229, and 57 terroirs cultivated for sorghum, millet, and cotton, respectively, in the semiarid zone, and 57, 54, and 38 terroirs under those crops in the subhumid zone. There is a clear North-South trend in crop yields across the country. High crop yields are observed in the subhumid zone of Burkina Faso, with means equal to 1122, 856 and 1276 (kg ha^{-1}) of sorghum, millet and cotton, respectively, as compared to yields of 901, 763 and 1160 (kg ha^{-1}) for these crops in the semiarid zone. This indicates that cropping conditions in the southern, subhumid zone are more favorable than the semiarid and arid zones of the North and the Northwest of the country.

Using the multi-temporal content of NDVI image series, the first three NDVI PCs, shown in Figure 3.4, accounted for 94.32% of the variability in the NDVI values. They explained the maximum reflectance variance during the crop growing period. The values of the variance inflation factor are significantly lower (< 5) for the explanatory variables (Table 3.2). This implies a negligible impact of multicollinearity in the CAR and GWR models on the precision of estimation.

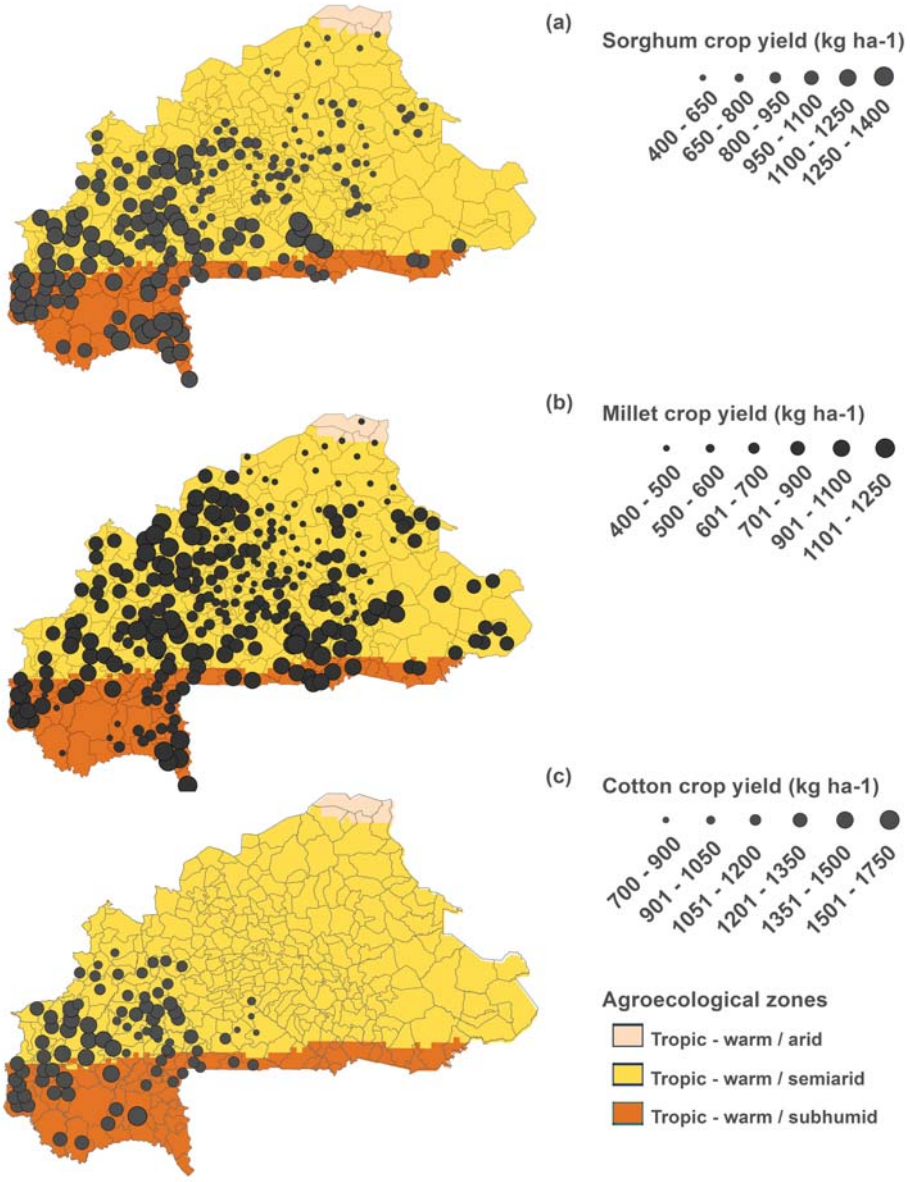


Figure 3.3 Spatial distribution of crop yield (kg ha^{-1}) observations in the semiarid and subhumid agroecological zones: sorghum (a), millet (b), and cotton (c)

Table 3.2 Parameter estimates from conditional autoregressive (CAR) models of sorghum, millet, and cotton in the semiarid and subhumid agroecological zones (AEZs) of Burkina Faso.

	Tropic / Semiarid						Tropic / Subhumid					
Crop	Variables	Est.[a]	SE[b]	P-value	VIF[c]	λ[d]	Parameters	Est.	SE	P-value	VIF	λ
Sorghum	Intercept	304.7	87.2	0.001	-	0.3	Intercept	-280.5	323	0.1 n/s[e]	-	0.3
	NDVI.PC1	249.9	37	<.00001	1.72		Rainfall	1.35	0.32	<.00001	1.3	
	Elevation	0.26	0.14	0.01	1.74		Slope	47.9	41	0.1 n/s	1.1	
	SoilLoam	0.02	0.29	0.1 n/s	1.43		SoilCalc	2.66	1.18	0.01	1.4	
	SoilSand	-0.66	0.37	0.05	1.46							
	RURPD	0.65	0.62	0.1 n/s	1.31							
Millet	Intercept	298.6	105.1	0.001	-	0.3	Intercept	614	288	0.01	-	0.4
	Rainfall	0.6	0.14	<.00001	1.05		NDVI.PC1	132	124	0.01	1.01	
	SoilCalc	0.7	0.49	0.01	1.03		Slope	36.5	69.7	0.01	1.01	
	SoilWL	1.72	0.96	0.01	1.02							
	RURPD	0.27	0.52	0.01	1.04							
Cotton	Intercept	660.9	221.7	0.001	-	0.3	Intercept	526	354	0.01	-	0.2
	NDVI.PC1	120.3	89.58	0.01	1.1		NDVI.PC1	109	120	0.01	1.01	
	Elevation	0.31	0.25	0.01	1.1		Elevation	0.58	0.3	0.05	1.01	

[a]Estimated coefficient
[b]Standard error
[c]Variance inflation factor
[d]Autoregressive coefficient
[e]Not significant at the 0.05 level

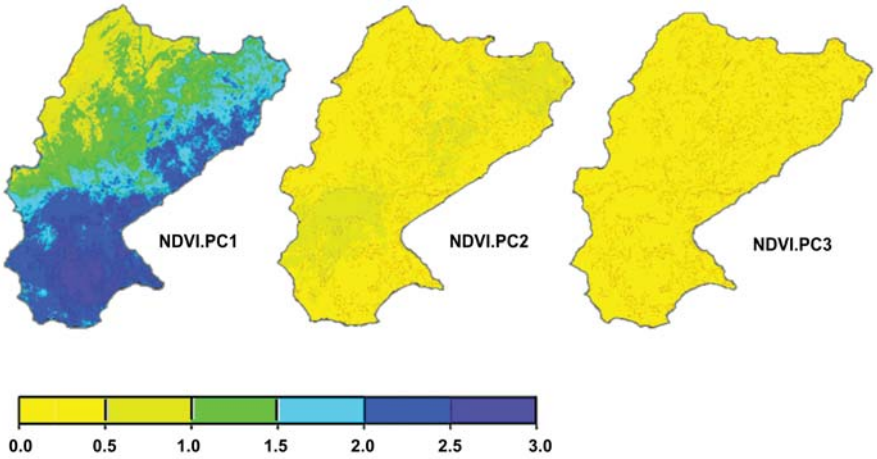


Figure 3.4 First three principal components (PCs) of 18 NDVI (10-days) composites

3.4.2 Spatial modeling of crop yields

Global regression Table 3.2 shows coefficients of the explanatory variables of the CAR models. Soil nutrients and water availability are generally major factors that limit crop yields in the country. Sorghum is generally grown on moist (occasionally wet) lands, while millet is cultivated on all remaining, dry land types. The percentage of area with carbonate, loam, sand, and water holding capacity in the topsoil proved to be significant in explaining the spatial variability of crop yields across the study area.

In the semiarid zone, the first principal component, NDVI.PC1, is positively related to sorghum and cotton yields. As expected, the percentage of loam in the topsoil has a significant positive relation with sorghum yield. Conversely, sorghum yield significantly decreases as the percentage of sand in the topsoil increases, the result of its decreasing water holding capacity. Higher crop yields can be observed for both upland sorghum and cotton. Millet yields are primarily determined by amount of rainfall and the percentage of carbonate in the topsoil (i.e. stony calcareous soils).

In the subhumid zone, NDVI.PC1 primarily shows a significant relation with millet and cotton yields, with elevation and slope variables distinguishing between yields of the two crops: upland cotton yields are high, whereas high sorghum and millet yields occur on parcels with relatively steeper slopes. Rainfall is the key determining factor for sorghum yields. In addition to good water retention and drainage capacity, soils with a higher percentage of carbonate in the topsoil provide the best conditions for sorghum cultivation on steep (30-40%) sloping land.

Tables 3.2 and 4 show the performance of CAR in dealing with spatial autocorrelation. Table 3.2 shows that the λ coefficient is positive and highly significant, indicating a strong spatial autocorrelation between yield counts at sample points located within each other's spatial neighborhood. Applying the relative neighborhood graph to reflect the spatial neighborhood of crop yield observations improves the CAR models, as reflected in an increase in R_a^2 and a decrease in AIC value (Table 3.4). The results of the CAR models indicate that inclusion of the spatial autoregressive term significantly improves estimation of coefficients for the models. A high Moran's I value of the CAR residuals indicates that the models may suffer from spatial non-stationarity.

Local regression As local spatial variation can only be partially explained, the CAR models are apparently not fully able to incorporate spatial non-stationarity. This is supported by Monte Carlo testing, which indicated a significant spatial variation ($p = 0.05$ to $p < 0.00001$) in the local parameter estimates (Table 3.3). As a consequence, GWR models became the focus of attention (Table 3.4). Rural population density and percentage of loam in the topsoil that were not significant in the CAR models were highly significant in the GWR model (Tables 3.2 and 3). For instance, rural population density was not significant in the semiarid zone, whereas Monte Carlo testing indicated spatial non-stationarity at the $p = 0.001$ significance level (Table 3.3). This can also be observed from the inter-quartile range of the local parameter estimate. The range (0.12 - 3.84) is larger than ± 1 standard deviations of the equivalent CAR parameters estimate (1.24), being $2 \times SE$ in Table 3.2. Similarly, the non-significant soil and topography variables in the CAR model showed significant spatial variability ($p = 0.001$) in the GWR model, as is specifically the case for sorghum in the semiarid and subhumid zones (Table 3.2).

3.4.3 Model comparison

Comparisons of the models are shown in Table 3.4. Compared with CAR models, the explanatory power of the GWR models increased for all crops in both AEZs, with R_a^2 values equal to 0.85, 0.70 and 0.78 for sorghum, millet and cotton crops, respectively, in the tropic semiarid zone, and 0.76, 0.67, and 0.65, respectively, in the tropic subhumid zone. The ANOVA F-test suggests that GWR gave a significant improvement ($p = 0.001$)over the CAR models for the AGRISTAT crop yield observations. Improvement as presented by R_a^2 should be interpreted jointly with the AIC value. It decreased for the GWR model of cotton yield in the semiarid zone, whereas for the models of millet yield there is a negligible difference in AIC values between CAR and GWR (i.e. AIC ≤ 3). For sorghum, however, the AIC value is lower for the CAR model. In the subhumid zone, the decrease in AIC values of GWR models for sorghum, millet and cotton yield as compared to those of the CAR models is 15,

Table 3.3 Parameter estimates from geographical weighted regression (GWR) models of sorghum, millet, and cotton in the semiarid and subhumid agroecological zones (AEZs) of Burkina Faso.

AEZ	Crop	Variables	Min[a]	I[st] Q[b]	Med[c]	3[rd] Q[d]	Max[e]	MC[f]	Adapt[g]
Tropic-warm / semiarid	Sorghum	Intercept	-901	205	473	792	1480	0.001	0.05
		NDVI.PC1	-512	19.3	134	251	507	0.001	-
		Elevation	-0.3	-0.05	0.24	0.8	2.96	0.001	-
		SoilLoam	-5.77	-1.35	-0.44	0.45	4	0.001	-
		SoilSand	-5.46	-1.8	-0.7	0.32	3.25	0.001	-
		RURPD	-1.89	0.12	1.44	3.84	10.4	0.001	-
	Millet	Intercept	-2170	-265	166	866	2150	0.001	0.03
		Rainfall	-1.81	-0.1	0.68	1.26	3.76	0.001	-
		SoilCalc	-3.33	-0.1	0.66	2.5	10.2	0.05	-
		SoilWL	-12.4	-2.1	1.04	3.94	19.3	0.05	-
		RURPD	-11.9	-1.5	0.85	2.65	9.05	0.001	-
	Cotton	Intercept	226	646	1050	1700	2600	0.001	0.1
		NDVI.PC1	-154	114	185	288	477	0.1	-
		Elevation	-2.3	-1.45	-0.43	-0.05	0.79	0.001	-
Tropic-warm / subhumid	Sorghum	Intercept	-2730	-1160	308	914	1900	0.001	0.1
		Rainfall	-0.73	0.26	0.94	2.18	3.7	0.001	-
		Slope	-60.8	30.6	55.5	139	285	0.05	-
		SoilCalc	-4.1	1.15	2	3.81	8.1	.1	-
	Millet	Intercept	-458	396	1050	1480	4670	0.05	0.08
		NDVI.PC1	-1620	-289	-97.3	200	538	0.05	-
		Slope	-141	49.7	97.5	279	522	0.05	-
	Cotton	Intercept	-665	20.2	700	1170	1200	0.1	0.15
		NDVI.PC1	-46	47.6	112	214	496	0.01	-
		Elevation	-1.3	0.01	0.39	0.8	1.44	0.01	-

[a]Minimum
[b]First Quartile
[c]Median
[d]Third Quartile
[e]Maximum
[f]Monte Carlo significance test for spatial non-stationarity of parameters
[g]Adaptive quartile proportion of observations to include in weighting scheme

Table 3.4 Comparison of conditional autoregressive (CAR) and geographical weighted regression (GWR) models in the semiarid and subhumid agroecological zones (AEZs) of Burkina Faso.

AEZ	Crop	Model	n	$R_a^{2\,a}$	AIC^b	Moran's I^c	σ^d	CV^e	ANOVA^f
Tropic-warm / semiarid	Sorghum	CAR	153	0.76	1916	0.04	101	1396456	-
		GWR	-	0.85	1933	0.12	107	904675	0.001
	Millet	CAR	229	0.56	2955	-0.27	135	4127845	-
		GWR	-	0.70	2957	0.22	137	3044658	0.001
	Cotton	CAR	56	0.50	740	-0.15	159	1186721	-
		GWR	-	0.78	728	0.19	129	679660	0.001
Tropic-warm / subhumid	Sorghum	CAR	57	0.54	698	-0.43	93	579190	-
		GWR	-	0.76	683	0.16	71	326428	0.001
	Millet	CAR	54	0.32	715	-0.55	147	1434193	-
		GWR	-	0.67	711	-0.01	127	764121	0.01
	Cotton	CAR	38	0.30	489	-0.23	143	902721	-
		GWR	-	0.65	483	-0.28	108	610207	0.001

[a] Adjusted R^2
[b] Akaike's Information Criterion
[c] Spatial autocorrelation of residuals
[d] Standard error of residuals.
[e] Cross-validation sum of squared errors of prediction (mean over 10-fold)
[f] ANOVA F-test (P-value)

4, and 6, respectively. This indicates that, using the same variables, GWR improved upon CAR after accounting for differences in degrees of freedom. This is further supported by a decrease in the Moran's I values of GWR residuals. The residuals of GWR models showed little or no spatial autocorrelation, suggesting that the GWR coefficients do not suffer from local dependence.

3.4.4 Crop yields spatial prediction and validation

Maps of predicted sorghum, millet and cotton yields in the tropic semi-arid and subhumid zones are presented in Figures 3.5.

Local R^2 values for all crops in both AEZs were mapped to mark the areas where GWR resulted in a non-significant local relationship (Figure 3.6). Consequently, the predicted crop yields for those areas have lower levels of accuracy and should be interpreted with care. In the semiarid zone, the local R^2 values ranged from 0.36 to 0.76 for the sorghum yield model, from 0.18 to 0.74 for the millet yield model, and from 0.1 to 0.88 for the cotton yield model. In the subhumid zone, they ranged from 0.1 to 0.86 for the sorghum model, from 0.2 to 0.79 for the millet model and from 0.1 to 0.77 for the cotton model. Low R^2 values may point to insufficient samples in a particular area or missing significant explanatory variables in local models. This may affect the quality of crop yield estimates in such areas.

The results of the 10-fold cross-validation are provided in Table 3.4. Comparison of the SSE values for the model predictions based on this cross-validation show that GWR provides a more accurate model for predicting crop yields in the study area. This is further supported by visual interpretation of maps of predicted yields and a comparison with the AGRISTAT observed crop yields in Figure 3.3. Clearly, local patterns of AGRISTAT sampled crop yields are reflected in the GWR maps.

3.5 Discussion

This study deals with nation-wide assessment of crop-yield relationship. Because of the high spatial variability of agroecological conditions in Burkina Faso, particularly in its semiarid zone, we designed a multilevel spatial stratification procedure based on criteria related to crop acreage and/or crop yield. As a concrete step, we linked AGRISTAT data collected from representative terroirs to the semiarid and subhumid agroecological zones (AEZs). This stratification together with the use of GWR resulted in a significant improvement over commonly applied CAR models to identify local patterns of crop yield.

Our study confirms that topography affects crop performance in Burkina Faso, under both low and high rainfall conditions (West *et al.*, 2008). In the semiarid zone, central and northern Sahelian farmers tend to abandon marginal uplands and elevated borders of lowland fields, because these land types perform poorly in the low rainfall conditions.

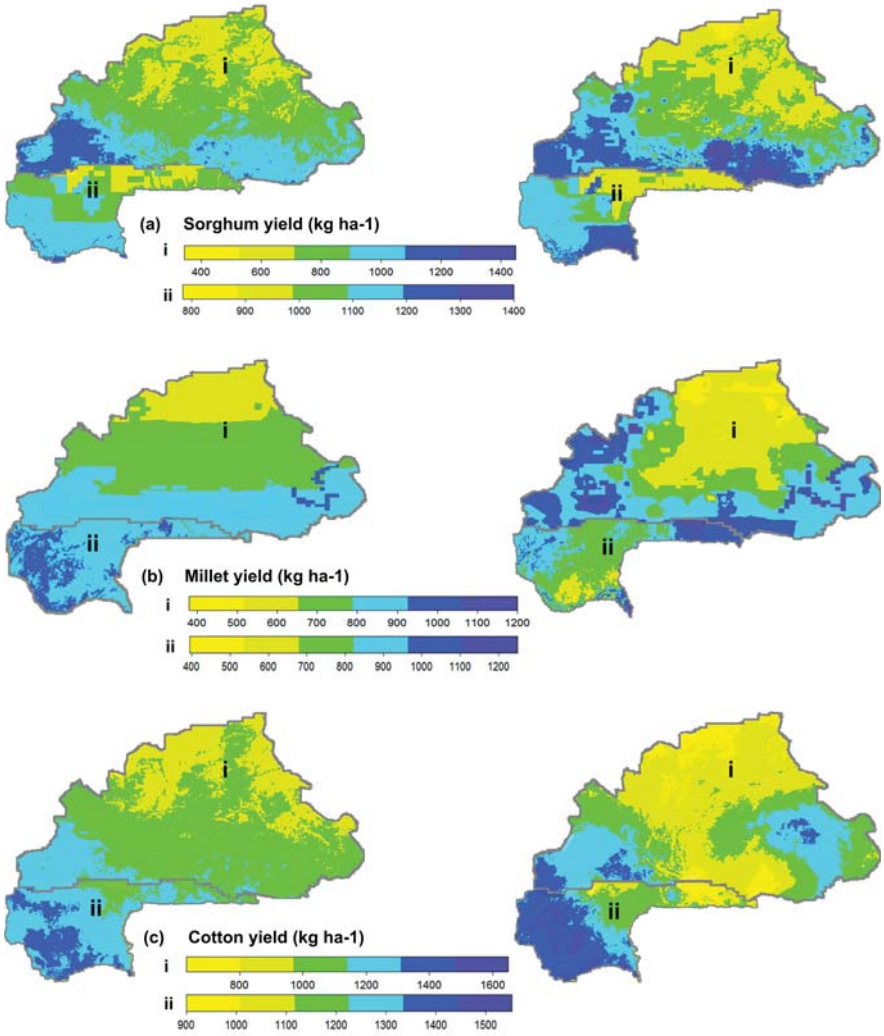


Figure 3.5 Crop yield (kg ha^{-1}) maps from the conditional autoregressive (CAR) model (left) and the geographical weighted regression (GWR) model (right) of sorghum (a), millet (b), and cotton (c); Semiarid zone (i) and Subhumid zone (ii)

Figure 3.6 Local R^2 values from the geographical weighted regression (GWR) model of crop yield: sorghum (a), millet (b), and cotton (c)

In these areas, we observed higher sorghum yields as loam content in soils increased, whereas high millet yields occurred on calcareous soils, for which carbonate concentrations and water holding capacities are higher. In the eastern subhumid zone, lowland fields with steeper slopes are often flooded due to run-off to and from neighboring lands. In these lowlands, we observed increasing yields of sorghum and millet with increase in slope, whereas high cotton yields were observed on wet uplands in the western subhumid zone.

The study also shows the effect of weather conditions. Spatial variability of rainfall is high in the semiarid zone. Consequently, agroclimatic conditions vary more rapidly in this zone as compared to the subhumid zone. To cope with this high variability, farming practices vary considerably in the semiarid zone. Frequent application of semi-permanent planting basins such as *zai* by farmers in the semiarid zone and residue management to reduce losses of soil and water are, however, labor intensive components in subsistence farming (Lal, 1991).

Data on labor availability were not available on a national scale. With the aim in mind of being able to extend the models used for Burkina Faso to other West African countries, we chose to use rural population density as a proxy variable for labor availability, the assumption being that labor availability is positively related to the density of the rural population. Using this proxy, the study shows that in the semiarid zone, rural population density is a significant factor in explaining the variability of millet and sorghum yields.

Two properties of spatial data, spatial non-stationarity (Fotheringham *et al.*, 2002) and multicollinearity (Wheeler & Tiefelsdorf, 2005) require careful use of diagnostic parameters before applying GWR. Monte Carlo significance testing for spatial non-stationarity confirmed that the spatial covariates varied in the study area. We realized that the selection of an appropriate GWR kernel bandwidth is vital for the calibration of the spatially varying relationship between crop yields and factors affecting crop yields. Minimizing values of both the AIC criterion and the spatial autocorrelation of GWR residuals showed that the adaptive bandwidth in GWR is effective and that the local relationship is reliable. It was found that Moran's I statistics on the residuals from GWR were significantly reduced. A high degree of multicollinearity tends to inflate the variance of predicted values in a regression analysis, particularly in GWR. By applying a PCA to the NDVI image series we were able to eliminate local multicollinearity between consecutive image bands. Analysis of variance of inflation factors confirmed that the correlations between the principal components of NDVI and other spatial covariates were not significant.

Data are irregularly distributed in the study area. For instance, as the northern and eastern parts of the semiarid zone lack appropriate growing conditions for cotton, cotton samples are unevenly distributed. As a result, predicted cotton yields in these parts disagreed with the AGRISTAT data. In particular, in the eastern part of the zone, comparatively less significant local models of cotton yields (i.e. lower values of local R^2)) were observed and, consequently, cotton yield predictions were

poor in this part. Interestingly, the CAR model performed comparatively better than the GWR model in such poorly sampled areas.

Also, crop-specific high-resolution land cover and land-use maps were missing for the study area. Those would be useful for accurately delineating areas where crops are actually grown. This lack may result in uncertainty when estimating local crop yield relations in parcels on which mixed cropping is practiced. Additional explanatory variables may further be obtained from cultural or socioeconomic characteristics, investment capacity or policy factors. For Sahelian terroirs in the North, lower values of local R^2 already show less significant local models for millet yields (see Figure 3.6b). This may in turn result in poor millet-yield prediction in the area.

The NDVI time series data are often used for regional crop monitoring purposes since this index can capture variations of canopy cover, especially in semiarid areas (Budde *et al.*, 2004). However, the nonlinear behavior of this and other vegetation indices has been widely demonstrated in the literature (Dorigo *et al.*, 2007). NDVI might therefore be more sensitive to differences in background soil contamination that to biophysical parameters such as canopy cover or the amount of chlorophyll present in the canopy. To cope with this, we included in our analysis inputs that account for color, texture, or moisture related differences in soil reflectance (Huete & Tucker., 1991). The maps of soil types (e.g. clay, sand, calcareous) and topography have the same spatial resolution as SPOT NDVI (i.e. 1 km). Future study might address the sensitivity to the variation of brightness contrast between vegetation and soil background, or the use of alternative vegetation indices that are less influenced by confounding factors.

An important motivation for this study was to model crop yields for a BEFM application over large areas in West Africa. By using the spatially varying parameters within a GWR, we were able to reliably predict crop yields for a regular grid with a cell size of 1 km^2. In this way, several spatial and temporal instabilities could be overcome in the quantitative analysis of crop yields at regional scales. The study demonstrates the clear advantage of GWR for obtaining a wall-to-wall spatial coverage for generating maps of relatively large areas. Further extension of this approach to the whole of West Africa remains challenging, however.

3.6 Conclusions

Crop yield modeling at a regional scale in West Africa is challenging because of (i) complex nonlinear and locally variable relationships between crop yields and factors affecting crop yields, and (ii) lack of sufficient data. Remote sensing provides timely and synoptic coverage, while national statistical databases, such as AGRISTAT in Burkina Faso, provide field observations of crop yields for selected georeferenced locations. We tested the use of global CAR and local GWR models to account for

spatial dependency and local variability of millet, sorghum, and cotton yields in the semiarid and subhumid zones of Burkina Faso.

The results showed that modeling crop yields over the highly heterogeneous landscape of Burkina Faso require d incorporating the spatial variability of rainfall, topography, labor availability, and selected soil properties, including carbonate, loam and sand content, and water holding capacity. By applying spatial regression models, we observed that the SPOT NDVI, elevation, slope, and rainfall significantly affected crop yield spatial variability in both the semiarid and subhumid zones. In the semiarid zone, soil properties and labor availability also influenced sorghum and millet yields. For terroirs with higher values of these variables, we observed a significant increase in sorghum yield, to a maximum of 350 kg ha^{-1}, and an increase in millet yield, to a maximum of 275 kg ha^{-1}. In the subhumid zone, the maximum increase in cotton yield was 210 kg ha^{-1}, on the Southwest uplands; the maximum increase in millet yield was 275 kg ha^{-1}, on downstream terroirs with steeper slopes; and the maximum increase in sorghum yield was equal to 200 kg ha^{-1}, on areas having a higher frequency of rainfall and carbonate content in the soil.

More spatial variability of crop yields was observed with local models, showing that the effect of explanatory variables was highly localized in the study area. Monte Carlo analysis further confirmed the spatial variability of the observed relationships. Thus, accounting for spatial non-stationarity was essential for improving the quality of crop yield predictions in Burkina Faso.

By incorporating the extent of the spatial non-stationarity into the relations, GWR gave higher R^2 values than the CAR models. Thus, the improvement of the GWR model over the CAR model suggested that the spatial covariates varied spatially in their effects across Burkinabé terroirs. However, we observed that CAR models performed better than GWR models in areas with less than adequate crop yield observations. All the same, GWR models for areas with a low local R^2value could be improved further by including land cover maps showing cropping areas on grid cells.

Geographically weighted regression (GWR) models can be calibrated to generate timely and accurate crop yield maps in Burkina Faso. Such maps can subsequently be used to build spatial decision-support systems, to map net primary production, or to parameterize bio-economic farm models.

4

Using Geographical Weighted Regression Kriging for crop yield mapping in West-Africa

[1]This chapter is revised and resubmitted in the *International Journal of Geographic Information Science* for publication as "Using geographically weighted regression kriging for crop yield mapping in West-Africa"

4. Using Geographical Weighted Regression Kriging for crop yield mapping in West-Africa

Abstract

Predicting crop yields on large regional areas increasingly allows running wall-to-wall applications of agricultural decision-making and food security. Empirical approaches apply regression models, relating environmental and management factors to observed crop yields, followed by using the estimated relations to predict yields at unvisited locations. Observed crop yields are, however, spatially dependent and, assumptions in those approaches may not be always satisfied in modelling crop yield on a regional scale. This paper presents Geographically Weighted Regression Kriging (GWRK) as a novel approach that combines geographically weighted regression (GWR) with kriging. GWRK is applied on sorghum crop yield in Burkina Faso. The regression is calibrated using the crop yield data derived from subnational surveys. Crop yields are related to the crop biophysical conditions derived from the SPOT 4 NDVI data, precipitation, and topographic grids and, to local crop management factors. Ordinary kriging predicted the GWR residuals at unvisited grid cells. GWRK improved estimates of uncertainty for the sorghum yield predictions, reducing the lower uncertainty range value with 20%, and the upper uncertainty range value with 40%. Moreover, GWRK reduced the prediction error variance as compared to ordinary kriging (20.4 versus 36.8) and to regression kriging (20.4 versus 34.1). Results indicate that climate and topography have a major impact on the sorghum yield in Burkina Faso. The financial ability of farmers influences the crop management and thus the sorghum crop yield. The paper concludes that GWRK effectively utilized information present in the external covariate datasets, improving the accuracy of sorghum yield predictions.

4.1 Introduction

Sustainable farming strategies aim at maintaining yield levels while protecting the environment. They benefit from technologies for increased crop production, thus achieving food security and economic profitability. To accomplish this, bio-economic farm models (BEFMs) were developed (van Ittersum *et al.*, 2008). BEFMs can optimize agricultural farm responses by means of simultaneously assessing the plethora of factors of sustainable farming and the resulting trade-offs among the bio-physical, economic, and socioeconomic goals. Commonly BEFMs are used as a site-specific application on which all input datasets have been prepared beforehand for a particular farm or village. This contrasts to a wall-to-wall application, for which no specific site has yet been identified, and the model is applied to any site in sub-regions of a large region. Providing crop yield estimates for a wall-to-wall application of BEFM is challenging, as it requires timely and accurate mapping of crop yields over large-areas, considering local factors of crop yields (Vossen, 1999). Large-area crop models and GIS databases are indispensable to map crop yields at national and regional scales. For this, often empirical models are used with low input data requirements, thus avoiding site specificity (de Wit *et al.*, 2008). GIS databases developed from national surveys and other global datasets allow deriving external covariate datasets that determine location-specific factors of crop yields. The accuracy of the empirical relations and of the covariate datasets leverages the accuracy of crop yield mapping and, increases the efficiency of wall-to-wall applications. Despite this, analyzing, quantifying, and reducing the uncertainties have so far received little attention in the course of developing and reporting empirical models of crop yield prediction.

Uncertainty in prediction models is hardly avoidable because the prediction of an unsampled location is always associated with error. Specifically in crop yield mapping, a number of environmental and management factors makes yields highly variable in space. Failure to incorporate such high spatial variability contributes to the uncertainty ranges of crop yield models. In recent studies, crop yield assessments observed inaccuracies, when crop models were applied at the national and regional scales (Reidsma *et al.*, 2007b; Challinor *et al.*, 2009). These studies further observed that inaccuracies primarily resulted from the inability of the models to utilize information present in environmental and management datasets. Other studies observed that such inaccuracies resulted not only from the quality of the models itself, but also from quality of acquisition methods of the input covariate datasets (Faivre *et al.*, 2004). Moreover, ignoring completely the covariates determining spatial variability of crop yields has been considered a major reason of unsatisfactory performance of crop models at regional scales (Reidsma *et al.*, 2007a). This is particularly true in data scarce regions where limitations in data quality and data quantity make the quantitative assessment of the models difficult.

Geostatistical methods, e.g. kriging, allow one to create gridded

yield maps, by interpolating observed crop yields based upon the variogram (Faivre *et al.*, 2004; Castoldi *et al.*, 2009). Kriging predicts a target variable by taking into account the spatial dependence between data, and the kriged estimate at each unsampled location includes the standard deviation of the prediction error. Ordinary kriging (OK) considers the spatial configuration of observed data (Lloyd, 2011), whereas co-kriging and regression kriging (RK) provide ways to incorporate the variability in the external covariate datasets into the kriged predictions (Papritz & Stein, 1999). RK allows the variance to vary in space by fitting a trend for the external covariate datasets with multiple linear regression (MLR) (Diggle *et al.*, 1998).

For large regional areas, however, global variograms may not incorporate effectively the spatial variability. To deal with this, the nonstationary kriging methods stratify the space and use local variogram models for each stratum (Stein *et al.*, 1988; Harris *et al.*, 2010a; Lloyd, 2011). Spatial regression methods like geographically weighted regression (GWR) use kernel functions to weigh both observations and covariate datasets directly at each calibration point (Fotheringham *et al.*, 2002). Here we use a novel approach to model crop yields that applies GWR in the RK framework. Such a hybrid GWRK predicts an attribute at unsampled locations by modelling the locally varying trend using covariates of observed attribute. It then applies kriging to predict regression residuals.

This study focuses on yield mapping in West Africa. Here, sorghum and millet comprise the staple cereal of people living in the drought-prone tropical regions (Maunder, 2002). In Burkina Faso, sorghum is the major cereal for food consumption i.e. 34% of all cereals (FAO, 1998). The Famine Early Warning System Network (FEWSNET, 2012) monitors yields of cereal crops countrywide, including the sorghum to mitigate food crises.

The main objective of this study is to model sorghum crop yield and its uncertainty at a regional scale in West African cropping system, using the crop yield observations obtained from countrywide georeferenced surveys. The study is applied to Burkina Faso that has relatively high spatial variability of sorghum crop yields due to strong variation of environment. We explored the use of GWRK to model the sorghum crop yield and compared it with other geostatistical methods.

4.2 Materials and methods

4.2.1 Study area

Burkina Faso has dominant rainfed agriculture. Spatial variability of climate in the country is characterized by a strong North-South annual rainfall gradient. The average annual rainfall decreases 1 mm per km and, besides this latitudinal trend, a difference of 200 – 300 mm may occur

in any direction within a radius of only 100km (Graef & Haigis, 2001). Rainfall occurs during the 3 – 5 summer months, with half the rainy days occurring in July and August (Graef & Haigis, 2001). Crops are grown during the rainy season. Periodic droughts and strong spatial rainfall and land variability restrain the farmers to adapt agricultural management strategies at the local level than at the national level (West *et al.*, 2008). The farmers apply those strategies commonly on their individual farm lands in a farmer's community, the so-called 'terroirs'. A terroir is the basic unit of communal agricultural management and, it is led by a traditional chief. The next administrative level is the district. Conventional subsistence or small-scale farming is the mainstay agricultural activity in Burkina Faso, covering approximately 85% of cultivated lands (UNCCD, 2000). This type of farming is typically associated with a low level of inputs, low financial ability of farmers, manual labor, local cultivars, little or no fertilization, no crop protection, and small-area farms run by households and having less accessibility to markets. Sorghum is typically grown in subsistence farming.

4.2.2 Crop yield sampling and analysis

Crop yield data were collected from the Statistiques Agricoles (AGRISTAT) Burkina Faso (AGRISTAT, 2010). AGRISTAT compiles household survey data for representative terroirs in order to reduce the operational costs. AGRISTAT measures the areas and the geographical locations of the household parcels, and gets the weight (kg) of the sorghum grains from all those measured parcels in a representative terroir. In this study, the AGRISTAT (georeferenced) household surveys for the year 2009 were used to obtain observed sorghum crop yield (kg ha^{-1}) per representative terroir, providing a total of 221 observations in the country.

Inadequacy in the configuration of observation sites contributes to erroneous assessment of the relations between observations and external covariate datasets (Challinor *et al.*, 2009). A factor that may contribute to this uncertainty is the spatial clustering of the input data and it is usually assumed that the representative terroir locations are not significantly clustered. To test the assumption, we compared its pattern with Complete Spatial Randomness (CSR) (Cressie, 1991; Diggle, 2003; Baddeley & Turner, 2005). CSR measures the distribution of the distances from an arbitrary representative terroir to its nearest neighbor. The theoretical mean value of the nearest neighbor distance function G(r) was obtained by computing the mean of 100 simulated values. Simulation envelopes were derived by computing the maximum absolute difference between the simulated and theoretical curves. The second assumption to be tested is that the yields are normally distributed. A Kolmogorov–Smirnov (K–S) two-side testing was used to test for difference in shape between the distribution of the observed yields and the normal distribution. The null hypothesis is that the two distributions do not differ.

4.2.3 Preparing external covariates of crop yield

Crop yields are primarily influenced by local climate, land characteristics, and the levels of inputs and management applied to the land. The following external covariate datasets were used to model the sorghum crop yield: Remote sensed-based vegetation indices, rainfall, topography, rural population density, poverty head count ratio, and market access.

The Normalized Difference Vegetation Index (NDVI) is a RS-based vegetation index, calculated as: $NDVI = (NIR - R)/(NIR + R)$, where NIR is the spectral reflectance in the near-infrared where top-of-the-canopy reflectance is dominant, and R is the reflectance in the red portion of the spectrum where chlorophyll absorbs strongly (Dorigo *et al.*, 2007). NDVI was used as a biophysical indicator of the vegetation productivity during the sorghum growing season. NDVI data were obtained from the SPOT 4 Vegetation (VGT) satellite images. We used a georeferenced temporal series of 10-day SPOT VGT NDVI composites with a spatial resolution of 1 km^2. These products are a combination of daily atmospherically corrected data of all vegetation measurements of the given decade into a single image, using the maximum value composite algorithm. In total 18 images were obtained for the sorghum growing season from June 2009 to November 2009. A factor analysis was performed to obtain the standardized principal components of NDVI (NDVI.PCs), to reduce the dimensionality of data and to remove multicollinearity.

Climatic conditions are important determining factors for sorghum crop yield (Graef & Haigis, 2001). The Tropical Applications of Meteorology using Satellite (TAMSAT) merges the satellite data with rain gauges observations to derive rainfall estimates over the African region. The TAMSAT rainfall estimates have been validated for West Africa and the Sahel region using a dense rain gauge network covering area of 1° square (Grimes *et al.*, 1999). For 10-day TAMSAT rainfall estimates, 85% of the estimated and measured values agree to within 1 standard error for 1° square. We obtained 10-day TAMSAT rainfall data for the year 2009. To reduce the data dimensionality, a factor analysis was performed to obtain the principal components of rainfall time series (PREC.PCs).

Topography affects the crop performance, in both low and high rainfall conditions (West *et al.*, 2008). For instance, Central and Northern Sahelian farmers mostly abandon the marginal uplands and elevated borders of lowland fields, because those land types perform poorly in low rainfall conditions. Similarly, in the regions of higher rainfall conditions towards Southern Burkina Faso, the lowland fields are prone to get flooded, due to run-off to from neighboring lands. Elevation was selected as a topographical variable. Elevation data (ELEV) were obtained from Hydro 1 km Africa datasets, developed at the US Geological Survey (USGS) Earth Resources Observation Systems Data Center (http://eros.usgs.gov).

Extensive data on characteristics of individual farms are used to relate location-specific management data and crop yields. For instance, (Reidsma *et al.*, 2007a) using the Farm Accountancy Data Network (FADN) data observed that levels of inputs and capital intensity of farmers

influenced wheat yields at the European scale. Such extensive data were not available in Burkina Faso. To access management applied to terroirs, we therefore used regional grids of socioeconomic data as proxy indicators, which are explained below.

Financial ability of farmers influences input intensity e.g., fertilization use, pest control use, and use of improved crop cultivars. In subsistence farming, the levels of inputs may be marginalized depending on where the population is living. In areas where a larger population is living below poverty farmers may have less capital to invest in improved levels of inputs. A commonly used metric is the poverty head count ratio (PHCR), being the percentage of the population living below the established poverty line. We used the 1.25 poverty line i.e. less than 1.25 US dollar purchasing power parity (PPP) per day. For this, the gridded data of sub-national PHCR (1 km^2) were obtained from HarvestChoice (http://harvestchoice.org), expressed in 2005 international equivalent PPP dollars.

Availability of labor in a terroir is a crop management factor in subsistence farming. Depending on labor availability, the farmers use intensive land management to increase yields per hectare of cropped area (West *et al.*, 2008). We used rural population density as proxy for labor availability. Rural population density (RURP) i.e. number of rural people per km^2 (2005) were obtained from HarvestChoice (http://harvestchoice.org/).

Accessibility to markets is a reliable estimator of crop areas at a regional scale (Ramankutty, 2004). Crop areas can be used as proxy to determine capital intensity of farms (Reidsma *et al.*, 2007a). Low capital intensity prohibit farmers of remote terroirs to transport their crops to the market, or modern inputs back to their farms (Fortanier, 2006). Larger crop areas were expected close to markets, having more capital available for investments in new technologies. Market access (MARK) was calculated based on the map of major markets of Burkina Faso, obtained from FEWS NET (FEWSNET, 2012). A simple model was applied based on the distance of a terroirs to its nearest markets. The obtained gridded dataset has a cell size equal to 1 km^2.

4.2.4 Statistical Analysis

The first step in the statistical analysis is the definition of a regularly spaced grid covering the study area with a cell size equal to 1 km^2. We define our spatial model to predict crop yield at visited and unvisited grid cells using the values of the covariates, as

$$\begin{aligned} Y(s) &= f(NDVI.PC(s), PERC(s), ELEV(s), RURP(s), PHCR(s) \\ &\quad MARK(s)) + H(s) \end{aligned} \tag{4.1}$$

where the notation $Y(s)$ represents the sorghum yield at a location s that is modelled in two components: a trend function f, and the model error $H(s)$ denoting a small-scale fluctuations around f with variance $VAR\{H(s)\}$. The function f determines the overall influence of

$NDVI.PC(s)$, $PERC(s)$, $ELEV(s)$, $RURP(s)$, $PHCR(s)$, $MARK(s)$ covariates, following the notations in the previous Section. It was modeled as a linear function using MLR and GWR, when applying MLRK and GWRK, respectively.

4.2.4.1 MLR and GWR

MLR and GWR were applied to model the sorghum yield trend. The coefficients of these regression models were applied next to predict the yield at unvisited locations using the grid cell values of the covariates at those locations.

MLR assumes that $H(s)$ is a stationary random field with $E\{H(s)\} = 0$ and $VAR\{H(s)\} = G$, where the elements of the $n \times n$ diagonal matrix G reflect a pure nugget variance i.e. zero spatial autocorrelation, with $G = \sigma^2 I$. In the case of MLR, the regression coefficients are estimated as $\hat{\beta}_{MLR} = C_{MLR}Y$, where $C_{MLR} = (X^T X)^{-1} X^T$, and X being the matrix with covariates (Chambers & Hastie, 1992). The MLR prediction at s is

$$\hat{Y}_{MLR}(s) = x_\circ(s)^T \hat{\beta}_{MLR} \tag{4.2}$$

where $x_\circ(s)$ is a vector of covariates at a grid cell s. The MLR prediction error variance at s is estimated as

$$VAR[\hat{Y}_{MLR}(s) - Y(s)] = [1 + x_\circ(s)^T (X^T X)^{-1} x_\circ(s)]\hat{\sigma}^2_{MLR} \tag{4.3}$$

where $\hat{\sigma}^2_{MLR} = RSS/(n - np)$, RSS is the residual sum of squares, and the denominator term is known as effective degrees of freedom of the residual, and np is equivalent to the number of parameters in a global linear model.

Similar to MLR, also GWR assumes that $H(s)$ is a stationary random field. GWR parameters are estimated locally at all observed locations, as $\hat{\beta}_{GWR}(s) = C_{GWR}Y(s)$, where $C_{GWR} = (X^T W(s) X)^{-1} X^T W(s)$ (Fotheringham *et al.* (2002), pp. 55). Here, $W(s)$ is a $n \times n$ diagonal matrix of spatial weights. The weighting matrix is specified by means of a kernel function. We used a Gaussian function as the kernel function to specify a weighting matrix considering the neighboring terroir locations,

$$W_{st} = \exp[-0.5(d_{st}/b)^2] \tag{4.4}$$

where b is a (non-negative) kernel bandwidth, and d_{st} are the distances between s and neighboring terroir locations t. The weighting will be small for terroir locations far from s, excluding these observations from parameter estimation at location s.

The GWR estimated local relations were then used to predict the crop yield at unvisited locations as

$$\hat{Y}_{GWR}(s) = x_\circ(s)^T \hat{\beta}_{GWR}(s), \tag{4.5}$$

with the GWR prediction error variance at s (Leung *et al.*, 2000) as,

$$VAR[\hat{Y}_{GWR}(s) - Y(s)] = [1 + x_\circ(s)^T (C_{GWR} C_{GWR}^T) x_\circ(s)]\hat{\sigma}^2_{GWR} \quad (4.6)$$

where np in the above expression of $\hat{\sigma}^2_{MLR}$, can be termed as the effective number of parameters in the expression of $\hat{\sigma}^2_{GWR}$ for GWR model.

4.2.4.2 OK and KED

OK allows to predict a spatially dependent attribute, using a variogram that relates the autocorrelation to the separation distance between two sample locations. OK was used to interpolate observed crop yields and also the residuals from MLR and GWR. The OK kriging predicts crop yield at an unvisited location s (Cressie (1991), pp. 119) as

$$\hat{Y}(s) = \sum_{s=1}^{n} W_s . Y(s_i), \qquad \sum_{s=1}^{n} W_s = 1 \quad (4.7)$$

Here, W_s is the weight assigned to each of the observed crop yields and, is yet unknown. The condition that the sum of weights is equal to 1, makes the $VAR[\hat{Y}_{OK}(s) - Y(s)]$ minimal among all linear unbiased predictors.

When applying OK, we can write $VAR\{H(s)\} = G'$, where G' is a $n \times n$ matrix, whose elements are found from variogram $\gamma(h)$ and reflect covariances of the sample locations. The variogram model used in this study was the spherical model:

$$\gamma(h) = \left\{ \begin{array}{ll} 0 & \text{if } h = 0 \\ C_0 + C_1 . [\frac{3h}{2a} - \frac{1}{2}(\frac{h}{a})^3] & \text{if } 0 \leq h \leq a \end{array} \right\} \quad (4.8)$$

where C_0 is small-scale nugget variance, C_1 is large-scale structural variance, and a is the correlation range.

The spatial covariances between the values at the prediction and the sample locations are contained in the vector σ'. The unbiased predictor with minimal variance of the prediction error is given by

$$\hat{Y}(s) = x_\circ(s)^T \hat{\beta}_{GLS} + \sigma'^T G'^{-1} (Y - X\hat{\beta}_{GLS}), \quad (4.9)$$

where the trend parameters $\hat{\beta}$ are estimated from generalized least squares (GLS) as, $\hat{\beta}_{GLS} = (X^T G'^{-1} X)^{-1} X^T G'^{-1} Y$. In the case of OK, X consists of a vector with values 1 only, whereas for KED it extends with an additional column of the values of the external covariates. The prediction error variance for KED is

$$\begin{aligned} VAR[\hat{Y}(s) - Y(s)] \quad = \quad & \{\hat{\sigma}^2 - \sigma'^T G'^{-1} \sigma'\} \\ & + (x_\circ(s) - X^T G'^{-1} \sigma')^T . (X^T G'^{-1} X)^{-1} \\ & \times (x_\circ(s) - X^T G'^{-1} \sigma') \end{aligned} \quad (4.10)$$

where $\hat{\sigma}^2$ is the estimate of the residual variogram sill $C_0 + C_1$. Again, the OK prediction error variance is obtained on setting a constant trend in (Eq 4.10) (see for details, Papritz & Stein (1999), pp. 95–98).

4.2.4.3 MLRK and GWRK

An explicit solution of (Eq 4.9) and (Eq 4.10) is regression kriging (RK), which combines a separately fitted regression on external covariates with kriging of the regression residuals (Diggle *et al.*, 1998). (Chilès & Delfiner, 1999), and (Rivoirard, 2002) showed that both KED and RK are mathematically equivalent and give the same predictions and error estimates if a global neighborhood is specified. RK however allows to specify a local, nonparametric, and nonlinear trend component (Harris *et al.*, 2010b). Here we specified both MLR and GWR, to fit a model in MLRK and GWRK, respectively. In the both cases $VAR\{H(x)\}$ includes the covariance between point pair observations. This allows including nonzero spatial autocorrelation of the regression residuals. The procedure we followed was that first variograms of the regression residuals from MLR and GWR were determined, followed by ordinary kriging (OK) of the regression residuals towards the unvisited grid cells. The interpolated residuals were then added to the respective predicted drift surfaces. The MLRK prediction was then derived as,

$$\hat{Y}_{MLRK}(s) = x_\circ(s)^T \hat{\beta}_{MLR}(s) + \hat{H}_{OK}(s), \tag{4.11}$$

and, the GWRK prediction as,

$$\hat{Y}_{GWRK}(s) = x_\circ(s)^T \hat{\beta}_{GWR}(s) + \hat{H}_{OK}(s), \tag{4.12}$$

In this way both heteroskedasticity of the external trend using the covariates and spatial dependency of residuals were included and compared. Note that $\hat{Y}_{GWRK}(s) = \hat{Y}_{GWR}(s)$, if the GWR residual variogram is determined as a pure nugget variogram. The additive relationship of predictions from (Eq 4.11) and (Eq 4.12) continues to prediction variances as well. Hence, the MLR prediction variance (Eq 4.3) and the GWR prediction variance (Eq 4.6) were added to the OK prediction error of residuals at s (derived from Eq 4.10), respectively to obtain the MLRK and GWRK prediction error variances at s.

4.2.5 Multicollinearity analysis

The variance inflation factor (VIF) was applied to analyze multicollinearity between explanatory variables (Kutner (2005), p. 408). A tolerance value between $(0 - 1)$ determines the acceptable level of multicollinearity between two variables. A tolerance value above 0.4 is recommended, but

since some correlation between variables, especially between NDVI and rainfall, may occur, we choose a tolerance level of 0.2. The full analysis then comes down to the following:

1 For the factor analysis of NDVI and TAMSAT series, a minimum tolerance level of 0.6 was applied.

2 A matrix scatter plot was used to visualize mutual relationships between response and explanatory variables.

3 MLR was applied, and all non-significant variables were dropped from the analysis.

4 Variables with a VIF value greater than 5 were omitted from the analysis.

4.2.6 Validation

Cross validation was used to compare predicted and observed values. Cross validation omits part of the observations for validation and interpolates the remaining dataset towards the locations of the omitted data, followed by minimizing the RMSE between the predicted data with the omitted data. Cross validation was also used to calibrate the kernel bandwidth in GWR (Fotheringham *et al.* (2002), pp. 212), and to compare the accurate variogram model and kriging type (Pebesma, 2004). Using cross validation of kriged residuals from MLR and GWR, the prediction accuracy of MLRK and GWRK was compared, by computing the frequency distribution of z-score, variance explained, mean error (ME), and root mean square error (RMSE). Variograms were re-calculated for the predicted sorghum crop yield values at the sample location, from OK, KED, MLRK, and GWRK, and were compared to that of OK interpolation set (i.e. from sorghum crop yield observation data).

To compare MLR and GWR, we used common accuracy and precision statistics such as the residual sum of square (RSS), the standard error of the estimate (SE), the Akaike Information criterion (AIC), ANOVA, and Monte Carlo test (Fotheringham *et al.* (2002), pp. 212–216). ANOVA tests the null hypothesis that GWR model represents no improvement over a global model. Monte Carlo test tests the significance of the spatial variability in the local parameter estimates.

The GWRK model was compared with other kriging models, including OK, KED, and MLRK models. The mean absolute error (MAE) and mean square error (MSE) were used as comparison criteria between predicted and actual values, which define their optimality. They were calculated as:

$$MAE = \frac{\sum_{n} |\hat{Y}(s) - \hat{Y}_m(s)|}{n} \tag{4.13}$$

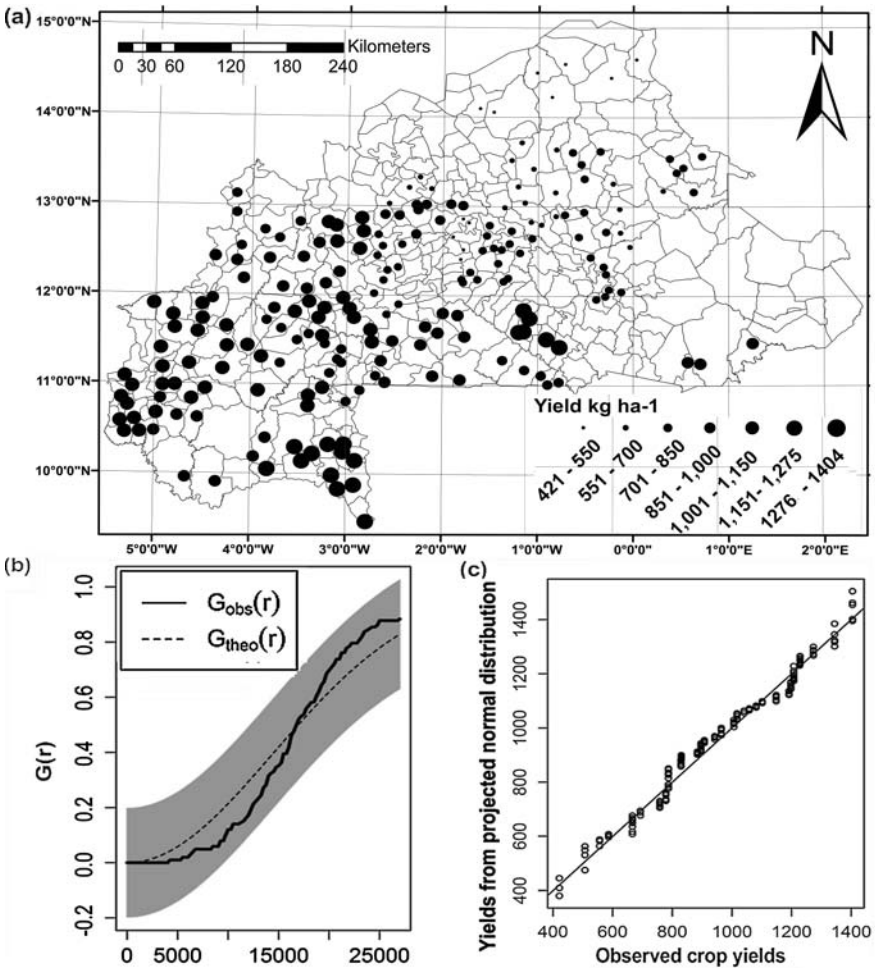


Figure 4.1 Observed sorghum yield (kg ha^{-1}) for year 2009 in the study area (a) – Comparison of the confidence bands for G function theoretical and observed distributions in complete spatial randomness (CSR) (b) – Q-Q plot comparing the observed sorghum yield (horizontal axis) to the yields form projected normal distribution with the standard deviation and mean values of observed sorghum crop yield (vertical axis) (c).

and,

$$MSE = \frac{\sum_{n}(\hat{Y}(s) - \hat{Y}_m(s))^2}{n} \tag{4.14}$$

where $\hat{Y}(s)$ are the prediction values and $\hat{Y}_m(s)$ are the mean values and n is the length of the grid cells.

The MAE gives the local accuracy (unbiasedness) and the MSE the prediction accuracy (minimum error).

The R package spgwr (Bivand *et al.*, 2008) was used to perform GWR, and the R package gstat (Pebesma, 2004) to automatically fit the variograms and to perform the kriging of GWR residuals. In geostatistical applications, spatial predictions that involve change of support generally cause over- or under-estimation. To deal with this, the entire mapping procedure was carried out by applying block kriging on 1 km^2 grid cell. This size is equal to the grid cell size of explanatory data sets and also, to the crop yield prediction support.

4.3 Results

4.3.1 Statistical analysis

Fig 4.1(a) shows spatial distribution of the observed sorghum crop yields (kg ha^{-1}) in Burkina Faso. The mean observed sorghum yield (961 kg ha^{-1}) in 2009 is slightly less than the 10 years (2000–2009) county average yield (969 kg ha^{-1}). There is a clear North-South trend of sorghum yields across the country. High yield values occur towards the Southern part of Burkina Faso, with a subhumid agroecological gradient. This zone has more favorable cropping conditions as compared to the semiarid and arid zones towards the Northern parts of the country.

Fig 4.1(b) shows that the line of observed value of G(r) for data pattern is within 80% of distance of the theoretical value of G(r) for CSR. The CSR test confirms that the representative terroirs locations are not clustered and are representative relative to the geographical space of the study area. The K–S test (D= 0.11, p-value= 0.1) rejects normality assumption of the distribution of observed sorghum crop yield. A normal distribution was assumed with the standard deviation and mean values of observed sorghum crop yield, which was then plotted against the distribution of observed sorghum crop yield (Fig 4.1c). The plot shows that the highest yields in the distribution of observed sorghum yield are lower than for the corresponding normal distribution, with the maximum difference equal to 200 kg ha^{-1}. This is typically observed in the case of crop yield observations that are obtained from large-scale agricultural surveys. Such data are frequently spatially dependent and may not satisfy the assumptions of traditional statistics. For instance, MLR assumes that independent data follow normal distribution and homoscedasticity,

Table 4.1 Parameter estimates for the sorghum yield model fitted using the multiple linear regression (MLR) regression.

Target variable	Regression Params.	Est.	SE[a]	P-value	VIF[b]	Tolerance
Yield	Intercept	171.20	91.70	0.050	-	-
	NDVI.PC1	117.20	38.50	0.001	4.59	0.22
	PREC.PC1	34.10	9.80	<0.001	3.97	0.30
	PREC.PC2	38.40	8.70	<0.001	1.65	0.60
	ELEV	0.37	0.09	<0.001	2.25	0.45
	PHCR	3.64	1.05	<0.001	1.64	0.60
	MARK	0.80	0.25	0.001	1.43	0.70

[a] Standard error
[b] Variance Inflation Factor

i.e. constant variance. Here we obtained a high degree of spatially autocorrelation (Moran's I statistic= 0.83, p< 0.001) of the observed sorghum crop yields.

Figure 4.2 shows maps of external covariates to predict the sorghum crop yield. Factor analysis of 10-day SPOT VGT NDVI composite images, covering the sorghum growing season, captured 99% variance of the entire image series into the first three principal components. The first principal competent explained more than 91% of the total variability in the image bands and, this component (NDVI.PC1) was therefore included in the analyses. Similarly, factor analysis of 10-day TAMSAT rainfall estimates yielded approximately 93% of the total variability into the first two principal components (PREC.PC1 and PREC.PC2).

4.3.2 Regression analysis

MLR showed a relation ($R_a^2 = 0.64$, $p < 0.001$) between the sorghum yield and those explanatory variables. The coefficients obtained from MLR regression are given in Table 4.1. A Moran's I value of 0.51 ($p < 0.001$) for the MLR residuals showed that the model can be improved by (I) using spatial regression methods, and (II) kriging the residuals in the case of regression kriging.

GWR was performed to determine the environmental and management factors represented by explanatory variables are responsible for local spatial variability of the sorghum crop yield. Accuracy statistics of the performed MLR and GWR models are shown in Table 4.2. The results of ANOVA showed a statistically significant ($p < 0.001$) improvement of GWR over the global MLR approach. The RSS and SE values in the sorghum yield estimations were reduced in the GWR model. The adjusted coefficient of determination, R_a^2, increased from the global MLR to the GWR model, although an increase can be expected relating to the difference in degrees of freedom (DF). The reduction in the AIC from the MLR model suggests, however, that the GWR approach is better even after considering the differences in DF.

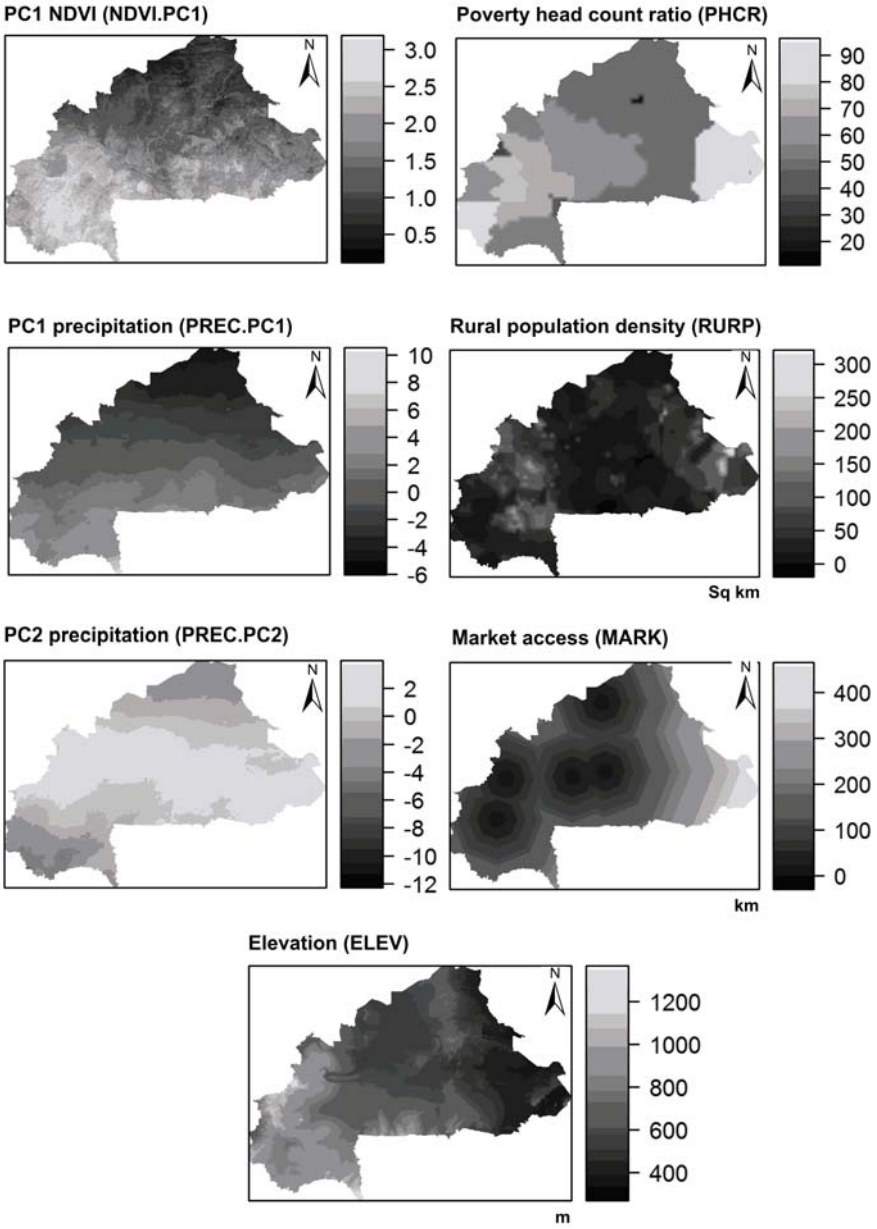


Figure 4.2 Maps of external covariates to predict sorghum crop yield in Burkina Faso.

Table 4.2 Accuracy and precision statistics for sorghum regression models fitted using both the global multiple linear regression (MLR) and geographically weighted regression (GWR) approaches.

Model	RSS[a] (sqrt)	SE	DF	F	AIC[c]	$R_a^{2\,d}$
MLR	2017	141.0	7	-	2686	0.64
GWR	1099	91.6	148	5.58	2611	0.90

[a] Residual sum of square
[b] Standard error
[c] Akaike Information criterion
[d] Adjusted coefficient of determination

To inspect the regression parameters in terms of spatial non-stationarity, the summary comparison of the MLR and GWR models is presented in Tables 4.1 and 4.3. This comparison shows a considerable amount of non-stationarity in all parameter estimates i.e. the ranges of local estimates between the lower and upper quartiles are greater than ±1 standard deviations of the MLR equivalent parameters estimates, being 2× S.E of each global estimate in Table 4.1. For instance, examining the NDVI.PC1 parameter estimate, we observed that the inter-quartile range of the GWR local parameter estimate ranging from 1.24 to 203 is well beyond the ±1 standard deviation (89) of the MLR parameter estimate. Moreover, Monte Carlo significance test on the local estimates indicates that there is significant spatial variation in the local parameter estimates for all variables (Table 4.3). A low Moran's I value of 0.14 ($p < 0.001$) was observed for the GWR residuals, compared to that of 0.51 ($p < 0.001$) for the MLR residuals. Moreover, the spread of GWR residuals is smaller, as is seen from the density histograms of MLR and GWR residuals in Figs 4.3(b) and 4.3(c). This suggests that the GWR accounted for most of the spatial variability in the sorghum yields from the spatial variability of the external covariate data. Next, spatial structure observed in the GWR and MLR residuals is to be incorporated to construct GWRK and MLRK regression kriging models, respectively.

4.3.3 Geostatistical analysis

No anisotropy in the data was observed. In the case of OK interpolation of sorghum yields, the parameters of the spherical variogram showed that the average distance up to which the variogram increases is approximately 182 km (Fig 4.4a). This distance covers approximately $8 - 10$ neighboring terroirs, as the mean shortest inter-terroir spacing is 18 km. The variograms of MLR and GWR residuals (Figs 4.4b and 4.4c) showed a decrease both in the total sill and range estimates. For regression kriging, this indicates that more variability is taken into the trend component, leaving less for spatial autocorrelation of residuals. For instance, a lower sill of the variogram of GWR residuals shows the strength of the GWR trend, which resulted into a comparatively low smoothing effect in the sorghum yield map from GWRK as compared to OK (Fig 4.7). Again the

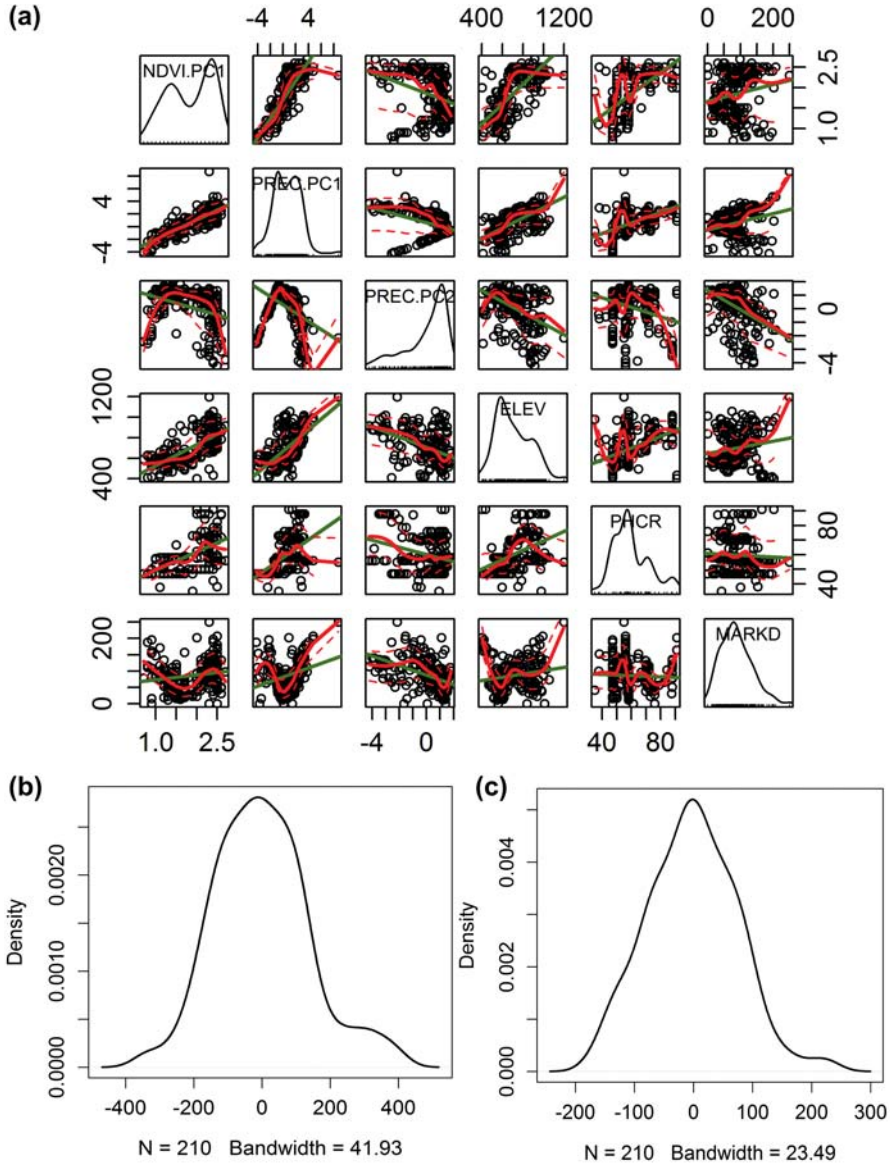


Figure 4.3 Matrix scatterplot to visualize mutual relationships between independent and dependent variables (a) – Kernel density plot of MLR residuals (b) – Kernel density plot of geographically weighted regression (GWR) residuals (c).

Table 4.3 Parameter estimates for the sorghum yield model fitted using the geographically weighted regression (GWR) approach.

Parameters	Min[a]	1st Q[b]	Med[c]	3rd Q	Max[d]	MC[e] Test
Intercept	-1310	-79.6	448	713	1740	0.001
NDVI.PC1	-555	1.24	72.8	203	506	0.001
PREC.PC1	-192	10.9	52.9	96.5	362	0.001
PREC.PC2	-529	-38.6	25.2	102	473	0.001
ELEV	-1.5	-0.20	0.23	0.52	2.56	0.001
PHCR	-19.2	-0.03	3.03	7.55	35.2	0.001
MARKD	-3.29	-0.32	0.94	2.29	6.75	0.001

[a]Minimum
[b]Quartile
[c]Median
[d]Maximum
[e]Monte Carlo significance test for spatial non-stationarity of parameters

Table 4.4 Cross validation (residuals) results of the geostatistical prediction models of sorghum crop yield – ordinary kriging (OK), kriging with external drift (KED), multiple linear regression kriging (MLRK), and geographically weighted regression kriging (GWRK).

Model	Mean (Z-score)	SD (Z-score)	Variance explained	Mean error	RMSE[a]
OK	-0.005	0.89	0.88	-0.88	81.6
KED	-0.005	0.83	0.89	-1.14	79.7
MLRK	-0.007	0.86	0.88	-1.58	81.1
GWRK	-0.004	1.01	0.90	-0.66	71.2

[a]Root mean square error

variogram of GWR residuals exhibits a clear spatial structure (Fig 4.4c). The variogram distance increases up to approximately 41 km. The GWR residuals have been predicted using ordinary kriging. Fig 4.7d shows the GWRK map of sorghum crop yield, obtained from combining the GWR trend and the predicted GWR residuals.

The results of cross validation (Table 4.4) show that GWRK performed better than both OK and MLRK. In GWRK, the mean error value was close to 0 (−0.66), and the correlation value between observed and predicted sorghum yields was close to 1 (0.9). For all the three kriging methods, the z-score values of cross validation showed that the variogram models reasonably accounted for the spatial autocorrelation to explain the variance. The RMSE value was observed lower in GWRK. The variograms were reproduced and fitted for the sorghum yield predictions from OK, KED, MLRK, and GWRK, see Figs 4.5(a,c,e,g). The reproduced GWRK variogram model is approximately identical to the OK variogram model (Fig 4.5(g) and Fig 4.4a), whereas the GWRK considerably minimized the local error variance as compared to OK, KED, and MLRK (Figs 4.5(b,d,f,h) and 4.7(d)).

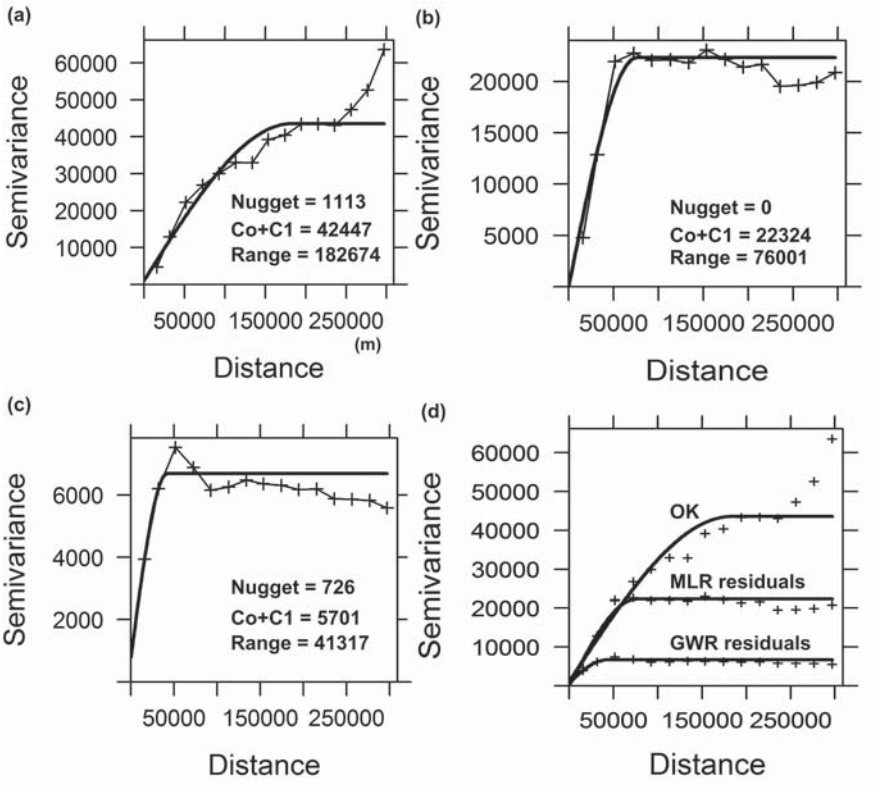


Figure 4.4 Variograms of ordinary kriging (OK) (a), of residuals of multiple linear regression (MLR) (b) and of residuals of geographically weighted regression (GWR) (c), and their comparison (d).

4.3.4 Sorghum yield prediction and uncertainty of prediction

The maps of sorghum crop yield produced by MLR and GWR and the variance of prediction error are shown in Fig 4.6. The MLR prediction map presented an overall North-South global trend of sorghum yield. MLR, however, failed to show the short-range spatial variability of sorghum yields, that is clearly present on the Eastern part of maps obtained with geostatistical methods (see Fig 4.7). The MLR depicted trend shows the strength of explanatory variables used in the analysis. The summary comparisons of prediction methods (see Table 4.5), gave 2.23% lower mean for MLR, and 0.63% lower mean for GWR, compared to the mean sorghum yield at primary locations. MLR reduces the prediction variances and tends to make them smooth across the area. For example, MLR overestimated 9.25% of the upper observed sorghum yields and, underestimated 7.67% of the lower crop yields. GWR gave 18.8% over-fit of the maximum observed yield values, however, the predicted mean and median yield values considerably improved compared to the observed mean and median yield values. The MLR prediction error variance was

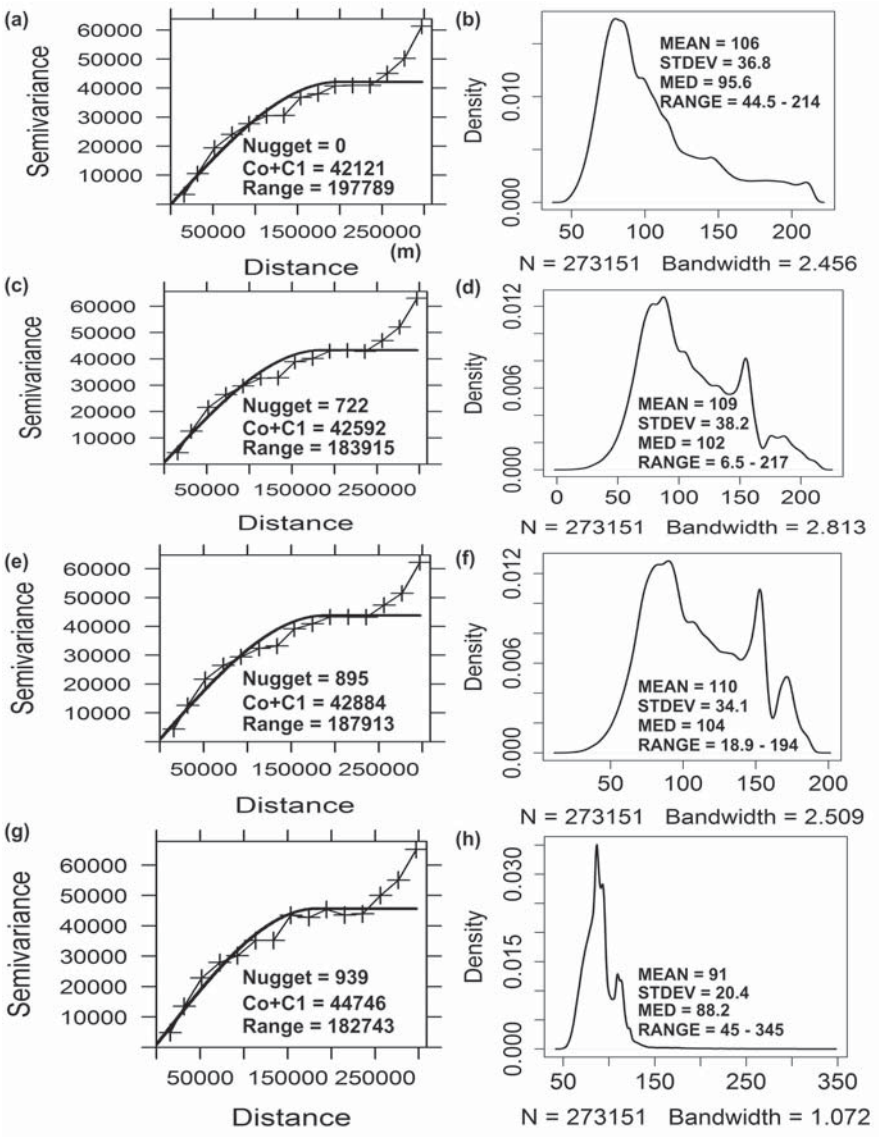


Figure 4.5 Variograms reproduced from the predicted sorghum yield values at sampled locations from models: ordinary kriging (OK) (a), kriging with external drift (KED) (c) multiple linear regression kriging (MLRK) (e), and geographically weighted regression kriging (GWRK) (g) – Corresponding kernel density plots for local prediction error variances of OK interpolation (b), KED (d), MLRK (f), and GWRK (h).

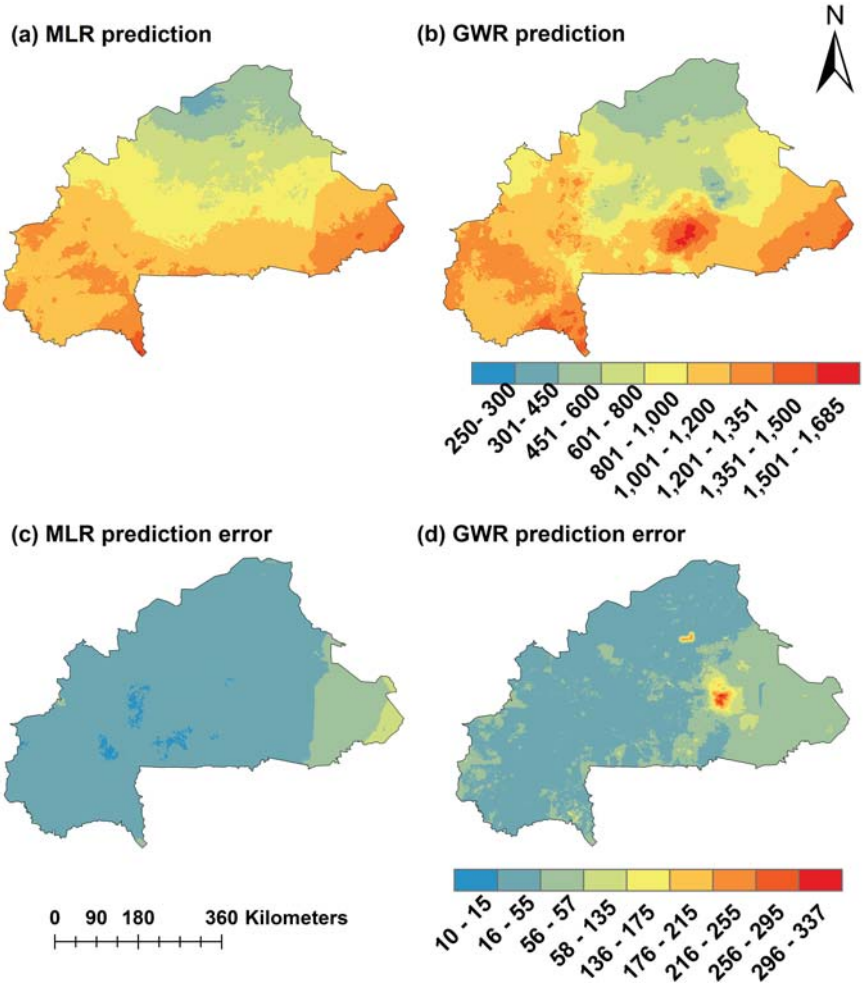


Figure 4.6 Sorghum crop yield prediction (kg ha^{-1}) from multiple linear regression (MLR) (a), and geographically weighted regression (GWR) (b) – Estimates of the crop yield prediction uncertainty from MLR (c), and GWR (d).

more uniform over the study area than the GWR.

Comparison of OK, KED, MLRK, and GWRK for mapping sorghum yield is shown in (Fig 4.7). Higher values of the sorghum crop yield (kg ha^{-1}) were observed towards the Southern part of Burkina Faso. This shows that spatial variability of NDVI, precipitation, and elevation successfully explained the favorable cropping conditions along agroecological gradients and between sites. The maps of OK, KED, and MLRK prediction errors showed high residual variability in the Northern and NE areas, with 75% values in the range of 56 – 215. This can be traceable from Fig 4.1(a), which shows the observed sorghum data are relatively absent

Table 4.5 Histogram descriptive statistics, mean absolute errors (MAE), mean square errors (MSE), and prediction error variance for the sorghum yield predictors – Observed sorghum yield (kg ha^{-1}) sampled at 210 terroirs – Models of global multiple linear regression (MLR) and local geographically weighted regression (GWR) – Interpolating sorghum yield observations, using ordinary kriging (OK)– Predicting sorghum crop yield with external covariate data, using kriging with external drift (KED), multiple linear regression kriging (MLRK), and geographically weighted regression kriging (GWRK).

Model	n	Min[a]	1st Q[b]	Med[c]	3rd Q	Max[d]	Mean	MAE	MSE	Error variance
Observed	210	421	787	964	1191	1404	961	-	-	-
MLR	273151	391	787	995	1118	1534	940	482	287217	19.2
GWR	273151	268	759	1009	1133	1685	953	486	288295	25.8
OK	273151	424	741	942	1069	1412	913	392	204315	36.8
KED	273151	383	727	937	1078	1475	907	523	329203	38.2
MLRK	273151	385	747	977	1127	1487	935	503	307817	34.1
GWRK	273151	267	762	980	1140	1676	955	480	272669	20.4

[a]Minimum
[b]Quartile
[c]Median
[d]Maximum

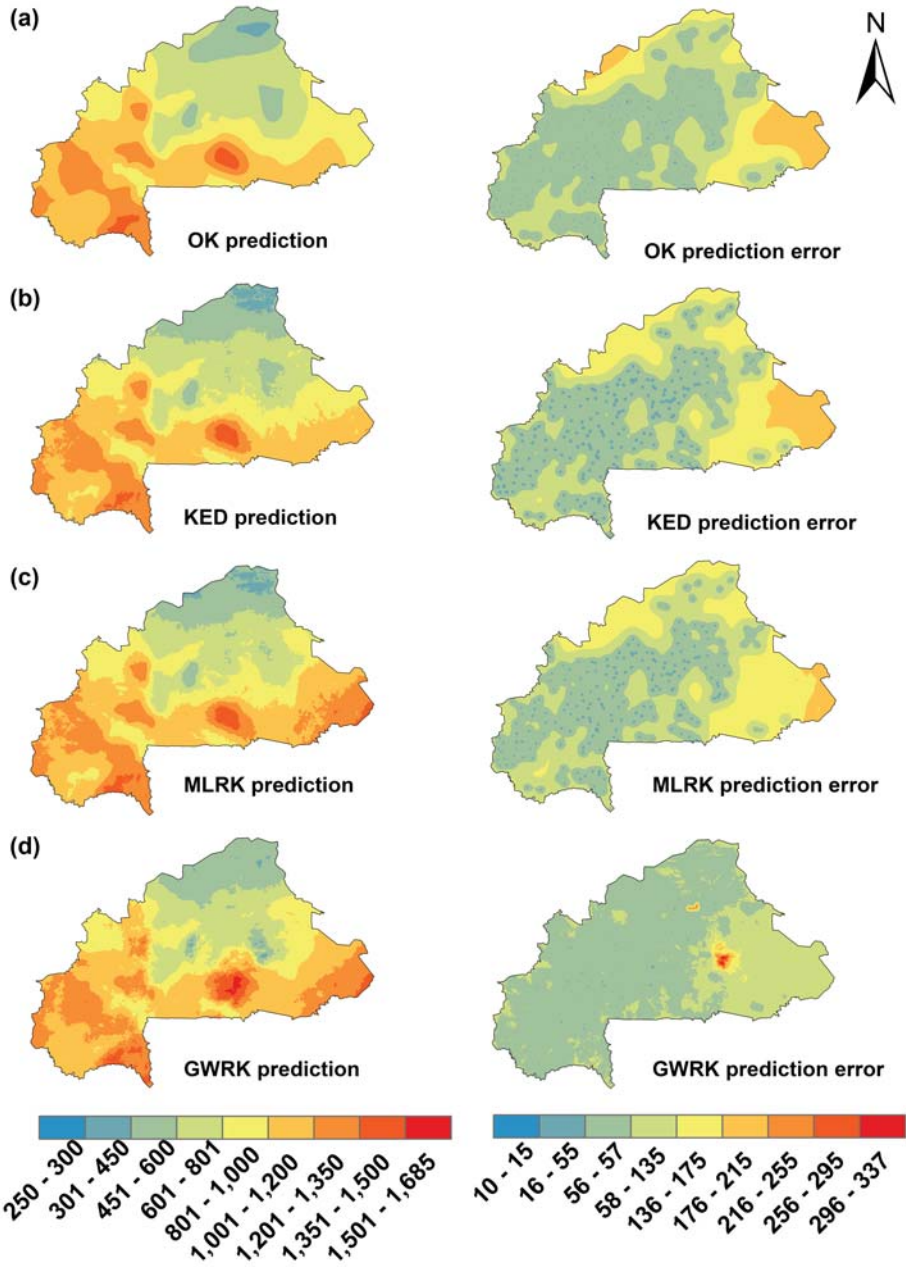


Figure 4.7 Sorghum crop yield prediction (kg ha^{-1}), and estimates of prediction uncertainty from: ordinary kriging (OK) (a), kriging with external drift (KED) (b), multiple linear regression kriging (MLRK) (c), and geographically weighted regression kriging (GWRK) (d).

in those areas. GWRK considerably reduced this residual variability to 95% values in the range of 56 – 135. This gives another evidence that GWRK effectively utilized information present in the environmental and management datasets to improve accuracy of crop yield predictions.

OK interpolation over-smoothed the predicted sorghum yields compared to the observed yields (5.25% lower mean). MLRK improved this smoothing effect, but local error variance become closer to that found with OK (see Table 4.5). Moving from GWR to GWRK, however, did not significantly improve prediction (mean) and accuracy of prediction (MAE and MSE). This indicates that the GWR trend fits are strongly local and captured most of the variation in the data, leaving a relatively reduced spatial autocorrelation of residuals. In GWRK, however, the prediction error variance was reduced considerably compared to MLRK (from 34.1 to 20.4) and, compared to GWR, it was reduced from 25.8 to 20.4. Moreover, the GWRK predicted mean and median yield values were accurately estimated compared to OK, KED, and MLRK. GWRK outperformed all other models both for sorghum yield prediction and mean variance of prediction error.

4.4 Discussion

This study explored the utility of various global and regional gridded datasets, to predict the sorghum crop yield in Burkina Faso. Results showed that climate, topography, financial ability of farmers and labor availability are important explanatory factors for sorghum crop yield. The GWR analysis indicated a large spatial variation in local parameter estimates. Financial ability of farmers is better characterized by the sub-national poverty head count data than by market access.

The sorghum yield response provided by the NDVI was related to environmental and management factors. Likewise, information on soil conditions is also important to explain variation in sorghum crop yields (West *et al.*, 2008). To explore these relations in Burkina Faso, reliable soil data provide point or polygon based information of individual soil profiles. Whereas, they were required for individual grid cells. The Harmonized World Soil Database (HWSD) provide such data for many Sub-Saharan countries, but they are not yet available for Burkina Faso (FAO & IIASA, 2012). The HWSD gridded soil data may be used in the future to better represent soil properties at the level of detail that is demanded by the GWRK model as in this study.

GWRK has the potential to predict accurately over larger areas, using external covariate datasets, specifically in situations of nonstationary relations that could not be properly modeled by GWR or MLRK. The prediction accuracy of GWRK is high, provided that sources of the model uncertainty are properly analyzed. In particular, here we analyzed the following:

1. Prediction accuracy in regression-based approaches largely depends upon the accuracy of estimated relationships. In GWR, (Harris *et al.*, 2010b) showed that the choice of both kernel type and bandwidth optimization method affects the local sample sizes and thus the prediction uncertainty. An optimal bandwidth determines the extent of the spatial neighborhood, to accurately calibrate each locally weighted regression. There is a well-known trade-off between a small bandwidths that may lead to large standard errors, and a large bandwidth that may lead to small standard errors (Fotheringham *et al.*, 2002). Here we used a spatially adaptive kernel i.e. a small bandwidth for large observations in a given area, and a larger bandwidth otherwise. Presently, there is no well-established method to quantify the uncertainty induced by improper kernel types or improperly calibrated bandwidths of the kernel and thus the resulting local parameter uncertainty. (Harris *et al.*, 2010b) accessed such uncertainty by calibrating and competing different kernel types, and different optimization methods. These comparisons however would soon become rather tedious. As an alternative, we compared the prediction maps from GWRK to other models and, analyzed the model efficiency based on (I) how well the model reproduced the sample variogram, and (II) how much the model minimized the local error variance. GWRK reproduced the sample variogram model while minimizing the local error variance as compared to OK, KED, and MLRK.

2. In MLRK, the relation between the dependent variable and its covariates is assumed to be linear. This allows us to simply sum the drift prediction error variance and the kriging error variance of the predicted residuals, to obtain the MLRK prediction error variance. However, the correlations between the GWR and the kriging components of GWRK may be more complex and a simple addition of the two variances might result into a considerable bias in the GWRK prediction error variance. Presently, there is no empirical method to estimate such bias. (Harris *et al.*, 2010b) applied a pragmatic approximation to the GWRK variance by subtracting the residual variogram sill estimate from the sum of prediction variance and the residual OK variance. This approximation, however, gave negative GWRK variances. We therefore applied the additive relationship of predictions from (Eq 4.12) to the GWR prediction variance as well, which gave a significantly reduced (non-negative) GWRK variance.

3. In a multivariate framework, correlation of explanatory variables may lead to multicollinearity, resulting in instable results in the estimated GWR regression coefficients. Multicollinearity is often observed when analyses include many environmental factors as covariates. For instance, it is likely that NDVI images present a high correlation with weather variables like precipitation. To prevent such issues, (Reidsma *et al.*, 2007a) retained a minimum set of environmental covariates, i.e. one for temperature and one for

precipitation. We adopted a similar strategy and performed factor analysis for NDVI and TAMSAT data to select minimal inter-band correlation and maximal information content. Thus, we included only the most significant principal components for the entire time series. In this way, we allowed GWRK to avoid multicollinearity, and also to determine local anomalies. For example, NDVI revealed vegetation stress that may not be due to the lack of precipitation, but due to crop specific conditions, such as pests and diseases.

In literature, kriging (e.g. OK) is established to model crop yields at a within-field scale for precision agriculture. There is a demand for synergistic approaches to predict location-specific crop yields timely and accurately over wider regions, a yield map that in turn can be used in developing wall-to-wall agricultural information services. In this respect, to the best of our knowledge this is the first attempt to apply GWRK to predict crop yields at a regional scale and to quantify the prediction uncertainty, particularly, in West Africa. Using GWRK and the external covariate datasets, we interpolated the sorghum crop yield over a regular grid covering the country of Burkina Faso, having grid cell size equal to 1 km^2. We further operationalized this crop yield interpolation approach for spatializing a BEFM as a spatial decision support. The system will help farmers to formulate on-terroir agricultural policies, which we think can be transferred to extension systems throughout the country. We hypothesize that the developed method is of interest to decision-makers and information specialists in the agricultural domain. Approaches to quantify uncertainties in large-area crop models can help to improve the sources of uncertainty given by the sampling design, the model structure and available covariate datasets, and to increase the confidence of decision-makers by taking into account the accurately estimated prediction uncertainty.

4.5 Conclusion

This research investigates GWRK, to generate estimates of the sorghum crop yields in Burkina Faso. This hybrid approach applies geographically weighted regression (GWR) to model the local drift, and kriging to interpolate the GWR residuals.

GWRK performance is compared with ordinary kriging (OK), kriging with external drift (KED), and regression kriging (MLRK). The accuracy is compared both for prediction of the sorghum crop yields and for estimation of uncertainty surrounding those predictions. Accuracy of the crop yield prediction is evaluated using mean absolute error (MAE), mean square error (MSE), and the adjusted coefficient of determination, R_a^2. The accuracy of uncertainty estimation for those models is evaluated from the prediction error variances and the root mean square error of residuals cross validation. The results indicate that GWRK is superior to all other kriging-based approaches, with the improved values of R_a^2 equal

to 90% and RMSE equal to 71.2. Compared to KED and MLRK, both the MAE value and the prediction error variance are reduced in GWRK (480 versus 523 and 503) and (20.4 versus 38.2 and 34.1). GWRK effectively utilized information present in the external covariate datasets, improving accuracy of the sorghum crop yield predictions.

5

Investigating rural poverty and marginality in Burkina Faso using remote sensing-based products

[1]This chapter is based on: Imran, M, Stein, A., Zurita-Milla, R. (2014). Investigating rural poverty and marginality in Burkina Faso using remote sensing-based products. *International Journal of Applied Earth Observation and Geoinformation*: Volume 26, Pages 322–334.

5. Investigating rural poverty and marginality in Burkina Faso using remote sensing-based products

Abstract

Poverty at the national and sub-national level is commonly mapped on the basis of household surveys. Typical poverty metrics like the head count index are not able to identify its underlaying factors, particularly in rural economies based on subsistence agriculture. This paper relates agro-ecological marginality identified from regional and global datasets including remote sensing products like the Normalized Difference Vegetation Index (NDVI) and rainfall to terroir agricultural production and food consumption in Burkina Faso. The objective is to analyze poverty patterns and to generate fine resolution poverty map at a national scale. A new indicator was composed from a range of welfare indicators quantified from detailed household surveys of terroir communities. This resulted into a spatially varying set of welfare and poverty states. Next, a geographically weighted regression (GWR) was used to relate each welfare and poverty state to the agro-ecological marginality. This helped to show how environmental factors affect living conditions in terroir communities. We found that the poverty patterns thus obtained agreed with poverty incidence obtained from national surveys. Agro-ecological stress and related marginality vary locally between terroir communities within each region. About 58% variance in the welfare indicator is explained by the factors of terroir agricultural production and 42% is explained by the factor of food consumption. GWR exploits well the spatial variation of environmental variables to explain poverty patterns at the regional and communal level. We conclude that the spatially explicit approach based on multi-temporal remote sensing products effectively summarizes information on poverty and facilitates further interpretation of the newly developed welfare indicator.

5.1 Introduction

Subsistence farming is an important agricultural practice in many African states. For instance, in Burkina Faso approximately 92 percent of the country workforce is actively associated with the agricultural sector, of which 80 percent are small holder farmers who live in rural areas (aka terroirs) and have less than 1 ha of land (USAID, 2009). Agricultural production is largely constrained by a range of biophysical factors related to soil properties, rainfall and water availability (West *et al.*, 2008). The agro-ecological conditions vary spatially and respond to a highly local physical environment. In Burkina Faso, more than 80 percent of the total population lives in terroirs, of which 94 percent is considered poor (USAID, 2009). The lack of local infrastructure often restrains terroir households to apply sustainable farming practices since it limits the farmer's access to market and services (Alasia *et al.*, 2008; Gatzweiler *et al.*, 2011). This suggests that rural poverty in Burkina Faso can be related to the agricultural productivity and that it can be characterized from the spatial distribution of agro-ecological potential.

Traditionally, poverty as opposite to welfare is mapped by analyzing a range of socioeconomic factors obtained from targeted household surveys. Such surveys assess household capital assets, e.g. income, expenditure, food consumption, and other living conditions. Using these, indices are obtained to estimate the incidence of poverty. For example, the headcount index (HCI) is the percent of the population in an area living below an established poverty line, i.e., a normative level of income or expenditure. To extrapolate these surveys towards an entire region, various small area estimation techniques have been developed (Hoddinott & Quisumbing, 2003; Benson *et al.*, 2005). These techniques make predictions by relating the household welfare status from targeted household surveys to the household characteristics from national census, and apply the relation to households with same characteristics. A clear insight into the likely causes of the situation is often missing, because factors of marginality are not included during poverty mapping (Hyman *et al.*, 2005; Robinson *et al.*, 2007). Also, these techniques depend on the availability of national censuses that take place only once in several years due to their high operational costs.

To locate marginal areas, alternative approaches analyze environmental constraints (e.g., soil erosion, droughts) using remote sensing (RS) data and products. Being able to acquire up-to-date data over a large area by utilizing the high spatial and temporal coverage provided by RS (Parkins & MacKendrick, 2007; Alasia *et al.*, 2008), these approaches can quantify the increased susceptibility of specific areas to become marginal due to extreme events of environmental constraints. However, the environmental approaches are primarily concerned with marginality and they rarely quantify its impact on livelihood status. Following this, (Nelson *et al.*, 2012) related RS products with household level expenditure obtained from survey data to explain the poverty patterns in

Uganda. Although this approach advanced traditional poverty mapping, it is insufficient to interpret the observed relations, because a single aspect of poverty (e.g., income, expenditure, and other living conditions) is usually not enough to explain welfare and marginality, particularly in terroir economies in Burkina Faso, which is based on subsistence agriculture (Gatzweiler *et al.*, 2011).

In this paper, a geographically explicit approach is presented for studying poverty and marginality at a fine resolution and over a larger area. We investigate both agro-ecological marginality from RS-based products and welfare and marginality from household conditions. By studying these conditions over a large area, we aim at better understanding the factors that determine household marginality. In this way, this paper advances current environmental procedures of poverty mapping creating a more dynamic method that can be effectively utilized by policy-makers to reduce poverty (Nelson *et al.*, 2012).

In practice, our main objective is to use regional and global datasets including RS products for extrapolating poverty quantified from the targeted household surveys. The study is illustrated using data from Burkina Faso where agricultural surveys are collected annually targeting only representative communities countrywide. We developed a composite index from several welfare aspects observed from household surveys. This index and the RS products are used to map poverty at the national scale.

5.2 Background

5.2.1 Study area

This study is conducted using data from Burkina Faso, which is ranked among the poorest countries of the world (USAID, 2009). Agriculture contributes to 31% of the GDP and to 60% of the exports that are the main source of growth of the national economy. The livestock subsector accounts for 25% of agricultural GDP and 8% of total national GDP (USAID, 2009). Several environmental and socio-economic factors affect agricultural production like the spatial variation in both frequency and intensity of rainfall during the crop growing season (West *et al.*, 2008). Administratively, the country is divided into 13 regions and 351 districts, which are split in about 7000 terroir communities. The term terroir community refers to a group of adjacent households of agro-pastoralist farmers. One representative terroir community per district was surveyed to collect targeted household surveys (AGRISTAT, 2010). In this study, we used terroir community as the level to quantify poverty and marginality in Burkina Faso.

The head count index (HCI) is available for 1994, 1998, 2003 and 2009. In 1994, the country's first surveys for household living conditions were conducted on the basis of agro-climatic regions. Later in 1998, they shifted to the administrative regions. The HCI was compiled based

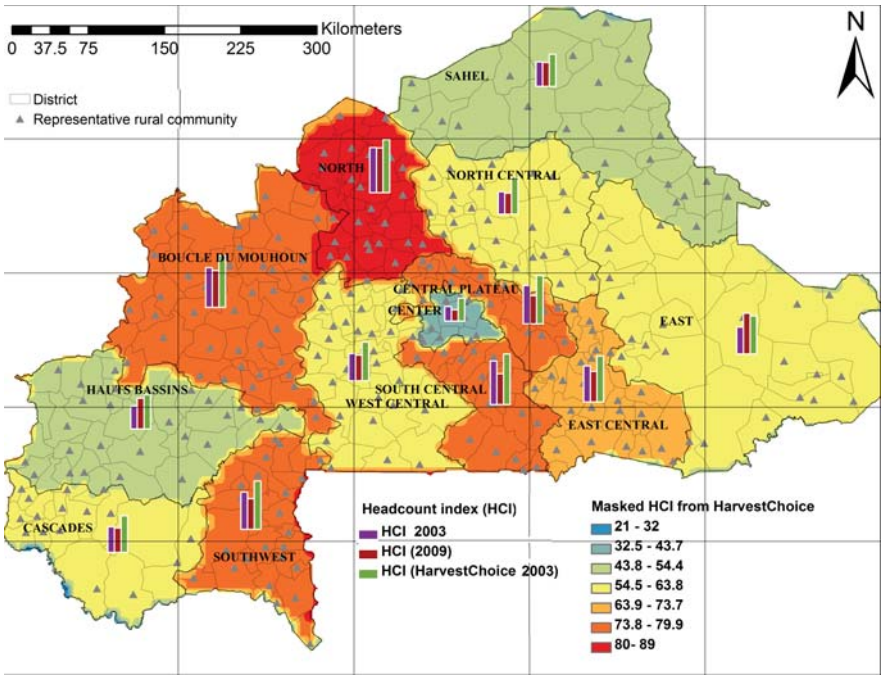


Figure 5.1 Mean Headcount index (HCI) for the 13 administrative regions of Burkina Faso, calculated from country's national surveys of 1994, 1998, 2003 and 2009; and from HarvestChoice data.

on the poverty line of 1 USD (United States Dollar) adult^{-1} day^{-1}. In 2010, HarvestChoice compiled HCI maps (gridded) from surveys carried out between 1998 and 2003 by establishing a poverty line of 1.25 USD adult^{-1} day^{-1} (Wood *et al.*, 2010). These studies consistently show that the North, South Central, Central Plateau, Boucle du Mouhoun, East Central, and Southwest regions are typically affected by poverty, with a rate of incidence well above the national average (Figure 5.1). The HCI is close to the national average in the West central, Eastern, and Cascades regions, whereas the other regions are relatively less affected by poverty.

5.2.2 Mapping communal welfare in Burkina Faso – our approach

This study defines marginality as a function of cause-effect relations between stressors and assets. Stressors are often exogenous factors that directly or indirectly affect the agricultural production of terroir communities (Alasia *et al.*, 2008). Assets are proxies to represent the impact of stressors on household welfare status. For our purposes, we focus on assets that are related to the agricultural outcomes of

terroir households obtained from field surveys. We consider a terroir community to be marginal if it encounters high stress on agricultural production, eventually resulting into low assets. By linking assets to stressors, poverty and marginality can be quantified and determining factors can be identified. In order to do so, we made the following two assumptions:

1. A high production in rural Burkina Faso helps to increase farmers welfare in that they can meet daily food requirement: poor farmers will suffer from food insecurity and food insecure farmers will be poor farmers (de Graaf *et al.*, 2001).
2. Households in a terroir community combine individual lands for cultivation and face similar agro-ecological and socioeconomic conditions (Bigman *et al.*, 1999). Farmer's poverty and marginality therefore vary considerably between the terroir communities and only to a lesser degree result into income differences between individuals within communities.

5.3 Materials and methods

5.3.1 Extraction of asset variables

Asset variables were extracted from the country's agricultural surveys carried out in 2009. AGRISTAT surveys all households in the representative terroir communities (AGRISTAT, 2010). Asset variables were obtained from data of all households belonging to a representative terroir community. In total 3540 households were surveyed to cover the 303 districts of the country. The following five asset variables were derived:

- Percentage of household members employed in farming activities (HME). Both paid and non-paid works were considered. Paid household members work in various farm and livestock activities and get wages in the form of food or cash, whereas non-paid members participate in activities without any compensation, e.g. women/child as family or collective labor.
- Agricultural production of each household (AGPROD), obtained as the projected crop grains (kg) for the current crop season. AGRISTAT asks farmers to make this projection considering the vegetative performance of the crops at the household parcel level.
- Household stocks (STOCKS) left from the previous crop season, obtained as observed crop grains (kg).
- Number of animals (e.g. bulls, donkeys) owned by each household (NA).
- Household food consumption (CONSUM), calculated as the minimum dietary energy consumption (kcal) per household member per day. AGRISTAT records number of food servings consumed by a household member in the last seven days. We considered that the

consumed food was obtained from any source, e.g. self-produced, purchased, obtained as wage compensation, or donated.

5.3.2 Developing a composite communal asset index

A weighted combination of the five chosen asset variables was made to compute a composite communal asset index (CAI). As these assets have a skewed distribution and are potentially highly correlated, the following procedure was applied.

First, we transformed the asset variables using the logarithmic function to remove skewness from the raw data. Second, to account for their different measurement units, this transformation was followed by a normalization to a common measurement scale using the Min-Max method (Ebert & Welsch, 2004):

$$I_y = \frac{y - y_{min}}{y_{max} - y_{min}}; \qquad y_{min} < y < y_{max} \tag{5.1}$$

Here I_y is the normalized variable of the log-transformed asset variable y, y_{min} and y_{max} are the minimum and maximum of y across all terroir communities. Extreme minimum and maximum values were examined for outliers in order to avoid negative effects on the subsequent analysis. Third, a minimum residual factor analysis was applied to capture non-overlapping information between the correlated asset variables (Berlage & Terweduwe, 1988). This analysis groups the asset variables according to their degree of correlation. Subsequently, an Ordinary least squares (OLS) regression was applied to adjust the eigenvalues of the correlation matrix to minimize the off diagonal residual correlation matrix. The minimum number of factors to retain for the factor analysis was decided based on Horn's parallel analysis (Horn, 1965). In this analysis, the minimum residual solution was transformed into an oblique solution using the oblimin method of rotation. Within each of these factors, all asset variables were weighted to reflect the proportion of their variance over the study area which is explained by the factor. The weights were obtained by squaring and normalizing the estimated factor loadings.

5.3.3 Extraction of stressor variables

The following stressor variables were derived from RS data:

The Normalized Difference Vegetation Index (NDVI), calculated as $(NIR - R)/(NIR + R)$, where NIR is the spectral reflectance in the near-infrared where canopy reflectance is dominant, and R is the reflectance in the red portion of the electromagnetic spectrum where chlorophyll absorbs strongly (Dorigo *et al.*, 2007). NDVI was used as a biophysical indicator of vegetation stress. For 2009, a time series of SPOT VEGETATION NDVI composite (S10) products were obtained from (VGT4Africa,

2012). This product is derived from 10-day data and mapped onto a 1 km latitude-longitude grid.

The intensity and spatial distribution of rainfall poses a significant climatic stress on the agricultural production. A series of 10-day Tropical Applications of Meteorology using Satellite (TAMSAT) images was acquired to extract the climatic stress on agricultural production for 2009. The TAMSAT rainfall estimates (RFE) have been validated for West Africa and the Sahel region using a dense rain gauge network covering area of 1° square (Grimes *et al.*, 1999). For 10-day TAMSAT RFE, 85% of the estimated and measured values agree to within 1 standard error for 1° square.

To capture the agro-ecological marginality from RS, we applied the Harmonic ANalysis of Time Series (HANTS) algorithm to the 2009 SPOT and TAMSAT image series (Verhoef, 1996). HANTS describes the main characteristics of the temporal vegetation and rainfall behaviors by considering only significant frequencies present in the time series profiles. To such frequencies, it applies a least squares curve fitting procedure based on harmonic components (sines and cosines) (Roerink *et al.*, 2000). More details on HANTS parameterization are provided in Appendix B.

In addition RS-based gridded products were analyzed as potential long-term stressors on food production:

Length of growing period (LGP, days) characterizes agro-climatic constraints that relate potential productivity of lands with the average daily temperature and surface water balance. The areas of shorter LGP bear a long-term high stress from dry conditions. LGP data obtained from (HarvestChoice, 2012) is based on 1960-1995 data from (IIASA/FAO, 2012). Data showing the degree to which soil properties exert stress on agricultural production (LASC, land areas with soil constraints) were obtained from (FAO, 2012b). Topographic data of slope (SLOPE, degrees) and elevation (ELEV, meters) were obtained from (USGS, 2012).

Besides these short and long term environmental factors, population density and market access are considered known stress factors of per capita food and agricultural production and consumption in sub-Saharan Africa (Dreschel *et al.*, 2001). We obtained population density data (PD, people per km^2) from (HarvestChoice, 2012). We calculated the market access as a Euclidean distance (MARKD, meters) from terroir communities to the major trade markets of food commodities (cereal and livestock) in Burkina Faso. Furthermore, most of people living in the northern half of Burkina Faso are agro-pastoralists. Poor households, particularly women, generally contribute labor to keep poultry and small livestock (e.g. goat, sheep) (USAID, 2009). We obtained data on poultry and small livestock (Livestock per km^2) from (HarvestChoice, 2012). All spatial data were clipped and/or resampled to a common grid of a 1 km spatial resolution.

5.3.4 Linking CAI and the stressor variables

We used Geographic weighted regression (GWR), a spatial regression technique, for which the 303 CAI values were the independent variable

and the values of the stressors were the explanatory variables. Being an extension of global regression techniques such as ordinary least square (OLS) (Fotheringham *et al.*, 2002), GWR identifies and models spatial non-stationarity, i.e., spatially varying relationships to present a significant improvement over a global regression (Leyk *et al.*, 2012). We therefore, first, computed OLS as a 'baseline' global model to test statistical significance of the coefficients for each explanatory stressor variable and to test the model residuals for spatial autocorrelation and clustering.

Let a set of observations of CAI be denoted as $CAI(s_1)$, $CAI(s_2)$, ..., $CAI(s_n)$, where s_i is location of a terroir community (i.e. representative community per district), and n is the number of observations. The global regression can be expressed as,

$$CAI(s) = \beta_0 + \sum_k \beta_k X_k(s) + \epsilon_s, \tag{5.2}$$

where β_0 is the intercept, β_k represent the estimated coefficients for explanatory stressor variables X_k, $X_k(s)$ is the value of the variable X_k at location s and ϵ_s denotes the random error term for location s. We selected stressor variables that were significant ($p < 0.05$), tested them for impact of multicollinearity on the estimation precision of regression coefficients (Kutner *et al.*, 2005), and calculated the Akaike Information Criterion (AIC) (Hurvich *et al.*, 1998). Clusters of high and low residual values at the representative community level may indicate spatial variation in the CAI relationship. We tested for significant local clusters in model residuals based on local indicators of spatial associations (LISA) (Anselin, 1995) using Rook contiguity (i.e. two representative terroir communities are neighboring if their districts share common borders, see Fig 5.1) for creating the spatial weights matrix. Moreover, we computed the bivariate LISA to check the co-variation of the value of CAI at a given representative community with the average of neighboring values of each of the selected stress factors.

Geographic weighted regression establishes separate models for each sampled location (Fotheringham *et al.*, 2002), and therefore allows for estimating locally the regression coefficients to account for spatial variation of these coefficients across a given study area (Gao *et al.*, 2012). This changes the model in Eq 5.2 to,

$$CAI(s) = \beta_0(s) + \sum_k \beta_k(s) X_k(s) + \epsilon(s), \tag{5.3}$$

where $\beta_0(s)$ and $\beta_k(s)$ represent the model local estimates of intercepts and coefficients at a location s, and $X_k(s)$ are stressor variables. The model estimates local coefficients from,

$$\hat{\beta}_{GWR}(s) = (X^T W(s) X)^{-1} X^T W(s).CAI(s), \tag{5.4}$$

where $W(s)$ is a $n \times n$ diagonal matrix of spatial weights specified with a spatial kernel function. The kernel centers on a terroir community location s, and weights the surveyed values on neighboring locations t subject to a distance decay. We used the Gaussian weighting as kernel function such that,

$$W_{st} = \exp[-0.5(d_{st}/b)^2], \tag{5.5}$$

where d_{st} is the distance between the sth and Sth terroir community locations (i.e. a spatial neighborhood), and b is the kernel bandwidth that can be specified as global, i.e., the same kernel size at each location, or as adaptive bandwidths that vary in size.

The spatially non-stationary relationships in GWR may vary according to the spatial scales of each stressor variable on CAI. This scale-dependency largely demands for finding an optimal kernel size for local estimation in GWR. For this, the most common approaches, i.e., cross-validation or minimizing the Akaike's information criterion (AIC) (Hurvich *et al.*, 1998) may not sufficiently determine an effective bandwidth for model fitting and performance (Leyk *et al.*, 2012). We tested the effect of spatial scale of local relationships on the stability and multicollinearity of GWR coefficients. By increasing stepwise the GWR adaptive bandwidth measures (0.05 = 15n, 0.1 = 30n, 0.15 = 45n, 0.2 = 60n, 0.25 = 75n, and 0.3 = 90n, where n is the number of neighboring terroir communities), we investigated the effect of increasing spatial neighborhood on local estimation. For each increment, we recorded AIC value and spatial stationarity index (Fotheringham *et al.*, 2002), which is a ratio between the interquartile range for GWR coefficients and twice the standard error (SE) of the same variables from the equivalent global model. We selected adaptive proportion of terroir communities, for which (i) the AIC score was minimum, and (ii) the variance of the stationary index for all stressor variables were larger within an effective spatial scale (Gao *et al.*, 2012).

We compared the performance of the local and global models based on AIC and adjusted-R^2 values, and performed ANOVA F-test. Moreover, Moran indexes (Moran's I) of residuals were computed to compare the ability to deal with spatial autocorrelation between OLS and GWR.

We applied GWR as a local spatial prediction model to predict CAI at unsampled locations as,

$$\hat{CAI}(s_0) = {X_0}^T.\hat{\beta}_{GWR}(s), \tag{5.6}$$

where X_0 is the vector of p stressor variables at an unsampled location s_0, $\hat{\beta}_{GWR}$ is the vector of $p + 1$ estimated drift model coefficients. We evaluated the GWR performance by using the following methods that quantify differences between the observed and predicted CAI values:

1. We computed histogram descriptive statistics to describe the observed and predicted CAI distributions.

2. We calculated mean absolute errors (MAE), mean square errors (MSE), and root mean square error (RSME) to compare the differences between the CAI observed in the AGRISTAT data and the GWR predicted CAI.
3. Although the HarvestChoice HCI data have a sufficient spatial resolution for validating the predicted CAI, these were however based on the country's 1998-2003 surveys (Wood *et al.*, 2010). Alternatively, the HCI data obtained from the country's 2009 national surveys of household living conditions were available only for the administrative regions. We used this latter choice as an independent data source for validation and calculated the CAI averages for the 13 Burkinabé regions. Furthermore, the predicted CAI values vary from 0 to 1 such that the lower index values represent a low assets level and/or a high stress level. Whereas, HCI ranges 0 to 100 such that the lower index values represent a low poverty level. We therefore transformed CAI into what we called the composite poverty index (CPI) as, $CPI(s) = (1 - CAI(s)) \times 100$, where s is a location.

5.4 Results

5.4.1 Extraction of asset variables

Table 5.1 shows average community assets (raw data) aggregated from the household survey data of representative terroir communities belonging to the 13 Burkinabé administrative regions. The asset variables AGPROD, STOCKS, CONSUM, and NA show high variation among the different regions. The average HME however is not significantly varying for the different regions and is approximately equal to the country mean (78%). The STOCKS variable shows that only a low percentage (5-10) of the total surveyed households is able to meet consumption requirements from their food stocks. The scatterplots of all with all asset variables show a high level of correlation among the asset variables (Figure 5.2). The logarithmic transformation successfully reduced the observed skewness from the asset variables, from the observed skewness in the (1.63 – 3.08) range to the (-0.09 – -0.2) range.

5.4.2 Developing a composite communal asset index

The assets variables were combined into a composite communal asset index (CAI) using a minimum residual factor analysis. The results of such an analysis are presented in Figures 5.3(a and b) and in Table 5.2. These results show that the 5 asset variables are correlated with 2 minimum residual factors with proportion variances equal to 0.33 and 0.24, respectively, thus accounting for 57% of the total variance. Using the rotated factor loadings, the asset variables were aggregated into factor-specific scores (Figure 5.3b). The first minimum residual factor (MR1)

Table 5.1 Community average assets (raw data) aggregated from the household data of 303 surveyed terroir communities belonging to the 13 Burkinabé regions.

Region	n[a]	Mean HME[b]	Mean AGRPROD[c]	Mean STOCKS[d]	Mean NA[e]	Mean CONSUM[f]
Boucle du Mouhoun	43	77.7	41420.3	179.8	40.53	1206
Cascades	15	78.2	49574.5	266.6	33.66	1545.8
Center	9	84.1	10900.8	147.3	17.30	574.33
East Central	23	75.3	23352.7	243.7	46.00	1018.3
North Central	27	67.6	14748.6	161.6	31.50	639.1
West Central	30	82.1	32836.2	295	35.00	1190.7
South Central	17	81.5	16984.1	250	43.20	977.1
East	22	78.2	26248.4	198.8	50.90	1189.3
Hauts Bassins	28	76.3	54198.6	193	35.10	1326.5
North	25	76.2	21212.1	302.2	32.70	988
Central Plateau	18	76.4	21940.5	336.8	43.80	741.5
Sahel	22	77.9	11629.7	50.4	34.90	1000.2
Southwest	24	81.4	46052.7	194.9	38.80	1199.8

[a]Number of surveyed terroir communities in region.
[b]Household members employed in farming activities (%).
[c]Crop production (kg of grains) of households for the current crop season.
[d]Household stocks (kg of grains) left from the previous crop season.
[e]Number of animal owned by household.
[f]Minimum dietary energy consumption (kcal) per household member per day.

Table 5.2 Rotated factor loadings and factor-specific scores for individual assets in the composite asset index (CAI)

Interpretation	Factor 1 Crop and livestock production		Factor 2 Food consumption	
Variables of individuals assets	Factor loadings	Weights[a]	Factor loadings	Weights
STOCKS[b]	0.81	0.42	-0.12	0.01
NA[c]	0.74	0.35	0.13	0.02
AGRPROD[d]	0.56	0.20	0.27	0.07
HME[e]	0.18	0.03	-0.12	0.01
CONSUM[f]	0.02	0	0.98	0.88
Weight of factors in CAI[g]		0.58		0.42
Selection criteria:				
Eigenvalues		1.67		1.20
Test-statistics:				
	Chi-square			$p < 0.05$

[a]Normalized squared factor loadings.
[b]Household stocks (kg of grains) left from the previous crop season.
[c]Number of animal owned by households.
[d]Crop production (kg of grains) of households for the current crop season.
[e]Household members employed in farming activities (%).
[f]Minimum dietary energy consumption (kcal) per household member per day.
[g]Normalized sum of squared factor loadings.

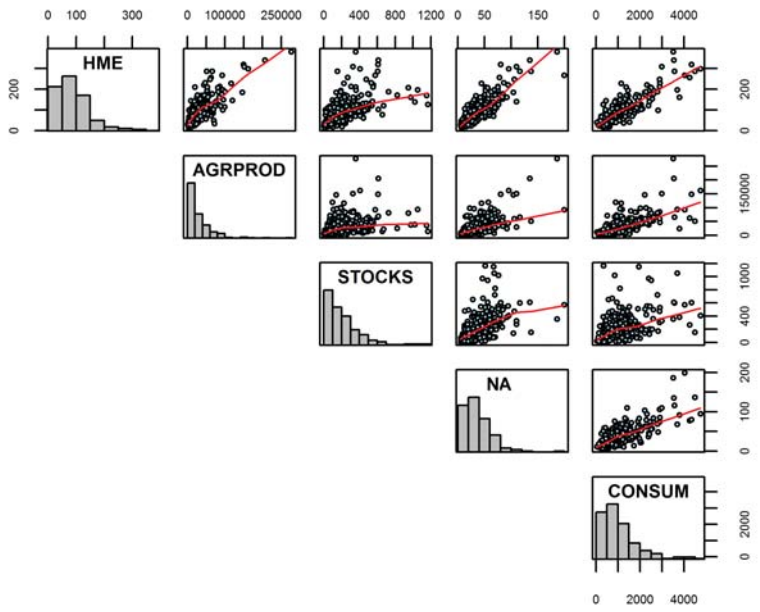


Figure 5.2 AGRISTAT data by household assets: household members employed (HME), households crop production (AGRPROD), household stocks (STOCKS), number of animal owned by household (NA), and minimum dietary energy consumption (kcal) per household member per day (CONSUM).

has loadings on the asset variables that present patterns of crop and livestock production, whereas rotated factor loadings from the second minimum residual factor (MR2) are projected mainly on the household food consumption patterns across the terroir communities. The squared factor loadings represent the proportion of the total unit variance of the assets which is explained by the factor. MR1 accounted for 42%, 35%, and 20% of the variance in the values of STOCKS, NA, and AGRPROD assets, whereas MR2 accounted for 88% and 0.01% of the variance in the CONSUM and HME values, respectively. A small contribution of HME can also be justified on the bases of summary statistics of raw data in Table 5.1, showing a low variation of the asset variable over the entire country. The resulting factor-specific scores are aggregated into the CAI by weighting each factor according to its relative contribution to the overall variance with MR1 and MR2 explaining the 58% and 42%,

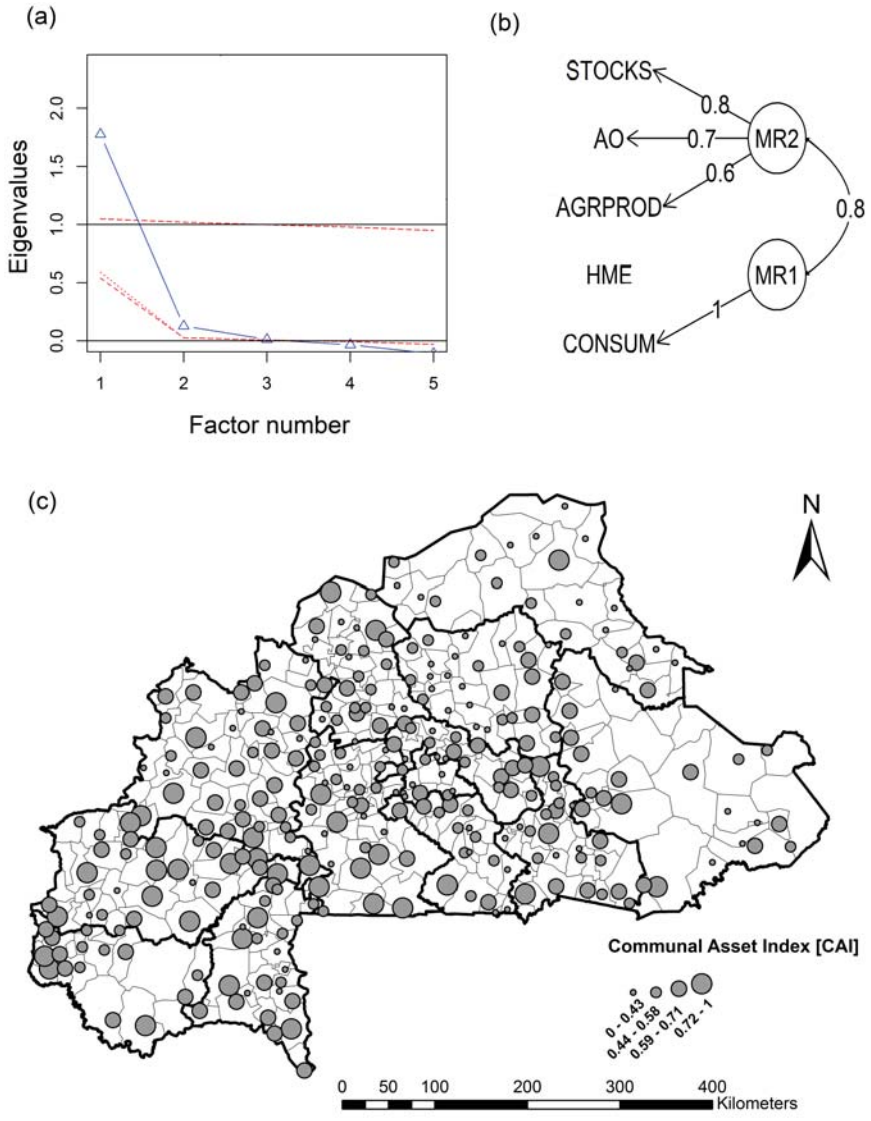


Figure 5.3 Minimum residual factor analysis – (a) eigenvalues (on vertical axes) express the proportion of the total variance in the data explained by each factor, and (b) Minimum residual factors (MR1 and MR2) standardized values of the individual assets multiplied by their individual weights – (c) spatial distribution of the composite asset index (CAI) observations at 303 surveyed terroir communities.

respectively. Crop and livestock production obtained a slightly higher weight than food consumption.

Figure 5.3(c) shows the spatial distribution of CAI. For cartographic purposes, the resulting CAI values were classified into four welfare intervals. In general, marginal communities fall within the first two levels where the first level (CAI=0–0.43) represents a severe marginality and the second level (CAI=0.43–0.59) represents a high marginality, whereas the next two levels (CAI=0.58–0.71) and (CAI=0.71–0.98) represent low marginal and high welfare communities, respectively. High marginal communities mostly correspond to the regions of Boucle du Mouhaun, North, North Central, South Central, East Central, Center, Central Plateau, and Sahel. In the Cascades and Haut Bassins areas, most communities fall within the low marginal range, whereas the Southwest, West Central, and Eastern regions have both low and high marginal communities (see Figure 5.1 for the names of the regions).

5.4.3 Extraction of stressor variables

Only RS-based products were processed with HANTS since we are interested in obtaining main characteristics presented in these time profiles. Figure 5.4 shows the output of HANTS algorithm that contributed significantly to explain CAI. The HANTS algorithm reduced the data from 36 decadal images of an image series to 7 amplitude-phase datasets, i.e., single amplitude and phase images for each frequency, where the zero-frequency (mean) is without a phase. Based on RFE phase datasets we differentiated the three seasons: wet season May-September (78 – 179), post-wet season October-November (77 – 129), and dry season December-April (1 – 51). High inter-season difference of rainfall indicates an extreme dry period for vegetation during which households usually depend on stocks. The RFE amplitude datasets showed a high spatial variability of rainfall intensity, with a low and declining rainfall in the North as compared to the higher but more homogeneous rainfall in the South. Consequently, the mean NDVI signal (Figure 5.4a) shows a North-South directed increasing trend of vegetation performance. Given the limited use of irrigation in Burkina Faso, the northern communities have low potential for household food production and stocks.

5.4.4 Linking CAI and the stressor variables

Table 5.3 shows results from the global OLS model using the stressor variables that significantly ($p=0.05$ to $p<0.0001$) contributed to explaining the CAI variation. Agro-ecological stressor variables, both short-term (i.e. NDVI, RFE during the 2009 crop growing season) and long-term (i.e. LGP, LASC, SLOPE) consistently showed a significant agro-ecological stress ($p < 0.05$ to $p < 0.001$) on the agricultural production potential in Burkina Faso. We observed no significant impact of multicollinearity (i.e. VIF ≤ 5 for all the stressor variables). We found highly significant global spatial autocorrelation in the model residuals (Moran's I = 0.22;

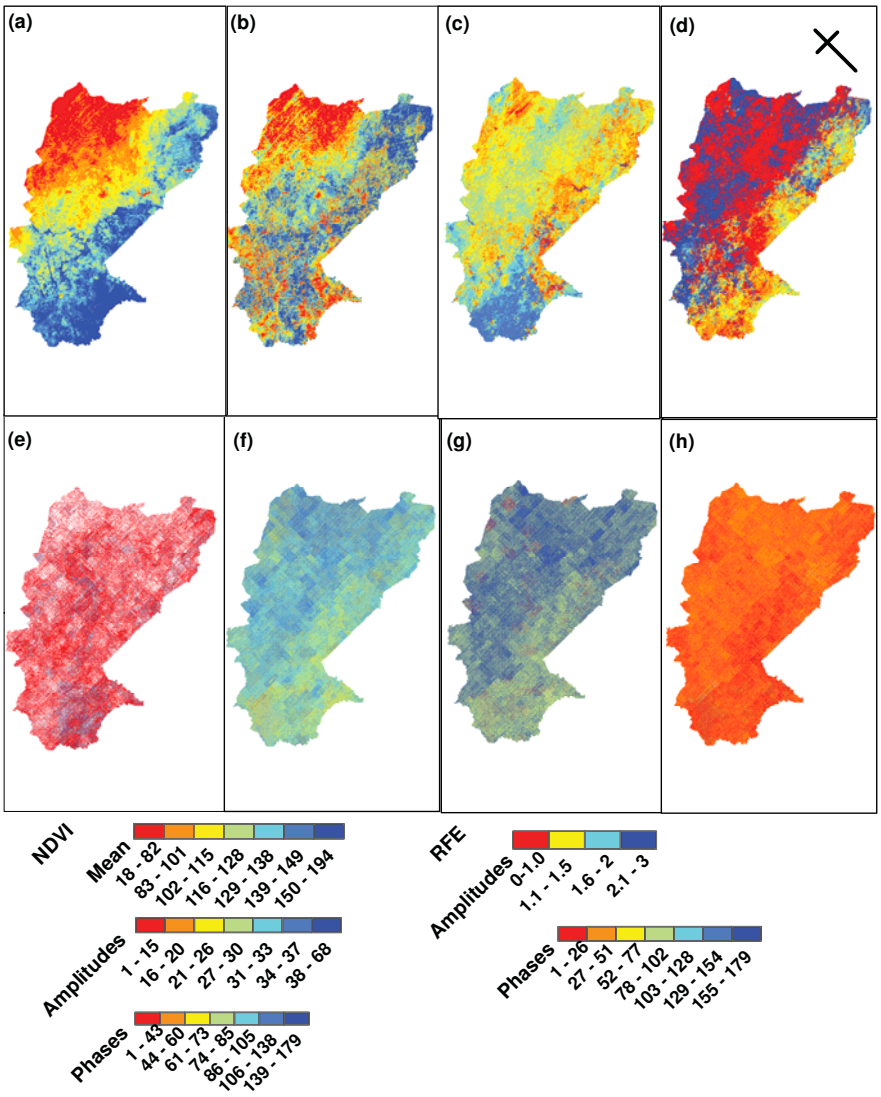


Figure 5.4 Output of HANTS algorithm applied to the Normalized Difference Vegetation Index (NDVI) image series – (a) mean; (b) first amplitude; (c) second phase; (d) third phase, and the Tropical Applications of Meteorology using Satellite (TAMSAT) image series – (e) third amplitude; (f) first phase; (g) second phase; and (f) third phase.

Table 5.3 Properties of the global and local estimates of stressor variables to explain composite asset index (CAI) using ordinary least square (OLS) and geographically weighted regression (GWR).

Parameters	(r, p)[a]	OLS Ests.[b]	OLS Std. Error[c]	Significance	VIF[d]	GWR Local Est. Range[e]
Intercept		0.476^{+0}	0.122^{+0}	0.0001	-	$+0.411^{+0}$ - $+0.756^{+0}$
NDVI (mean)	(0.283, <0.0001)	-0.242^{-2}	0.846^{-3}	0.001	5	-0.454^{-2} - -0.168^{-2}
NDVI (amplitude 1)	(−0.033, 0.1)	-0.244^{-2}	0.106^{-2}	0.01	1.21	-0.334^{-2} - -0.105^{-2}
NDVI (phase 2)	(0.201 , <0.001)	0.125^{-2}	0.498^{-3}	0.01	1.15	$+0.510^{-3}$ - $+0.360^{-2}$
NDVI (phase 3)	(0.148 , <0.001)	0.235^{-3}	0.133^{-3}	0.05	1.11	$+0.110^{-4}$ - $+0.338^{-3}$
RFE (amplitude 2)	(0.061, 0.1)	0.168^{+0}	0.884^{-1}	0.05	1.38	$+0.115^{-1}$ - $+0.209^{+0}$
RFE (amplitude 3)	(−0.091, 0.1)	-0.113^{+0}	0.431^{-1}	0.001	1.4	-0.180^{+0} - -0.850^{-1}
LGP	(0.332, <0.0001)	0.304^{-2}	0.714^{-3}	<0.0001	5	$+0.137^{-2}$ - $+0.444^{-2}$
LASC	(−0.265, <0.0001)	-0.357^{-1}	0.160^{-1}	0.01	2.27	-0.665^{-1} - -0.189^{-1}
SLOPE	(0.119 , 0.01)	0.525^{-3}	0.272^{-3}	0.05	1.11	$+0.287^{-3}$ - $+0.920^{-3}$
TPD	(−0.116, 0.01)	-0.192^{-2}	0.100^{-2}	0.05	1.38	-0.427^{-2} - $+0.390^{-3}$
LIVESTOCK	(−0.159, 0.001)	-0.607^{-2}	0.137^{-2}	<0.0001	1.56	-0.797^{-2} - -0.257^{-2}
MARKD	(−0.018, 0.1)	-0.820^{-6}	0.450^{-6}	0.05	1.15	-0.218^{-5} - -0.926^{-6}

[a]Pearson correlations between the CAI and the stressor variables of Normalized Difference Vegetation Index (NDVI), rainfall estimates (RFE), length of growing period (LGP), land areas with soil constraints (LASC), total population density (TPD), poultry and small livestock (LIVESTOCK), and distance to major trade markets (MARKD).
[b]Parameter estimates from OLS.
[c]Standard error.
[d]Variance Inflation Factor (VIF).
[e]Inter-quartile range of GWR local coefficients.

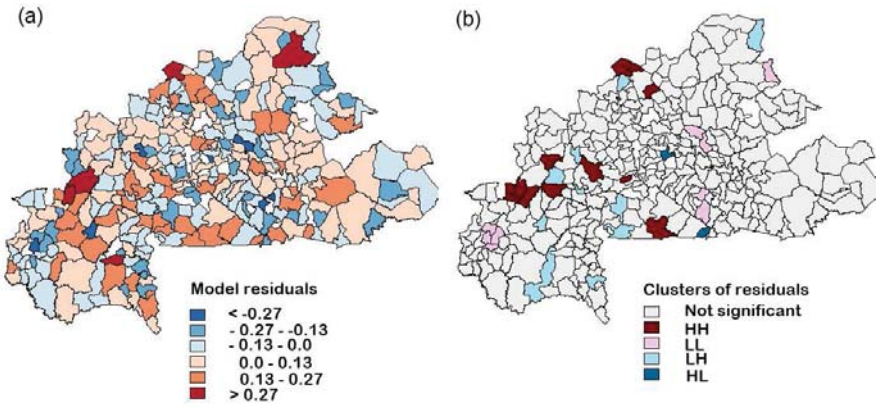


Figure 5.5 (a) Spatial distribution of Ordinary Least Square (OLS) residuals – (b) statistically significant local clusters of model residuals based on LISA (HH – high values; LL – low values; HL and LH – outliers).

$p < 0.001$) (Figure 5.5a). We also observed significant local clusters of low and high model residuals based on LISA – suggesting statistically significant clusters of over- and underestimations in the South, Center and in the North Burkina Faso, respectively (Figure 5.5b).

The LISA maps in Figure 5.6(a–f) illustrate spatial associations between CAI and the six selected stressor variables, NDVI, RFE, SLOPE, TPD, LIVESTOCK, and MARKD. We found statistically significant clusters of high CAI and high NDVI values, in the neighboring terroir communities in southern districts of Burkina Faso, and clusters of low CAI and low NDVI values in the neighboring terroir communities in the Sahel and Central North districts (Figure 5.6a). However, in these districts and also in the East, we observed a considerable number of clusters of low CAI and low values of livestock in the neighboring terroir communities (Figure 5.6e), as well as clusters of high CAI and surrounding high values of livestock in the Center and Central Plateau districts. Similarly, there are significant clusters of high CAI and high rainfall (Figure 5.6b). Significant clusters of high CAI and high TPD in the southwest (Figure 5.6d) can be observed. We also observed statistically significant clusters of low CAI and high MARKD, and high CAI and low MARKD in neighboring terroir communities in the East and Central East districts (Figure 5.6f).

We investigated the effect of different adaptive bandwidths (proceeding with proportions from 0.3 to 0.05) on local estimation in GWR. We observed that the variations in the stationary index for all predictors were

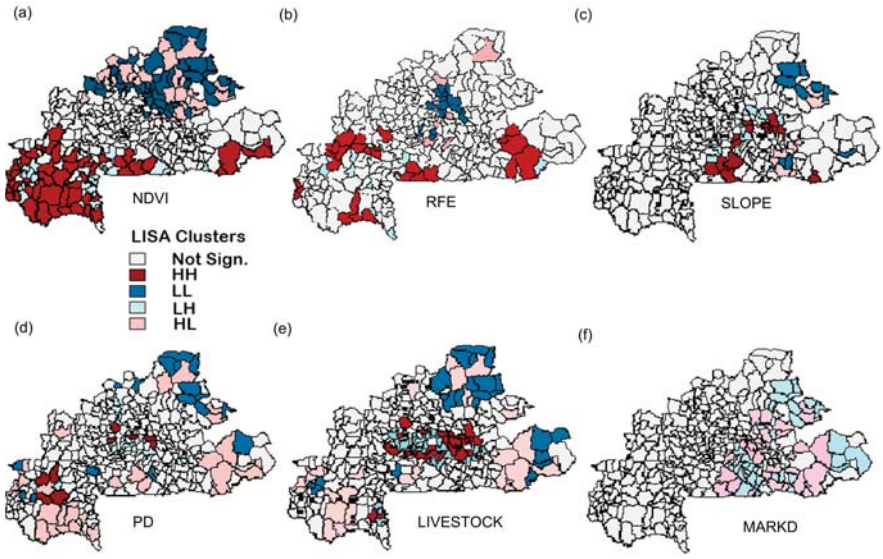


Figure 5.6 Statistically significant ($p < 0.001$) spatial clusters from a bivariate LISA analysis: using composite communal asset index (CAI) and (a) NDVI, (b) rainfall, (c) length of growing period (LGP), (d) population density (PD), (e) poultry and small livestock (LIVESTOCK), and (f) market distance (MARKD) (HH high–high values; LL low–low values; HL and LH – outliers; first letter indicates CAI, second one the stress factor).

larger for small bandwidths (proportions of 0.1 and 0.05). The stationary index became quite flat with increasing bandwidths (for greater than 0.1). At the smallest adaptive bandwidth (proportion of 0.05, i.e., on average 15 neighboring terroir communities), we observed that, compared to the global model, the AIC value of GWR decreased from 410 to the highest minimum value of 340 and, the $R^2{}_a$ value increased from 0.24 to highest maximum value of 0.50. The ANOVA F-test suggests that GWR was a significant improvement (p = 0.01) over the global model. Moreover, we observed a decrease in Moran's I value of GWR residuals close to zero (-0.04).

The maps of local coefficients from the GWR models for NDVI (mean), RFE (amplitude 3), SLOPE, TPD, LIVESTOCK, and MARKD variables are shown in (Figures 5.7a-f). We observed high local variability of these coefficients in the study area. For example, Table 5.3 shows that a significant negative correlation exists between CAI and TPD (r = 0.116, p = 0.01). Figure 5.7(d) however shows that both negative and positive correlations occur in the study area. Stronger negative correlations show that a decrease in the population density may cause a higher increase in

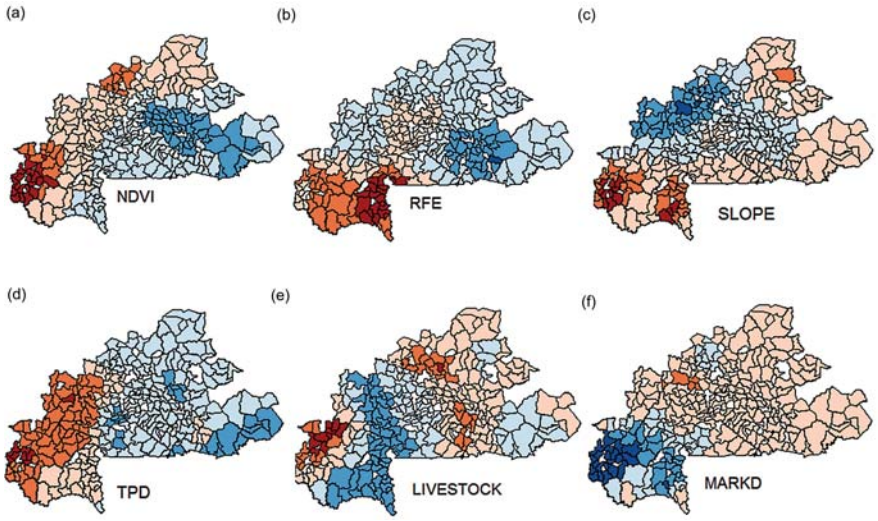


Figure 5.7 Classification of the Geographically Weighted Regression (GWR) coefficients for communal asset index (CAI) using proportion of terroir communities (adaptive bandwidth = 0.05). Light blue = Min; Dark brown = Max. Using six natural class breaks on the GWR coefficient values ranges in Table 5.3.

CAI in the North, Centre, and in the East of study area. While positive correlations are mainly located in the South and southwest of the study area. Similarly other spatially varying local coefficients show the spatial non-stationarity of the relationship between CAI and related stressor variables at 303 surveyed locations.

The GWR predicted CAI (Figure 5.8) show less marginal terroir communities in the southern half of the country as compared to in the eastern and northern regions. This indicates that the poverty remains pronounced in the North, South Central, Central Plateau, Boucle du Mouhaun, and East Central (Figure 5.1), while the terroir communities in the Center and in the Sahel regions become more poor. In these regions, the predicted CAI belonged to the high marginality range (i.e. CAI=0.43–0.58). RS-based products (Figure 5.5) revealed high agro-ecological stress in these regions. In Cascades and Haut Bassins, the predicted CAI fell within the low marginal range (i.e. CAI=0.59–0.71). While both low and high marginal terroir communities can be found in the Southwest, West Central, and in the East regions.

Table 5.4 compares the accuracy of the observed and the GWR predicted CAI. Compared to the histogram of the observed CAI (i.e. based on AGRISTAT data), GWR slightly overestimated the minimum values and underestimated the maximum values. For each of the Burkinabé regions,

Table 5.4 Histogram statistics, mean absolute errors (MAE), mean square errors (MSE), and root mean square error (RSME) to compare the differences between the original and the predicted composite asset index (CAI) using Geographically Weighted Regression (GWR).

Model	n	Min[a]	1[st] Q[b]	Med[c]	3[rd] Q	Max[d]	Mean	MAE	MSE	RMSE
Observed	303	0.177	0.426	0.547	0.665	0.983	0.549	-	-	-
GWR	273151	0.342	0.491	0.557	0.613	0.794	0.550	0.139	0.0283	0.168

[a]Minimum.
[b]Quartile.
[c]Median.
[d]Maximum.

Table 5.5 Comparisons of the average Communal Poverty Index (CPI) with the Headcount Index (HCI) in 13 regions of Burkina Faso.

Region	Average HCI (1998-2009)	HCI (2009)[a]	Mean CPI ($CPI = (1 - CAI)X100$)[b]	
			Observed[c]	Predicted (GWR)
Boucle du Mouhoun	55.1	56	43	46.2
Cascades	37.1	37.3	37.9	33.1
Center	18.7	17.3	59.6	57.2
East Central	50.9	46.6	44.8	45.1
North Central	41.3	31.9	54.2	49.5
West Central	41.7	38.8	44.3	45.9
South Central	57.1	46.7	45.9	45.2
East	49.9	62.2	43.4	44.7
Hauts Bassins	38.2	46.8	40.3	39.7
North	65.8	68.1	46.2	52.2
Central Plateau	50.5	42.9	42.6	47.7
Sahel	38.5	36.6	55.9	51.9
Southwest	49.4	46.8	41.9	39.7

[a]HCI obtained from the 2009 national surveys of household living conditions.
[b]Regional means of CPI based on the composite asset index (CAI) predictions from Geographically weighted regression (GWR).
[c]Regional means of CPI based on the CAI observed in the AGRISTAT data.

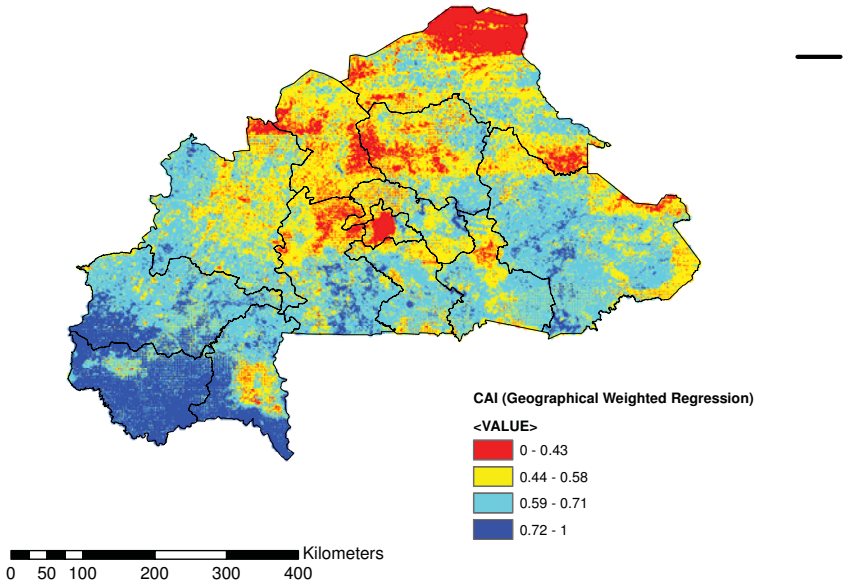


Figure 5.8 Interpolated communal asset index (CAI) using Geographically Weighted Regression (GWR).

Table 5.5 presents a comparison between the average CPI computed from the predicted CAI and the average HCI obtained from 2009 national surveys. No significant difference is observed between the two indices. CPI was lower for the North region (21%) and higher for the Sahel (24%), Central (70%) regions.

5.5 Discussion

To investigate terroir welfare in Burkina Faso, we composed an index, CAI, from variables representing household assets in the 2009 AGRISTAT survey data. Weights were assigned to the correlated asset variables by using a sound statistical procedure, so that CAI reflects the contribution of each individual asset. It was shown that the CAI effectively characterizes the differing welfare levels of terroir communities. The CAI observations at surveyed community locations are related to the collocated values of stressor variables. By exploring these relationships locally, the geographically weighted regression (GWR) gave a sufficiently varying measure of poverty and marginality of terroir communities.

Well-justified variables are used to create the CAI so that it can provide a strong logical base for poverty mapping. For this we performed both an extensive review on poverty mapping in West Africa and an in-depth assessment of existing poverty patterns in the country's household survey data. We observed that poor households in rural Burkina Faso often have marginal food production (i.e. insufficient to meet their consumption requirements). This is also reflected in the AGRISTAT data where food insecure households consistently fail to attain an adequate cereal production for food consumption (AGRISTAT, 2010). To compose the CAI we, therefore, selected asset variables that were directly related to household food production and consumption. The factor analysis confirmed our choices as the Chi-square significance ($p < 0.05$) suggested that the common factors can sufficiently explain the intercorrelations among the variables included in the analysis. This analysis also showed that the first factor of CAI has loadings on the household asset variables, including expected cereal production, cereal stocks, and number of animals owned. Obviously, these variables can be related to the levels of household food production in the study area. Also, the variance of second CAI factor was significantly (88%) explained by the asset variable on household food consumption.

The calculation of the CAI required a logarithmic transformation of the asset data, followed by a Min-Max transformation to the [0,1] interval. In this way we could arrive at a common measurement scale not affected by the units of each asset variable. To justify this, we explored several other methods as well. Those included ranking, standardizing the data towards values with zero mean and a standard deviation of one, and use of a categorical scale level. The Min-Max proved the most robust one in terms of taking into account the data properties and being closest to a normal distribution.

The CAI is related to various regional and global spatial datasets from a range of domains, including (i) RS-based products depicting agro-ecological stresses related to weather, soil and topography, (ii) maps of urban, rural and total population densities, and (iii) maps showing distance to markets and travel time maps indicating degree of geographic accessibility to urban cities of population size 20, 50, 100, and 250 thousands. Spatial variation of the CAI in terroir communities is, however, mainly affected by the variables belonging to the agro-ecological domain. These variables indicate a strong environmental stress on the households' food production potential. Because of this stress, the rainfed cereal production is destined mainly for household consumption, with only 10-20 % of the cereals brought to market (USAID, 2009). For the arid and semiarid regions in the North, a highly positive local relation between CAI and livestock (Figure 5.3e) indicates that marginal communities tend to have more livestock to counter less favorable agro-ecological conditions.

The proposed CAI has strong local correlation with the stressors from RS-based products. Spatial prediction models incorporate local dependence into the process of prediction, for which GWR tends to calibrate local models on each surveyed location. The GWR technique has been

widely used to investigate spatial non-stationary present in geographical relationships (Fotheringham *et al.*, 2002). Local coefficient estimation in GWR largely depends on selecting an optimal kernel bandwidth (Gao *et al.*, 2012; Leyk *et al.*, 2012). Furthermore, the GWR technique is more sensitive to the effect of multicollinearity (Wheeler & Tiefelsdorf, 2005) than the OLS regression. A careful use of diagnostic statistics was, therefore, accomplished to confirm reliable GWR experimentation. Both the diagnostic testing and the visual interpretation of local coefficient distributions in study area showed significant spatial non-stationarity of the stressor variables.

The reason of positive GWR relationships between CAI and TPD may be that a higher population density is usually related to more labor availability. Areas having a higher rate of arable land growth often face labor constraints (West *et al.*, 2008). About 18% of the country's 46% arable land is concentrated in the West and South of the study area (USAID, 2009). Negative GWR relationships between CAI and population density may be attributed to a relatively high population growth as compared to increase in cultivation land per household, which often exerts greatest pressure on land nearby urban centers, and in the Central Plateau. The negative correlation of MARKD with CAI means that a decrease in the distance to market is associated with an increase in welfare. One possible explanation for this negative relationship is that distant locations are usually accompanied with a decline of community's ability to access food from markets. Particularly in livestock-dominant eastern and northern regions, a low cereal production level and being far from market together put a terroir community at risk of getting high food prices (USAID, 2009).

Compared to the HCI in the Center, a high CPI average is observed in some terroir communities nearby the urban areas. This indicates that some non-farm activities are contributing heavily to the communal income and employment. Slightly above-average CPI is observed in Sahel. It is a livestock dominant region because lack of favorable agro-ecological conditions often poses a low cereal production. Slightly below-average CPI value in the North region may be due to the fact that the welfare in CAI is evaluated in the view of households food production potential based on both the current cereal stocks and the expected cereal production. Whereas, HCI calculated from the national surveys indicates proportion of household whose current expenditure level is under the established poverty line (i.e. 1 USD adult^{-1} day^{-1}).

In addition to the link to agricultural and livestock production system, further experimentation with this approach to poverty extrapolation is needed. This should include the following: (i) evaluating welfare variables that represent non-farm activities, e.g. fishery, handicrafts, mining, particularly for communities residing near the urban areas or for considering terroir employment during the dry season, (ii) other stresses to food productivity like limited access to inputs, credit, and land, and (iii) evaluating the effect of food utilization like limited access to adequate health services, potable water and sanitation.

We intend to use this study output in our ongoing research to spatial-

ize the bio-economic farm model (BEFM) over large areas in West Africa. This BEFM focusses on Burkinabé subsistence terroir communities and helps formulate sustainable farm policies, by effectively associating the production potential of their land parcels and their marginality status and food consumption requirement. The marginality assessment is necessary to further asses the communities capability of applying modern inputs like fertilizers, pest control, and crop varieties. Serving this input, the poverty and marginality maps will allow parameterizing the BEFM for all 7000 terroir communities in Burkina Faso. Further extending this approach to the whole of West Africa will require a careful selection of asset variables to compose welfare index for terroir communities in the entire region.

5.6 Conclusion

This study shows the performance of the composite asset index (CAI) for poverty mapping in Burkina Faso. The index replaces common indices, and is based on survey data. CAI is interpolated and mapped towards the entire country by using the RS-derived external covariates that represent agro-ecological stress.

The study shows that 58% of the variance of CAI is explained by the factor representing variables of food production and 42% is explained by the factor representing variables of food consumption. The composite asset index thus well represents the variation of welfare and marginality in terroir communities. This variation is significantly explained by the stressor variables of NDVI, rainfall, length of growing period, soil nutrients, and topography. Spatial dependency between CAI and the stressor variables is incorporated into a geographically weighted regression (GWR) model that is able to identify areas where poor agro-ecological conditions constrain terroir communities from attaining an adequate level of welfare.

We conclude that level of household food production and consumption is directly related to a welfare and poverty level in the Burkinabé terroir communities. Relationship between CAI and the stressor variables varies considerably between the terroir communities. The composite poverty index (CPI) based on the predicted CAI showed similar patterns as compared to the commonly applied Headcount Index. We thus conclude that agro-ecological marginality and poverty incidence are positively related in terroir areas of Burkina Faso.

Timely, cost-effective, and fine resolution poverty maps of CAI were generated targeting terroir areas. These maps can be applied for decision making related to food security and poverty. This study has thus highlighted the potential of the proposed method to identify causes of poverty that may help formulate better policies.

6 Reflections, Conclusions and Further Recommendations

Data scarcity poses obstacles to apply agricultural models at regional scales, and, thus, to develop location based wall to wall services. In particular, in developing countries like Burkina Faso, lack of spatially-explicit data demands upscaling a range of variables from the existing datasets such as ground surveys or remote sensing (RS) products. Statistical models can be applied for upscaling variables estimated at the field/household level to farm and even to whole countries. However, spatial upscaling requires including various environmental and socioeconomic factors that drive spatial variability of the variables over large heterogeneous areas. Spatial statistics can be applied to upscale the variable estimates to regional scale through modeling their high spatial variability. Moreover, spatial data infrastructure (SDI) technology can be used for linking the existing databases to models, but its deployment requires the models to be adapted for spatially-explicit datasets and upscaling procedures. Focusing on these issues, this research is undertaken:

1. To investigate SDI technology to propose a flexible framework to link spatial upscaling to simulation models at regional scale for deploying wall to wall services.
2. To model the relationship between the observed crop yields and their collocated explanatory variables at the terroir level and to upscale the yield estimates to the national-scale of Burkina Faso.
3. To model uncertainty in the regional modeling and upscaling of crop yields in Burkina Faso.
4. To model the farmers welfare and marginality status at the terroir level using targeted household surveys, and to investigate regional and global datasets including RS products for upscaling the terroir-level marginality estimates to the national-scale of Burkina Faso.

Methods from spatial data infrastructure (SDI) technology were used for providing a flexible environment for integration and spatial statistics were applied for upscaling. This research resulted into the following reflections, conclusions and further recommendations for the different research objectives.

6.1 Reflections

6.1.1 Reflection on the use of SDI technology for data and model integration

A particular objective of this study is to propose a flexible framework to link upscaling models to farm simulation models at regional scale. Regarding this, the study explores how such a framework can take benefit from the SDI base technology, and how the regional models can be adapted for wall to wall SDI services. By model adaptation as a wall to wall service, we mean

1. the model is a web service which can essentially be instantiated for any terroir location in the country.
2. the model web service can be initialized with the orchestration of other web services for spatially-explicit data, either upscaled with the spatial statistical models (e.g. regional models of crop yields and farmers marginally) or provided by third-party datasets (e.g. ground surveys, RS data).
3. the framework system is robust for exposing quality of data and model services through metedata, and, thus, it allows for services discovery and binding for a terroir location.

Deploying such wall to wall services is challenging, particularly for a variety of data and models at regional scale. To do so, several methods from SDI base technology are explored, including geospatial metadata, standards, semantics, visualization, and web mapping (i.e. catalogues). Based on this, a framework system is proposed (cf. Chapter 2) to achieve an adequate level of technical and conceptual interoperability in data and models integration as,

6.1.1.1 Technical integration

To overcome the technical barriers, it is proposed to implement open geospatial consortium (OGC) standard interfaces and data encoding. This research demonstrates the use of OGC standard interfaces for deploying data and models as geospatial web services. The models deployed with OGC web processing service (WPS) standard and the datasets with OGC web feature service (WFS) standard can expose their quality (i.e. fitness-for-use) through metadata, which was found useful to discover and integrate appropriate resources in providing wall to wall agricultural services. The use of OGC web services and of a lightweight web browser interface is adequate for linking data and model web services published in the catalogues. This interface is user-friendly so that even less expert end-users can easily discover and link web services for data and models and devise effective farming practices.

The model web service deployed in this study is static, e.g., the model results represent the equilibrium situation for a single crop season

in a year in Burkina Faso. Consequently, the linking of model web services with other services may be relatively straight forward. But, simulations may require linking outcomes across many crop seasons in a year. Such interdependent outcomes may not be well encapsulated by the OGC web standards as for instance in OpenMI framework for integrating legacy models. The OpenMI framework, however, causes dependencies on framework-specific libraries that may be difficult to resolve when using the model elsewhere. This kind of dependency can be overcome by publishing model components following the publish-find-bind paradigm of service oriented architecture (i.e. based on XML standards like UDDI or ebXML), as adapted in current SDIs, for example, in SOA-based INSPIRE architecture. Moreover, the applicability of the OGC and OpenMI standards can be used in conjunction for service oriented environmental modeling, to overcome limitations of the two, as adopted in (Castronova *et al.*, 2013).

6.1.1.2 Conceptual integration

A paramount obstacle to provide wall to wall services is the conceptual and semantic mismatching of data and models. To overcome this, it is proposed to design an integrated conceptual schema. This, on the one hand, essentially conceptualizes concepts in the third-party databases. Whereas on the other hand, it explicates the parametric space of models. The target schema can reflect the model formalisms that express model inputs/parameters and their associations to model assumptions. The classes and their relationships (i.e. terminology) in the integrated conceptual schema are proposed to be derived from a shared ontology. Following the target integrated schema, the database schemas can be transformed, and the results can be materialized into an integrated database or can be linked to a model through deploying a geospatial data service. The established mapping of source and target schema in this setting can provide the bases for conceptual and semantic integration and models in an SDI environment.

6.1.1.3 Transformation services

For deploying wall to wall services, the SDI-based framework proposes to make models spatially-aware at regional scale and to achieve geospatial data and models integration loosely through OGC-compliant wrapper implementations. National SDIs however mostly adopt a top-down approach for sharing spatial datasets to a broad range of applications, whereas, a specific application or end-user needs are unforeseen. Therefore, SDIs often provide the implementation specifications for transformation services to bring diverse spatial datasets published in SDI catalogs to the requirements of applications (e.g. spatial analysis unit, level of detail). In this regards, this research discusses the prospect for transformation

services, in the context of adapting regional models of crop yield and marginality and sharing survey data to wall to wall services. Based on these implementations, two types of transformations were identified: scale-related and schema-related. The scale-related transformations can be accomplished through applying spatial statistical models, such as the models applied for upscaling crop yield and marginality to regional scale (cf. Chapters 3 and 5). The schema-related transformations concern to spatial mapping from source schema to target schema in the thematic attribute space, e.g., spatially aggregating all household members in a terroir to compute available family labor (i.e. management data). Schema-related transformations can be accomplished through schema mapping operators, such as designed in Appendices D. In SDI framework, these operators can be deployed through GIS functionality offered by spatial database systems (e.g. PostGIS) or by geospatial web processing service (i.e. OGC WPS).

With the provision of versatile transformation services in a SDI-based framework for environmental applications, an important issue is to communicate uncertainty/error information through the participant Geospatial services. This is particularly the case of scale-related transformations that use spatial upscaling procedures. The quantified uncertainty in upscaling (cf. Chapter 4) should be communicated when a geospatial data service is linked to the participant model service, particularly if the model input is sensitive to uncertainty.

6.1.2 Reflection on the use of spatial statistics for upscaling and modeling

6.1.2.1 A data quality perspective

Applying spatial upscaling to crop yield and marginality is challenging, particularly from field/household to terroir level and to the national-scale of Burkina Faso, because

- lack of sufficient data on the factors of spatial variability in crop yield and marginality.
- complex locally-varying farming and cropping conditions.
- uncertainties inherent in the spatial data and/or in the spatial upscaling, and/or in their integration.

For the first issue, national statistical databases such as AGRISTAT in Burkina Faso provide field observations of a range of biophysical and socioeconomic variables. However, due to high operational costs, the measurements are taken for selected georeferenced locations at a relatively coarse spatial and temporal resolution. In this research, we used AGRISTAT data from the field surveys conducted in 2009. These surveys were conducted in 351 representative terroirs out of the approximately

7000 terroirs in the country. Several socioeconomic and management variables such as labor and input costs may be obtained from the survey data, which may be used as representatives or proxies in models at regional scale. This is more difficult for the biophysical variables as they need to represent areas with high spatial variability. One solution could be to use agroecological zones as representatives for agroecological suitability, but these zones are too coarse to represent the terroir level farming and cropping conditions at regional scale. Alternatively, remote sensing can be used as it provides timely and synoptic coverage. With the application of relatively high resolution remote sensing, the spatial models has shown clear advantages to upscale a range of surveyed variables from household level to terroir and to the national-scale of Burkina Faso.

Second, the variables of land and crop performance may be spatially related to their underlying factors in terroirs. However, these relationships may exhibit spatial non-stationarity due to heterogeneous landscape in Burkina Faso, which may not be fully captured by global spatial models. Alternatively, local models such as the geographically weighted regression (GWR) may be applied to model spatial non-stationarity in the relationships. In doing so, they are able to upscale the variables surveyed at representative terroirs to the country. To model spatial dependency in data, GWR uses kernel functions while geostatistical methods use variograms. Therefore, the spatial models were compared for precisely taking into account the locally-varying farming and cropping conditions in the process of upscaling variables to regional scale, and, thus, resulting more accurate representative data for wall to wall services in Burkina Faso.

Third, uncertainty may be induced when the spatial models of upscaling are linked with several spatial datasets of different supports. Uncertainties may also be associated with field measurements, sampling and monitoring networks, and flaws in the upscaling procedures or in the statistical analysis itself. Uncertainty may further be introduced as a result of inadequacy of models to incorporate the spatial variability in large heterogeneous areas, in particular in West Africa. The quantification and propagation of uncertainties should therefore be an integral part of research on spatial modeling and upscaling. Therefore, a solid statistical methodology is applied for uncertainty assessment in this research. It quantifies jointly and systematically the uncertainties manifested in the spatial models for upscaling. The methodology proceeds as,

- first, to test the sampling network of representative terroirs in Burkina Faso, it applies point pattern analysis to check whether the field measurements are not clustered in the country.
- second, to quantify uncertainty in upscaled estimates, it presents a hybrid approach of geographical weighted regression kriging (GWRK), which applies geographically weighted regression (GWR) to model the local drift, and kriging to interpolate the GWR residuals.
- third, to test inadequacies in the model itself, it compares the

GWRK performance of uncertainty quantification with ordinary kriging (OK), kriging with external drift (KED), and regression kriging (MLRK).

The developed methodology for uncertainty quantification was applied on the regional crop yield modeling and upscaling in Burkina Faso. The spatially-explicit results of the quantified uncertainty in regional model can be obtained on gridded uncertainty maps, which can then be communicated for incorporation when the upscaled estimates are linked to the model application.

6.1.2.2 The case of regional crop yield modeling

Different environmental (e.g. rainfall, elevation, slope, soil type and water holding capacity) and management (e.g. labor availability) factors affect the yields of both food crops (i.e. sorghum and millet) and commercial crops (i.e. cotton) in the country. By establishing local relationships of crop yields with those factors, crop yield models were developed to fit the heterogeneous conditions found at a regional scale in West-Africa. Prediction accuracy of regression models in general and in the GWR model in particular largely depends upon the accuracy of the estimated relationships. Because the estimates of local coefficients in GWR are influenced by the spatial scales of crop yield observations and factors affecting crop yields. Different diagnostic statistics were used for accurately incorporating spatial non-stationary into the GWR local relationships. However, finding a precise spatial scale at which local relations can accurately represent the modeled processes deserves a separate research. Remote sensing-based vegetation indices such as the Normalized Difference Vegetation Index (NDVI) derived from SPOT-VEGETATION data can efficiently reveal the spatial variability of crop yields in the study area. Crop yield values show a positive relation with NDVI values as well as with rainfall estimates that can be obtained from remotely sensed imagery. Modeling crop yields further shows that the influence of hydrological processes related to slope and elevation, and soil properties can be evaluated to estimate yield spatial variability of main crops in Burkina Faso.

Statistical testing was found useful to check if spatial-stationarity is present in the crop yield relationship with the explanatory variables. This however is not sufficient to show improved performance of a model to explain crop yields after incorporating spatial non-stationarity information. Comparing global CAR and local GWR models is important to test the effect of spatial non-stationarity in the relationships between crop yields and the explanatory variables. In doing so, the performance of GWR was found better than that of CAR models for crop yields in Burkina Faso. For areas with a low performance of local prediction, GWR can be improved further by including land cover maps showing cropping areas on grid cells. Moreover, GWR models can be calibrated with RS data to generate timely and accurate crop yield maps in Burkina Faso, which can

subsequently be used to initialize the bio-economic farm model for crop yield estimates in individual terroirs. These methods can also be applied to quantify other variables of farming systems for applying bio-economic farm models at national or regional scales.

6.1.2.3 The case of uncertainty modeling in crop yield upscaling

Observations of crop yields obtained from ground surveys were upscaled from terroir level to the scale of Burkina Faso. The upscaling accuracy however depends upon the accuracy of the modeled relations between observed crop yield and its collocated explanatory variables, including SPOT-VEGETATION (NDVI) data, elevation, slope, and rainfall. GWRK performance is compared for accuracy with OK, KED, and MLRK. The accuracy is compared both for crop yield upscaling and for estimation of uncertainty surrounding those upscaling outcomes.

Accuracy of the crop yield upscaling can be effectively evaluated using absolute error (MAE), mean square error (MSE), and the adjusted coefficient of determination, R_a^2. This however essentially be validated from independent datasets. But the unavailability of independent validation data in West-African countries like Burkina Faso is common and this may lead to a lack of credibility of the results. One solution to this problem may be to divide the available data into the experimental and validation datasets, but data availability and representativeness is an important limiting factor at regional scale. Alternatively, the cross-validation statistic is applied in this research, which omits part of the observations for validation and upscales the remaining dataset towards the locations of the omitted data, followed by minimizing the root mean square error (RMSE) between the upscaled data with the omitted data. Following this, the RMSE of residuals cross validation can be used to evaluate accuracy of uncertainty estimation in crop yield upscaling to regional scales in West-Africa.

6.1.2.4 The case of regional marginality modeling

Quantifying socioeconomic variables is highly challenging, because factors influencing such variables are complex. Most terroir communities do subsistence farming and the marginality modeling estimates their welfare status to assess their capability of applying modern inputs like fertilizers, pest control, and crop varieties. To estimate farmers welfare, indices are often obtained from targeted household surveys. For example, the headcount index is the percent of the population in an area living below an established poverty line, i.e., a normative level of income or expenditure. However, a clear insight into the likely causes of welfare and poverty is usually missing, because their factors are not included during their modeling. In terroir economies based on subsistence agriculture, factors such as income and expenditure may not be sufficient to map welfare

and poverty. This research used an approach in which, on one hand, a range of welfare and poverty aspects related to the terroir farming system were analyzed in the development of a welfare index. On the other hand, it quantified the increased susceptibility of specific areas to become marginal due to extreme events of environmental constraints. Thus, our welfare modeling approach based on the relationship between the welfare index and various stressors can generate not only the high resolution welfare and poverty estimates, but it can also reveal poverty patterns that can be effectively interpreted as compared to the commonly applied headcount index. Using field surveys and remote sensing products, these methods can be applied for sub-national welfare and poverty modeling in developing countries.

In developing a new welfare index for farmer communities in Burkinabé terroirs (in Chapter 5), the agro-ecological potential of their land parcels and food consumption requirement were associated to their welfare status. It was shown that asset variables of agricultural production, household stocks, household food consumption, and number of animals owned significantly contribute to the welfare index. These asset variables can be extracted from the household surveys. Spatially varying agroecological conditions, e.g., weather, soil and topography have a major impact on agro-ecological potential of terroirs. These stress factors can be quantified using RS products. The Harmonic ANalysis of Time Series (HANTS) algorithm applied on time series of SPOT-VEGETATION (NDVI) and of TAMSAT (rainfall) estimates was able to reveal patterns of agroecological suitability within terroirs. Additionally, factors such as length of growing season, soil nutrients, and topography have an immense impact on the agro-ecological potential of terroirs.

Geographical weighted regression can be applied to relate locally the welfare index to the agricultural production factors for each terroir. Subsequently, these relationships can be used for inferring levels of welfare indicator to whole of Burkina Faso. This modeling can provide a spatially detailed view of welfare and poverty of the farmers in the country. Yet, other factors such as limited access to credit may also be analyzed as agroecological stresses on food production. For this study, data on fertilizer applications, pesticides applications, and irrigation information were not available. The estimates of farmers welfare status can also be used as a proxy of farmers economic capacity to apply advanced inputs. Availability of detailed management data, however, can contribute more precise information to quantify the effects of different factors on the food production of the terroirs. Nevertheless, this study has highlighted the potential of the proposed method to identify causes of farmers marginality, which can be further linked with integrative studies to optimally allocate terroir resources for better agricultural production to securing food.

6.1.3 Implementing the proposed framework for wall to wall services

A case study implements the proposed framework design to deploy a wall to wall service in Burkina Faso. The service devises the optimized plans for on-terroir decision-making based on integratively assessing the several terroir resources. To do so, the service links the spatial upscaling of crop yields and farmers marginality to a bio-economic farm model, which is used as a farm simulation model for optimization. The basic purpose of using the bio-economic farm model is to demonstrate: (i) its adaptation in developing wall to wall agricultural services for on-terroir decision-making, (ii) its deployment as a geospatial standard web service, and (ii) its adaptation for regional modeling and upscaling in an SDI-based framework.

To deploy the farm simulation model as a web service, the study set out to provide a standard wrapper to allow the model to be exposed as an OGC web processing service. A source conceptual schema (Appendix C) was designed for the AGRISTAT survey data and a target integrated schema (cf. Figure 2.7) was designed for the integrated database. For the conceptual harmonization in these schemas, the concepts, classes and their relationships were derived from the SEAMLESS shared ontology (cf. Chapter 2). By means of providing both scale-related and schema-related transformations, the source conceptual schema was transformed into the integrated database in PostGIS/PostgreSQL. The scale-related transformations were provided by upscaling crop yield and marginality at the terroir level to the scale of Burkina Faso (cf. Chapters 3 and 5). In doing so, the uncertainty was quantified (cf. Chapter4), and the resulting uncertainty maps were stored in the integrated database. The schema-related transformations were provided by implementing schema mapping operators (Appendices D) by means of the GIS functionality in PostGIS. The OGC web feature services (WFSs) were deployed on top of the integrated database. Both the model service (i.e. WPS) and data services (i.e. WFSs) were published in catalogues. The end-user can discover and bind the data services to the model service for initialization at an individual terroir location. Communication of the uncertainty information (i.e. the quantified uncertainty in regional models), when the geospatial data services are linked to the model service, is postponed to future research due to limited time for this study.

6.2 Conclusions

1. Objective 1: *To investigate SDI technology to propose a flexible framework to link spatial upscaling to simulation models at regional scale for deploying wall to wall services.*

   The study concludes that the model deployed as a service through OGC standard wrapper implementations achieves an adequate level

of interoperability for interacting with geospatial data services, and it ensures end-user communication. The deployed framework allows accomplishing tasks of model Initialization through geospatial data services. The choice of terroir location determines which third-party datasets are going to be used, and the framework effectively accommodates this. Use of integrated database is found effective for conceptually linking various datasets and spatial models of upscaling to the bio-economic farm models. The integrated conceptual schema allows making explicit the model formalisms. The SEAMLESS ontology is useful to align the semantics between data and model formalisms in the agricultural domain. Using a lightweight web browser, farmers and extension workers can find data and model services published in catalogues to bind into executable workflows on a terroir location. This deployment, however, requires a usability test and proper feedback from the extension workers in Burkina Faso.

We conclude that the geospatial web services provide a scalable way of discovering and linking data and models for wall to wall agricultural services. The use of integrated schema and ontology, in the field of semantic extension of the web, has advantage for matching heterogeneous data and simulation model services, which is of high relevance to the future design of open service platforms for integrative assessments. Spatial upscaling is useful to accomplish scale-related transformations for wall to wall SDI services. Benefiting from SDI technology, the proposed framework has potential for enabling community participation in a common problem that results in the introspection of interoperable data and models as web services. Such a provision can enable wider exploitation of existing SDI catalogues to facilitate the integrated assessments of various resources in farming systems. Due to a relatively limited time for this study, integrating data and models web services did not take into account communicating uncertainty associated with those components but on a longer term, this should be investigated for precise assessments.

2. Objective 2: *To model the relationship between the observed crop yields and their collocated explanatory variables at the terroir level and to upscale the yield estimates to the national-scale of Burkina Faso.*

   The study concludes that modeling crop yields over the highly heterogeneous landscape of Burkina Faso requires incorporating the spatial variability of rainfall, topography, labor availability, and selected soil properties, including carbonate, loam and sand content, as well as water holding capacity. By applying the CAR and GWR spatial models, this study observes that SPOT NDVI, elevation, slope, and rainfall significantly affect crop yield spatial variability in both the semiarid and subhumid zones. By means of applying principal component analysis on vegetation indices

and calibrating GWR for local relations between crop yield and its affecting factors, several spatial and temporal instabilities could be overcome in the spatial analysis of crop yields at the national scale. In the semiarid zone, soil properties and labor availability also influence sorghum and millet yields. For terroirs with high values of these variables, the study observes a significant increase in sorghum yield, to a maximum of 350 kg ha^{-1}, and an increase in millet yield, to a maximum of 275 kg ha^{-1}. In the subhumid zone, the maximum increase in cotton yield is 210 kg ha^{-1}, on the Southwest uplands; the maximum increase in millet yield is 275 kg ha^{-1}, on downstream terroirs with steeper slopes; and the maximum increase in sorghum yield is equal to 200 kg ha^{-1}, on areas having a higher frequency of rainfall and carbonate content in the soil.

Spatial variability of crop yields was observed with local models, showing that the effect of explanatory variables is highly localized in the study area. Monte Carlo analysis further confirms the spatial variability of the observed relationships. Thus, accounting for spatial non-stationarity is essential for improving the quality of crop yield upscaling in Burkina Faso. By incorporating the extent of the spatial non-stationarity into the relations, GWR can perform better than the CAR models for upscaling crop yields. Thus, the improvement of the GWR model over the CAR model suggests that the spatial covariates vary spatially in their effects across Burkinabé terroirs. However, the study observes that CAR models perform better than GWR models in areas with less than adequate crop yield observations. The performance of GWR models in some areas could be improved further by including land cover maps showing cropping areas on grid cells.

3. Objective 3: *To model uncertainty in the regional modeling and upscaling of crop yields in Burkina Faso.*

   The study concludes that the sampling network of representative terroirs in Burkina Faso is not clustered, and, thus, it is not inducing uncertainty into the modeling and upscaling of crop yield at regional scale. A high spatial variation is observed in local parameter estimates of climate, topography, financial ability of farmers and labor availability. SPOT NDVI, precipitation, and elevation successfully explain the favorable cropping conditions along agroecological gradients and between terroir sites. Geographical weighted regression kriging (GWRK) improves the upscaling accuracy of crop yields by taking into account the spatial variability of local relations between crop yield and factors affecting crop yields. GWRK is superior to all other kriging-based approaches, with the improved values of R_a^2 equal to 90% and RMSE equal to 71.2. GWRK effectively utilized information present in the external covariate datasets, improving accuracy of the crop yield modeling and upscaling. Compared to KED and MLRK, both the MAE value and the

prediction error variance are reduced in GWRK (480 versus 523 and 503) and (20.4 versus 38.2 and 34.1). Moreover, estimation of error variances is more accurate compared to all other approaches. We conclude that GWRK has potential to upscale accurately over larger areas, using external covariate datasets, specifically in situations of spatial non-stationary relations that could not be properly modeled with non-spatial regression-based approaches.

4. Objective 4: *To model the farmers welfare and marginality status at the terroir level using targeted household surveys, and to investigate regional and global datasets including RS products for upscaling the terroir-level marginality estimates to the national-scale of Burkina Faso.*

   Remote sensing and spatial models were able to quantify the agro-ecological potential of terroirs and to relate it to the index values representing farmer's welfare status. This study shows that 58% of the variance of the welfare index can be explained by the factor representing variables of food production and 42% can be explained by the factor representing variables of food consumption. This study thus concludes that levels of agricultural production and household food consumption are directly related to welfare levels of farmers within the Burkinabé terroirs. The levels of agricultural production are significantly explained by the stressor variables of NDVI, rainfall, length of growing period, soil nutrients, and topography. The relationship between the stressor variables and the welfare index varies considerably between the terroirs. Spatial dependency between the stressor variables and the welfare index can be incorporated into a GWR model that is able to identify areas where poor agro-ecological conditions constrained terroirs from attaining an adequate level of welfare. GWR can be used to link well the spatial distributions of the welfare index and the agroecological suitability. Thus the welfare index represents well the variation of farmer's welfare in terroirs. Spatial variation of the farmers welfare shows that high marginal terroirs belong to the North, South Central, Central Plateau, Boucle du Mouhaun, and East Central regions.

   The study further concludes that interpreting GWR local coefficients can allow investigating the relationship between levels of welfare and food security at national and regional scales. The poverty index based on the upscaled welfare index showed similar patterns as compared to the commonly applied headcount index. This means that agro-ecological potential and poverty incidence are positively related within the Burkinabé terroirs. Timely, cost effective, and fine resolution poverty maps of farmers welfare and poverty can be generated targeting rural areas in other countries. These maps can be applied as representative data for assessing farmers capability of applying modern inputs like fertilizers, pest control, and crop varieties.

6.3 Recommendations

The above leads to the following recommendations for future research in providing the location based wall to wall services that support planning and decision-making based on the integrated assessment of agricultural resources:

1. Schema-related transformations were provided by means of implementing schema mapping operators. Their implementation requires analyzing metadata information in the integrated database. This analysis needs to be automated in a web service environment to allow data and models harmonization on-the-fly without user involvement. By doing so, geospatial data services could provide automatic transformations of spatial datasets when a model needs to initialize its parameters for application at a terroir location. Today's technology for metadata analysis, however, is too immature to expect automated solutions to solve this problem.
2. Presently the model formalisms are described in developing an integrated conceptual schema. However, the OGC WPS interface for a model should make explicit the model functionality based on model assumptions, inputs, and model spatial and temporal extent and support unit. This is difficult with the present WPS standard, being rather generic and not providing options to describe complex models in detail. Therefore, to make explicit the particular descriptions of a BEFM, the development of WPS application profiles may be investigated in future. Following this, the integrated schema can be described in a target WPS application profile, and spatial data services can be automatically replaced or perform transformations otherwise, if source data models do not meet the model quality requirements described in the target profile.
3. Crop-specific high-resolution land cover and land-use maps were not available for the study area. Such maps would have been useful for accurately delineating areas where crops are actually grown. This results into uncertainty when estimating local crop yield relations in parcels on which mixed cropping is practiced. Additional explanatory variables may further be obtained from cultural or socioeconomic characteristics, investment capacity or policy factors.
4. To effectively characterize the terroir land units in Burkina Faso, the methodology developed in this research needs to be applied for upscaling other input variables, such as crop lands, for integrative studies at regional scales. In the case study in Chapter 2, the crop land information obtained from 351 representative terroirs were used as proxy data. Future studies may improve the quality of this

input by applying high resolution remote sensing. Remote sensing has potential to delineate the crop lands.

5. For crop yield modeling at the regional scale, the normalized difference vegetation index (NDVI) might be more sensitive to differences in background soil contamination that to biophysical parameters such as canopy cover or the amount of chlorophyll present in the canopy. Future studies might address the sensitivity to the variation of brightness contrast between vegetation and soil background, or the use of alternative vegetation indices that are less influenced by confounding factors.
6. The Harmonized World Soil Database (HWSD) provides soil data for many Sub-Saharan countries, but they are not yet available for Burkina Faso. Presently, we used soil data from HarvestChoice. The HWSD gridded soil data may be used in the future to better represent soil properties at the level of detail that is demanded by the GWRK model of crop yield as performed in this study.
7. With the provision of versatile transformation services, for location based service at large scales, a vital issue that needs further investigation is how to communicate the quantified uncertainty in the datasets or in the models. Spatial models for scale-related transformations of datasets might induce some error or uncertainty. These must be communicated to the participant model services, particularly if the model input is sensitive to uncertainty. Moreover, uncertainty is associated with model assumptions and parameters. Cumulative uncertainty arising from the participant data and model services in web-based workflows must be communicated the decision-makers.
8. Extending the marginality upscaling approach to the whole of West Africa will require a careful selection of asset variables to compose welfare index for rural communities in the entire region. The present study considers only the welfare aspects related to the households food consumption and to the terroir agricultural production. However, in other countries, these assets may not be sufficient to characterize welfare. For instance, evaluating access to adequate health services, potable water and sanitation could also add useful information to the welfare index. In urban areas, the welfare variables that represent non-farm activities, e.g. fishery, handicrafts, mining must also be considered.
9. Future work should focus on examining solutions that will (1) close the conceptual gap in data type definition of model inputs/outputs and spatial datasets, and (2) describe model semantics (for example, model formalisms) to enable automatic composition of OGC data and model services in an SDI environment.

References

AGRISTAT. 2010. *Résultats Définitives Campagne (2008-2009), Burkina Faso*. Tech. rept. Statistiques sur l'Agriculture et l'Alimentation du Burkina Faso (AGRISTAT).

Alameh, N. 2003. Chaining geographic information Web Services. *IEEE Internet Computing*, 22–29.

Alasia, A., Bollman, R., Parkins, J., & Reimer, B. 2008. *An Index of Community Vulnerability: Conceptual Framework and Application to Population and Employment Changes (1981 to 2001)*. Tech. rept. Statistics Canada, Agriculture Division.

Anselin, L. 1995. Local Indicators of Spatial Association – LISA. *Geogr Anal.*, **2**7, 93–115.

Baddeley, A., & Turner, R. 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, **1**2, 1–42.

Baltenweck, I., van de Steeg, J., & Staal, S.J. 2004. *Farming systems characterisation in the Kenyan Highlands: Use of alternative methodologies*. Tech. rept. ILRI, Nairobi, Kenya.

Beare, M., Howard, M., Payne, S., & Watson, P. 2010. *Development of Technocal Guidanace for the INSPIRE Transformation Network Services, State Of The Art Analysis*. Tech. rept. RSW Geomatics LTD for EC JRC Contract Notice 2009/S 107–153973.

Belsley, D., Kuh, E., & Welsch, R. 1980. *Regression diagnostics: identifying influential data and sources of collinearity*. John Wiley & Sons, New York.

Benson, T., Chamberlin, J., & Rhinehart, I. 2005. An investigation of the spatial determinants of the local prevalence of poverty in rural Malawi. *Food Policy*, **3**0(5–6), 532–550.

Berlage, L., & Terweduwe, D. 1988. The classification of countries by cluster and by factor analysis. *World Development*, **1**6(12), 1527–1545.

Bernard, L., Kanellopoulos, I., Annoni, A., & Smits, P. 2005. The European Geoportal – one step towards the establishment of a European spatial data infrastructure. *Computers, Environment and Urban Systems*, **2**9, 15–31.

Bevan, A., & Conolly, J. 2009. Modelling spatial heterogeneity and non-stationarity in artifact-rich landscapes. *Journal of Archaeological Science*, **3**6(4), 956–964.

Bian, L. 2007. Object-Oriented Representation of Environmental Phenomena: Is Everything Best Represented as an Object? *Annals of the Association of American Geographers*, **9**7(2), 267–281.

Bigman, D., Dercon, S., Guillaume, D., & Lambotte, M. 1999. *Community Targeting for Poverty Reduction in Burkina Faso*. Tech. rept. Development Economics, Center for Economic Studies, Discussions Paper Series (DPS) 99.10.

Bivand, R. 2012. *spdep: Spatial dependence: weighting schemes, statistics and models.* R package version 0.5–51. http://CRAN.R-project.org/ (accessed 20 Nov. 2012).

Bivand, R., & Yu, D. 2012. *spgwr: Geographically weighted regression.* R package version 0.6–17. http://CRAN.R-project.org/ (accessed 20 Nov. 2012).

Bivand, R., Pebesma, E., & Rubio, V. 2008. *Applied Spatial Data Analysis with R*. Springer, Heidelberg.

Brimicombe, A. 2009. *GIS, Environmental Modelling and Engineering*. Second edn. Taylor & Francis Group.

Budde, M.E., Tappan, G., Rowland, J., Lewis, J., & Tieszen, L.L. 2004. Assessing land cover performance in Senegal, West Africa using 1-km integrated NDVI and local variance analysis. *Journal of Arid Environments*, **5**9(3), 481–498.

Castoldi, N., Bechini, L., & Stein, A. 2009. Evaluation of the spatial uncertainty of agro-ecological assessments at the regional scale: The phosphorus indicator in northern Italy. *Ecological Indicators*, **9**(5), 902–912.

Castronova, A.M., Goodall, J.L., & Elag, M.M. 2013. Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard. *Environmental Modelling & Software*, **4**1, 72–83.

Challinor, A.J., Ewert, F., Arnold, S., Simelton, E., & Fraser, E. 2009. Crops and climate change: progress, trends, and challenges in simulating impacts and informing adaptation. *Journal of Experimental Botany*, **6**0(10), 2775–2789.

Chambers, J., & Hastie, T. 1992. *Statistical Models in S*. Wadsworth & Brooks/Cole.

Chilès, J., & Delfiner, P. 1999. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, New York.

Chomitz, K.M., & Thomas, T.S. 2003. Determinants of land use in Amazonia: A fine-scale spatial analysis. *American Journal of Agricultural Economics*, **8**5(4), 1016–1028.

Cliff, A.D., & Ord, J.K. 1981. *Spatial Processes – Models and Applications*. Pion Ltd., London.

Cressie, N. 1991. *Statistics for spatial data*. John Wiley & Sons, Inc.

Cressie, N., & Wikle, C.K. 2011. *Statistics for Sptio-Temporal Data*. John Wiley & Sons, New Jersy.

CRUTS. 2006. *Climatic Research Unit (CRU) time-series datasets of variations in climate with variations in other phenomena. British Atmospheric Data Centre (BADC)*. http://badc.nerc.ac.uk/ (accessed 20 Nov. 2012).

de Beurs, K.M., & Henebry, G.M. 2010. *Phenological Research*. Springer. Chap. Spatio-Temporal statistical methods for modelling land surface phenology, pages 177–208.

de Graaf, J., Nikiema, R., Tapsoba, G., & Nederlof, S. 2001. *Agro-Silvo-Pastoral Land Use in Sahelian Villages*. Catena Verlag GMBH, Germany. Chap. Socio-economic Land Use Analaysis in Sahelian Villagest, pages 23–71.

de Wit, A.J.W., de Bruin, S., & Torfs, P.J.J.F. 2008. Representing Uncertainty in Continental-Scale Gridded Precipitation Fields for Agrometeorological Modeling. *Journal of Hydrometeor*, **9**, 1172–1190.

de Wit, C.T., van Keulen, H., Seligman, N., & Spharim, I. 1988. Application of Interactive Multiple Goal Programming Techniques for analysis and planning of regional agricultural development. *Agricultural Systems*, **2**6, 211–230.

Di, L. 2005. A Framework for Developing Web-Service-Based Intelligent Geospatial Knowledge Systems. *Annals of GIS*, **1**1(1), 24–28.

Diggle, P. 2003. *Statistical Analysis of Spatial Point Patterns*. Second edn. Oxford, University Press.

Diggle, P.J., Moyeed, R.A., & Tawn, J.A. 1998. Model-based Geostatistics. *Applied Statistics*, **4**7, 299–350.

Dixon, J., & Gulliver, A. 2001. *Farming Systems and Poverty*. Tech. rept. FAO and World Bank Rome and Washington D.C.

Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., & Schaepman, M.E. 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International journal of applied earth observation and geoinformation*, **9**, 165–193.

Dreschel, P., Gyiele, L., Kunze, D., & Cofie, O. 2001. Population density, soil nutrient depletion, and economic growth in sub-Saharan Africa. *Ecological Economics*, **3**8(2), 251–258.

Dutta, R., Stein, A., & Bhagat, R.M. 2011. Integrating satellite images and spectroscopy to measuring green and black tea quality. *Food Chemistry*, **1**27(2), 866–874.

Ebert, U., & Welsch, H. 2004. Meaningful environmental indices: a social choice approach, Journal of Environmental Economics and Management. *Journal of Environmental Economics and Management*, **4**7, 270–283.

Faivre, R., Leenhardt, D., Voltz, M., Benoît, M., Papy, F., Dedieu, G., & Wallach, D. 2004. Spatializing Crop Models. *Agronomie*, **2**4(4), 205–217.

FAO. 1998. *Wetland Characterization and Classification For Sustainable Agricultural Development.* Tech. rept. Food and Agriculture Organization - Regional Office for Africa (RAF).

FAO. 2005. *L'irrigation en Afrique en chiffres, Rapports Sur L'eau 29.* Tech. rept. Food and Agriculture Organization (FAO).

FAO. 2012a. *FAO statistical database.* Tech. rept. Food and Agriculture Organization of the United Nations (FAO).

FAO. 2012b. *Food Insecurity, Poverty and Environment Global GIS database.* Tech. rept. Food and Agriculture Organization (FAO).

FAO, & IIASA. 2012. *Harmonized World Soil Database (version 1.2).* Tech. rept. FAO, Rome, Italy and IIASA, Laxenburg, Austria.

FEWSNET. 2012. *Western Africa FEWS NET Food Security Outlook.* Tech. rept. Famine Early Warning System Network Burkina Faso (FEWSNET).

Foerster, T., Lehto, L., Sarjakoski, T., Sarjakoski, L.T., & Stoter, J. 2010. Map generalization and schema transformation of geospatial data combined in a Web Service context. *Computers, Environment and Urban Systems*, **3**4(1), 79–88.

Fortanier, F. 2006. *Multinational Enterprises, Commodity Chain Partnerships and Host Country Development Goals.* Tech. rept. Expert Centre for Sustainable Business and Development Cooperation (ECSAD).

Fotheringham, A.S., Brunsdon, C., & Charlton, M. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* John Wiley & Sons, New York, USA.

Francois, P., Catherine, R., Thomas, B., & Frédéric, V. 2009. The Use of UML as a Tool for the Formalisation of Standards and the Design of Ontologies in Agriculture. *Pages 1–17 of: Advances in Modeling Agricultural Systems*, vol. 25. Springer US.

Gao, J., Li, S., Zhao, Z., & Cai, Y. 2012. Investigating spatial variation in the relationships between NDVI and environmental factors at multiscales: a case study of Guizhou Karst Plateau, China. *International Journal Remote Sensing*, **3**3, 2112–2129.

Gatzweiler, F., Baumüller, H., Ladenburger, C., & von Braun, J. 2011. *Marginality: Addressing the Root Causes of Extreme Poverty.* Tech. rept. Center for Development Research, University of Bonn, Germany.

Geller, G.N., & Melton, F. 2008. Looking forward: Applying an ecological model web to assess impacts of climate change. *Biodiversity*, **9**(3&4), 79–83.

GeoNetwork. 2010. http://www.geonetwork-opensource.org/index.html, Accessed on Auguest 10, 2011.

GeoServer. 2010. http://geoserver.org/display/GEOS/Welcome, Accessed on Auguest 10, 2011.

Graef, F., & Haigis, J. 2001. Spatial and temporal rainfall variability in the Sahel and its effects on farmers management strategies. *Journal of Arid Environments*, **4**8(3), 221–231.

Granell, C., Díaz, L., & Gould, M. 2010. Service-oriented applications for environmental models: Reusable geospatial services. *Environmental Modelling & Software*, **2**5, 182–198.

Griffith, D. 1988. *Advanced Spatial Statistics*. Kluwer, Dordrecht.

Grimes, D.I.F., Pardo-Igúzquiza, E., & Bonifacio, R. 1999. Optimal areal rainfall estimation using raingauges and satellite data. *Journal of Hydrology*, **2**22, 93–108.

Groot, R., & McLaughlin, J.D. 2000. *Geospatial Data Infrastructure: Concepts, Cases, and Good Practice*. Spatial Information Systems Series. Oxford University Press.

Han, W., Yang, Z., Di, L., & Mueller, R. 2012. CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture*, **8**4, 111–123.

Harris, P., Charlton, M., & Fotheringham, A. 2010a. Moving window kriging with geographically weighted variograms. *Stochastic Environmental Research and Risk Assessment*, **2**4, 1193–1209.

Harris, P., Fotheringham, A., Crespo, R., & Charlton, M. 2010b. The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets. *Mathematical Geosciences*, **4**2, 657–680.

HarvestChoice. 2012. *Sub-national Sub-Saharan Africa data sets*. Tech. rept. HarvestChoice/International Food Policy Research Institute (IFPRI).

Hazell, P.B.R., & Norton, R.D. 1986. *Mathematical programming for economic analysis in agriculture*. Macmillan Publishing Company, New York, pp. 400.

Hirosawan, Y., Marsh, S.E., & Kliman, D.H. 1996. Application of standardized principal component analysis to land cover characterization using multitemporal AVHRR data. *Remote Sens. Environ*, **5**8, 267–281.

Hoddinott, J., & Quisumbing, A. 2003. *Methods for Micro economic Risk and Vulnerability Assessments. Social Protection Discussion Paper Series No. 0324*. Tech. rept. The World Bank.

Holloway, G., Shankar, B., & Rahman, S. 2002. Bayesian spatial probit estimation: A primer and an application to HYV rice adoption. *Agricultural Economics*, **2**7, 383–402.

Horn, J.L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika*, **3**0, 179–185.

Huang, N., & Diao, S.H. 2008. Ontology-based enterprise knowledge integration. *Robotics and Computer-Integrated Manufacturing*, **2**4(4), 562–571.

Huete, A.R, & Tucker., C.J. 1991. Investigation of soil influences in AVHRR red and near-infrared vegetation index imagery. *Int. J. Remote Sens*, **1**2, 1223–1242.

Hurvich, C.M., Simonoff, J.S., & Tsai, C-L. 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B*, **6**0(2), 271–293.

Hyman, G., Larrea, C., & Farrow, A. 2005. Methods, results and policy implications of poverty and food security mapping assessments. *Food Policy*, **3**0, 453–460.

IIASA/FAO. 2012. *Global Agro-ecological Zones (GAEZ v3.0)*. Tech. rept. IIASA, Laxenburg, Austria and FAO, Rome, Italy.

INSPIRE. 2008. *Infrastructure for Spatial Information in the European Community*. http://inspire.jrc.it, Accessed on Auguest 17, 2010.

ISO. 2005. *Geographic information Services - ISO Standard 19119*. Tech. rept. International Organization for Standardization.

ISO/IEC. 1996. *International Standard Information Technology, Open Distributed Processing, Reference Model (RM-ODP): Fundations, First Edition*. Tech. rept. Internatioal Organization for Standardization.

Jakeman, A.J., & Letcher, R.A. 2003. Integrated assessment and modelling: features, principles and examples for catchment management. *Environmental Modelling & Software*, **1**8(6), 491–501.

Janssen, S., & van Ittersum, M.K. 2007. Assessing farm innovations and responses to policies: A review of bio-economic farm models. *Agricultural Systems*, **9**4(3), 622–636.

Janssen, S., Andersen, E., Athanasiadis, I.N., & van Ittersum, M.K. 2009. A Database for Integrated Assessment of European Agricultural Systems. *Environmental Science & Policy*, **1**2(5), 573–587.

Janssen, S., Louhichi, K., Kanellopoulos, A., Zander, P., Flichman, G., Hengsdijk, H., Meuter, E., Andersen, E., Belhouchette, H., Blanco, M., Borkowski, N., Heckelei, T., Hecker, M., Li, H., Lansink, A. Oude, Stokstad, G., Thorne, P., van Keulen, H., & van Ittersum, M.K. 2010. A Generic Bio-Economic Farm Model for Environmental and Economic Assessment of Agricultural Systems. *Environmental Management*, **4**6, 862–877.

JRC. 2006. *VGT4AFRICA User Manual*. Tech. rept. Joint Research Centre of the European Commission (JRC), Ispra, Italy.

Kaza, S., & Chen, H. 2008. Evaluating ontology mapping techniques: An experiment in public safety information sharing. *Decision Support Systems*, **4**5, 714–728.

Kiehle, C. 2006. Business logic for geoprocessing of distributed geodata. *Computers & Geosciences*, **3**2(10), 1746–1757.

Knapen, M.J.R, Athanasiadis, I.N., Jonsson, B., Huber, D., Wien, J.J.F., Rizzoli, A.E., & Janssen, S. 2009. Use of OpenMI in SEAMLESS. *Pages 330–331 of: AgSAP conference 2009*. Wageningen University and Research Centre, The Netherlands.

Kutner, M., Nachtsheim, C., J.Neter, & Li, W. 2005. *Applied linear statistical models*. Boston, MA: McGrawHill Irwin.

la Rosa, D.D., Mayol, F., Diaz-Pereira, E., Fernandez, M., & de la Rosa Jr., D. 2004. A land evaluation decision support system (MicroLEIS DSS) for agricultural soil protection: With special reference to the Mediterranean region. *Environmental Modelling & Software*, **1**9(10), 929–942.

Lal, R. 1991. Tillage and agricultural sustainability. *Soil Tillage Res*, **2**0, 133–146.

Lambin, E.F. 2003. *People and the environment: Approaches for linking household and community surveys to remote sensing and GIS*. Kluwer Academic Publishers, Norwell, USA. Chap. Linking socio-economic and remote sensing data at the community or at the household level: Two case studies from Africa.

Lambin, E.F., Cashman, P., Moody, A., Parkhurst, B.H., & Pax., M.H. 1993. Agricultural production monitoring in the Sahel using remote sensing: present possibilities and research needs. *J. Environ. Manage.*, **3**8, 301–322.

Lee, W.S., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D., & Li, C. 2010. Sensing technologies for precision specialty crop production. *Computers and Electronics in Agriculture*, **7**4(1), 2–33.

Leung, Y., Mei, C., & Zhang, W. 2000. Statistical tests for spatial non-stationarity based on the geographically weighted regression model. *Environmental Planning*, **3**2, 9–32.

Leyk, S., Norlund, P.U., & Nuckols, J.R. 2012. Robust assessment of spatial non-stationarity in model associations related to pediatric mortality due to diarrheal disease in Brazil. *Spat. Spatiotemporal. Epidemiol*, **3**(2), 95–105.

Lloyd, C.D. 2011. *Local models for spatial analysis*. FL, USA: Taylor & Francis.

Louhichi, K., Kanellopoulos, A., Janssen, S., Flichman, G., Blanco, M., Hengsdijk, H., Heckelei, T., Berentsen, P., Lansink, A.O., & van Ittersum, M. 2010. FSSIM, a bio-economic farm model for simulating the response of EU farming systems to agricultural and environmental policies. *Agricultural Systems*, **1**03(8), 585–597.

Maué, P., Stasc, C., Athanasopoulos, G., & Gerharz, L. 2010. Geospatial Standards for Web-enabled Enviromental Models. *Article under review for the International Journal of Spatial Data Infrastructures Research*, **5**.

Maunder, A. 2002. *Sorghum worldwide*. Ames, IA, USA: Iowa State Press. Chap. Sorghum and millet diseases, pages 11–17.

Molenaar, M. 1998. *An introduction to the theory of spatial object modelling for GIS*. London: Taylor & Francis.

Nagelkerke, N. 1991. A note on a general definition of the coefficient of determination. *Biometrika*, **7**8, 691–692.

Ncube, B., Dimes, J.P., Wijk, M.T.V, Twomlow, S.J., & Giller, K.E. 2009. Productivity and residual benefits of grain legumes to sorghum under semi-arid conditions in southwestern Zimbabwe: unravelling the effects of water and nitrogen using a simulation model. *Field Crops Res.*, **1**10(2), 173–184.

Nebert, D. 2004. *SDI Reference Manual - Cookbook*. Tech. rept. The Federal Geographic Data Committee (FGDC).

Nelson, A., Rogers, D., & Robinson, T. 2012. *Poverty mapping in Uganda: Extrapolating household expenditure data using environmental data and regression techniques*. Tech. rept. Food and Agricultural Organization of the United Nations (FAO).

OGC. 2005. *OGC – Web Feature Service Specification*. Tech. rept. Open Geospatial Consortium Inc.

OGC. 2007a. *OGC – Catalogue Services Specification*. Tech. rept. Open Geospatial Consortium Inc.

OGC. 2007b. *OGC – Web Processing Service Specification*. Tech. rept. Open Geospatial Consortium Inc.

OGC. 2008a. *OGC – Web Coverage Service Specification*. Tech. rept. Open Geospatial Consortium Inc.

OGC. 2008b. *OGC – Web Map Service Specification*. Tech. rept. Open Geospatial Consortium Inc.

OGC. 2008c. *OGC 08-062r4 – OGC Reference Model*. Tech. rept. Open Geospatial Consortium Inc.

Overmars, K.P., de Koning, G.H.J., & Veldkamp, A. 2003. Spatial autocorrelation in multi-scale land use models. *Ecological Modelling*, **1**64, 257–270.

Ozdogan, M. 2010. The spatial distribution of crop types from MODIS data: Temporal unmixing using independent component analysis. *Remote Sens. Environ*, **1**14, 1190–1204.

Papritz, A., & Stein, A. 1999. *Spatial Statistics for Remote Sensing*. Ames, IA, USA: Springer. Chap. Spatial prediction by linear kriging, pages 83–113.

Parker, P., Letcher, R., Jakeman, A., Beck, M.B., Harris, G., Argent, R.M., Hare, M., Pahl-Wost, C., Voinov, A., Janssen, M., Sullivan, P., Scoccimarro, M., Friend, A., Sonnenshein, M., Barker, D., Matejicek, L., Odulaja, D., Deadman, P., Lim, K., Larocque, G., Tarikhi, P., Fletcher, C., Put, A., Maxwell, T., Charles, A., Breeze, H., Nakatani, N., Mudgal, S., Naito, W., Osidele, O., Eriksson, I., Kautsky, U., Kautsky, E., Naeslund, B., Kumblad, L., Park, R., Maltagliati, S., Girardin, P., Rizzoli, A., Mauriello, D., Hoch, R., Pelletier, D., Reilly, J., Olafsdottir, R., & Bin, S. 2002.

Progress in integrated assessment and modelling. *Environmental Modelling & Software*, **3**(17), 209–217.

Parkins, J.R., & MacKendrick, N.A. 2007. Assessing community vulnerability: A study of the mountain pine beetle outbreak in British Columbia. *Canada Global Environmental Change*, **1**7, 460–471.

Pebesma, E.J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **3**0(7), 683–691.

Peeters, A., Ben-Gal, A., Hetzroni, A., & Zude, M. 2012. *2012 International Congress on Environmental Modelling and Software: Managing Resources of a Limited Planet, Sixth Biennial Meeting*. International Environmental Modelling and Software Society (iEMSs). Chap. Developing a GIS-based Spatial Decision Support System for Automated Tree Crop Management to Optimize Irrigation Inputs, pages 101–118.

Prasad, A.K., Chai, L., Singh, R.P., & Kafatos., M. 2006. Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, **8**(1), 26–33.

Ramankutty, N. 2004. Croplands in West Africa: A geographically explicit dataset for use in model. *Earth Interactions*, **8**, 1–22.

Reichardt, M. 2010. Open standards-based geoprocessing Web services support the study and management of hazard and risk. *Geomatics, Natural Hazards and Risk*, **1**(2), 171–184.

Reidsma, P., Ewert, F., & Lansink, A.O. 2007a. Analysis of farm performance in Europe under different climate and management conditions to improve understanding of adaptive capacity. *Climatic Change*, **8**4, 403–422.

Reidsma, P., Boogaard, F. Ewert H., & van Diepen, K. 2007b. Regional crop modelling in Europe: The impact of climatic conditions and farm characteristics on maize yields. *Agricultural Systems*, **1**00(1–3), 51–60.

Rivoirard, J. 2002. On the structural link between variables in kriging with external drift. *Math Geol*, **3**4, 797–808.

Rizoli, A.E., Leavesley, G., Ascough, J.C., Argent, R.M., Athanasiadis, I.N., Brilhante, V., Claeys, F.H.A, David, O., MDonatelli, Gijsbers, P., Havlik, D., Kassahun, A., Krause, P., Quinn, N.W.T, Scholten, H., Sojda, R.S., & Villa, F. 2008. *Environmental Modelling, Software and Decision Support: State of the art and new prospectives*. Elsevier. Chap. Integrated Modelling Frameworks For Environmental Assessment and Decision Support, pages 101–118.

Robinson, T., Emwanu, T., & Rogers, D. 2007. Environmental Approaches to Poverty Mapping: an example from Uganda. *Information Development*, **2**3, 205–215.

Roerink, G.J., Menenti, M., & Verhoef, W. 2000. Reconstructing cloudfree NDVI composites using Fourier analysis of time series. *int. j. remote sensing*, **2**1(9), 1911–1917.

Roncoli, C., Ingram, K., & Kirshen, P. 2001. The costs and risks of coping with drought: Livelihood impacts and farmer's responses in Burkina Faso. *Climate Res.*, **1**9, 119–132.

Roncoli, C., Jost, C., Kirshen, P., Sanon, M., Ingram, K.T., Woodin, M., Somé, L., Ouattara, F., Sanfo, B.J., Sia, C., Yaka, P., & Hoogenboom, G. 2009. From accessing to assessing forecasts: an end-to-end study of participatory climate forecast dissemination in Burkina Faso (West Africa). *Climatic Change*, **9**2, 433–460.

Rothman, D.S., & Robinson, J.B. 1997. Growing pains: A Conceptual Framework for Considering Integrated Assessments. *Environmental Monitoring and Assessment*, **4**6(1), 23–43.

Schäffer, B. 2009. 52North Open Source WPS and SEXTANTE. *In: FOSS4G, 2009 Free and open source software for geospatial conference, Sydney, Australia.*

Shaner, W.W., Philipp, P.F., & Schmehl, W.R. 1982. *Farming Systems Research and Development: Guidelines for Developing Countries.* Boulder, Colorado, USA: Westview Press.

Sharma, V., Irmak, A., Kabenge, I., & Irmak, S. 2011. Application of GIS and geographically weighted regression to evaluate the spatial non-stationarity relationships between precipitation Vs. irrigated and rainfed maize and soybean yields. *Trans. ASABE*, **5**4(3), 953–972.

Staal, S.J., Baltenweck, I., Waithaka, M.M., de Wolff, T., & Njoroge, L. 2002. Location and uptake: Integrated household and GIS analysis of technology adoption and land use, with application to smallholder dairy farms in Kenya. *Agricultural Economics*, **2**7, 295–315.

Stein, A., Hoogerwerf, M., & Bouma, J. 1988. Use of soil map delineations to improve (co)kriging of point data on moisture deficitsn. *Geoderma*, **4**3, 163–177.

Steinman, J.S., Lammers, C.N., & Valinski, M.E. 2009. A Proposed Open Cognitive Architecture Framework. *Pages 1345–1355 of: Winter Simulation Conference 2009.*

Therond, O., Hengsdijk, H., Casellas, E., Wallach, D., Adam, M., Belhouchette, H., Oomen, R., Russell, G., Ewert, F., Bergez, J-E, Janssen, S., Wery, J., & Ittersum, M.K. Van. 2011. Using a cropping system model at regional scale: Low-data approaches for crop management information and model calibration. *Agriculture, Ecosystems & Environment*, **1**42(1–2), 85–94.

UNCCD. 2000. *Burkina Faso: Programme d'action national de lutte contre la désertification.* Tech. rept. United Nations Convention to Combat Desertification (UNCCD).

USAID. 2009. *USAID office of food for peace Burkina Faso Food Security country framework FY 2010-2014.* Tech. rept. USAID, Washington, D.C.

USGS. 2012. *Hydro Africa datasets of Earth Resource Observation and Science Center (EROS).* Tech. rept. United States Department of the Interior Geological Survey (USGS).

van Ittersum, M.K., Leffelaar, P.A., van Keulen, H., Kropff, M.J., Bastiaans, L., & Goudriaan, J. 2004. On approaches and applications of the Wageningen crop models. *European Journal of Agronomy*, **2**4, 201–234.

van Ittersum, M.K., Ewert, F., Heckelei, T., Wery, J., Olsson, J. Alkan, Andersen, E., Bezlepkina, I., Brouwer, F., Donatelli, M., Flichman, G., Olsson, L., Rizzoli, A., van der Wal, T., Wien, J.E., & Wolf, J. 2008. Integrated assessment of agricultural systems - A component-based framework for the European Union (SEAMLESS). *Agricultural Systems*, **9**6(1–3), 150–165.

Verhoef, W. 1996. *Application of Harmonic Analysis of NDVI Time Series (HANTS). In Fourier Analysis of Temporal NDVI in the Southern African and American Continents.* Tech. rept. Winand Staring Centre for Integrated Land, Soil and Water Research, Wageningen, The Netherlands.

VGT4Africa. 2012. *Distribution of VEGETATION data in Africa through EUMETCast.* Tech. rept. Joint Research Centre (JRC), Ispra, Italy.

Vossen, P. 1999. *Data and Models in Action.* Kluwer Acdamic Publishers, The Netherlands. Chap. Finding and using data for small scale applications of agrometeorological such as yield forecasting at a European scale, pages 49–64.

Wall, M.M. 2004. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, **1**21(2), 311–324.

West, C.T., Roncoli, C., & Ouattara, F. 2008. Local perceptions and regional climate trends on the Central Plateau of Burkina Faso. *Land Degradation & Development*, **1**9(3), 289–304.

Wheeler, D., & Tiefelsdorf, M. 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, **7**(2), 161–187.

Wood, S., Hyman, G., Deichmann, U., Barona, E., Tenorio, R., Guo, Z., Castano, S., Rivera, O., Diaz, E., & Marin, J. 2010. *Sub-national poverty maps for the developing world using international poverty lines: Preliminary data release.* http://povertymap.info (accessed 20 Nov. 2012).

Zhanng, T., & Tsou, M-H. 2009. Developing a grid-enabled spatial web portal for Internet GIServices and geospatial cyberinfrastructure. *International Journal of Geographical Information Science*, **2**3(5), 650–630.

Mathematical formulations of objective function and constraints of the BEFM

A

A.1 Objective function

E_CINCOME(f) . . . OBJ =E= REV(f) - PRODCOST(f) - FOODCOST(f);

A.2 Constraints

1. **Production constraint**
   E_PRODUCT(f, i, aez) . . . OUTPUT(f, i, aez) =E= yield(f,i,aez) ∗ A(f, i, aez);
2. **Revenue constraint**
   E_REVENUE(f) . . . REV(f) =E= SUM(aez,(SUM(i,price(f, i) ∗ OUTPUT(f, i, aez))));
3. **Capital constraint**
   E_CAPITAL(f, i, aez) . . . C(f,i,aez) =E=inputs(f , i, aez) ∗ A(f, i, aez);
4. **Food consumption/security constraint**
   E_FOOD_SECURITY(f, i)$food(i) . . . SUM(aez,OUTPUT(f, i, aez)) + PURCH(f, i) =G= cons(i) ∗ size(f);
5. **Cost constraints**
   - E_PRODCOST(f) . . . PRODCOST(f) =E= SUM(aez, SUM(i$crops(i), K(f, i, aez) + wage ∗ TLABOUR(f, i, aez)) + SUM(i$past(i),K(f, i, aez)+ wage ∗ TLABOUR(f, i, aez)));
     Where wage is a scalar number for daily wage.
   - E_FOODCOST(f) . . . FOODCOST(f) =E= SUM(i$food(i), PURCH(f, i) ∗ price(f, i));
6. **Budget constraint**
   E_BUDGET(f) . . . PRODCOST(f) + FOODCOST(f) =L= capstart(f);

7. **Labor constraints**

   - F_LABOUR(f) ... SUM(i, SUM(aez, FLABOUR(f,i,aez))) =L= workdays ∗ famlab(f);
     Where workdays=260 is a scalar number for working days per year.
   - E_LABOUR(f,i,aez) ... A(f, i, aez) ∗ labreq(i, aez) =L= FLABOUR(f, i, aez) + TLABOUR(f, i, aez);

8. **Land constraints**

   - E_TOLAND(f, aez) ... SUM(i$luse(i),A(f, i, aez)) =E= TOTLAND(f, aez);
   - E_GRLAND(f, aez) ... A(f, 'animal', aez) =L= initland(f, 'animal', aez);
   - E_PERLAND(f, i, aez)$per(i) ... A(f, i, aez) =L= initland(f, i, aez);
   - E_LANDVAL(f) ... LANDVAL(f) =E= SUM(aez, SUM(i$past(i), A(f, i, aez)));

Parameter values to apply HANTS

B

The curve fitting procedure in each HANTS run was controlled by setting the following parameters:

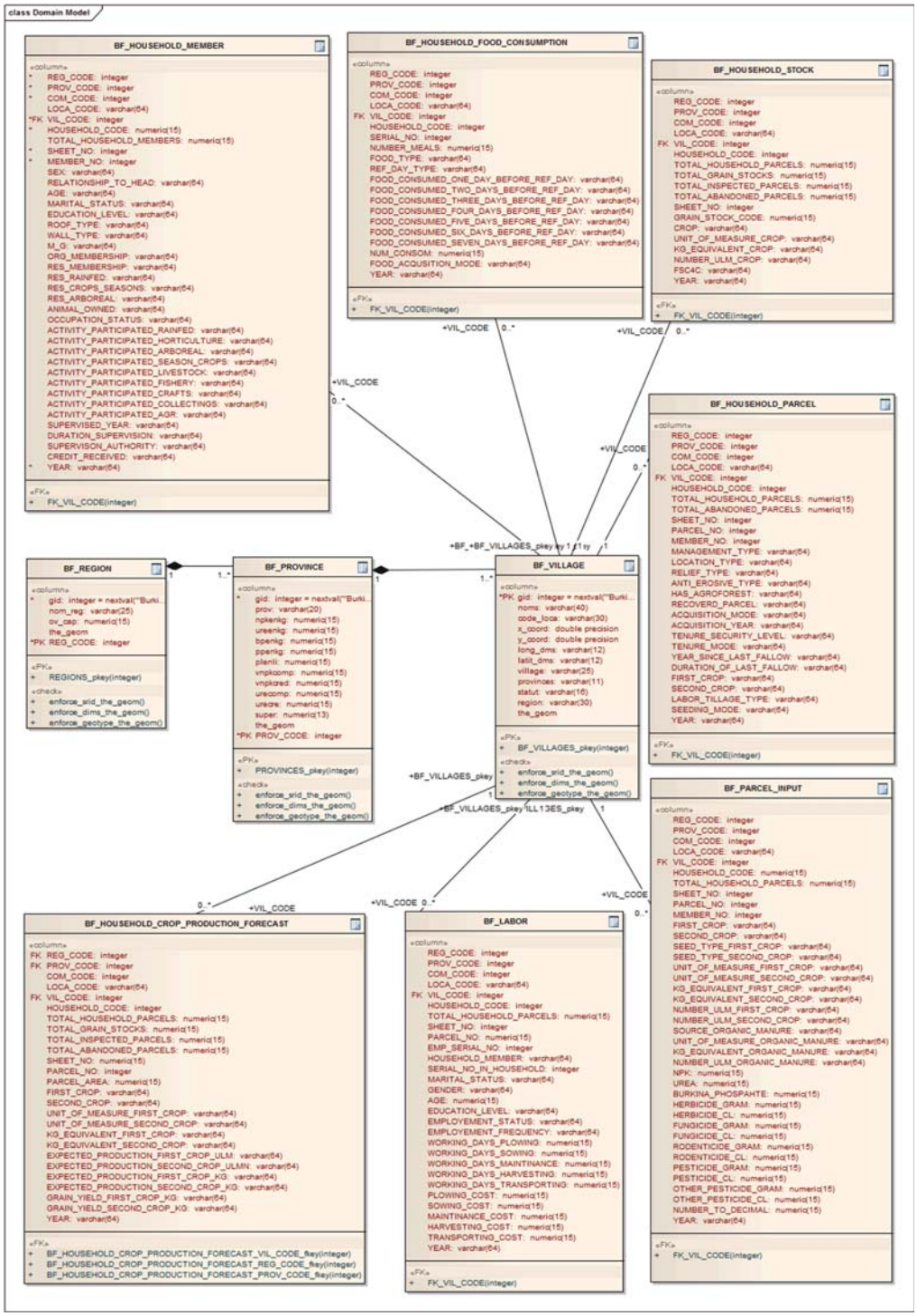
1. Number of frequencies (NOF): 3, i.e., annual, semi-annual (6 months) and seasonal (3 months) frequencies.
2. Hi/Lo suppression flag (SF): low, i.e., the low values are rejected during curve fitting, because cloud contamination always corresponds to low or negative values. Particularly in the case of NDVI, cloud affected observations are remained even after applying the classical maximum value composting algorithm.
3. Fit error tolerance (FET): 0, i.e., curve fitting continued until all invalid values were removed.
4. Invalid data rejection threshold (IDRT): 1-254
5. Degree of overdeterminedness (DOD): 13, i.e., curve fitting was applied such that the 13 observational data points remain available in addition to the minimum data points.
6. Delta: 1 for the NDVI time series and 100 for the RFE time series.

For further details on these parameters see (Roerink *et al.*, 2000).

Source conceptual schema based on AGRISTAT surveys in Burkina Faso

C

C. Source conceptual schema based on AGRISTAT surveys in Burkina Faso



Schema mapping operators

D

This appendix provides details on the basis of schema-related transformations that were implemented for the integrated database in PostGIS. They transform the source conceptual schema (cf. Appendix B) to the integrated conceptual schema (cf. Chapter 2). For the mathematical formulations of operators, I acknowledge the contribution of Dr. Rolf A. de By and H.B. (Hiwot Berhane) Gidey (MSc student).

A schema transformation has a set of operations that help mapping the data sets to the FSSIM inputs by translating the source schema to the integrated schema. Operations need to be performed in sequence to produce a correct and appropriate result. In this study, a formal language basis for several transformation operators has been developed. These transformation operators are implemented in the database. A schema transformation might be expressed as,

The source dataset SD has as its type the source schema SS, and thus it contains tables (i.e., tha data models of input spatial datasets):

$$SD : \langle T_1 : I\!P\tau_1, \ldots, T_m : I\!P\tau_m \rangle .$$

The integrated dataset TD, likewise, has as its type the integrated schema TS, and will eventually also contain tables (i.e., the data model of BEFM parametric space):

$$TD : \langle T_1 : I\!P\tau_1, \ldots, T_n : I\!P\tau_n \rangle .$$

In the sections below, we define and discuss a small language of powerful expressions that we specifically developed to define schema transformations. Our intuition on that language is based on three important design principles:

- Schema transformations are functions that construct a dataset from a given dataset. Such functions may affect the structure but also the data content.
- We want to achieve a *modularized* transformation language, in which separate transformations can be combined to define an overall transformation; technically speaking, this means that we want to achieve functional *compositionality* of transformations. Simply put, if f_1 and f_2 are two independent transformations of a source

dataset SD, then the following equation should hold most of the time:

$$f_1 \circ f_2(SD) = f_2 \circ f_1(SD),$$

where $\circ$ denotes functional composition. We are writing "most of the time" because two transformations may not always be as independent from each other as we would wish. We return to discuss this matter in Section D.1.

- The formal semantics of the expression forms that we propose is defined in such a way that it allows us to express the intuition "take the whole dataset, make the indicated local change, and return the dataset thus obtained."

D.0.1 Formal semantics described intuitively

We follow the classical set-up of type theory. Our language has expressions, members of a set E, and types, members of a set T. Examples of expressions are: SD, 2, $25 + 12$, $SD.T_1$ and so forth. Examples of types are: integer, real, $\langle a : \text{integer}, b : \text{real}\rangle$, and so forth. Observe that record types are explicitly included, as we need them to formally describe datasets and tables in datasets.

There is an important relationship between E and T, known as the well-typedness relationship, and denoted as ":". When we write $44 :$ integer, we are saying that the expression 44 is well-typed, and has type integer. We thus know that $: \subseteq E \times T$. Obviously, not all expressions have all types, but some expressions may have multiple types. No examples given here.

Any formal language that uses type theory for its precise definition does not only define E, T and the well-typedness relation : between them, but also defines a formal semantics, often expressed as $[[\ldots]]$ of expressions $e \in E$ and types $\tau \in T$. So, $[[e]]$ is the formal semantics of e, and $[[\tau]]$ is the formal semantics of τ. An expression semantics is normally a value, while a type semantics is normally a set of values. Whenever the rules of our language say that $e : \tau$ holds, one is assured that we also have $[[e]] \in [[\tau]]$, i.e., the semantics of an expression is a member of the semantics of its type.

Important is to understand that a type, like integer, has a formal semantics written as [[integer]]. In this case, that $[[\ldots]]$ semantics of the type is a set of values. With integer, that set of values is $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. One can assume that all basic types have such semantics.

We need to say something specific about the semantics of records and record types. Above, we used as an example record type

$$\langle a : \text{integer}, b : \text{real}\rangle,$$

a type that identifies two attributes a and b, with respective types integer and real. An example record of that type is the following record r:

$$r \stackrel{\text{def}}{=} \langle a = 44, b = 3.14159265\rangle.$$

Observe that these are valid statements in our formal language.

What now is the semantics of that record r, and what is the semantics of its type? We assume that the attribute names for records and their types are drawn from a previously defined name set A, and so in the example case we have $a \in A$ and $b \in A$. the trick to look at records semantically, is by considering them functions that map attribute names nicely onto values. In the case of record r, that function is $[[r]]$, for which we know the domain:

$$dom[[r]] = \{a, b\},$$

which is a subset of A, obviously. The function maps to the semantics of expressions associated with the attribute names:

$$[[r]](a) = [[44]] \text{ and } [[r]](b) = [[3.14159265]].$$

This discussion leads us intuitively to the semantics of record *types*. Such a semantics can only be a collection of functions of the nature of $[[r]]$. And we can thus define the semantics of r's type as follows:

$$[[\langle a : \text{integer}, b : \text{real}\rangle]] \stackrel{\text{def}}{=} \left\{ \begin{array}{lll} f & | & f \text{ is a function} \wedge \\ & & dom(f) = \{a, b\} \wedge \\ & & f(a) \in [[\text{integer}]] \wedge \\ & & f(b) \in [[\text{real}]] \end{array} \right\}$$

These examples naturally generalize to arbitrary records and record types. We needed to discuss this because below we are going to define additional expression formats that can only be precisely defined once we understand the set-up sketched above.

D.0.2 Basic expressions needed

This section lists a number of expressions that are needed to define schema transformation expressions. They provide the fundamental 'building blocks' to construct larger, end-user-oriented expressions. Assume $r : \langle a_1 : \tau_1, \ldots a_n : \tau_n\rangle$ is a record, assume that R_1 and R_2 are tables of type $\mathbb{P}\langle a_1 : \tau_1, \ldots a_n : \tau_n\rangle$ with primary key $\{a_{i_1}, \ldots, a_{i_h}\}$, and that S, T and U are union-compatible value sets. We will use the notation keys(R_i) to denote the set of primary key values present in table R_i.

1. The expression

$$r \textbf{ except } a_j = e_j, \ldots, a_k = e_k \quad \text{where } \{a_j, \ldots, a_k\} \subseteq dom[[r]] \tag{D.1}$$

denotes a record r' obtained by copying r and replacing its attribute value for a_j by the value of e_j, and so on, until the attribute value for a_k, which value is replaced by the value of e_k. The except expression allows local attribute value overwriting. The formal semantics of this expression is a function shaped like $[[r]]$ that maps onto different semantic values for the listed attributes.

2. The expression

$$r \textbf{ dropatt } a_j, \ldots, a_k \quad \text{where } \{a_j, \ldots, a_k\} \subseteq dom[[r]] \tag{D.2}$$

denotes a record r' obtained by copying r and removing from it the listed attributes. The formal semantics of this expression is a function shaped like $[[r]]$ but with a reduced domain, mapping onto the same values for remaining attributes.

3. The expression

$$r \textbf{ addatt } a_j = e_j, \ldots, a_k = e_k \quad \text{where } \{a_j, \ldots, a_k\} \cap dom[[r]] = \emptyset \tag{D.3}$$

denotes a record r' obtained by copying r and adding to the record the named attributes with associated values. The formal semantics extends that or r by adding to the domain of the function, and mapping onto the values of the respective expressions.

4. The expressions

$$S \textbf{ union } T, S \textbf{ setminus } T, S \textbf{ intersection } T \tag{D.4}$$

denote the regular and well-known set-theoretic expressions.

5. The expression

$$R_1 \textbf{ exceptset } R_2 \quad \text{where keys}(R_2) \subseteq \text{keys}(R_1) \tag{D.5}$$

denotes a set R' obtained by copying R_1 and replacing those records that have a key in keys(R_2) by the corresponding records in R_2. Observe the similarity with the except expression, and the fundamentally different semantics.

6. The expression

$$R_1 \textbf{ addset } R_2 \quad \text{where keys}(R_1) \cap \text{keys}(R_2) = \emptyset \tag{D.6}$$

denotes a set R' obtained by copying R_1 and adding to it the records in R_2. Observe that the addset expression has a semantics similar to that of union, however, that its use comes with a precondition of non-overlapping key sets.

7. The expression

$$R_1 \textbf{ dropset } R_2 \quad \text{where keys}(R_2) \subseteq \text{keys}(R_1) \tag{D.7}$$

denotes a set R' obtained by copying R_1 and removing from it those records that have a record in R_2 with matching primary key. Observe that the dropset expression has a semantics similar to that of setminus, however, that its use comes with a precondition of key set inclusion.

8. Assume U denotes a set, u is a variable, ϕ is a predicate and μ an arbitrary expression. The predicative set expression

$$\{\ u \in U | \phi(u) \bullet \mu(u)\ \}, \tag{D.8}$$

denotes the set obtained from U, letting u range over its members, and producing those $\mu(u)$ values for which $\phi(u)$ tests true. This expression is a combination of a filter (ϕ) and a map (μ), both of which are only optional parts of the expression.

The filtermap expression naturally generalizes to the use of multiple sets, as follows:

$$\{\ u \in U, \ldots | \phi(u, \ldots) \bullet \mu(u, \ldots)\ \} \tag{D.9}$$

We may be needing more forms later, but we will leave it here for now. What we might be needing is a bag extension to the above where we have only provided notation and expressions for handling sets.

The reader should observe that many of the above expression forms have an approach of copying the original r or R, and leaving it as much intact as possible, applying only local change.

D.0.3 Renaming of attributes

Renaming is an important schema transformation step. It can take place at least at two levels, but possibly more. We propose two expression forms to express renaming, of which the first is the more generic.

Renaming of table names is an important schema transformation operation. It allows us to rename a source table into a integrated table. Let us assume that a is a table name in SD, but that b is not. The general form is then

$$SD \textbf{ renametable } a \textbf{ to } b. \tag{D.10}$$

The meaning of this expression can be given as a combination of our base expressions:

$$(SD \textbf{ addatt } b = SD.a) \textbf{ dropatt } a.$$

Renaming attributes of a table is also an important schema transformation operation. It allows us to rename an attribute of a source

table. Let us assume that T is a table name in SD, that a is one of its attributes, but that b is not. The general form is then

$$SD \textbf{ renameatt } T \textbf{ from } a \textbf{ to } b. \tag{D.11}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$SD \textbf{ except } T = \{\ t \in SD.T \bullet (t \textbf{ addatt } b = t.a) \textbf{ dropatt } a\ \}.$$

D.0.4 Adding and dropping attributes

Adding tables as attributes in our dataset, let us assume that $a_j, ..., a_k$ are table names and are not in SD. The general form is then,

$$SD \textbf{ addatt } a_j = e_j, \ldots, a_k = e_k$$

Dropping tables as attributes from our dataset, let us assume that $a_j, ..., a_k$ are table names in SD. The general form is then,

$$SD \textbf{ dropatt } a_j, \ldots, a_k.$$

Adding table attributes in one of the tables in our dataset, let us assume that T is a table name in SD and $a_j, \ldots, a_k$ are not attributes of T. The general form is then

$$SD \textbf{ addatt } a_j = e_j(t), \ldots, a_k = e_k(t) \textbf{ on } T \textbf{ as } t \tag{D.12}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$SD \textbf{ except } T = \{\ t \in SD.T \bullet (t \textbf{ addatt } a_j = e_j(t), \ldots, a_k = e_k(t))\}.$$

The $e_j(t), \ldots, e_k(t)$ are tuple variables to have a different value for each tuple in the table T.

Dropping attributes in one of the tables in our dataset, let us assume that T is a table name in our dataset SD and $a_j, \ldots, a_k$ are attributes of table T. The general form is then

$$SD \textbf{ dropatts } a_j, ..., a_k \textbf{ from } T \tag{D.13}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$SD \textbf{ except } T = \{\ t \in SD.T \bullet (t \textbf{ dropatt } a_j, \ldots, a_k)\ \}.$$

D.0.5 Filtering operation

Filtering is a conditional statement applied to a source dataset to extract subsets. The integrated dataset may require data that meets a certain condition and this condition must be provided on the integrated schema. Thus filtering allow us to extract dataset that meets the predefined conditions.

Filtering on a single table let us assume that T is a table name in SD and $\phi(t)$ is a filtering condition. The form of the operator is then

$$SD \textbf{ filter } T \textbf{ as } t \textbf{ on } \phi(t) \tag{D.14}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$SD \textbf{ except } T = \{\ t \in SD.T | \phi(t)\ \}.$$

Filtering on multiple tables is necessary to filter the tables in our dataset to produce a integrated dataset that meets all of the predefined conditions. Let us assume that $a_j, \ldots, a_k$ are table names in SD and $\phi(t_j, \ldots, t_k)$ are filtering conditions on the individual tables. The form of the operator is then

$$SD \textbf{ filter } a_j \textbf{ as } t_j, \ldots, a_k \textbf{ as } t_k \textbf{ on } \phi_j(t_j), \ldots, \phi_k(t_k) \tag{D.15}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$SD \textbf{ except } a_j = \{\ t_j \in SD.a_j | \phi_j(t_j)\ \}, \ldots, a_k = \{\ t_k \in SD.a_k | \phi_k(t_k)\ \}$$

D.0.6 Merging operator

The merging operator allows us to combine or concatenate two or more attribute values in the source dataset into one attribute value in the integrated dataset. Let us assume that T is a table name in SD and $a_j, \ldots, a_k$ are attributes of table T that are going to be concatenated, and the attribute b is not in T and thus the general form is then

$$SD \textbf{ mergeatt } a_j, \ldots, a_k \textbf{ to } b = e(a_j, \ldots, a_k) \textbf{ in } T \tag{D.16}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$SD \textbf{ except } T = \{\ t \in SD.T \bullet (t \textbf{ addatt } b = e(t.a_j, \ldots, t.a_k)) \textbf{ dropatt } a_j, \ldots, a_k\ \}.$$

The function e takes the attribute names $a_j, \ldots, a_k$ and combines or concatenates them.

D.0.7 Type conversion operator

The type conversion operator converts the data type of source dataset attribute to a data type that is specified on the integrated schema. The conversion could be a spatial type conversion or non-spatial type conversion. Let us assume that T is a table name in SD and a is an attribute of table T,

$$SD \textbf{ converttype } a \textbf{ to } \gamma \textbf{ in } T \tag{D.17}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$((SD \textbf{ addatt } a' = conv(a, \gamma) \textbf{ on } T \textbf{ as } t) \textbf{ dropatt } a \textbf{ from } T) \\ \textbf{renameatt } T \textbf{ from } a' \textbf{ to } a.$$

The function $conv(a, \gamma)$ takes expression a and changes the data type of a by γ, where γ is the data type of the integrated dataset required by the integrated schema.

D.0.8 Reclassification operator

The reclassification operator maps values of a source attribute value to a fixed integrated attribute value. The integrated dataset may need a certain number of classes and the integrated schema should contain information on the allowable classification values;

$$SD \textbf{ reclassify } a = f(a) \textbf{ in } T \tag{D.18}$$

The meaning of this expression can be given as a combination of our base expressions, through the following equivalent expression:

$$SD \textbf{ except } T = \{\ t \in SD.T \bullet (t \textbf{ except } t.a = f(a))\ \}.$$

The function $f(a)$ takes the attribute a and returns the classified value of a.

D.0.9 Augmentation operator

The augmentation operator is used to define values for attributes in the integrated dataset properties missing in source dataset. The operator assign default value of the attribute in the integrated dataset since the attribute was not available in the source dataset. The default values of the attributes should be defined in the integrated schema. This operator is the same as adding a table attributes in equation D.12.

D.1 Compositionality

As we have seen in Section D, functional compositionality may not exist between the CST operations. We are interested in $f \circ g = g \circ f$, this generally gives us a "yes" or "no" answer but more often it is a no answer and we want to know specifically under what condition it is a yes answer. In the matrix below we will see the condition where functional compositionality exist and where it does not.

$f \circ g = g \circ f$		f							
		C.10	**C.11**	**C.12**	**C.14**	**C.15**	**C.16**	**C.17**	**C.18**
g	**C.10**	Yes	No	No	No	No	No	No	No
	C.11	No	Yes	No	Yes	No	No	No	No
	C.12	No	No	Yes	No	No	No	Yes	No
	C.14	No	Yes	No	Yes	No	No	Yes	No
	C.15	No	No	No	No	Yes	No	No	No
	C.16	No	No	No	No	No	Yes	No	Yes
	C.17	No	No	Yes	Yes	No	No	Yes	No
	C.18	No	No	No	No	No	yes	No	Yes

Table D.1 Compositionality matrix

D.1.1 Compositionality case proof

- Case 1
  In general, we want to know of the many expression forms introduced above, whether they can be used as composed functions in complex schema transformations. If we have two schema transformations f and g, we are interested in finding out whether $f \circ g = g \circ f$, and if not which are the conditions under which that equality holds. We look at a single case here first. The case we have in mind is that, $f \equiv 3.11$ and $g \equiv 3.14$ where, $f \equiv SD$ **renameatt** T **from** a **to** b and $g \equiv SD$ **filter** T **as** t **on** $\phi(t)$. To study the compositonality, we need to look at two expressions. The first is

  $$f \circ g = (SD \textbf{ filter } T \textbf{ as } t \textbf{ on } \phi(t)) \textbf{ renameatt } T \textbf{ from } a \textbf{ to } b.$$

  The second is where f is applied to the data set before g:

  $$g \circ f = (SD \textbf{ renameatt } T \textbf{ from } a \textbf{ to } b) \textbf{ filter } T \textbf{ as } t \textbf{ on } \phi(t).$$

  It is fairly obvious to see that these two expression are equivalent, *provided* that predicate $\phi(t)$ does *not* express its condition on t in terms of $t.a$. If it does, then the use of $\phi(t)$ in the second expression is nonsensical, and it should there be replaced by $\phi[a \rightarrow b](t)$. The latter itself would need to be defined as "just $\phi(t)$, but with any use of $t.a$ having been substituted by $t.b$."

- Case 2
  Let us see two schema transformations f and g, and $f \equiv 3.16$ and $g \equiv 3.18$,
  $f \equiv SD$ **mergeatt** $a_j, \ldots, a_k$ **to** $b = e(a_j, \ldots, a_k)$ **in** T and
  $g \equiv SD$ **reclassify** $a = f(a)$ **in** T. To study the compositonality, we need to look at two expressions. The first is

$$f \circ g = (SD \textbf{ reclassify } a = f(a) \textbf{ in } T) \textbf{ mergeatt } a_j, \ldots, a_k \textbf{ to } b = e(a_j, \ldots, a_k) \textbf{ in } T.$$

  The second is where f is applied to the data set before g:

$$g \circ f = (SD \textbf{ mergeatt } a_j, \ldots, a_k \textbf{ to } b = e(a_j, \ldots, a_k) \textbf{ in } T) \textbf{ reclassify } a = f(a) \textbf{ in } T.$$

  We can say that these two expressions are equivalent, *provided* that the merging function $e(a_j, \ldots, a_k)$ does *not* use the attribute a. If it does, then the reclassification in the second expression does not make sense.

- Case 3
  Let us see two schema transformations f and g, and $f \equiv 3.17$ and $g \equiv 3.12$ where $f \equiv SD$ **converttype** a **to** γ **in** T and $g \equiv SD$ **addatt** $a_j = e_j(t), \ldots, a_k = e_k(t)$ **on** T **as** t. To study the compositonality, we need to look at two expressions. The first is

$$f \circ g = (SD \textbf{ addatt } a_j = e_j(t), \ldots, a_k = e_k(t) \textbf{ on } T \textbf{ as } t) \textbf{ converttype } a \textbf{ to } \gamma \textbf{ in } T.$$

  The second is where f is applied to the data set before g:

$$f \circ g = (SD \textbf{ converttype } a \textbf{ to } \gamma \textbf{ in } T) \textbf{ addatt } a_j = e_j(t), \ldots, a_k = e_k(t) \textbf{ on } T \textbf{ as } t.$$

  The above two expressions are equivalent, *provided* that the the two equation works on different attribute but even if they work on the same attribute the output is the same due to the *where* clause that restrict the equations to work on already existing attribute for *convettype* and *addatt* make sure the attribute names does not exist.

Farmer communities and AGRISTAT data collection in Burkina Faso

E

Figure E.1 Interviewing the farmer communities during fieldwork studies in Burkina Faso.

Figure E.2 Questionnaire forms collected during country-wide agricultural surveys by the Statistiques Agricoles du Burkina Faso (aka AGRISTAT).

Samenvatting

De voorwaarden voor gewasgroei in West Afrika variëren sterk in ruimte en tijd. Om hier goed mee om te gaan is een grote variatie aan berekeningsmodellen ontwikkeld. Deze zijn gebaseerd op landbouwkundige processen op verschillende schalen. Boeren en adviseurs hebben wetenschappelijke hulpmiddelen nodig die het mogelijk maken om toegang te krijgen tot gegevens, deze te combineren en te schatten met als doel om duurzame oplossingen te bereiken op de schaal van een boerenbedrijf. Idealiter zijn deze hulpmiddelen onderdeel van een landbouwkundig technologische infrastructuur van ruimtelijke gegevens (SDI). Hiermee wordt het mogelijk adviezen te maken voor een heel continent. In dit proefschrift worden vier studies gecombineerd die het maken van zo'n landbouwkundig SDI ondersteunen voor Burkina Faso.

De eerste studie suggereert en ontwikkelt een flexibel kader voor het opschalen van gegevens bestanden en voor het verbinden van dergelijke bestanden met simulatiemodellen. Dit kader is gebaseerd op een SDI. Een advies georiënteerde architectuur van een SDI stelt ons in staat op gegevens bestanden en modellen te gebruiken als herbruikbare web diensten. De studie onderzoekt hoe een open structuur voor een SDI omgeving kan worden gebruikt die door meerdere gebruikers kan worden benut om gegevens en modellen te integreren voor het ontwikkelen van continent-brede adviezen. Daarnaast onderzoekt de studie hoe een dergelijke omgeving er voor kan zorgen dat modellen worden aangepast aan opgeschaalde gegevens die via veld inventarisaties zijn verzameld. De toegang tot gegevensbestanden en modellen wordt ontwikkeld via standaard verpakking implementaties. Het voorgestelde kader is toegepast bij het nemen van beslissingen op boerderijen in Burkina Faso. Om dit te realiseren maken de verpakking implementaties gebruik van een landbouwkundig simulatiemodel die het 'Model als een dienst' paradigma volgen, alsmede van gegevensbestanden als ruimtelijk beschikbare diensten. Het orkestreren van dergelijke diensten staat een gebruik toe van participatieve integratie van de verschillende landbouwkundige middelen. Bij het toetsen van de diensten binnen het studiegebied ontdekte de studie dat het kader baat heeft van de verschillende gegevensdiensten in de huidige SDI implementaties. Om de variabelen aan te passen die verzameld zijn uit landbouwkundige inventarisaties in Burkina Faso voor het toepassen van SDI diensten, vond de studie dat het bovendien nodig

was om ruimtelijk statistische methoden en om satellietwaarnemingen te gebruiken om de gegevens op te schalen naar het landelijk niveau.

De tweede studio gebruikt gegevens aan biofysische, socio-economische en menselijke hulpmiddelen verzameld binnen terroirs in Burkina Faso om gewasoogsten te schatten en deze op te schalen naar het landelijke niveau. De studie onderzoekt de toepassing van satellietgegevens om ruimtelijke variatie in oogstgegevens te analyseren. Een samengestelde SPOT-VEGETATION (NDVI) tijdreeks van 1 km resolutie, 10-daagse composieten is gebruikt voor de periode die de gewasgroei omsluit. Veldwaarnemingen aan gewassen zijn verkregen vanuit veldinventarisaties die gepubliceerd zijn in nationale statistische gegevensbanken. Daarnaast zijn gegevens gebruikt die zijn verkregen uit online gegevensbestanden die voor Afrika beneden de Sahara zijn verkregen vanuit satellietbeelden. Geografisch gewogen regressiemodellen zijn gebruikt om de gegevens te interpoleren vanaf het veldniveau naar het landelijke niveau. De schattingen die zo verkregen zijn zijn opgeslagen in een geodatabase. De ruimtelijke gegevensdiensten die naast deze geodatabase zijn gebruikt kunnen op een geschikte manier een boerderijmodel calibreren op een terroir locatie. Onzekerheid die voortkomt uit een beperkte beschikbaarheid van de gegevens heeft een voor de hand liggend effect op de stabiliteit van de statistische modellen bij het schatten van de volledige ruimtelijke variatie van oogsten in een sterk ruimtelijk heterogeen landschap. Hiervoor moest de modelonzekerheid voor gewasopbrengsten op de landelijke schaal gemodelleerd worden. De studie concludeert dat statistische methoden en satelliettechnologie gebruikt kunnen worden voor het opschalen van gewas oogsten. Op die manier konden schattingen verkregen worden voor het hele land.

De derde studie kwantificeert de onzekerheid in gewas opbrengst modelleren op de nationale schaal. Hiervoor zijn gewas opbrengst observaties gebruikt die over het hele land zijn verkregen via geografische enquétes en ruimtelijke statistische opschaling. De studie presenteert een hybride aanpak waarbij ordinary kriging en geografisch gewogen regressie worden geïntegreerd. Deze geografisch gewogen regressiekriging aanpak is toegepast in Burkina Faso. De studie toont aan dat het kwantificeren van onzekerheden in gewasmodellen voor grotere gebieden bijdraagt aan een verbetering van de bronnen van onzekerheid die worden gegeven door het bemonsteringsontwerp en de model structuur. Bovendien kunnen de onzekerheidskaarten die op deze manier verkregen zijn het vertrouwen van eindgebruikers vergroten door rekening te houden met de onzekerheid in de nauwkeurig geschatte opbrengsten voorspellingen.

De vierde studie onderzoekt regionale en mondiale datasets, inclusief RS producten, voor het modelleren van de marginaliteit status van terroir gemeenschappen. Opgeschaalde gegevens worden gebruikt die verzameld zijn via huis-aan-huis enquétes in Burkina Faso. Schattingen van de marginaliteit worden opgeschaald naar aan de nationale schaal. Voor dit doel wordt ervan uitgegaan dat de sociaaleconomische status van de terroir gemeenschappen grotendeels afhankelijk is van

het agroklimaat potentieel van de landbouwsystemen. Dit potentieel kan worden vastgesteld uit regionale en mondiale datasets. Informatie over biofysische factoren die invloed hebben op het agroklimaat potentieel van terroirs is verkregen uit SPOT-VEGETATION NDVI-waarden en uit regenval schattingen geëxtraheerd uit TAMSAT gegevens. Een indicator is ontwikkeld die menselijke, sociale en financiële vaste activa van terroir gemeenschappen binnen landbouwsystemen kwantificeert. Via een statistische analyse is de relatie was opgeschaald om marginaliteit te schatten op nationale schaal. Geografisch gewogen regressie was in staat de landbouwsystemen in kaart brengen. Op die manier is een beter inzicht te verkrijgen over de status van de marginaliteit van boeren binnen terroirs. Hierbij zijn integrerende modellen gekalibreerd op de nationale en regionale schaal. Deze kalibratie vereiste aanpassing van de marginaliteit status om het vermogen van terroirs te kunnen beoordelen bij het toepassen van meststoffen, ongediertebestrijding en gewas variëteiten.

Samenvattend, ruimtelijke statistiek is als kernmethodologie toegepast voor het opschalen van opbrengsten en de gemeenschappelijke marginaliteit status op het terroir niveau naar de landelijke schaal van Burkina Faso. Kwantificering en voortplanting van onzekerheden zou vanaf nu een integraal onderdeel moeten zijn bij het onderzoek naar ruimtelijke modellering en opschaling. Om die reden is deze methodologie gebruikt voor onzekerheid evaluatie. Het kwantificeert de onzekerheden die aanwezig zijn in de bemonstering van representatieve terroirs in Burkina Faso en in de ruimtelijke modellen voor opschaling. Het gebruik van het SDI kader kan hierbij een robuuste omgeving bieden voor het integreren van datasets en ruimtelijke opschaling voor boerderij simulatie modellen voor de ontwikkeling landbouwdiensten op de nationale schaal.

Biography

Imran (Muhammad) was born on 1st of August in Khushab, an agricultural-based rural district in Punjab, Pakistan. From 1998 to 2000, he studied in University of the Punjab, Pakistan, and obtained Master degree in Physics. In 2003, he joined Center for Information Technology Pakistan Railways (CITPR). At CITPR he led to establish computerized and ticketing system in the entire country. Later at CITPR much effort was invested in developing the freight container tracking system based on GIS and remote sensing. His works were highly appreciated in the realm of efficient transportation services across the country. During this period, he also received master degree in IT (2004-2006) from University of the Punjab. This and his background motivated his research in agricultural systems with emphasis on GIS and remote sensing. In Sep 2007, he started his research under MS leading towards PhD program with a joint scholarship provided by the Higher Education Commission (HEC) of Pakistan and the Netherlands Organization for Cooperation in Higher Education (NUFFIC). In March 2009, he received master degree in Geo-informatics from the faculty of Geo-information Science and Earth Observation (ITC). Subsequently he pursued the doctoral study in the same faculty. With emphasis on spatial statistics, remote sensing and spatial data infrastructure technology, his PhD research aimed to design a framework for wall to wall agricultural data, models and services at the regional scale. Besides this, his main research interests include modeling Geodata, integrating Geodata and models for location-based web services and developing spatial decision-support systems for environmental applications.

During the course of the PhD study, the research output was presented in the form of a number of research papers published in reputed journals, and in international conferences. They are listed here for

reference:

- Imran, M, Zurita-Milla, R, Stein, A. (2013). Modeling crop yield in West-African rainfed agriculture using global and local spatial regression. *Agronomy Journal*:105(4).
- Imran, M, Stein, A., Zurita-Milla, R. (2014). Investigating rural poverty and marginality in Burkina Faso using remote sensing-based products. *International Journal of Applied Earth Observation and Geoinformation*: Volume 26, Pages 322–334.
- Imran, M. (2010) SDI - based architecture for integrated agricultural assessments and decision - making by farmer communities in sub - Saharan Africa. In: Proceedings of the GIScience 2010 doctoral colloquium, Zurich, Switzerland, September 2010 / J.O. Wallgrün, A.-K. Lautenschütz. - Heidelberg: Akademische Verlagsgesellschaft, 2010. - 86 p. ; 24 cm. ISBN 978-3-89838-640-1. pp.45–50.
- Imran, M., Zurita-Milla, R. and de By, R.A. Integrated environmental modeling: an SDI - based framework for integrated assessment of agricultural information. Presented at AGILE 2011: the 14th AGILE International Conference on Geographic Information Science, 18-21 April 2011, Utrecht, Netherlands. 9 p.
- Imran, M., Zurita-Milla, R. and de By, R.A. Uncertainty in agricultural integrated assessment workflows in the SDI framework: abstract. In: Spatial statistics 2011 : mapping global change, abstracts for the conference, 23-25 March, 2011, University of Twente, Enschede, The Netherlands / editor A. Stein, , E. Pebesma, , G.B.M. Heuvelink. - Enschede: University of Twente Faculty of Geo-Information and Earth Observation ITC, 2011. 1 p http://intranet.itc.nl/papers/2011/pres/imran_unc.pdf.

ITC dissertations

A complete list of ITC dissertations is online on the ITC website: www.itc.nl/research/phd/phd_graduates.aspx.

This dissertation has number 234.