Aalto University
School of Engineering

An Exploration of Water Poverty in Lao People's Democratic Republic

Master's Thesis
Department of Built Environment
School of Engineering
Aalto University

Espoo, 8 August 2016

Bachelor of Engineering in Environmental Engineering Marko Kallio

Supervisor: Professor Kirsi Virrantaus
Advisor: Professor Kirsi Virrantaus

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076
AALTO www.aalto.fi

Diplomityön tiivistelmä

Tekijä Marko Kallio	
Työn nimi An Exploration of Water Poverty in Lao People's Democratic Republic	
Koulutusohjelma Geomatiikka	
Pääaine Geoinformation Technology	**Koodi** IA3002
Työn valvoja Professori Kirsi Virrantaus	
Työn ohjaaja(t) Professori Kirsi Virrantaus	

Päivämäärä	**Sivumäärä**	**Kieli**
06.08.2016	88+27 sivua	Englanti

Tiivistelmä

Yhdistyneet Kansakunnat julkaisi äskettäin uudet kestävän kehityksen tavoitteet seuraaville 15 vuodelle. Vähintään kuusi uusista 17 tavoitteesta kytkeytyy suoraan veteen, ja vesi vaikuttaa epäsuorasti useaan muuhun tavoitteeseen. Vesiresurssit ovat täten keskeisessä asemassa tavoitteiden saavuttamisessa. Tämän aseman lisäksi Maailman Talousfoorumi on arvioinut että veteen liittyvät ongelmat ovat eräitä ihmiskunnan suurimpia haasteita tulevaisuudessa. Näihin ongelmiin puuttuminen vaatii tehokkaita ja integroituja keinoja, kuten Vesiköyhyysindeksi (VKI). VKI on kokonaisvaltainen työkalu jolla voidaan arvioida vesiresurssien tilaa niin veden riittävyyden, saavutettavuuden kuin ympäristönkin kannalta.

Tämä diplomityö tutkii vesiköyhyyden maantieteellistä ja ajallista jakautumista Laosissa VKI:ä hyödyntäen. Laos sijaitsee alueella jonka vesiresursseja hallitsee monsuunisateet aiheuttaen suuren eron kuivan ja sadekauden välille. Tämän eron vuoksi työssä kehitetään VKI jolla voidaan ottaa huomioon sade- ja kuivakausi ja vertailla niitä mielekkäästi. Molemmille vuodenajoille lasketaan oma VKI ja joita vertaillaan erilaisilla eksploratiivisen (spatiaalisen) data-analyysin ja spatiaalisen tietojenlouhinnan keinoin. Lisäksi tutkitaan mitkä syyt aiheuttavat vesiköyhyyttä Laosissa, ja muuttuvatko ne maantieteellisesti tai vuodenajoittain. Tutkimus perustuu avoimesti saatavilla olevaan dataan, ja analyyseihin käytetty R-koodi julkaistaan avoimesti internetissä osoitteessa http://markokallio.fi/waterpoverty/.

Tutkimuksessa löytyi merkittäviä eroja vesiköyhyyden maantieteellisessä ja kausittaisessa jakautumisessa. Vesiköyhyys on suurta vuoristoisissa ja vaikeasti saavutettavissa paikoissa maaseudulla, mutta pienenee mitä lähemmäs Mekong-jokea kuljetaan. Vesiköyhimmät provinssit ovat Xekong, Oudomxai ja Phonsaly, kun rikkaimmat löytyvät pääkaupungin ympäristöstä. Suuria eroja löytyi myös kuivan- ja sadekauden välillä; mitä suurempaa vesiköyhyyttä kuivalla kaudella esiintyy, sitä vähemmän tilanne paranee sadekaudelle tultaessa. Joillakin seuduilla vesiköyhyys on jopa suurempaa sadekaudella kuin kuivalla kaudella. Pääsyyt köyhyydelle löytyvät VKI:n käyttö-, saatavuus- ja suhteellisesti kapasiteetti-osaindeksistä. Veden riittävyys on ongelma pääasiassa vain läntisissä ja luoteisissa provinsseissa kuivan kauden aikaan.

Avainsanat Vesiköyhyysindeksi, WPI, köyhyys, datan louhinta, spatiaalinen datan louhinta, eksploratiivinen data analyysi, Laos, Lao PDR, spatiaalinen klusterointi, maantieteellisesti painotettu regressio, paikalliset tilastot, PCA

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076
AALTO
www.aalto.fi

Abstract of master's thesis

Author	Marko Kallio
Title of thesis	An Exploration of Water Poverty in Lao People's Democratic Republic
Degree programme	Degree Programme in Geomatics
Major	Geoinformation Technology
Code	IA3002
Thesis supervisor	Professor Kirsi Virrantaus
Thesis advisor(s)	Professor Kirsi Virrantaus

Date	Number of pages	Language
06.08.2016	88+27 pages	English

Abstract

The United Nations recently revised and published new Sustainable Development Goals for the next 15 years. At least six of the 17 goals are directly linked to water, and several others are indirectly affected by water issues. Water is central to achieving the goals and water-related issues have been identified by the World Economic Forum as some of the biggest risks the world is facing in the future. Efficient measures to address the issues require integrated approaches, such as the Water Poverty Index (WPI). WPI is a holistic tool to assess water resources in an integrated way, combining water resource availability, social dimensions of access and capacity to manage water resource as well as the environmental requirements for utilization of water.

This thesis examines the spatio-temporal distribution and the causes of water poverty in Lao PDR through WPI. Laos is located in Monsoon Asia with extreme seasonal differences in water availability. Due to this seasonality, WPI is developed in a manner that allows computing dry and wet season WPI separately and comparing them in a meaningful way. Exploratory (spatial) data analysis as well as spatial data mining methods are employed to investigate the distribution and causes of water poverty. The research is based on freely available data, and R code used in the analyses are openly published at http://markokallio.fi/waterpoverty/.

Significant spatial and temporal differences are found. Water poverty is high in the rural areas and in the mountains, while the low-lying lands near the Mekong river exhibit relatively low water poverty. Three provinces; Xekong, Oudomxai and Phongsaly are very poor, while the area around Vientiane Capital show least water poverty. Major difference is found also between seasons with WPI increasing in the water-rich more than in the water-poor areas as the wet season starts. In addition, it was found that in some locations, water poverty is higher during the wet season than in the dry season. The main causes driving water poverty are found to be Use and Access related, and in relative terms, Capacity related (especially village road access). Resource availability is problematic mainly in the western and northwestern provinces during the dry season.

Keywords Water Poverty Index, WPI, poverty, data mining, spatial data mining, exploratory data analysis, Laos, Lao PDR, spatial clustering, geographically weighted regression, local statistics, PCA, geographically weighted PCA, Monsoon Asia

Foreword

This thesis is the end (a temporary end, I might add) of a long road that started when I enrolled to study Environmental Engineering and decided to specialize in water instead of renewable energy. Ever since, water in one form or another has been an increasing theme in the beginning of my academic career as well as in private life. The final "nudge" came when I started my engineering thesis studying water quality of Nam Ngum Watershed in Lao PDR. The work on Nam Ngum introduced me to the world of modelling and GIS. In fact, this introduction was so powerful that it led me to choose Geoinformatics as the major of my Master's Degree.

The exact topic of this study took over a year to form, and included a trip to Laos and Cambodia to meet with experts and to get a feel for the region – something that was lacking from the previous thesis. After the first trip, I spent two months in Vientiane modelling water resources in a small catchment of Nam Xong. During this time, I decided that the topic shall be about water and its connection to society. Still, it took six more months before water poverty was settled as *the* issue I'd study – from candidates such as poverty per se or migration due to water. This social dimension was something entirely new to me, and extremely interesting because it allowed me to combine my skills in water and geoinformatics with this entirely new field. I have not regretted the choice – instead, this topic has efficiently cleared my mind and helped me realize what topic's I'd like to continue to study.

I have always been puzzled by the reason why people thank their family members in forewords and acknowledgements of a Master's Thesis (or any other book or a study for that matter). It always seemed to me like another assignment, a piece of work for a university that is just a requirement to finish the degree. Working with this thesis has shown this image I held entirely false.

This thesis is also a fulfilment of a childhood dream: to become a researcher. At least, it is an important step in truly becoming one. There are many different people who I should thank that made this journey possible. I mention here the most influential ones.

First I would like to thank the two persons who made it possible for me to do meaningful work in the context of water resources, and without whom I would not be on this path: Juha Sarkkula and Jorma Koponen. Meeting these two extraordinary persons several years ago set me to the path that would lead me to study the Mekong Region. In addition to Juha and Jorma, Alex Smajgl and John Ward from the Mekong Region Futures Institute inspired the final topic of this thesis. These four persons are also to thank for the chance to work for a few months in Lao PDR in the summer of 2015.

A number of other people have also influenced my work, especially in the beginning when the topic and research questions were still taking shape. Matti Kummu from the Water and Development Research Group in Aalto University provided invaluable help and comments along the research process. Without a lengthy discussion with Aura Salmivaara, the research would have probably gone to an unnecessarily difficult direction.
Kirsi Virrantaus, the supervisor and advisor to this thesis is of course an important person whose helpful comments and direction helped to shape the theory and methodology that provide the backbone of the work. She also took the trouble of reading through my drafts while on a vacation in order for me to reach my deadlines.

And the family? My wife Federica had to endure the stress I experienced over the work. She kept on backing me up even on the occasions when the thesis and my work consumed all the available time I had in an overwhelming manner. She's a treasure.

A special thank also belongs to Maa- ja Vesitekniikan Tuki ry, whose two scholarships first allowed me to travel to Southeast Asia to plan the thesis and finally helped to make this thesis a reality.

8th August 2016 in Helsinki

Marko Kallio

Contents

Abstract
Tiivistelmä
Foreword

Contents
List of Figures
List of Tables
Abbreviations

1 Introduction 1
2 Background 3
2.1 Introduction to Lao PDR 3
2.2 Water Poverty Index 5
2.2.1 Water Poverty Research in Lao PDR 6
3 Theory 8
3.1 Spatial is Special 8
3.1.1 Spatio-Temporality 9
3.1.2 Spatial Autocorrelation 10
3.1.3 Modifiable Area Unit Problem 10
3.2 Exploratory Data Analysis 12
3.2.1 Univariate Exploration 13
3.2.2 Multivariate Exploration 15
3.2.3 Exploratory Spatial Data Analysis 16
3.2.4 Geographically Weighted Summary Statistics 17
3.3 Spatial Data Mining 19
3.3.1 Spatial Clustering 20
3.3.2 Spatial Regression 21
3.3.3 Geographically Weighted Principal Component Analysis 22
4 Materials and Methods 24
4.1 Data and Data Sources 24
4.2 Developing the Water Poverty Index 25
4.2.1 Resources Index 25
4.2.2 Access Index 26
4.2.3 Capacity Index 27
4.2.4 Use Index 27
4.2.5 Environment Index 28
4.2.6 Calculating the Water Poverty Index 29
4.3 Analysis Methodology 30
4.4 Implementation Tools 32
5 Results 34
5.1 Exploring the Variables 34
5.1.1 Resources 35
5.1.2 Access 38
5.1.3 Capacity 39
5.1.4 Use 41
5.1.5 Environment 43
5.2 Spatial Dimensions of Water Poverty 44
5.2.1 Dry Season 44
5.2.2 Wet Season 48

5.3 Seasonal Water Poverty 51
5.3.1 Weighting Schemes 51
5.3.2 Difference in Seasonal Index Scores 54
5.4 Mining the Causes of Water Poverty 60
5.4.1 Cluster Analysis 60
5.4.2 Geographically Weighted Principal Component Analysis 64
5.4.3 Geographically Weighted Regression 71
6 Discussion 77
6.1 Weaknesses 78
6.2 Strengths 79
7 Conclusion 80
7.1 Spatial Variation in Water Poverty 80
7.2 Spatio-Temporal Variation in Water Poverty 81
7.3 Causes of Water Poverty 81
7.4 The Way Forward 82

References 83

Appendix 1. Vmod Model Description 89
Appendix 2. Additional Data for Initial Dataset Exploration 92
Appendix 3. Additional Data for Water Poverty Index Exploration 96
Appendix 4. Additional Data for Cluster Analysis 101
Appendix 5. Additional Data for Geographically Weighted Principal Component Analysis 102
Appendix 6. Additional Data for Geographically Weighted Regression 105

List of Figures

Figure 2.1. Hillshaded Digital Elevation Map of Laos. 3
Figure 2.2. General map of Laos with important economic corridors (roads), major rivers and administrative capitals. 4
Figure 3.1. Illustration of different spatial data types. (O'Sullivan & Unwin, 2010) 9
Figure 3.2. Illustration of the Modifiable Area Unit Problem and its effects on regression. (O'Sullivan & Unwin, 2010) 11
Figure 3.3. An example of a stem and leaf plot. (Steltman, 2015) 14
Figure 3.4. An annotated Box plot. IQR stands for Interquartile range, Q1 and Q3 stand for the first and the third quartile. (Steltman, 2015) 14
Figure 3.5. QQ-plots of normally (left) and non-normally (right) distributed samples. (Steltman, 2015) 14
Figure 3.6. Parallel Coordinate Plot of automotive data. Source: GGobi; http://homes.cs.washington.edu/~jheer//files/zoo/ex/stats/parallel.html. 15
Figure 4.1. The exploratory method used in the study. 31
Figure 5.1. Number of villages in the dataset for each province of Laos. 34
Figure 5.2. Dot density map of villages in the dataset. Coordinate system used is UTM Zone 48N (EPSG: 32648). 35
Figure 5.3. Resources component variability in a) dry and b) wet season. Hinges of the boxes signify 25% and 75% of the sample and whiskers cover 1.5 times the interquartile range. 36
Figure 5.4. Villages with water scarcity (water availability score less than 100) in a) dry and b) wet seasons. 36
Figure 5.5. Scatterplot matrix for Resources component. The upper matrix is for dry season and the lower for wet season. 37
Figure 5.6. Access component variability. Hinges of the boxes signify 25% and 75% of the sample and whiskers cover 1.5 times the interquartile range. 38
Figure 5.7. Capacity component variability for the dry season. Hinges of the boxes signify 25% and 75% of the sample and whiskers cover 1.5 times the interquartile range. 40
Figure 5.8. Villages without road access in a) dry and b) wet season. 40
Figure 5.9. Scatterplot matrix for dry season Capacity component. 41
Figure 5.10. Use component variability for wet season. 42
Figure 5.11. A scatterplot and a boxplot between wet and dry season irrigation. 42
Figure 5.12. a) Environment component variability for the dry season and b) disaster scoring difference between the seasons. 43
Figure 5.13. WPI components and WPI for dry season. The WPI score is calculated using PCA derived weighting scheme from the components and combined using a multiplicative function. 46
Figure 5.14. Stacked density (left) and normal density plots for dry season WPI in each province. 47
Figure 5.15. Rank plot of dry season WPI. Ranks are ordered so that the highest ranks are given to villages with the highest WPI. 47
Figure 5.16. WPI components and WPI for wet season. The WPI score is calculated using PCA derived weighting scheme from the components and combined using a multiplicative function. Note that the colouring is relative to the component, not over the entire range 0-100. 49
Figure 5.17. Stacked density (left) and normal density plots for wet season WPI in each province. 50
Figure 5.18. Wet season WPI ranks, ordered so that higher rank is given to the villages with higher WPI. 50

Figure 5.19. Loadings of the first three principal components from scheme using data from both seasons. 52
Figure 5.20. WPI calculated from the three objective weighting schemes: a) dry season WPI with weights derived from dry season only, b) wet season WPI from weights derived from wet season only, c) dry season WPI from weights derived using both seasons, and d) wet season WPI from weights derived using both seasons. 53
Figure 5.21. WPI difference between single-season and both-seasons weighting schemes. Calculated by subtracting single-season weighted WPI from the both-season weighted WPI. 54
Figure 5.22. Difference between wet and dry season WPI. Calculated by subtracting dry season from the wet season WPI (both seasons weighting). 55
Figure 5.23. Mean provincial a) WPI and b) WPI rank for dry and wet seasons. The provinces are ordered according to the wet season. 56
Figure 5.24. Local mean WPI and local standard deviation for dry and wet season. Bandwidth for the calculation is 400 nearest neighbours using Gaussian weighting scheme. 57
Figure 5.25. Scatterplot of the WPI between wet and dry seasons. 58
Figure 5.26. Local correlations for the components which (on average) significantly correlate with seasonal WPI's. Dry season WPI correlates with a) RES, b) CAP and c) USE. Wet season correlates with d) CAP, e) USE and f) ENV. 59
Figure 5.27. Selected spatial *k*-means clustering schemes for a) dry and b) wet season. 61
Figure 5.28. Boxplot of WPI components for each cluster in dry and wet season. 62
Figure 5.29. Selected Rank based clusters for a) dry and b) wet season. 64
Figure 5.30. The highest loading (”winning”) dry season WPI components for the first three Principal Components and bar plots of their frequencies. 65
Figure 5.31. The highest loading (”winning”) wet season WPI components for the first three Principal Components and bar plots of their frequencies. 66
Figure 5.32. Boxplot of GWPCA derived weights for individual villages. 66
Figure 5.33. Spatial variation of the locally derived component weights. Subplots a-e show the dry season weighting scheme while f-j present the wet season weighting. Neutral colour signifies an equal weighting scheme where all components are weighted at 0.2. 67
Figure 5.34. WPI calculated using locally derived weights for a) dry and b) wet season. The plots on the second row show the difference between WPI calculated from locally and globally weighted WPI for c) dry and d) wet season. The difference is calculated by subracting globally weighted WPI from the locally weighted one. 69
Figure 5.35. Dry season WPI subtracted from wet season WPI, both calculated using GWPCA weighting scheme. 70
Figure 5.36. GWR model R2 for a) dry and b) wet seasons. 72
Figure 5.37. Dry season model prediction (a) and residuals (b) and wet season model prediction (c) and (d) residuals. 74

List of Tables

Table 2.1. Variables used by Lawrence et al (2002) to calculate an international comparison of WPI. 7
Table 3.1. Main ESDA techniques according to Anselin (1998). 17
Table 3.2. Typical descriptive statistics for the univariate probability density function f(x) (Brunsdon et al, 2002) 18
Table 3.3. Data Mining methods and techniques according to Miller and Han (2009). . 20
Table 4.1. Summary of the variables and scoring used for calculating the components. 25
Table 4.2. Disaster types and the index calculations they are used for. 28
Table 5.1. Moran's I for Resource component variables. 38
Table 5.2. Moran's I for Access component variables. 39
Table 5.3. Moran's I for Capacity component variables. 41
Table 5.4. Moran's I for Use component variables. 43
Table 5.5. Moran's I for Environment component variables. 44
Table 5.6. Objective weights for the components derived using Principal Component Analysis. 51
Table 5.7. Mean local correlations between seasonal WPI and their corresponding components. 58
Table 5.8. Number of clusters proposed by the 24 indices used by NbClust package. .. 60
Table 5.9. Share of villages assigned in rich clusters for each province. 63
Table 5.10. Model goodness statistics for dry and wet season global and local models. 72
Table 5.11. (Step-wise) selected model variables in the order of selection and the p-values of monte carlo test for spatial homogeneity of the coefficients. 73
Table 5.12. Variables with collinearity problems according to VIF and VDP diagnostics. 75

Abbreviations

ACC	Access Component
AGNES	Agglomerative Nesting
CAP	Capacity Component
CDA	Confirmatory Data Analysis
CV	Cross-Validation
DIANA	Divisive Analysis
DM	Data Mining
EDA	Exploratory Data Analysis
ENV	Environment Component
ESDA	Exploratory Spatial Data Analysis
GW	Geographically Weighted
GWPCA	Geographically Weighted Principal Component Analysis
GWR	Geographically Weighted Regression
GWSS	Geographically Weighted Summary Statistics
HDI	Human Development Index
IFAD	International Fund for Agricultural Development
IQR	Inter-Quartile Range
Lao PDR	Lao People's Democratic Republic
MAUP	Modifiable Area Unit Problem
Moran's I	Moran's Index
MRC	Mekong River Commission
PC	Principal Component
PCA	Principal Component Analysis
PCP	Parallel Coordinate Plot
RES	Resources Component
SAR	Spatial Autoregressive Model
SDG	Sustainable Development Goal
SDM	Spatial Data Mining
USE	Use Component
VDP	Variance Decomposition Factor
VIF	Variance Inflation Factor
WPI	Water Poverty Index

1 Introduction

The United Nations recently revised and published new Sustainable Development Goals (SDG's) for the next 15 years. At least six of the 17 goals are directly linked to water, and several others are indirectly affected by water issues. (The United Nations Department of Economic and Social Affairs, 2016) The sustainable development goals are not separate entities, instead, they are connected to each other in a high degree. This calls for integrated approaches in dealing with the issues the SDG's attempt to address. Water is central to achieving the goals. In addition, water-related issues have been identified by the World Economic Forum as some of the biggest risks the world is facing in the future (World Economic Forum, 2016).

Access to water and poverty have been linked for a long time in research, and it is understood that water occupies a central role in poverty alleviation. (Sullivan, 2002; Perez-Foguet & Garriga, 2011) Water and poverty are linked through water management, not water scarcity, which is primarily related to food security due to agriculture being the dominant water user globally. For poverty, water management issues relate to drinking water access, cooking and sanitation through policy failure, lack of infrastructure and low capacity. (Perez-Foguet & Garriga, 2011) These factors have led to development of indicator approaches in water resource research. To answer the need, Water Poverty Index (WPI) was developed as holistic tool to assess water resource in an integrated manner, combining resource availability, social access to water and the environmental water requirements. (Sullivan, 2002)

Water Poverty Index has been tested in many case studies on different scales across the world. However, it has only been applied in a single study (to the author's knowledge) in Mainland Southeast Asia, on Srepok River Basin in Cambodia (Ty, et al., 2010) and in an international comparison (Lawrence, et al., 2002). The Mekong Region is currently undergoing an accelerating dam-building phase, with 72 new large scale dams in the plans or under construction (International Rivers, 2015). Building of the dams is problematic due to its effects on food security, biodiversity and flooding (Ziv, et al., 2011).

The big challenges faced by the riparian countries of the Mekong in the wake of building the infrastructure provide the motivation for this study. WPI is a useful tool to assess a multidimensional issue such as poverty, water and the social and environmental change the new infrastructure causes. The primary goal of this study is to investigate the seasonal and spatial differences of water poverty prior to the accelerated dam building. It is meant as a tool aiding decision making in the work towards the new SDG's laid out in 2015. In addition, the author thinks of it as a first step in a comprehensive application of WPI in Southeast Asia. A secondary goal of the thesis is to investigate whether the water poverty index could (and should) be extended to include a temporal dimension – water is not only spatially highly varying, but also temporally. Lao People's Democratic Republic (Laos) is chosen as the area of study for two reasons. First, majority of all the new dam infrastructure is planned and constructed in Laos. Second, the author is familiar with the country through other research and having worked in Vientiane for a few months.

Three research questions guide the investigative process:

1. Are there distinct differences between areas in their water poverty?
2. Are there distinct spatio-temporal differences in water poverty?
3. What are the causes of water poverty in Laos? Do the causes differ across space and seasons?

The first question seeks to determine spatial distribution of water poverty and whether it is possible to identify areas that are relatively poor in the context of water. The second question seeks to inspect the spatio-temporal dimension of water poverty. Laos is located in the area with Monsoon rains and it may be expected that differences in water poverty can be find across dry and wet seasons. The third question tries to find causes of water poverty, and whether the cause varies in different areas or in different seasons.

To achieve the objectives laid out for this thesis, A Water Poverty Index is developed and calculated for Lao PDR for the dry and wet seasons separately. The indices are then subjected to a number of Exploratory Data Analysis and (Spatial) Data Mining methods to uncover information about the spatial and seasonal water poverty. Global and local exploratory statistical analysis as well as spatial clustering are used to answer the first and second research questions. The third research question is addressed with datamining using Geographically Weighted Principal Component Analysis and Geographically Weighted Regression in addition to spatial clustering.

The analyses in this thesis were done using R version 3.3.0 (R Core Team, 2016), a free and open source statistical programming language. In addition to the analyses, all illustrations (unless otherwise stated) were created in R using either the base package or "ggplot2", a package which is based on Leland Wilkinson's Grammar of Graphics (Wickham, 2009). Since the graphics in this thesis are somewhat complex and detailed, in order to provide a better view, interactive version of the thesis is published in the author's home page in http://markokallio.fi/waterpoverty/. In addition, the R code and the data are openly published in the home page as well as in GitHub repository mkkallio/waterpoverty.

This thesis is structured in the following manner: In the second chapter, background information on Laos is given to place the work in its geographical context. In addition, information on current knowledge of the water poverty in Laos as well as general description of Water Poverty Index (WPI), the selected method used in the analyses, is given. The third chapter introduces the theoretical framework on spatial phenomenon, exploratory data analysis and spatial data mining. Following that, the fourth chapter extends the theory into a methodology this thesis follows to build a WPI for Laos and its exploration. The fifth chapter, Results, begins by looking at the selected variables used in WPI calculation, followed by a detailed exploration of WPI across seasons and space. The chapter continues to report the data mining results via spatial clustering, Geographically Weighted Principal Component Analysis and Geographically Weighted Regression. The Sixth chapter discusses the limitations and problems encountered in the research process, and finally, the last chapter concludes the thesis with a summary of the findings and answers to the research questions.

2 Background

The background chapter provides basic introduction to Lao PDR to set a context to the geographic area of this study. Following that, an introduction to Water Poverty Index (WPI) is given to familiarize the reader to the main concept of this study. Finally, Water poverty is examined through previous studies in Lao PDR.

2.1 Introduction to Lao PDR

Lao People's Democratic Republic (Laos for the remainder of the paper) is a land locked country in Mainland Southeast Asia located between latitudes 13°-22.5° North and longitudes 100° and 108° East (WGS84. However, UTM Zone 48N (EPSG:32648) is the coordinate system used in the thesis). It is bordered by Cambodia to the south, Thailand to the west, Myanmar in the northwest, China to the north and Viet Nam to the east. The total land area of Laos is 236 800 km^2 with 80% of its land surface classified as mountains (see Figure 2.1). Cultivable land is considered to account for only 25% of the total land surface. The lowlands of Laos accommodate 56% of the total population of approximately 6.8 million, which is young; half of the population is under the age of 22 with life expectancy of 65.8 years. Laotians are also rural; current estimates place the share of rural population to 68-71%. (United Nations in Lao PDR, 2015)

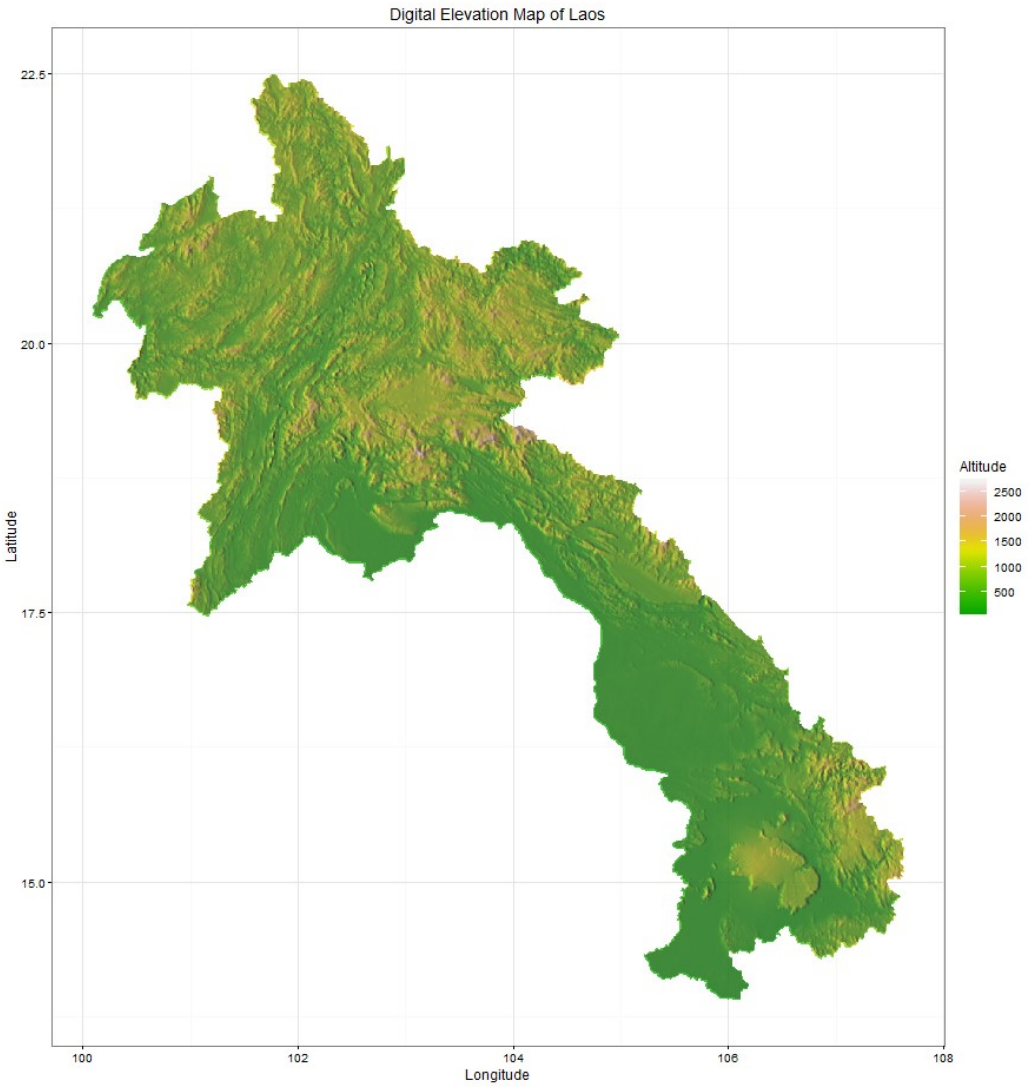


Figure 2.1. Hillshaded Digital Elevation Map of Laos.

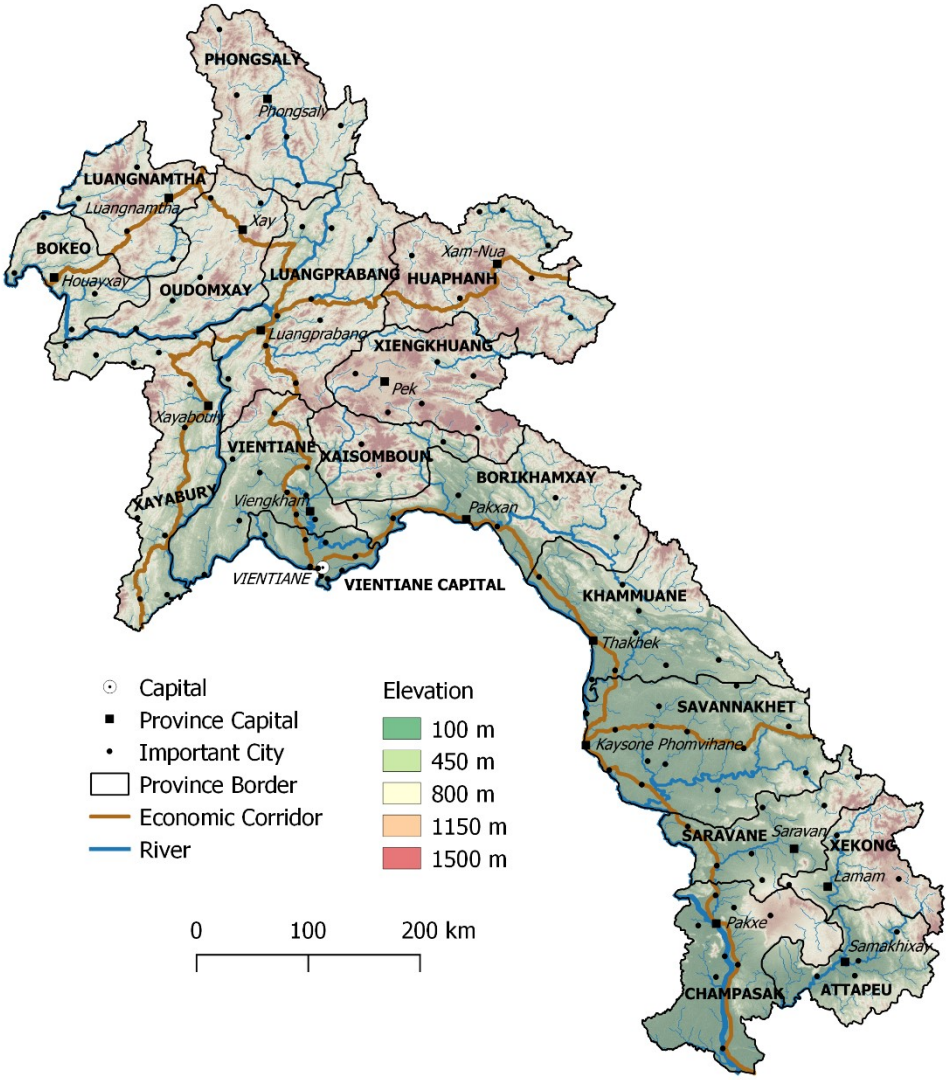


Figure 2.2. General map of Laos with important economic corridors (roads), major rivers, provinces and administrative capitals.

Currently, Laos is divided into 18 provinces after Xaysomboun Special Region was approved as a province in 2013 (however, this thesis uses the previous province division where Xaysomboun is a part of Vientiane Province because all used census data was collected prior to 2013). The provinces (shown in Figure 2.2) in turn are divided into 145 districts. The country is currently carries the status of a Least Developed Country, however the government aims to graduate from the status in the 2020s. Economic growth in the current decade has been rapid and steady at approximately 8%. A majority of the growth comes from the natural resources industry, namely mining sector, hydropower construction and forestry industry, and they contribute 18% of the entire GDP of Laos. (United Nations in Lao PDR, 2015) In fact, 72 new major hydropower dams are planned or already under construction, nine of which are located in the Mekong River mainstream (International Rivers, 2015).

Laos is located in the Mekong River Basin, one of the world's great rivers, nearly entirely. Mekong's source lies in the Tibet in the Himalayas and flows through China, Myanmar, Thailand, Laos, Cambodia and Vietnam into the South China Sea. Overall, it spans for

almost 4350km and measured by discharge, it is the 8th largest river in the world. River basins in Laos contribute the highest volumes of all riparian countries; 35% in the dry and up to 60% in the wet season of the entire annual discharge (Mekong River Commission, 2007). Due to its location, Mekong's flow is influenced by the Southwest Monsoon which results in very large difference between wet and dry season flow. In fact, on average the wet season contributes over 85% of the annual precipitation. In Lao PDR, annual precipitation varies from less than 1000mm near Louang Prabang to more than 3000mm in some mountainous areas. (Babel & Wahid, 2009; Mekong River Commission, 2011)

Human impacts on the water resources are relatively low. Water pollution is not alarmingly high except in the Delta area in Vietnam, albeit local hotspots of water pollution can be found. (Babel & Wahid, 2009; Mekong River Commission, 2007) In addition, only 0.9% of the discharge is withdrawn for utilization in Lao PDR. Agriculture is the main water user in the entire Mekong River Basin, with up to 99% of withdrawn water used in agriculture. Despite a high share of agriculture in water use, 90% of rice crops in Laos are rainfed. (Babel & Wahid, 2009)

2.2 Water Poverty Index

Water Poverty Index (WPI) has been developed to answer to the need of incorporating other factors to the prevailing convention (at least, at the time of development) of thinking water from a purely resource based point-of-view. WPI is intended as a holistic policy tool which combines physical (the resource-based view) with social sciences to better address the requirements for alleviating water-related poverty. It is known that, without sufficient water (in areas experiencing water poverty), any poverty alleviation measures are likely to be unsuccessful. There are several ways a person may be water poor. One may not have enough water for basic needs because it is not available, or the access to water may be limited because it is only available at a distance. Water poverty may also be due to income poverty; a person not being able to afford the price of safe water. (Sullivan, 2002; Lawrence, et al., 2002)

Sullivan (2002) lists several pressing needs for a holistic view on water. Increase in the living standards of populace is known to increase water consumption. WPI helps in identifying the regions and communities where water is needed, and to aid in the equitable distribution of the resource. Another important need for WPI is the link between poverty and water. Poverty in general is a topic which has been researched from many points of views, however, though many research papers touch water, there are not many attempts that link poverty to water explicitly. The key issues in constructing a meaningful WPI are (as in any other composite index) are the choice of components, sources of data, choice of formula and the choice of a reference period. The problem is quantifying a phenomenon that cannot be directly measured (who is water poor, who is not?). In addition, the choice of the scale of analysis is important, and all other choices should reflect the scale. Water environment is heterogenous by nature with water availability changing dramatically over short distances. Access to the same water source may also vary from community to community, or even within family groups. Such inherent variation in the domain of water poverty adds to the challenge in presenting it as an index. A country level WPI may tell nothing about the regional differences in water poverty, and a regional index may not be able to represent both rural and urban population. (Sullivan, 2002)

Sullivan's original paper (2002) describes several ways how WPI *could* be calculated. However, one of the earliest applications of WPI developed a methodology similar to Human Development Index (HDI), dividing it into five distinct components: Resource, Access, Capacity, Use and Environment. (Lawrence, et al., 2002) Each of the components are further broken into sub-components.

Resources (RES) attempts to measure the availability of water resource, taking into account both, the internal water resource as well as the water inflow from an external source. Access (ACC) measures not only the access to safe water for drinking, cooking and sanitation, but also for agricultural and other uses. Capcacity (CAP) involves education, health, income and the ability to influence the managing of the shared water resource. Use (USE) involves domestic, agricultural and other water use. Environment (ENV) includes the environmental factors which are important in relation to the capacity for the community's (or country's) ability to utilize the resource. (Lawrence, et al., 2002)

2.2.1 Water Poverty Research in Lao PDR

Water Poverty Index has not been applied, according to the author's knowledge, to Lao PDR except on a whole country basis in international comparisons. Lawrence et al (2002) found that Lao PDR had a WPI of 58.5, which places it in the middle range among the countries of the world. As a comparison, the least and most water poor country was found to be Finland with a score of 79.9, and Ethiopia with a score of 34.0. The study used variables presented in Table 2.1 to calculate the individual components. Individual component scores for Laos (from a maximum of 20) were 13.9 for Resources, 5.4 for Access, 12.0 for Capacity, 16.8 for Use and 10.4 for Environment. According to this study, biggest problems regarding water poverty in Lao PDR is in Access component and the best situation in Use component. (Lawrence, et al., 2002)

However, despite water poverty *per se* has not been widely studied, water-related issues have been widely researched – namely poverty and agricultural issues. International Fund for Agricultural Development (2014) places the current (2010) poverty rate at 27.6%, mentioning that Laos is one of the poorest and least developed countries in the region. As a reference, incidence of poverty in the main dataset of this study (Population Census 2005) is approximately 35%. Poverty (and especially water-related poverty) causes malnutrition with 44% of children under the age of 5 being chronically malnourished. Farming is mainly practised for subsistence with farmers having poor conditions for economic production of crops. The report by IFAD (2014) places majority of the poor population in the mountainous and rural areas. 70% of Population in rural Laos lack access to sanitation (Babel & Wahid, 2009) and are geographically and istitutionally isolated. They are isolated from markets, education and health services and administrative services. (International Fund for Agricultural Development, 2014)

Table 2.1. Variables used by Lawrence et al (2002) to calculate an international comparison of WPI.

WPI Com-ponent	Data Used
Resources	Internal Freshwater Flows External Inflows Population
Access	% of population with access to clean water % of population with access to sanitation % of population with access to irrigation adjusted by per capita water resources
Capacity	ppp per capita income Under-five mortality rates Education enrolment rates Gini coefficients of income distribution
Use	Domestic water use in litres per day Share of water use by industry and agriculture adjusted by the sector's share of GDP
Environment	Indices of: • Water quality • Water stress • Environmental regulation and management • Informational capacity • Biodiversity based on threatened species

3 Theory

The third chapter of the thesis introduces the reader to spatial data and why a lot of emphasis is put on location. First, spatial data is briefly introduced including a description of its special nature and issues related to that. Once the special nature of spatial data is covered, theory about Exploratory Data Analysis (EDA), which is the main approach of the investigative work conducted in this work, is discussed. In fact, spatial data analysis is descriptive and exploratory in their nature (O'Sullivan & Unwin, 2010). Finally, the principles of selected Data Mining (DM) methods and their spatial variants are introduced as an extension to the classical EDA.

3.1 Spatial is Special

The term spatial data includes all data that have a spatial component; the data has connection to a location on Earth. Spatial data can be divided among two main types; objects and fields. An object is a digital representation of an *entity*, a real world phenomenon that ”*is not subdivided into phenomena of the same kind*” (Zhang & Goodchild, 2003, p. 31). A feature is a defined entity and its object representation. Objects are represented by vectors, which are collections of location coordinates (either a point, a line or an area) together with *attributes*, specific qualities or quantities of the entity that is represented by the object. The boundaries of objects are *crisp*, meaning that the object contains information that is within the borders of the object, but tells nothing of the outside. However, in the real world, objects seldom are crisp with the exception of man-made objects (e.g. houses, roads or fences which be precisely defined, or abstract entities such as cadastres). This is a difference between a ”smooth” field and a crisp object. A field can be represented by a mathematical function of space and time (for example in the case of gravitational or magnetic fields). However, a mathematical representation may not always be possible or necessary. Fields can also be represented by irregular or regular points, rasters, area objects in the form of a triangulated irregular network (non-overlapping triangles) or contours as in topographic maps (e.g. elevation of land surface or depth of water). A raster is a data model of a field which includes equal size cells which are arranged in rows and columns. Each cell (a pixel; the smallest non-divisible unit of the raster) contains a single or multiple values for attributes as well as location coordinates. Rasters are often used due to the ease of processing them. An illustration of the main types of objects and a raster representation of a field are shown in Figure 3.1. (O'Sullivan & Unwin, 2010; Zhang & Goodchild, 2003)

In addition to defining spatial data in the form of coordinates, position of an entity can be expressed with its *relationship* with other entities. Measures if spatial relationships are distance, direction, proximity, adjacency and connectivity. Several distance metrics exist, but most commonly Euclidean distance is used: e.g. the length of a connecting line. Direction also refers to the connecting line. Proximity of an entity is defined by a circle or another shape that is drawn around the object. These three are *metric* relationships; they can be measured and quantified. Adjacency and connectivity on the other hand are *topological* relationships: They remain unchanged when the spatial reference is altered. Adjacency is defined e.g. by a common line between areas, shared line or face (volumes) or in the case of lines, common end points. Connectivity is similar to adjacency and proximity, however, the two connected entities do not need to touch each other directly (there

can be intermediate objects) nor does it need to be within a certain distance as with proximity. A natural example is a road or a river network: A river delta is connected to the all the streams that flow to the sea through the same outlet. Similarly, a road in Foggia, Italy is connected to a small residential street in Helsinki, Finland through a complex network that spans through Europe. (O'Sullivan & Unwin, 2010)

The data type and model appropriate to represent the real world entity is dependent on scale. (O'Sullivan & Unwin, 2010) An illustrative example a village for which there are several options. First, one can represent each individual building in a village using area features. When looking at the village in a smaller scale, the buildings shrink and a point becomes the most convenient representation. Similarly, one can represent the village as a polygon which delimits the areal extent of it. On a smaller scale, a point will the most efficient object type.

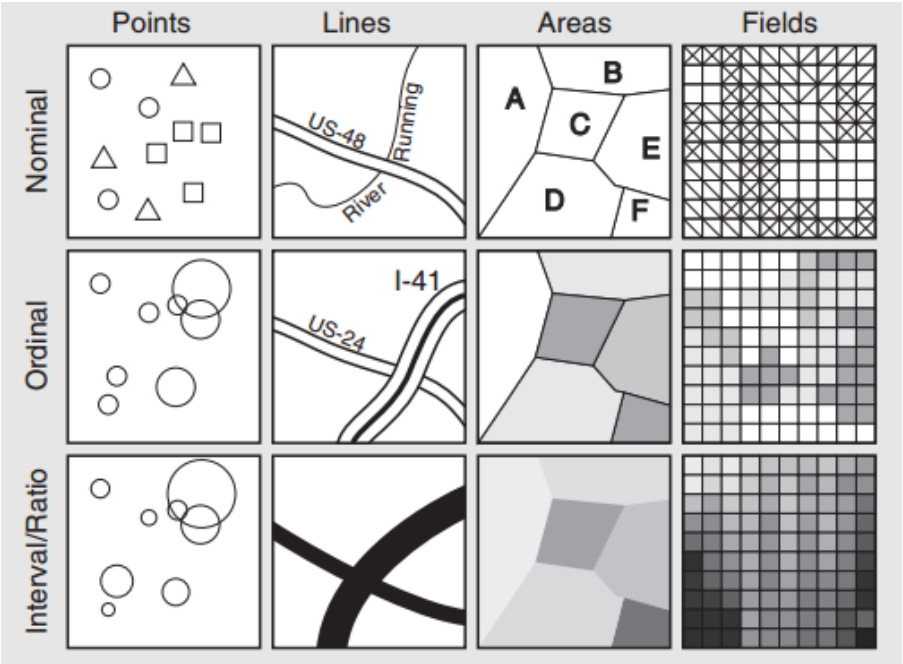


Figure 3.1. Illustration of different spatial data types. (O'Sullivan & Unwin, 2010)

3.1.1 Spatio-Temporality

Most spatially distributed phenomena are complex and multivariate processes, which vary in *time* in addition to *space*. (Zhang & Goodchild, 2003) Often these processes can be stored in data only as a set of discrete representations of infinite number of different states. In other words, many datasets only express a ”snapshot” in time. (Erwig, et al., 1999; Cressie & Wikle, 2011) Geometries change over time, in which case the term is either *moving object* or *moving regions* (fields). Some examples of moving objects are cars, or river boats. A moving field could be a precipitation pattern; the field is omnipresent, only the region of rainfall changes over time. In addition to the geometry, qualitative components (attributes) of the entities may also change in time. In the case of rainfall, an attribute could describe the changing intensity of the rain or its acidity. According to Isard (1970), time can be classified into four types: Linear (absolute time), cyclic (recurring time), ordinal time (relative order) and time as a distance, where spatial dimension is used to represent time. Linear change is constant and trending (long term change), cyclical

time is a time period which starts again when last period ends (e.g. day, week, year). Shifting change is an abrupt, sudden change which can be short or long term.

Many standard statistical analysis techniques and methods work poorly when applied on spatial (or spatio-temporal) data. The main reasons for the problems are Spatial Autocorrelation and Modifiable Area Unit Problem (and the related issues of scale and edge effects) and ecological fallacy. These are the causes of why *Spatial is Special.* (O'Sullivan & Unwin, 2010)

3.1.2 Spatial Autocorrelation

Spatial autocorrelation is a technical term is well described by Waldo Tobler's famous First Law of Geography:" Everything *is related to everything else, but near things are more related than distant things"* (Tobler, 1970) In other words, things that are near another are more likely to be similar in their properties than things that are far apart (Note: This applies to the temporal dimension as well). Take for instance a random building in a city centre: you are much more likely to find another building close by than by taking a random building in the rural countryside. Likewise, standing on a mountain you are likely closer to other mountains than to a tropical jungle. In fact, O'Sullivan and Unwin (2010) argue that geography as a science would not exist if spatial autocorrelation would not exist. The non-randomity described above is the root cause why standard statistical techniques perform poor on spatial data; most of the assume random sampling, which generally is not true in geography. Parameter estimates made from non-random samples will be biased toward the regions with largest numbers of sample points. (O'Sullivan & Unwin, 2010)

Despite the problems caused by spatial autocorrelation, techniques have been developed which can be used to describe it. Having a mathematical description helps to decide whether or not there truly is a spatial pattern, and how unusual it is. The most common measure of spatial autocorrelation is Moran's Index (Moran's I), which can be defined as a translation of non-spatial correlation to a spatial context. (O'Sullivan & Unwin, 2010) Moran's I is calculated with Equation 1

$$I = \left[\frac{n}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right] x \left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right] \quad (1)$$

where *i* and *j* are different zones, or areal units, *y* is the data value and *w* is a weight assigned for each zone or unit based on their spatial relationship. The value of the index is positive if most nearby data points are above or below the mean, and negative if they are on the opposite sides. Generally, Moran's I above 0.3 or below -0.3 can be considered as relatively strong autocorrelation. (O'Sullivan & Unwin, 2010) For an excellent overview on spatial autocorrelation, and other measures than Moran's I, the reader is forwarded to Getis (2010).

3.1.3 Modifiable Area Unit Problem

Modifiable Area Unit Problem (MAUP) stems from the property of spatial data to aggregate information in to larger spatial units. An example is a population census which is

often collected at a household level, but reported in units of villages, districts, provinces, states or other similar entities. The problem is that these units are often arbitrary considering the phenomenon investigated while the choice of unit analysis affects statistics derived from them. (O'Sullivan & Unwin, 2010; Openshaw, 1983) The statistics may change when the unit area is changed, as seen in Figure 3.2. According to O'Sullivan and Unwin (2010), it is possible to show that using the same underlying data, it is possible to produce correlations whose strength is anywhere between -1 and 1! In water resources research, Salmivaara et al (2015) showed that changing the unit of assessment had a large effect on water shortage assessment in Monsoon Asia and concluded that water-related spatial studies are highly sensitive to changes in areal unit of analysis.

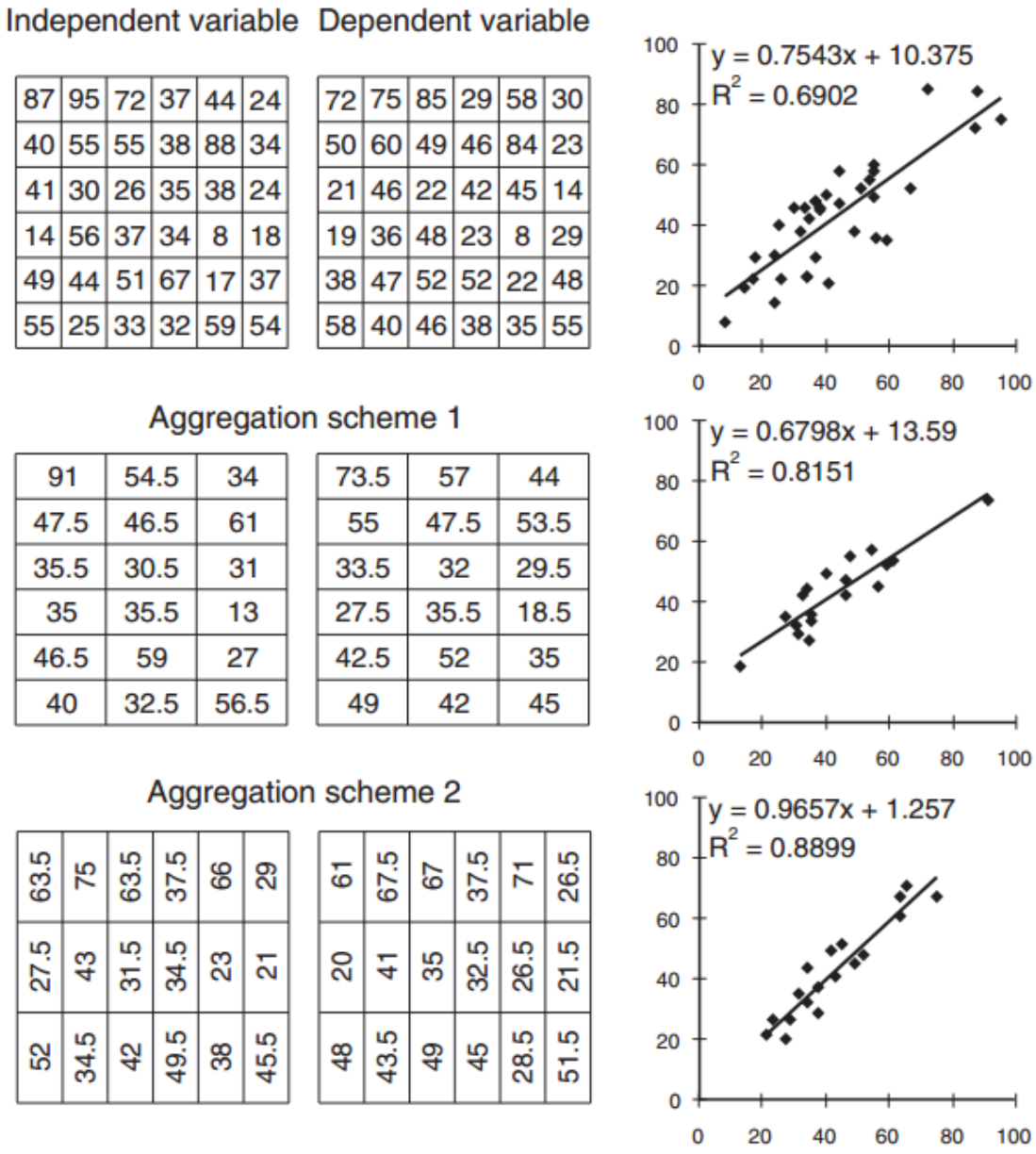


Figure 3.2. Illustration of the Modifiable Area Unit Problem and its effects on regression. (O'Sullivan & Unwin, 2010)

In practise, MAUP has been largely ignored due to the difficulty in selecting an appropriate unit analysis or due to a lack of understanding. Using aggregated data to address issues or to devise policies could lead to entirely different decisions whether alternative aggregation units were used. (O'Sullivan & Unwin, 2010) Openshaw (1983) states that the MAUP is an integral part of geography, and it should not be ignored, rather, it should be turned in to an exploratory tool and exploited. It has been suggested that zoning should be independent from the phenomenon under study, but Openshaw argues that, to truly investigate phenomena, zoning will need to be relevant. However, some techniques exist which can be used to address MAUP in certain situations, such as Geographically Weighted Summary Statistics developed by Brunsdon et al (2002). This method is described in detail in Section 3.2.4.

Spatial analysis is distinguished from traditional statistics also by the fact that *space is not uniform*. For instance, space in cities alternate between streets, parks, squares, industrial and commercial areas and residential suburbs. This type of non-uniformity must be considered in the spatial analysis. Another important problem associated with MAUP is edge effects which appear, as the name suggests, at the edges of a study area. Commonly, in the centre of a study area there are observations in every direction, while in the edges observations exist only in the direction of the centre. Often this does not reflect reality unless the study area is delimited bearing the edge effects in mind. (O'Sullivan & Unwin, 2010)

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a fundamental approach in statistics which includes all methods that are not formal statistical modelling or inference (Steltman, 2015). According to NIST/SEMATECH e-Handbook of Statistical Methods (2013), the aim of EDA is to

- maximize insight into a data set,
- uncover underlying structure,
- extract important variables,
- detect outliers and anomalies,
- test underlying assumptions,
- develop parsimonious models, and
- determine optimal factor settings.

In other words, EDA employs a variety of techniques which answer to a broad question of "what is going on here?" (Behrens, 1997) and can be described as *data-driven hypothesis* generation (Hand, et al., 2001). Roughly, the techniques fall under four categories over two axes – graphical and non-graphical and univariate and multivariate. (Steltman, 2015) In EDA the data is explored in a way that is not *confirmatory* as in traditional statistics. The emphasis on using statistical graphics (however, although the techniques are identical to those of statistical graphics, the approach is not) is due to human capabilities of visually identifying *patterns* in graphics. In particular, EDA process often consists of plotting raw data (e.g. using histograms, scatterplots, box plots…), plotting simple statistics (means, standard deviations) and to position such plots to maximize our pattern-spotting abilities. (NIST/SEMATECH, 2013; Hand, et al., 2001)

EDA differs from classical statistics in its process. Traditionally, a model is imposed on data, the model's performance is analysed and conclusions are drawn from there. In EDA the position of analysis and model is reversed; the data is first explored and a model is developed as *suggested* by the data. In addition, EDA process is subjective and depend on interpretation by the analyst, and thus they can differ from person to person. Traditional statistics is in a sense, more objective and formal. (NIST/SEMATECH, 2013) In addition, EDA can be characterized by the use of robust measures, re-expression of data and usage subset for further analysis. Moreover, EDA is flexible and the analyst is encouraged to scepticism and ecumenism when choosing which methods to apply. (Behrens, 1997) The explorative and confirmatory data analysis (CDA) processes, however, despite different approaches, are complementary and in practise should be used in conjunction. EDA is first employed to investigate variables and to develop hypotheses, looking at the data in every possible direction. The result of the process are models which are put to test using confirmatory techniques. EDA and CDA converge in certain methods, which are seemingly confirmatory, but are exploratory in their goal. These methods attempt to determine the best set of variables for a model instead of simply trying to confirm a predefined set or a model. One such method is stepwise regression, in which variables are assigned to a model one-by-one according to some criterion (e.g. cross-validation or Akaike Information Criterion). (Behrens, 1997)

Additionally, Behrens (1997) concludes that while documenting and publishing EDA process reduces the resources assigned for the advanced stages of modelling (and model building), the details it provides improve understanding the phenomenon under investigation in a way that simple summary statistics and tests cannot. EDA will also help prevent Type III errors: "Precisely solving the wrong problem, when you should have been working on the right problem". More recently, EDA has been extended by a newer concept called Data Mining (DM), which is another exploratory (and predictive) approach concerning extremely large databases. DM is discussed in Section 3.3.

3.2.1 Univariate Exploration

Some of the most important univariate non-graphical methods include calculating a central tendency (mean/median/mode), spread (standard deviation, variance, interquartile range), skewness and kurtosis of a sample (these are not explained here, however a good description for all of them can be found in any introductory statistical textbook, e.g. Steltman (2015)). Many of these distributional characteristics can be qualitatively seen in a histogram, which is one of the most important graphical univariate methods. Stem and leaf plot is less known variant of a histogram, in which bins are replaced by a whole number in the stem, and a sequence of decimals of all observations that fall in to the bin (see Figure 3.3; each zero behind the bar is an observation). (Steltman, 2015) Another very useful plot is the box plot (used extensively in this study). A box plot is useful for visualizing central tendency, symmetry and skewness of the distribution as well as identifying outliers. Figure 3.4 presents an overview of the components of a Box plot. Commonly the whiskers extend 1.5 times the interquartile range (IQR, the difference between first and third quartile), and any observations beyond is considered an outlier and they are plotted individually. This study uses the 1.5 IQR definition, however, other alternatives are sometimes used (such as ones based on standard deviations). It should be noted here that, in an ideal normally distributed sample, it can be expected that 0.7% of the sample would appear as outliers, meaning that interpretation is required. In general, box plots rely on robust statistics which makes them a great EDA tool. (Steltman, 2015)

```
The decimal place is at the "|".
1|000000
2|00
3|000000000
4|000000
5|00000000000
6|000
7|0000
8|0
9|00
```

Figure 3.3. An example of a stem and leaf plot. (Steltman, 2015)

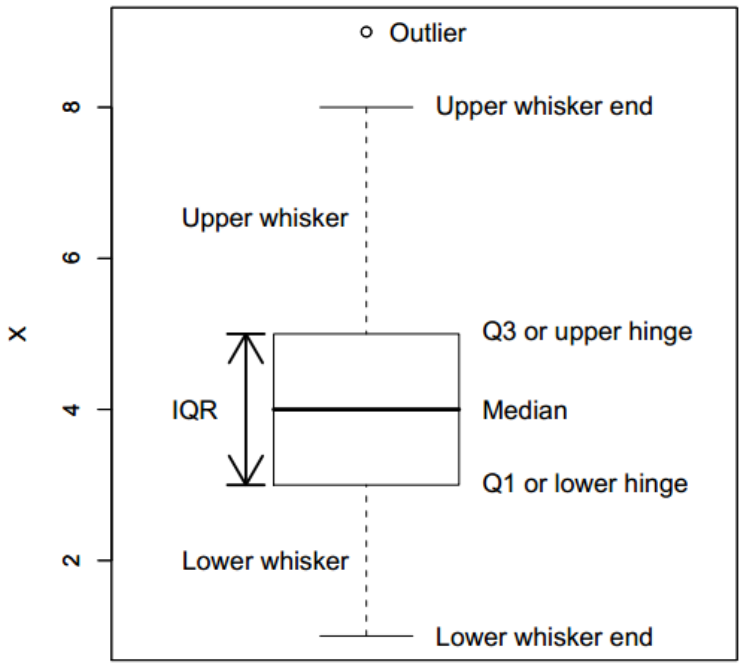


Figure 3.4. An annotated Box plot. IQR stands for Interquartile range, Q1 and Q3 stand for the first and the third quartile. (Steltman, 2015)

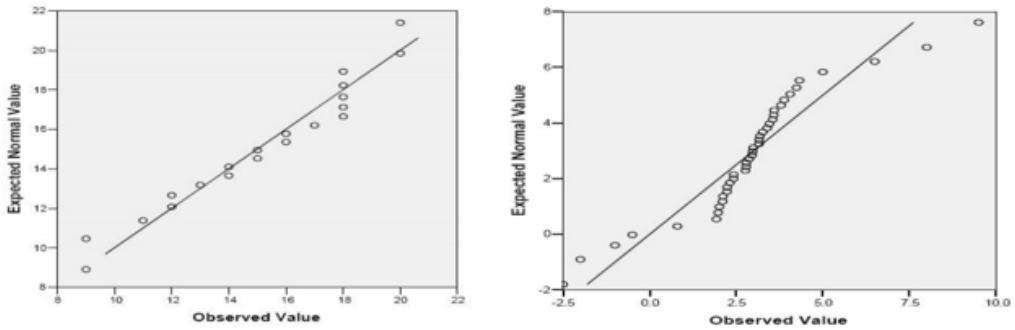


Figure 3.5. QQ-plots of normally (left) and non-normally (right) distributed samples. (Steltman, 2015)

A third univariate graphical method covered here is a quantile-normal plot (or a quantile-quantile plot, QQ plot). Many statistical tests assume that variables used to explain phenomena are approximately normally distributed. QQ plot is a way to assess the normality of a distribution. The plot is essentially a scatterplot where the values of the sample are on the x-axis and the expected normal values are plotted on y-axis. If the points fall approximately to a diagonal line, the sample is approximately normal. An example of normally and non-normally distributed plots are shown in Figure 3.5.

3.2.2 Multivariate Exploration

Multivariate techniques explore the relationship between two or more variables. There are far more techniques for multivariate data than there are for univariate samples. Only some of the methods are covered in this section, however, the reader is directed to read the NIST/SEMATECH e-Handbook of Statistical Methods (2013) for an extensive overview of different techniques. The most used non-graphical method in the thesis is correlation. Correlation is based on covariance, which is a measure of how much two variables vary together. Positive covariance means that when a measurement is above the mean, the one it is compared to probably is above the mean too. Negative covariance occurs when one is above and the other is below the mean. Zero covariance occurs when the variables vary independently of each other. Correlation is used because it is easier to interpret than covariance; a correlation of -1 means the variables are perfectly negatively correlated while +1 signifies a perfect positive correlation. A perfect correlation places observations on a straight line in a scatterplot. (Steltman, 2015)

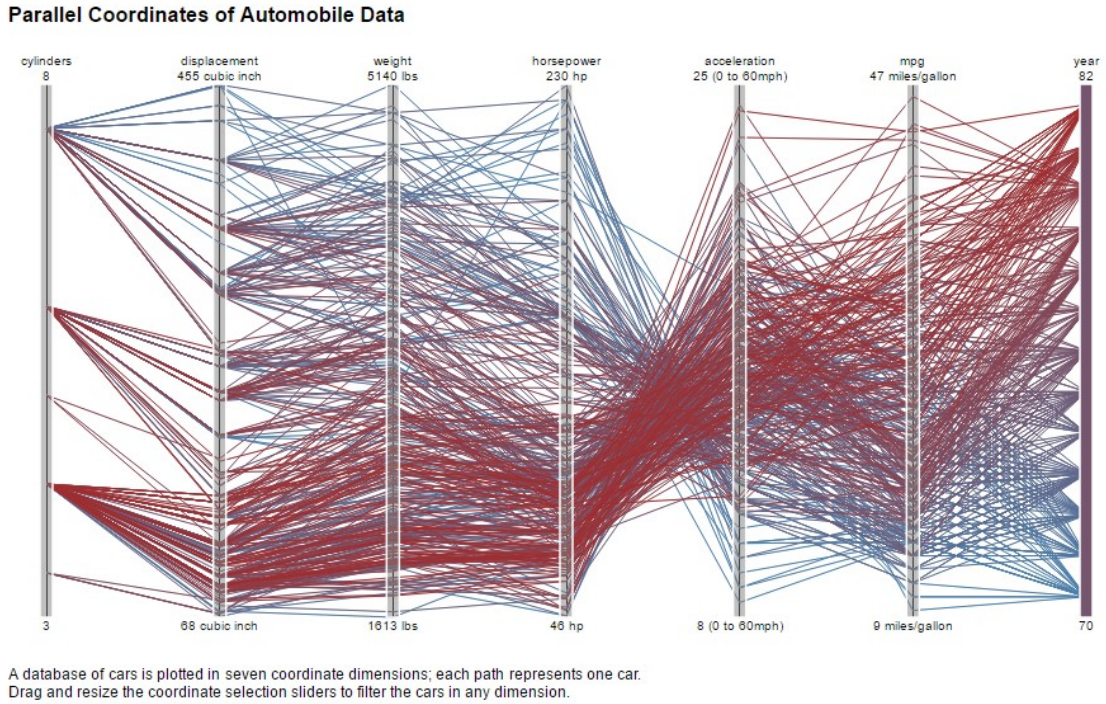


Figure 3.6. Parallel Coordinate Plot of automotive data. Source: GGobi; http://homes.cs.washington.edu/~jheer//files/zoo/ex/stats/parallel.html.

Box plots can be used in a multivariate exploration by placing them side-by-side in a plot. According to Steltman (2015) , they are the best graphical EDA technique to examine

relationship between categorical and quantitative variables. In addition, they perform well in visualizing the distribution of the quantitative variable at each level of a categorical value. Scatterplots can be used to visualize two or more variables at the same time. One variable is plotted on the x-axis, another on y-axis while size, colour and marker type can be used to map additional variables in the same view. (Steltman, 2015) Scatterplots can be placed in a matrix (creating a scatterplot matrix) providing a tool to visually inspect correlations between multiple variables. The scatterplot matrix makes use of the principle of *small multiples*, an approach where plots are shown side by side allowing for easy and quick comparison between different categories or variables. Small multiples can be constructed for nearly all types of visualizations; scatterplots, bar charts, pie charts, maps et cetera. (Heer, et al., 2010)

The final multivariate visualization method covered here is the Parallel Coordinate Plot (PCP, example shown in Figure 3.6). In PCP variables are placed side by side, and observations are represented by lines which connect the variables at their data points. An example of PCP is shown in Figure 3.6. PCP should be used in an interactive environment where reordering dimensions and filtering can help in pattern recognition. PCP's are excellent for compactly displaying many variables simultaneously. (Heer, et al., 2010)

3.2.3 Exploratory Spatial Data Analysis

Exploratory Spatial Data Analysis (ESDA) is based on the two central principals of EDA; the importance of data and the importance of analytical graphics (Bivand, 2010). ESDA is largely based on techniques that explicitly take spatial autocorrelation in to account, such as visualization of spatial distributions and associations. In addition to geovisualization techniques, ESDA employs robust statistical methods to detect spatial properties in the data. (Haining, et al., 1998; Anselin, 1998) Modern ESDA is characterized by visualization using dynamically linked displays which often include cartographic visualization, scatterplots, histograms, box plots or variograms among others. (Bivand, 2010) Anselin's (1998) division of ESDA techniques is given in Table 3.1. The tools are closest to the "spirit" of cartographic visualization, however the starting point differs. In ESDA, a standard statistical graphic is the centre point rather than the map. (Anselin, 1998) In addition to the methods in Table 3.1, Bivand (2010) mentions using spatial regression models as a source to explore spatial non-stationarity. Regression models can be extended to a spatial version in several ways, however, these are covered in the following section under Spatial Data Mining. For a thorough overview in different options for ESDA, the reader is directed to Bivand (2010).

Geostatistics is an additional ESDA method explicitly mentioned in Bivand (2010). In general, geostatistics includes a variety of interpolation methods, including inverse distance weighting and kriging. Geographically Weighted Summary Statistics (GWSS) is an exploratory method for deriving localized summary statistics (aka descriptive statistics), which is used in this study.

Table 3.1. Main ESDA techniques according to Anselin (1998).

Goal	Geostatistical Perspective	Lattice Perspective
Visualizing spatial distribution	Spatial cumulative distribution function	Box map Regional histograms Spatial exploratory analysis of variance
Visualizing spatial association	Spatially lagged scatterplot Vario-gram cloud plot Variogram boxplot	Spatial lag charts Moran scatterplot and map
Local spatial association	Outliers in variogram Outliers in variogram cloud plot	LISA maps Outliers in Moran scatterplot
Multivariate spatial association	Multivariate variogram cloud plot	Multivariate Moran scatterplot

3.2.4 Geographically Weighted Summary Statistics

GWSS allows for calculation of wide variety of statistics at geographical location revealing patterns that are not possible to be seen with global statistics. (Brunsdon et al 2002) The approach of GWSS is applying a weight on each observation based on their proximity to a point (*u, v*). It makes use of concepts in interpolation methods, such as moving window average or focal median function from Map Algebra. However, unlike the interpolation methods, GWSS has a more relaxed requirement for specifying a certain bandwidth (BW, the distance in which observations are taken in to account) prior to the analysis. In addition, probability densities can be utilized in GWSS. (Brunsdon et al 2002) Calculation of a localized statistic requires weighting of observations. One possible way to do this is inverse distance weighting, as shown in Equation 2

$$w_i = \exp\left(-\frac{d_i^2}{h^2}\right) \tag{2}$$

Where d_i is the Euclidean distance between observation *i* and (*u, v*), *h* is bandwidth and w_i is the given weight for observation *i*. Having specified a way to derive the weights, a locally weighted mean can be calculated with Equation 3

$$\bar{x}(u,v) = \frac{\sum x_i w_i}{\sum w_i} \tag{3}$$

where w_i is the weight of the *i*th observation and *x* is the value of the variable in question. (Brunsdon et al 2002) The equation above is a simple interpolation formula, however GWSS can be extended beyond the mean value. A local standard deviation (based on the localized mean) can be calculated with Equation 4

$$s_x(u,v) = \sqrt{\sum (x_i - \bar{x}(u,v))^2 w_i} \tag{4}$$

where s_x is the standard deviation of the variable *x*. In a similar manner, localized skewness can be calculated based on local mean and standard deviation. Table 3.2 shows a

collection of typical descriptive statistics used in univariate probability density functions. All of them can be localized. (Brunsdon, et al., 2002)

Table 3.2. Typical descriptive statistics for the univariate probability density function f(x). (Brunsdon et al, 2002)

Statistic name	Definition		Notation
	Continuous	Discrete	
Mean	$\int xf(x)\mathrm{d}x$	$\sum x\Pr(x)$	$\mathrm{E}(x)$
Standard deviation	$\int (x-\mathrm{E}(x))^2 f(x)\mathrm{d}x$	$\sum (x-\mathrm{E}(x))^2\Pr(x)$	$\mathrm{SD}(x)$
Skewness	$\frac{\int (x-\mathrm{E}(x))^3 f(x)\mathrm{d}x}{\mathrm{SD}(x)^{1.5}}$	$\frac{\sum (x-\mathrm{E}(x))^3\Pr(x)}{\mathrm{SD}(x)^{1.5}}$	$\mathrm{Sk}(x)$
p-Quantile	Solution for q of $\int_{-\infty}^{q} f(x)\mathrm{d}x = p$	Minimum solution for q of $\Pr(x<q)=p$	$\mathrm{Q}_p(x)$
Median	$\mathrm{Q}_{0.5}(x)$		$\mathrm{Med}(x)$
Inter-quartile range	$\mathrm{Q}_{0.75}(x)-\mathrm{Q}_{0.25}(x)$		$\mathrm{IQR}(x)$
Quantile imbalance	$\frac{2\mathrm{Med}(x)-(\mathrm{Q}_{0.75}(x)+\mathrm{Q}_{0.25})}{\mathrm{IQR}(x)}$		$\mathrm{QI}(x)$

Weighting scheme is at the core of GWSS as well as the other methods in the GW family. Examples of geographical weighting schemes are

- implicit weighting (global models),
- excluding observations beyond a certain distance,
- Gaussian weighting (kernel density function), and
- bi-square function (combination of the previous two).

Excluding observations beyond certain distance causes discontinuity, where including or excluding a single observation can have a big effect on the parameter estimate. Gaussian weighting applies a weight as a function of distance from the point of interest (kernel density function). Bi-square on the other hand is a combination of the two, excluding observation further than a certain distance, but applying a distance weighting to the remaining observations. (Brunsdon et al 1996, Brunsdon et al 1998) The selection of weighting function is critical as when the seeking distance grows, the parameter estimates get closer to a global model. Calibration is therefore important. A good solution to find an optimal bandwidth is least squares cross-validation Equation 5

$$\sum [y_i - y_{\neq i}^{*}(\beta)]^2 \tag{5}$$

where $y_{\neq i}^{*}(\beta)$ is the fitted value of y_i with the observation for point i omitted from the calibration procedure. This way of calibration counters a wrap-around effect of overfitting the model. (Brunsdon et al 1996) It should be noted that in some cases subjective choice of the distance can be more descriptive to the reality than a computed value. This is true especially when there is strong evidence in favour of some specific distance. (Brunsdon et al 1998) Similarly to the parameter estimations vary location to location, the optimal weighting function or distance may also vary by location. Edge effects is an obvious source of this kind of issue, but also the distance that the phenomena affect may vary where edge effects are a non-issue. (Brunsdon, Fotheringham, & Charlton, 1996) One

solution to this is to use an adaptive bandwidth which takes a specified n number of observations into account. Using adaptive weighting, areas of dense observations reduces the size of bandwidth as the number of local observations is high, and in areas of sparse observation population the bandwidth grows large.

3.3 Spatial Data Mining

Data mining (DM) is an exploratory approach by its nature and again, as in EDA, there is no a priori hypothesis in DM. Several definitions of DM exist. Luan (2002) defines it as *"the process of discovering hidden messages, patterns and knowledge within large amounts of data and of making predictions for outcomes or behaviours."* Larose (2005) on the other hand defines it as *"the process of automatically extracting useful information and relationships from immense quantities of data."* Alternatively, Hand et al (2001) provides the following definition: *"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."* The definitions have in common that DM is a *process*, like EDA, rather than a set of tools. In addition, DM deals with very large databases. Despite the word "automatically" in Larose's definition, human input is as essential to DM as it is to EDA (Larose, 2005). These facts – large databases and utilization of automatization and algorithms is what sets DM apart from classical EDA. It is also worthwhile to mention that DM is an interdisciplinary approach which includes statistics, database technology, machine learning, pattern recognition, artificial intelligence and information visualization. (Hand, et al., 2001; Shekhar & Chawla, 2003). The lack of predefined hypothesis allows DM to facilitate learning new and novel information and knowledge. In fact, DM is a non-deterministic and iterative process which aims to develop knowledge to be used in a decision making process. The end result of a DM process is a hypothesis, which can then be tested with statistical methods. (Miller & Han, 2009; Guo & Mennis, 2009) Hand et al (2001) define the outcomes of a DM process as *models* or *patterns*, which can be for example linear equations, rules, clusters, graphs, tree structures or recurrent patterns in time series.

Data mining and knowledge discovery can be divided into steps; data selection, data preprocessing, data enrichment, data reduction and projection, data mining and interpretation and reporting (note the absence of data collection – DM often deals with data that has already been collected (Hand, et al., 2001)). Data selection refers to determining the variables used for the data mining process. Data pre-processing is cleaning, noise reduction eliminating duplicate records and determining how to handle missing values. Data enrichment means combining datasets. Data reduction and projection involves dimensionality reduction to further reduce the number of variables in the data, or projecting the attribute space to a more efficient representation. Data mining is the application of different methods to the data to uncover new and interesting patterns (the selection of method depend on the type of knowledge to be mined – see Table 3.3), and finally interpretation and reporting involves visualization and evaluation of the data mining process. (Miller & Han, 2009)

Table 3.3. Data Mining methods and techniques according to Miller and Han (2009).

Knowledge Type	Description	Techniques
Segmentation or clustering	Determining a finite set of implicit groups that describe the data	Cluster analysis
Classification	Predict the class label that a set of data belongs to based on some training datasets	Bayesian classification Decision tree induction Artificial Neural Networks Support Vector Machines
Association	Finding relationships among itemsets or association/correlation rules, or predict the value of some attribute based on the value of other attributes	Association Rules Bayesian Networks
Deviations	Finding data items that exhibit unusual deviations from expectations	Clustering and other data-mining methods Outlier detection Evolution analysis
Trends and regression analysis	Lines and curves summarizing the database, often over time	Regression Sequential pattern extraction
Generalizations	Compact descriptions of the data	Summary rules Attribute-oriented induction

Extending DM into Spatial Data Mining (SDM) is not a trivial task, and, depending on the technique, may include several possible methods. The challenge is due to the special nature of the data – spatial autocorrelation. According to Shekhar and Chawla (2003) The goal of spatial data mining is to

1. identify spatial patterns,
2. identify spatial objects that are potential generators of patterns,
3. identify information relevant for explaining a spatial pattern, and
4. presenting the information in a way that is intuitive and supports further analysis.

The extensions of selected data mining methods in to spatial data mining methods are presented below.

3.3.1 Spatial Clustering

Clustering is a process where features are classified into groups of mutually *similar* features which are *dissimilar* with features in other groups in a way that minimizes intra-cluster and maximizes inter-cluster distances. Clustering is a method that has been used for a long time and for numerous different applications. According to Han et al. (2009), clustering methods can be divided into four groups: a) partitioning, b) hierarchical, c) density-based and d) grid based.

Partitioning methods divide the dataset of n observations into k partitions, where each partition represents a cluster. There are several algorithms to achieve the partitioning. One of the most used is k-means clustering, which is based on centroids of a cluster. In the beginning, all data points are assigned to a random cluster. The second step is to assign observations one by one to the cluster whose mean value is the closest to the value of the observation. The next observation is then assigned based on the new cluster means, and so on. The algorithm runs until a certain criterion is reached. This method requires the user to define the number of clusters in advance which may be a big disadvantage. In

addition, the method is sensitive to outliers and noise. The problem of outlier sensitivity in *k*-means clustering can be addressed by using *k*-medoid method. In *k*-medoid clustering, an observation is used as the object of reference instead of the mean of the cluster. The chosen *medoid* is the observation which is the closest to the cluster mean value. The algorithm works by minimizing the absolute error when observations are moved to other groups one by one. (Han, et al., 2009)

Hierarchical clustering produces a tree of clusters which can be formed by either agglomerative (bottom-up) of divisive (top-down) methods. AGNES (Aglomerative Nesting) is an algorithm in which at the beginning each object is in its own cluster. The process then merges the clusters into larger ones until all of the objects are in one cluster. DIANA (Divisive Analysis) does the opposite; the objects start in the same cluster and they are then divided in to subclusters until they are alone in their own clusters or another criterion is achieved. (Han, et al., 2009) Density based clustering produces clusters with arbitrary shape (*k*-means for example tends to produce spherical clusters) which produces clusters in regions of dense observations separated by regions of low density. Due to the nature of the data in this study, density based clustering is not used, however, for the interested, Han et al (2009) provides a useful overview of density-based clustering algorithms.

In a multidimensional point data set (such as the villages in this study), spatial clustering can be can be achieved in a number of ways called regionalization methods. The first option is to perform trial-and-error search. An example is the Automatic Zoning Procedure, which starts with random regions (clusters) and it iteratively improves it by switching boundary objects between regions. Second, one can perform a-spatial clustering on a dataset without information on the location, followed by spatial processing. The Third option is to do clustering with a spatially weighted dissimilarity measure. In practice, this is done by incorporating spatial information (e.g. coordinates) as variables in the clustering procedure. Finally, the last option is called contiguity constrained clustering and partitioning. (Guo, 2009)

3.3.2 Spatial Regression

The need for special techniques in regression rises again with the assumption of independency and normality in traditional statistics. Several methods of extending regression analysis to spatial domain exist. Two methods are covered here; Spatial Autoregressive Models (SAR) and Geographically Weighted Regression (GWR). SAR is a generalization of the linear regression model which accounts for spatial autocorrelation. SAR performs better than a-spatial regression models, however it is computationally heavy. SAR is, in fact, an extension of the linear regression model with an added spatial autocorrelation term ρ**Wy** in order to model the strength of spatial dependencies. Rho stands for a spatial autoregression parameter, **W** is a neighbourhood matrix representing spatial relationships in the data and **y** is the dependent variable. SAR is shown in Equation 6

$$\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \boldsymbol{x}\beta + \varepsilon \qquad (6)$$

where β is the regression coefficient and ε represents random error. (Kazar & Celik, 2012)

Another alternative is GWR, which is similar to GWSS in that observations are given a weight depending on their geographical proximity. It has been shown that parameter estimates for regression models can vary dramatically over geographical space if only a subset of the data is used. This may result in interesting relationships between variables being obscured if a global model is used. (Brunsdon et al 1996) The generalized form of the GWR shown in equation 5 can be calculated for any location in the geographical space, and not only in observation points. The Equation 7 for GWR is

$$y_i = a_{i0} + \sum a_{ik} x_{ik} + \varepsilon_i \quad (7)$$

Where y_i is the predicted value of the dependent variable at point i, a_{ik} is the value of kth parameter at location i, x_{ik} is the independent variable and ε_i is random error. (Brunsdon et al 1996, Brunsdon et al 1998)

Statistical testing of GWR (or any other local regression method) may be more difficult than testing a global model. If we assume that that the variables vary according to some distribution and that there is an error, it is natural that the parameter estimates will also have some error and distribution. Many different statistical tests have been developed for the GWR (Mei et al 2016), however, testing may be challenging using standard procedures due to spatial autocorrelation, which may produce misleading results. (Wei and Qi, 2012) Standard errors, t-values, goodness-of-fit measures etc. can be localized and assessing can done based on them (Demsar et al 2008), however, it has been found that goodness-of-fit measures alone are not adequate. This is because they assume that the standard errors are independent which usually is not the case with spatial processes (Laffan and Bickford, 2005). Monte Carlo simulation is a method often used in testing the results of GWR.

3.3.3 Geographically Weighted Principal Component Analysis

Geographically Weighted Principal Component Analysis (GWPCA) is not generally explicitly mentioned in the lists of spatial data mining methods, however it is included in here due to the fact that SDM processes often include dimensionality reduction and data processing as one of the steps.

Many datasets (and especially spatial datasets) are highly multidimensional data, which poses challenges for interpretation and visualization. Therefore, it is often desirable to reduce the number of dimensions. Principal Component Analysis (PCA) is a method to reduce dimensionality while capturing the maximum information present in the dataset. PCA captures the maximum variation of data and re-projects the original information in to an orthogonal space of n-dimensions. The first principal component (PC) represents the largest variation in data. The second PC then accounts for the largest amount of variation that is not captured in the first PC. The third captures the variation not accounted for by the first or the second PC, and so on. (Demsar, et al., 2013)

Wheeler and Tiefelsdorf (2005) and Mei et al (2016) have shown an important drawback of GWR that is local multicollinearity. Multicollinearity in the model variables may occur (even if they are not collinear in the global model) and this may have an adverse effect on the coefficient estimation. The local coefficients may become entirely interdependent. One possible remedy for this condition include using PCA. Like the previously introduced summary statistics and regression, PCA can be made spatially conscious. This can

be achieved in three ways; with a geographically weighted variant (GWPCA), adapted PCA taking spatial autocorrelation into account or by combining these two. (Harris, et al., 2011) The method used in this thesis is GWPCA, which is a natural extension of Locally Weighted PCA with the exception that the locally weighted variant gets its weights from *attribute* space instead of *geographical* space. (Harris, et al., 2011)

GWPCA is achieved by computing geographically weighted means, variances and covariances for each data observation (See section 3.2.4). GW covariance between variables y_1 and y_2 for location i is given by Equation 8

$$cov(y_{1i}, y_{2i}) = \sum_{j=1}^{n} w_{ij}(y_{1j} - \bar{y}_{1i})(y_{2j} - \bar{y}_{2i}) \quad (8)$$

Geographically weighted correlation coefficient can then be computed from the GW covariance and GW variances. GWPCA can also be calculated from a correlation matrix, which is required if the variable values are not in the same units. (Lloyd, 2010) Applications of GWPCA include addressing problems in GWR models as a means for local dimensionality reduction or local orthogonalization prior to applying GWR (Demsar et al 2013, Charlton et al 2010).

4 Materials and Methods

The fourth chapter starts with a description of the datasets and how they were used in the work. Following, a detailed description of the choice of variables and methods to calculate the Water Poverty Index is given. Third, the employed EDA and SDM methodology is explained along with the rationale of why the selected methods were used. Fourth, the tools to implement the selected methods are given.

4.1 Data and Data Sources

The data used in this thesis is mainly derived from household studies conducted by the Lao Statistics Bureau. The two main sources are Population and Housing Census 2005 (Lao Statistics Bureau, 2005) and Agricultural Census 2010/2011 (Lao Statistics Bureau, 2011). Both datasets are household level surveys conducted in face-to-face interviews of household heads. Some indicators are found from both datasets, and in these cases, the newer one was used. The datasets are available after registeration in the Lao DECIDE web service, http://www.decide.la/. In addition to the census data, a number of other datasets used are listed below:

- SEDAC Last of the Wild v2 is used in WPI Environment component to account for the human environment. Global Human Footprint (v2, 1995-2004) is an estimate of the anthropogenic influence created from nine global datasets: human population pressure (population density), human land use and infrastructure (built-up areas, nighttime lights, land use/land cover), and human access (coastlines, roads, railroads, navigable rivers). (Wildlife Conservation Society - WCS; Center for International Earth Science Information Network - CIESIN - Columbia University, 2005)
- Aqueduct Global Maps 2.1 (Gassert, et al., 2014) was used in the WPI Environment component to represent the state of the water resource. Specifically, the categories of threatened amphibians from the dataset was employd for this purpose.
- For modelling of water resources in Laos, baseline temperature and precipitation from the study of Lauri et al (2014) was used. The data is collected by Mekong River Commission and the national weather services of the MRC member states.
- Water consumption data was used for data mining purposes. The data used was Total water consumption from the Global Water System Project Digital Water Atlas, which is based on the Water GAP model version 2.1d. The spatial resolution of the dataset is 0.5º. (GWSP Digital Water Atlas, 2008c) In addition, Irrigation water consumption (GWSP Digital Water Atlas, 2008b) and Domestic water consumption (GWSP Digital Water Atlas, 2008a) was used for the same purpose.
- Harmonized World Soil Database (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) was used for data mining purposes. Specifically, elevation and slope classes were utilized from the dataset.

The use of the above datasets is explained in detail in the following sections under Developing the Water Poverty Index and Analysis Methodology.

4.2 Developing the Water Poverty Index

The datasets described in section 3.1 were used to calculate WPI separately for dry and wet seasons. The variables used are the same for both indices to allow easy comparison, however, their application differs slightly for what is relevant for the respective season. The variables used to calculate WPI are summarized in Table 4.1. It is good to note here what these variable selections represent specifically in this study. Resources represents water availability in a relatively straightforward manner. Access on the other hand describes the infrastructure present in the villages to make use of the available water. Capacity represents the ability of the villages to manage their water in a local context – it does not include institutional capacity. Use on the other hand describes the extent to which the water resource is being used and the dependency of the population on its availability. Lastly, Environment measures a mixture of variables from the state of water and soil and disaster occurrence to land use.

Table 4.1. Summary of the variables and scoring used for calculating the components.

		Scoring		
Component	Variable	Minimum (0)	Maximum (100)	Data Source
RES	Surface water availability	<500m3/cap/year	>1700m3/cap/year	Modelled (Vmod)
	Average daily precipitation	0 mm	Max wet season precipitation	Mekong River Commission / Lauri et al (2014)
	Annual longest consecutive drought days	Longest average dry period in the dry season	Shortest average dry period in the dry season	Mekong River Commission / Lauri et al (2014)
ACC	Irrigation type	No irrigation facilities	Maximum number of different irrigation methods	Agricultural Census 2010/2011
	Drinking water source(s)	Minimum score	Maximum score	Agricultural Census 2010/2011
	Toilet type	No toilet	Modern	Population Census 2005
CAP	Travel time to province and district capitals	>600min	Travel time 0min	Population Census 2005
	Village road access	No	Yes	Agricultural Census 2010/2011
	Literacy rate	0%	100%	Population Census 2005
	Incidence of poverty	100%	0%	Population Census 2005
USE	Share of irrigated crops from total crop area	0%	100%	Agricultural Census 2010/2011
	Agricultural area per capita	<0.1ha	>1 ha	Agricultural Census 2010/2011
	Share of population depending on aqua- or agriculture	100%	0%	Agricultural Census 2010/2011
ENV	Threatened amphibians	Category 4 (15-30% species threatened)	Category 1 (no treatened species)	Aqueduct project (Gassert et al, 2014)
	Disaster occurrence	No disasters	All disaster types occurring every 1-2 years	Agricultural Census 2010/2011
	Human Footprint	100	0	SEDAC Last of the Wild v2
	Soil degradation	Severe degradation	No degradation	Agricultural Census 2010/2011

A more detailed description of the variables as well as their processing is presented in the following sub-sections for each component.

4.2.1 Resources Index

Three variables were used to calculate the resources component. Surface water availability were simulated using a distributed physical hydrological model developed by the Environmental Impact Assessment Finland Ltd (Ympäristövaikutusten Arviointi Oy). The

main model used is comprises of the entire Mekong catchment in 5km resolution grid cells. Since the entire Laos is not contained in the Mekong Basin, three additional small catchments in northeast Laos were modelled with the same resolution to cover the entire country with model results. The model description and information on calibration is given in Appendix 1.

To include the effects of drought in the dry season, the length of the period with no rain (precipitation less than 1mm) in the dry season was calculated. For the WPI calculation, the average longest consecutive dry streak for the historical record (the length of the record varies from station to station) was used and interpolated to include the entire Laos. Surface water availability was scored according to Falkenmark indicator for water scarcity. According to Falkenmark et al (1989), water availability below 1700 m^3 per capita per year can be considered water scarce. Absolute scarcity occurs when less than 500 m^3 of water is available per capita per year. These two limits were used so that score 100 was applied when water availability was above the scarcity limit (1700 m^3 per capita per year), and score 0 was applied when less than absolute water scarcity limit (500 m^3 per capita per year) was available. Values in between were interpolated using the two limits. Scoring amount of precipitation was applied in a relative manner so that the maximum daily average precipitation in the wet season had a score of 100, and all other values scored relative to that. Average maximum duration of drought was calculated with an algorithm that counts the days in dry season with precipitation less than one millimetre, takes the maximum value for each dry season and averages them. Scoring was applied so that the best score was given to the shortest dry streak in the dry season, and 0 score to the longest dry period.

4.2.2 Access Index

The Access component were calculated using three variables; irrigation, drinking water source and toilet type. Irrigation and drinking water source are presented in the source data as Boolean values for different irrigation techniques and drinking water sources in the villages in question. Irrigation data is divided into eight categories;

1. permanent weir,
2. reservoir,
3. pump,
4. dyke,
5. temporary weir,
6. gabion,
7. other, and
8. not specified.

The different irrigation techniques were summed together for each village so that permanent weir, reservoir, pump and dyke were given double weight. The score was then calculated using a relative method where the village with the highest value was assigned with a score of 100, and the rest scored relative to this. A similar procedure was applied for the drinking water source, which is divided in six categories;

1. piped water,
2. protected well,
3. unprotected well,
4. surface water,
5. rain water, and
6. other.

Different to the irrigation calculation, some categories were given a positive value (Piped and protected and unprotected well) while others were given a negative value (surface water, rain water and other). The values were summed for each village, and a relative score was taken in the same way as in case of Irrigation. The third variable in ACC is Toilet type, which is divided into four categories in the source data: Modern, normal, other and no toilet. These categories were scored 100, 66, 33 and 0 respectively.

4.2.3 Capacity Index

Capacity component is calculated using four variables; sum of travel time to district and province capitals, road access to village (varies according to season), share of literate population from total population and incidence of poverty. The variables were chosen to reflect the ability of the village population to influence on the management of the water resource they are dependent on.

Literacy rate and incidence of poverty are presented in the source data as percentage values, and therefore they were used as they are for the index calculation (in the case of incidence of poverty, the score is 100-poverty rate). Travel time to district and provincial capital were summed together and scoring was made so that the village with shortest combined value got a score of 100 and villages with travel time of more than 600 minutes (10 hours) received a score of 0. The travel times in between were interpolated between these values. Road access was included in addition to travel time to represent additional challenges in reaching the administrative capitals from the village. The data set divided road access in to three categories; access in both seasons, access in dry season only and no road access at all. A village with road access got a score of 100 and a village with no road access was given a score of 0. Different scores were calculated for dry and wet seasons, as in some villages road access was not year round.

4.2.4 Use Index

Component of water use is calculated using three indirect variables due to direct water use data not being available. Share of irrigated cropland from the total cropland of the village is used as an indicator of the extent of used irrigation potential. Source data provides seasonal differences and therefore the irrigation scoring is different for the seasons. Second, agricultural area per person is used to indicate whether sufficient crop is produced. Third, the share of population that are dependent on either aqua- or agriculture is used to represent the population whose livelihood is dependent on water use capabilities and thus, are more vulnerable to water poverty. Scoring for the variables were applied differently for each variable. The first variable, share of irrigated area was used as it is given in the dataset (percentage). Agricultural area per person was scored so that a field area of more than 1 ha per capita received a score of 100, and less than 0.1 ha received score 0. The third variable, share of population dependent on either agri- or aquaculture

were presented in the dataset as percentages. The percentages were summed together and scoring was applied so that villages with zero percent dependent on either agri- or aquaculture received the best score, 100, while the worst score, zero, was given to villages with 100% dependency on either one.

4.2.5 Environment Index

The final component, Environment, is calculated using four variables; threatened amphibians to represent the general state of the water environment, disaster occurrence to represent the extremity of the climate and conditions surrounding the village, the state of soil degradation and fourthly, Human Footprint. Threatened amphibians was presented in the source data in four different categories;

1. low, 0%,
2. low to medium, 1-5%,
3. medium to high, 5-15%, and
4. high, 15-35% of amphibians threatened.

Scoring was applied so that a score of 100 was given to low category, 66 to low-to-medium, 33 to medium-to-high and zero score to high category. Soil degradation was likewise divided into four categories;

1. no degradation,
2. light,
3. moderate, and
4. severe degradation.

The scoring of soil degradation was applied in an identical way to the threatened amphibians. The third environmental variable, disaster, was divided into several subcategories in the source dataset. In addition to disaster occurrence in general, frequent disasters are represented in their own category (e.g. if flooding occurs frequently (every 1-2 years) in a village, the dataset value is true for both flood and frequent flood). Disasters were calculated separately for dry and wet seasons due to the nature of the disasters. It is assumed that flooding disasters do not occur during dry seasons, and that drought disasters are endemic to dry season. In addition, landslides are heavily related to strong rainfall events and due to that, they are not considered in the case of dry season.

Table 4.2 presents all the disaster categories and whether they are used in dry or wet season. The disasters were scored in a similar way as irrigation and drinking water source variables in the Access component. The occurrences were counted together, and compared to the situation where all disasters would be occurring frequently.

Table 4.2. Disaster types and the index calculations they are used for.

Disaster type	Used in
Flood	Wet season
Landslide	Wet season
Drought	Dry season
Pests	Both seasons
Other	Both seasons
Not specified	Both seasons

4.2.6 Calculating the Water Poverty Index

One of the strengths of the WPI is the ease of its application. The simplest and a common way of calculating the index is to calculate an average of the components, and to scale the sub-component values to minimum and maximum of the component range. Alternatively, one can use weighted average as shown in Equation 9

$$WPI = \frac{\sum C_i w_i}{\sum w_i} \tag{9}$$

where C_i is the component in question and w_i is the weight assigned for a specific component, and scaling with Equation 10

$$x_{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{10}$$

where x_i is the component *x* value *i* being scaled. (Lawrence, et al., 2002) However, research has shown that additive calculation of the index is not optimal due to an effect called full compensation, and therefore one should instead use multiplicative adding (geometric mean) (van der Vywer, 2013; Garriga & Foguet, 2010).

Weighting of the components has been a matter of discussion, and Sullivan et al (2006) suggest that researchers should not emphasize one component over the others because it is always a political decision. The problem of subjective weighting of the components has been addressed by Jemmali and Matoussi (2013), who used objective weighting by Principal Component Analysis (PCA). PCA is a traditional multivariate statistical method which re-projects multivariate data into principal components (PC) which represent variation in the data. The first principal component represents largest variation in the data, the second principal component represents the largest variation that is not described by the first component, and so on. The principal components can be used to derive weights for variables using Equation 11

$$w_i = \sum_{k=1}^{2} PC_k \frac{\sqrt{lambda_k}}{\sum_k \sqrt{lambda_k}} \tag{11}$$

where w_i is the weight assigned for *i*th component, PC_k is the characteristic vector (eigenvector) of the *k*th principal component and $lambda_k$ is the eigenvalue of *k*th principal component. WPI can then be objectively calculated using equation 12

$$WPI = \prod_{i=RES,ACC,CAP,USE,ENV} X_i^{w_i} \tag{12}$$

where X_i is the value of component *i*. (Jemmali & Matoussi, 2013)

4.3 Analysis Methodology

To answer the research questions presented in the Introduction chapter, a methodology specific for this study was developed. The methodology is based on the presented concepts of Exploratory (Spatial) Data Analysis and Spatial Data Mining; no prior hypotheses were developed prior to application and the data was approached in many different angles.

The methods employed, however, were selected specifically to answer the research questions. As a reminder, the research questions are:

1. Are there distinct differences between areas in their water poverty?
2. Are there distinct spatio-temporal differences in water poverty?
3. What are the causes of water poverty in Laos? Do the causes differ across space and seasons?

The methods to answer the research question are summarized in Figure 4.1. The approach is split in two; research questions 1 and 2 use the same methodology, while research question 3 uses additional methods from the SDM domain which does not concern the first two. However, the distinction is not strict; both "branches" provide complementary information to all of the research questions

The process to answer the first two research questions start with an exploration of the selected variables used to calculate the WPI. This step is an important one due to limitations in using composite indices to explain complex phenomena. (Lawrence, et al., 2002) Univariate global distributions as well as multivariate distributions of the variables are explored in addition to their spatial autocorrelations. Once the initial exploration is done, WPI is calculated for both, dry and wet seasons separately as a tool to assess the *temporal* dimension. The resulting two WPI datasets are then explored in detail using multivariate ESDA methods. This phase of exploration also includes the comparison of the two WPI's to determine whether the two seasons differ significantly from each other. Spatial dimension in the two WPI's is evaluated in two ways: First through GWSS in order to address the problem of MAUP described in Section 3.1.3. Second, WPI is explored in the context of Provinces in order to make interpretation easier and to aid in decision making in actions to alleviate water-related poverty in Laos. Finally, cluster analysis is performed on both seasons in order to make clear distinction between areas of different properties of the WPI components and their seasonal differences. Once the results of ESDA are formulated, they are tested through confirmatory data analysis. If the result is not satisfactory, ESDA is continued until acceptable results are found.

The methods to answer the causes employ SDM methods (although, it should be noted that this is an artificial distinction; SDM methods are also employed in answering the first two research questions as well as ESDA methods are employed in answering the third one). The used methods are clustering, GWR and GWPCA. Spatial clustering divides the area into distinct clusters which can be explored to answer characteristic drivers of water poverty in each cluster area. GWR is employed here to seek the variables which can be used to explain the computed WPI's. Here, additional variables are used in addition to the processed variables used for calculation of the indices, and more specifically, to investigate the local drivers. GWPCA on the other hand is used to explore the local variation in the attribute space. The results from the SDM procedure is subjected to statistical testing. As with the path to answer the first two research questions, the SDM process is continued until satisfactory results are found.

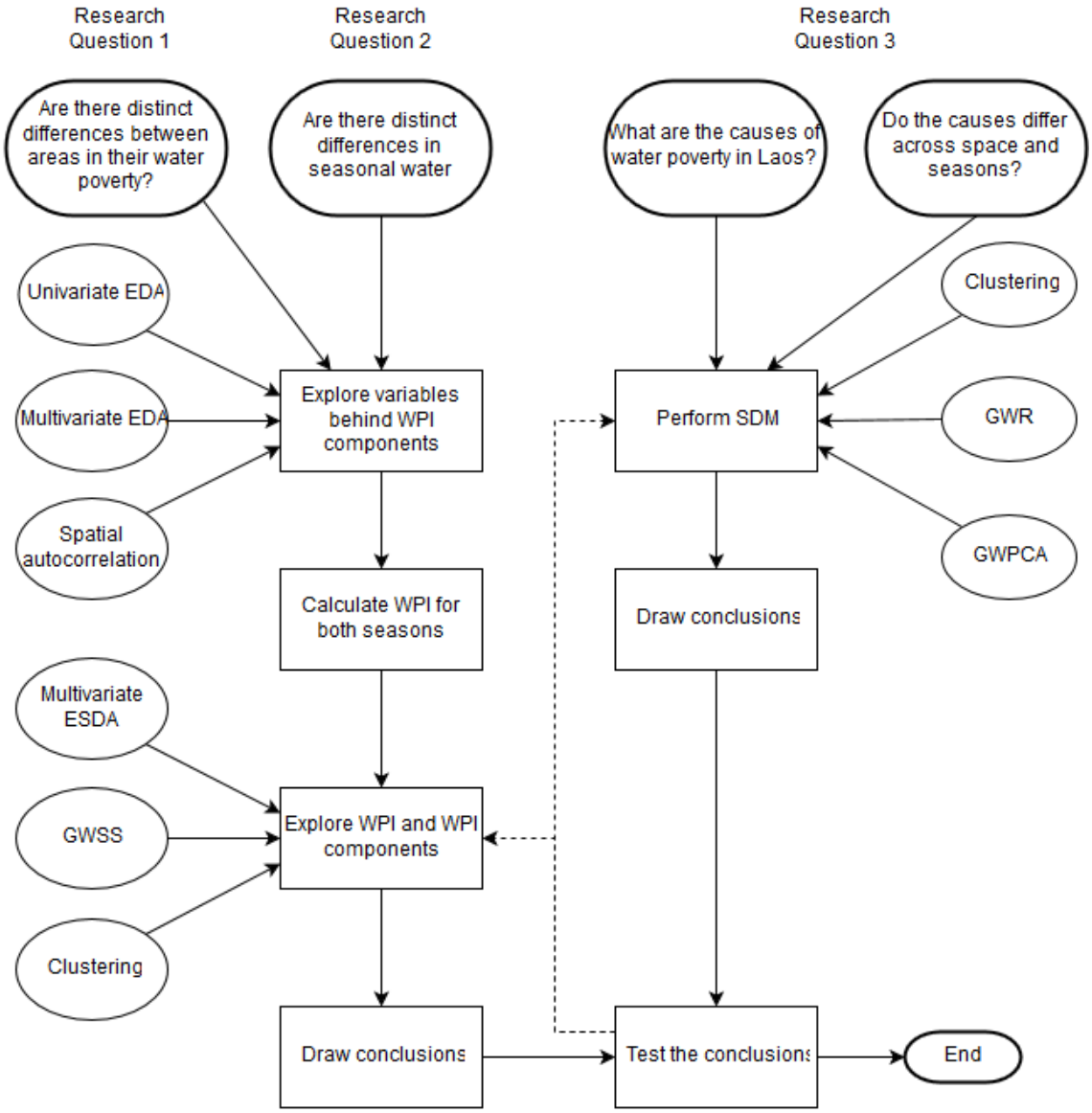


Figure 4.1. The exploratory method used in the study.

GWSS is used to get a local statistical view on the WPI components. The use of GWSS allows for addressing Edge Effects by using an adaptive bandwidth of 400 nearest neighbours (this is close to the optimal bandwidth of GWR) which accounts for approximately a sample of 5% from the total population. Additionally, using a local average of 400 villages addresses the problem of village level uncertainty reported in Epprecht et al (2008); the number of villages used takes the average to approximately provincial level, but evades the MAUP of crisp provincial borders. Using an adaptive bandwidth also ensures that the sample size in different locations does not vary.

Clustering in this study consists of several steps. The first one is to determine an optimal number of clusters via visual inspection and analytical solutions. Once the number of clusters is selected, spatial *k*-means clustering is applied to the WPI data supplemented with coordinates standardized to the WPI score range (0-100). In addition to the scores, spatial *k*-means clustering is applied for rank data to get an alternative view on the process. The clusters are then subjected to several EDA methods to study their characteristics, both in a state-wide and provincial perspectives.

Using GWR involves several steps to be taken; model and bandwidth selection, collinearity diagnostics, and statistical testing of the results. Model selection in this study is done using a step-wise selection using cross-validation (CV). The algorithm starts by calibrating a GWR model with a single independent variable. CV score is recorded for each variable, and the variable which produces the smallest CV score is selected. Then, the algorithm introduces the remaining independent variables one-by-one in addition to the already selected variable. These steps are repeated until CV score does not significantly improve. The input for the algorithm is the original values (i.e. not values processed to the scores) and consisted of all variables used to calculate WPI components, added with a number of additional, relevant variables (all of these are listed in Appendix 6). The model selection is performed with the same bandwidth as GWSS (400 Nearest Neighbours), and once the selection is done, bandwidth is optimised using CV. Basic and robust variants of GWR are used to estimate the regression coefficients to mine information about the local importance of the explanatory variables. The model is evaluated using local t-statistic, local R^2 and Monte Carlo simulation, as suggested by Demsar et al (2008) and by Brunsdon et al (1998). In addition, the three F statistics outlined in Leung et al (2000) are calculated. Local Multicollinearity is addressed in the vein of Wheeler and Tiefelsdorf (2005) and Wheeler (2006) suggestions.

GWPCA is used in the SDM process to gain additional information on the local differences in WPI components. It is run with the same bandwidth selection as GWSS and GWR step-wise model selection to ease interpretation in relation to the other methods. The analysis follows the recommendations in Demsar et al (2013), Charlton et al (2010) and Lloyd (2010). Spatial variation of the principal components is tested using Monte Carlo procedure.

All in all, the methodology is a combination of visual and computational exploratory methods with an emphasis on cartographic visualization and spatial variability.

4.4 Implementation Tools

Two main tools were used to implement the methodology outlined in the previous section. First, data manipulation and data collection were done using QGIS version Essen, 2.14 (QGIS Development Team, 2016). The variables used were collected to points representing the villages used in the study; in the case of raster data (modelling results, water consumption, elevation and slope class), the raster cell value in which the village point is taken as representative to the village. The raster data is resampled to 5km resolution prior to assigning it to the village. The dataset is then exported to a shapefile for analysis in R.

R is a free and open source statistical programming language (or programming environment) (R Core Team, 2016). R can be extended via user contributed "packages", which extend the functionality of base R. Currently (17th July 2016) there are 8775 packages available through the Comprehensive R Archive Network (CRAN), which is the largest repository for R packages. A fair number of different packages are utilized in the analysis work in this study. A list of the most important used packages is provided below (however, the list is not complete – packages are often linked to other ones. Only the main packages are included in the list):

- Analysis
  - Base R (R Core Team, 2016) for basic statistics and functionality.
  - sp (Bivand, et al., 2013) for handling spatial data in R.
  - spdep (Bivand & Piras, 2015) for calculating Moran's I.
  - GWmodel (Gollini, et al., 2015) to perform GWSS, GWR and GWPCA.
  - spgwr (Bivand & Yu, 2015) for an alternative implementation of GWR.
  - cluster (Maechler, et al., 2016) for clustering.
  - NbClust (Charrad, et al., 2014) for analytical choice of the number of clusters.
- Visualization
  - ggplot2 (Wickham, 2009) to create majority of the illustrations.
  - plotly (Sievert, et al., 2016) and ggiraph (Gohel, 2016) to create interactive versions of ggplot2 illustrations.
  - GISTools (Brunsdon & Chen, 2014) for some of the maps.

The source code for all the analyses and data manipulations is published under the author's personal website in http://markokallio.fi/waterpoverty as well as under a GitHub repository mkkallio/waterpoverty. An interactive version of the study can also be found in the author's website.

5 Results

The results chapter is organized so that for each analysis, first dry and wet season WPI are analysed separately, followed by a comparison of the two. First the initial data set and selected variables are explored. In the second part the first research question (the spatial dimension) is investigated. Once an answer is established, the second research question is addressed by pitting the seasonal WPI's against each other. Finally, in the last part, data mining on the causes of water poverty is performed in order to answer the third research question.

5.1 Exploring the Variables

The final dataset used for the analysis consists of a total of 8215 villages, which is the number of villages with data on both main datasets (population and agricultural censuses). The number can be considered extremely high, since the United Nations in Lao PDR country profile (2015) places the number of villages in Laos to approximately 8600. The number of villages in the dataset for each province is shown in Figure 5.1. Attapeu and Xekong contain the least numbers (145 and 226 respectively), while the biggest numbers are found in Houaphan, Luang Prabang and Savannakhet (710, 707 and 997 respectively). A dot density map of the villages in Figure 5.2 shows the province borders and the village locations. Villages in the data cover nearly all corners of the country and they form a representative sample.

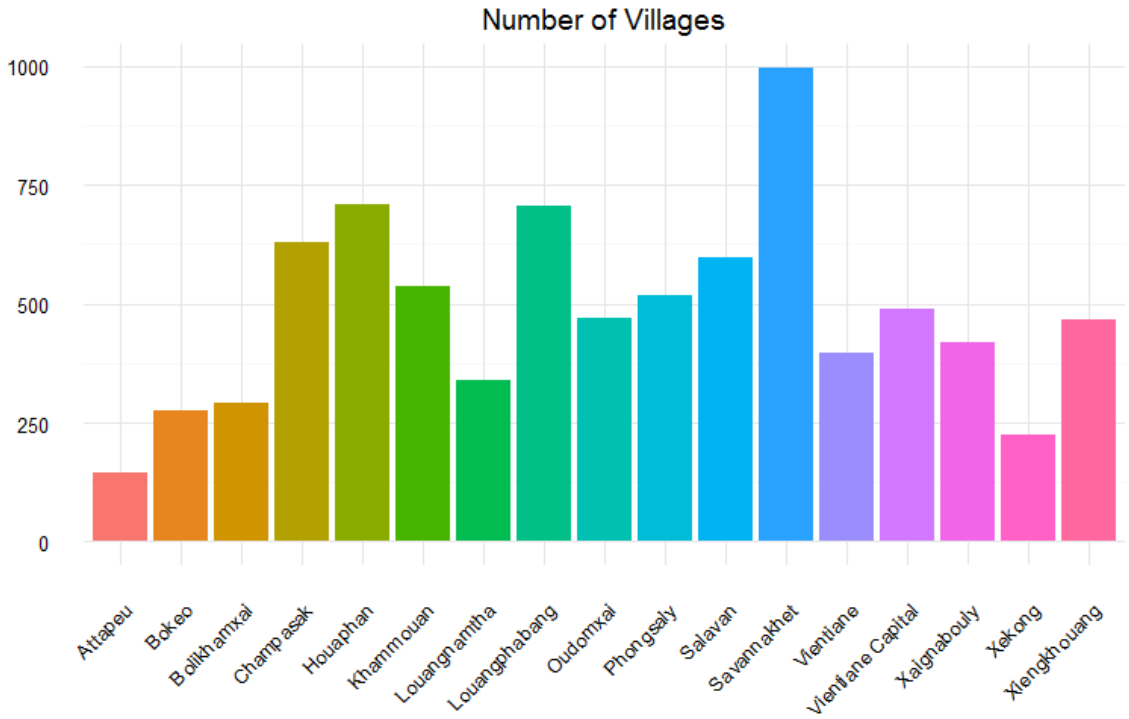


Figure 5.1. Number of villages in the dataset for each province of Laos.

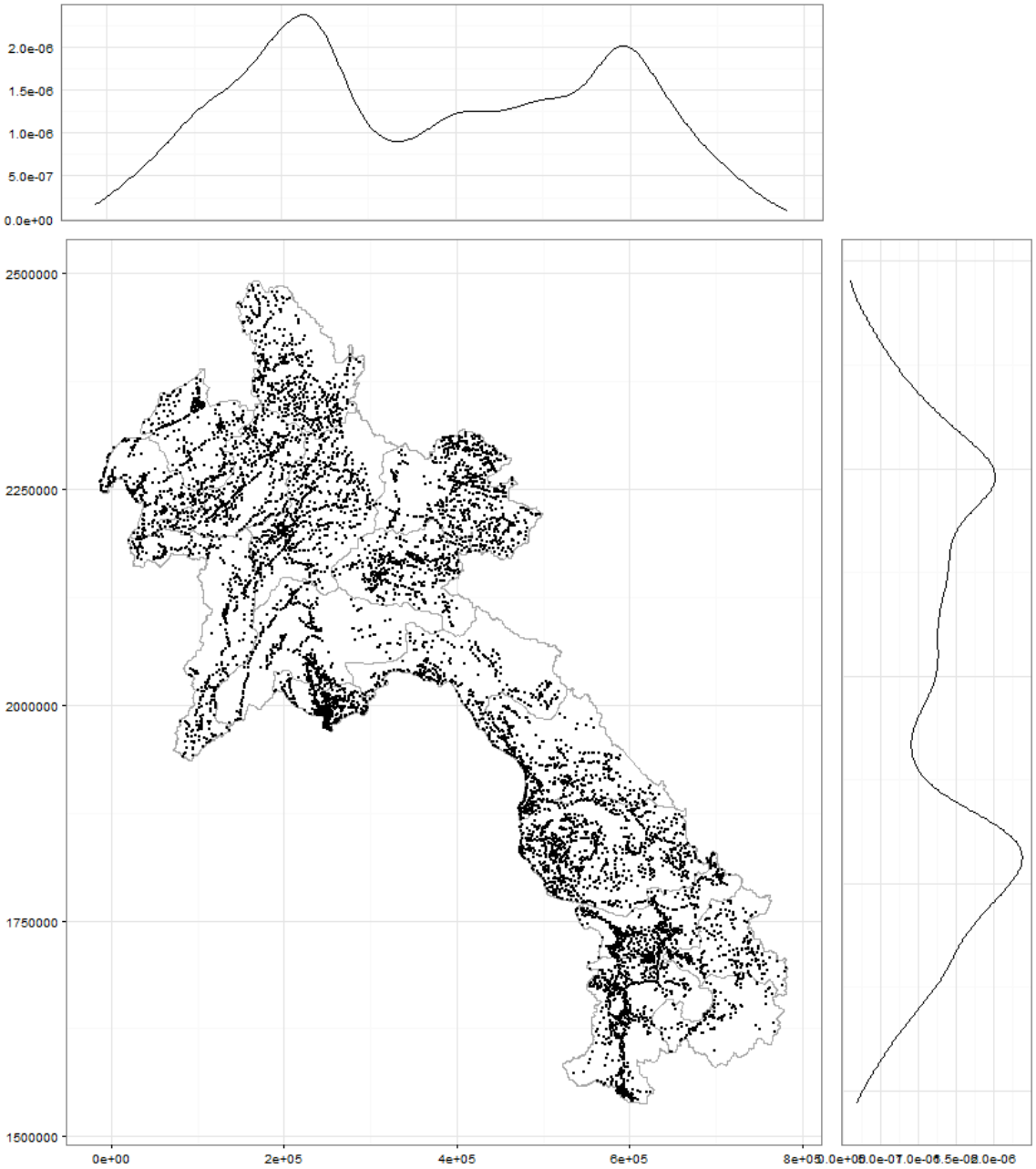


Figure 5.2. Dot density map of villages in the dataset. Coordinate system used is UTM Zone 48N (EPSG: 32648).

5.1.1 Resources

Resources Index consists of three variables which describe different aspects of the water availability. Box plots of the variables in dry and wet season are shown in Figure 5.3. Strikingly, most of the villages get a score of 100 for surface water availability even during the dry season due to the way the score is calculated (Falkenmark Stress Index, based on simulated discharge). On the other hand, a number of villages exhibit water scarcity even during the wet season, as we can see from the figure. These villages are not considered as outliers for it seems that the distribution simply has a long tail without major gaps between score values. Water scarce villages are presented in Figure 5.4. the total number of these villages is 1388 (16.9% of all villages) in the dry and 318 (3.9%) in the wet season, all in areas of least seasonal precipitation. Precipitation variability difference is

distinct between seasons as well as within wet season. Villages which score poorly on wet season precipitation are mostly located in the west and northwest of the country in provinces of Xayabouly, Vientiane, Louang Prabang, Oudomxai, Phongsaly and Louang Namtha. Small region in the south at the border between Khammuane and Savannakhet score low as well. The last variable, average longest consecutive drought day (days with precipitation less than 1mm) sequence is again extremely different between seasons. In the wet season, all villages get a score of 100, while in the dry season the entire range is occupied. Highest score is found in the northwest of the country, where the dry season is broken often by rainy days. In the south, however, the dry period can extend on average to more than 100 days. Maps of variable scoring are provided in **Error! Reference source not found.**.

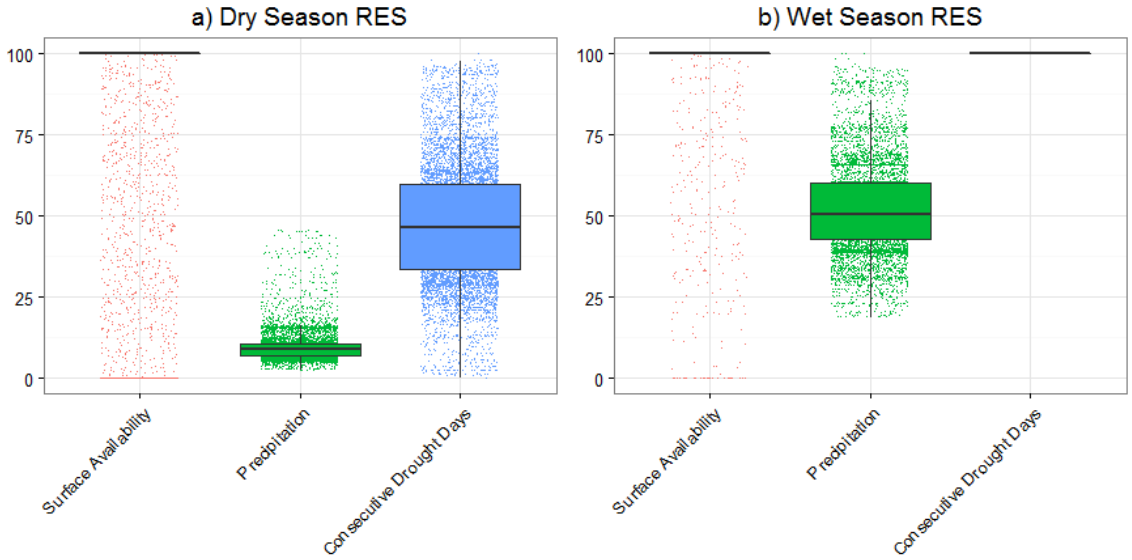


Figure 5.3. Resources component variability in a) dry and b) wet season. Hinges of the boxes signify 25% and 75% of the sample and whiskers cover 1.5 times the interquartile range.

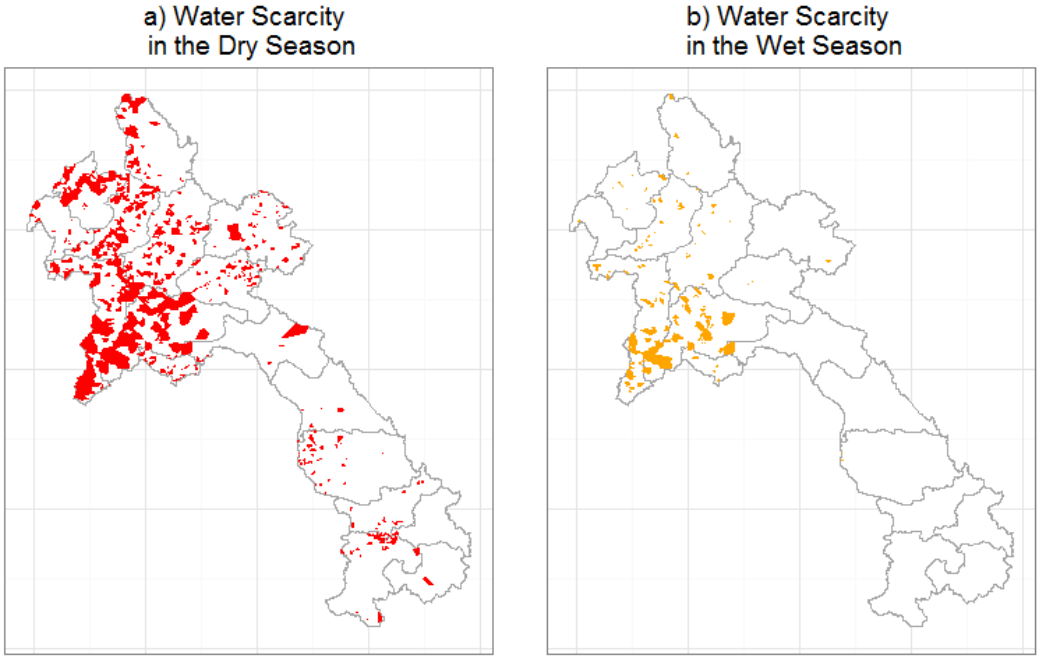


Figure 5.4. Villages with water scarcity (water availability score less than 100) in a) dry and b) wet seasons.

In addition to the box plots of the variables, their correlations were plotted in scatterplot matrices shown in Figure 5.5. The variables show only weak correlations among each other, which can be considered a desired property, according to Lawrence et al (2002). The dry season variable distributions are extremely skewed with the exception of drought days, which is approximately normally distributed. In the wet season, precipitation and soil water availability have a medium strength correlation, as can be expected, however the scatterplot shows that the correlation is not uniform across the range. Surface water availability is extremely skewed, but the other two distributed variables (precipitation and soil water availability) are more evenly spread out through the range.

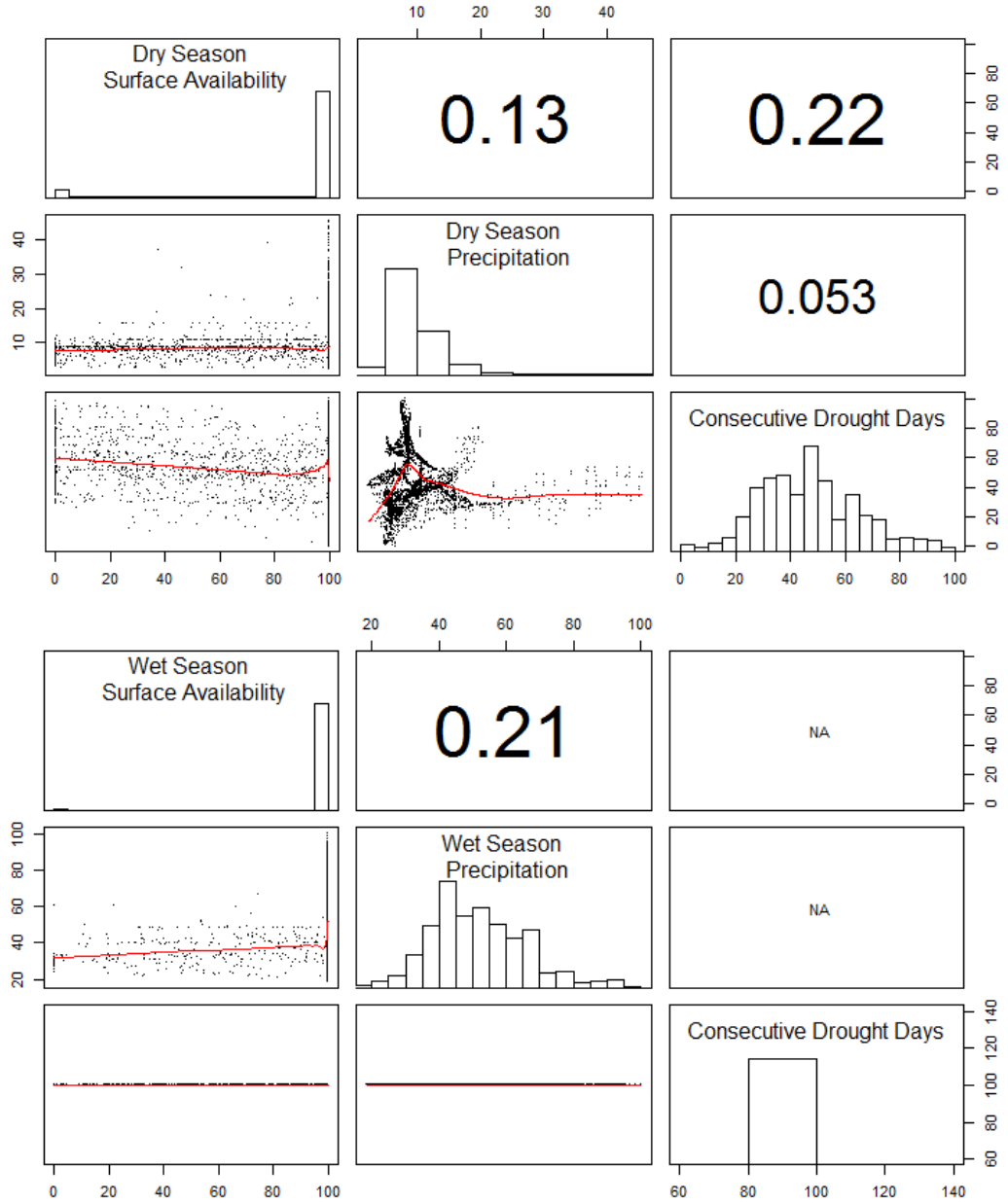


Figure 5.5. Scatterplot matrix for Resources component. The upper matrix is for dry season and the lower for wet season.

Spatial autocorrelation was analysed by calculating Moran's I (Table 5.1) for all of the variables. From the seven analysed variables (drought days for wet season were omitted

because the score is 100, meaning there is zero autocorrelation), all show substantial spatial autocorrelation. This is visually evident when looking at maps of the variable scores, provided in Appendix 2.Precipitation for both seasons and drought days in the dry season appear extremely autocorrelated.

Table 5.1. Moran's I for Resource component variables.

	Moran I statistic	Expectation	Variance
DryAvail	0.551	0.000	0.000
DryPrec	0.977	0.000	0.000
AvMaxDDay	0.996	0.000	0.000
WetAvail	0.418	0.000	0.000
WetPrec	0.989	0.000	0.000

5.1.2 Access

The Access component consists of three variables, which are identical for both, dry and wet season. Scores for all variables consist of steps, as can be seen from the box plot in Figure 5.6. Irrigation type scores are concentrated in the upper half of the range with only a handful of villages with scores below 50. Drinking water source and toilet type occupy the entire range with drinking water source centred around the score of 50 and toilet type skewed toward the lower half.

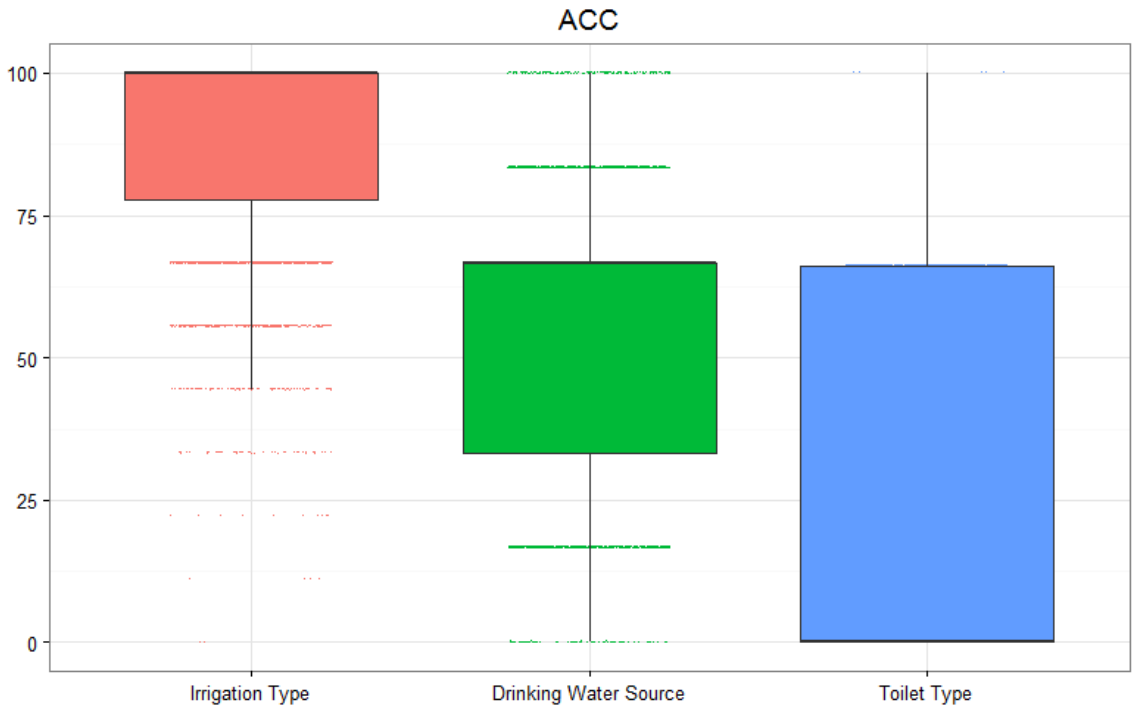


Figure 5.6. Access component variability. Hinges of the boxes signify 25% and 75% of the sample and whiskers cover 1.5 times the interquartile range.

The variables show only weak correlation with each other. Toilet type weakly correlates with both, drinking water source and irrigation type with correlation coefficients of 0.24 and 0.23 respectively. There is no correlation between irrigation type and drinking water source. Distribution are different for each of the variables. Irrigation type follows a Poisson distribution, while drinking water source follows a positively skewed normal distribution. Toilet type on the other hand exhibits two peaks in its distribution.

Moran's I was computed for all of the variables and shown in Table 5.2. All of the variables are spatially autocorrelated, however to a lower degree than the variables in Resources component.

Table 5.2. Moran's I for Access component variables.

	Moran I statistic	Expectation	Variance
Irrigation	0.383	0.000	0.000
Drinking Water Source	0.525	0.000	0.000
Toilet Type	0.510	0.000	0.000

5.1.3 Capacity

Four variables make out the Capacity component; travel time to (district + province) capitals, road access, literacy rate and incidence of poverty. The variables in this component are diverse and they occupy the entire range of index scores with small variations in the location and tails of the distribution. A box plot of the variables is shown in Figure 5.7. The variables all show similar pattern: high scores are found all along the Mekong River and where the landscape is generally flat. A clear trend is visual where the score gradually gets lower as distance from the Mekong increases. Interestingly, this pattern is visible in the travel time to administrative capitals as well: high scores are only found in small areas near the administrative capitals within the provinces located in the mountains. The low-lying areas near Mekong all score relatively high. Additionally, similar pattern can be seen in road access. Road access is the only one of the four which is changing between the seasons, and the only one that does not occypy the entire score range. In the dry season there are 716 (8.7%) villages with no road access (score 0), while in the dry season the number increases to 2777 (33.8%). The villages without road access are shown in Figure 5.8. Scatterplot matrix in Figure 5.9 reveals that the variables appear correllated in a considerably higher degree than the previous components of Access and Resources. Correlation between variables is not desirable, however, the scatterplot shows that while there is a medium strength correlation, the data points are spread out occupying the plot-space nearly entirely. The highest correlation can be found between literacy rate and incidence of poverty, which can be expected based on current knowledge of poverty. Travel time to capital correlates with literacy rate and poverty. This is also expected due the effect of better access to markets and education when a village is near an administrative capital. Surprisingly, road access does not have a strong correlation with travel time to capital.

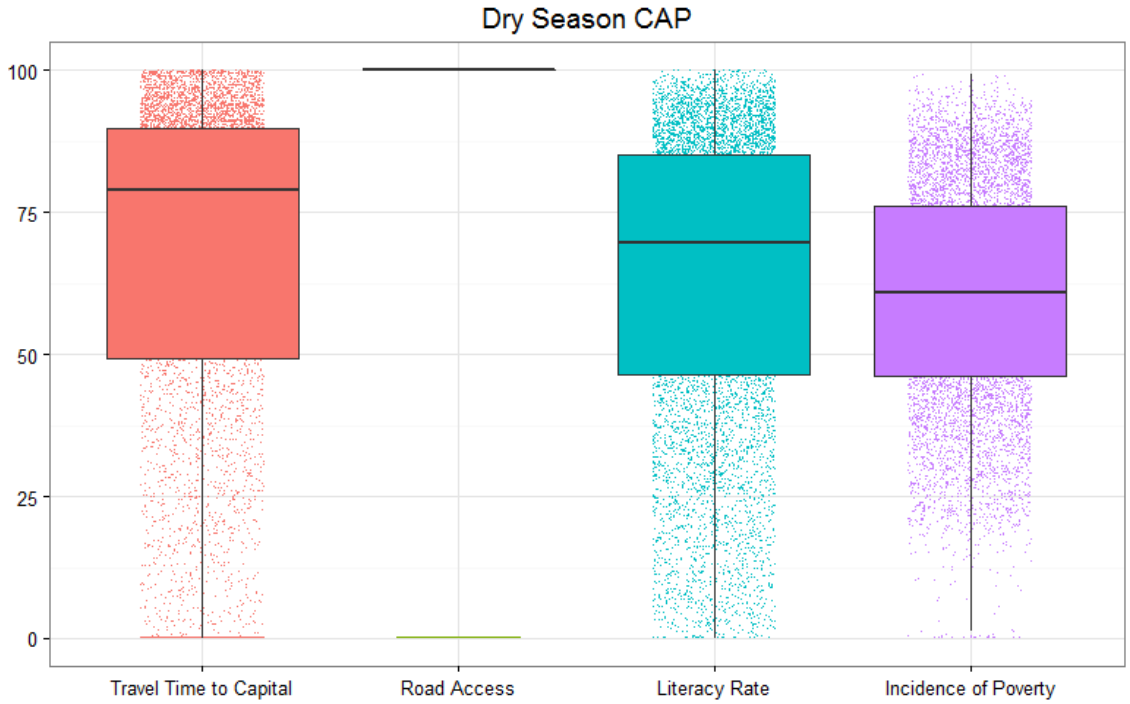


Figure 5.7. Capacity component variability for the dry season. Hinges of the boxes signify 25% and 75% of the sample and whiskers cover 1.5 times the interquartile range.

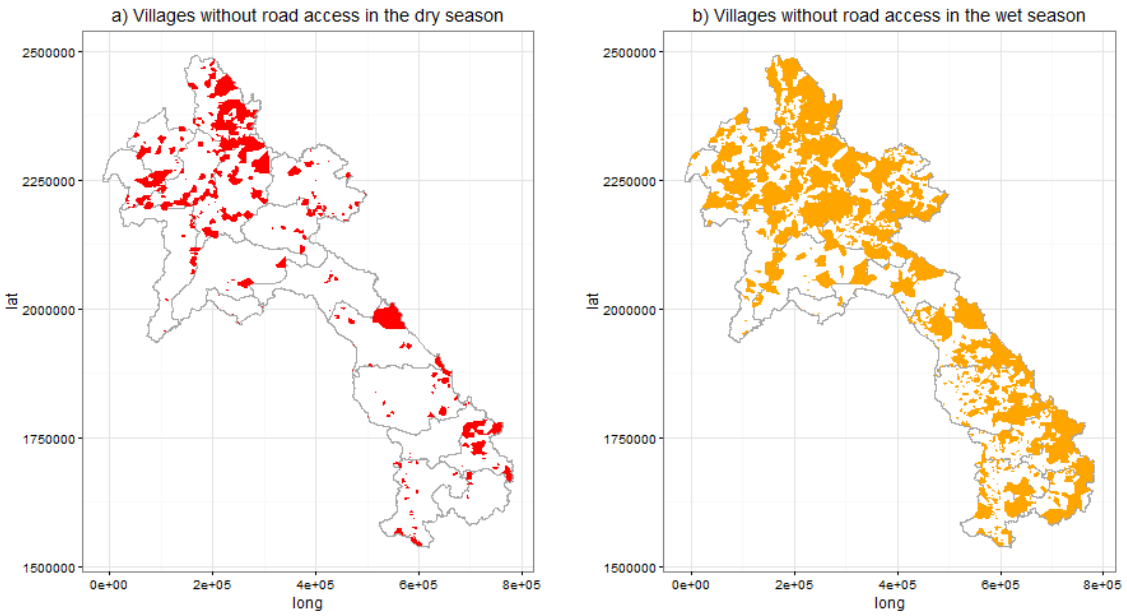


Figure 5.8. Villages without road access in a) dry and b) wet season.

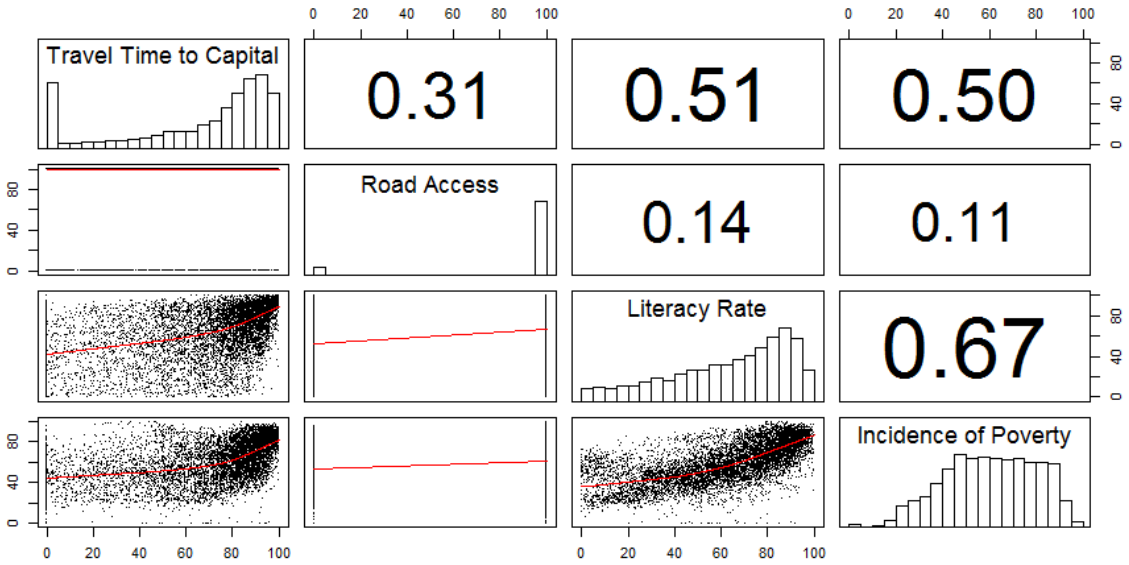


Figure 5.9. Scatterplot matrix for dry season Capacity component.

Table 5.3 summarizes Moran's I for the variables in Capacity component. Strong spatial autocorrelation can be found, unsurprisingly, in travel time to administrative capitals, literacy rate and the incidence of poverty. Likewise, road access is found to be autocorrelated and there is no significant difference between the seasons.

Table 5.3. Moran's I for Capacity component variables.

	Moran I statistic	Expectation	Variance
Travel Time to Capital	0.836	0.000	0.000
Dry Road Access	0.450	0.000	0.000
Wet Road Access	0.460	0.000	0.000
Literacy Rate	0.694	0.000	0.000
Incidence of Poverty	0.719	0.000	0.000

5.1.4 Use

Use component is made up from three variables; irrigation rate, agricultural area per person and rate of population depending on agri- or aquaculture. The variables are shown in a box plot in Figure 5.10. All of the variables occupy the entire range, however, only the rate of population depending on water is centred around score 50. The other two variables are skewed towards the bottom half and irrigation rate nearly entirely below score of 50. The only component that changes between the seasons is the irrigation rate. The relationship between dry and wet season is shown in Figure 5.11. The figure shows that surprisingly there are a number of villages in which a larger share of crops is irrigated in the dry season than in the wet season. The distribution of villages that score high in irrigation is another surprise: They appear to be away from the lowlands near Mekong and in the areas with higher slopes.

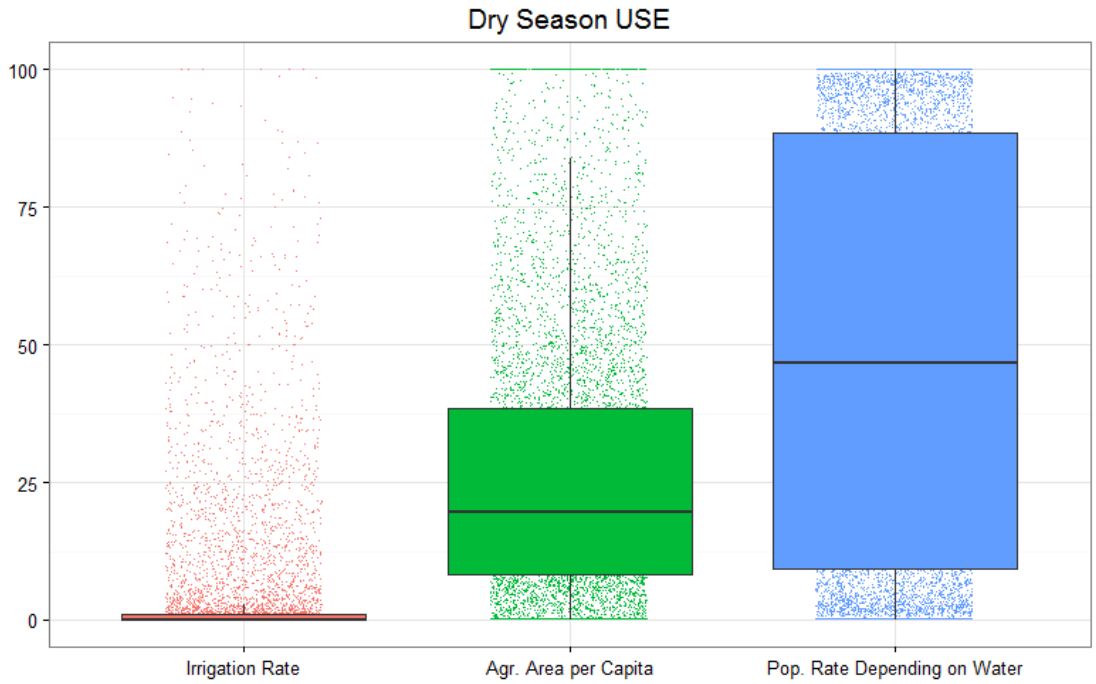


Figure 5.10. Use component variability for wet season.

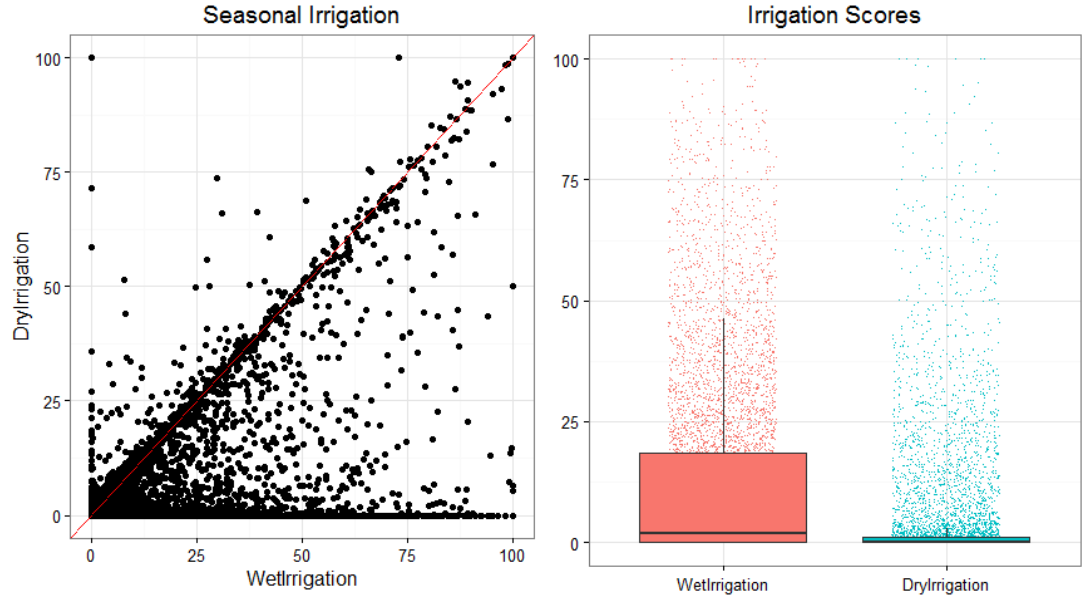


Figure 5.11. A scatterplot and a boxplot between wet and dry season irrigation.

Agricultural area per person scores generally low, however there are a few areas which seem to contain a large field area per person. These areas are the southern tip of Xayabouly, central and northern Louang Prabang, and the area near Mekong in Savannakhet, Saravane and Champasak. In Xayabouly, and in Champasak this area scores low on population depending on agri- or aquaculture for their main income. The pattern where this variable scores high and low is does not follow the geography of the country, with high and low scores found near and far from the major rivers and in an out of the mountainous areas. The variables show only weak correlations amongst each other with highest coefficient being 0.20. In addition, Moran's I was analysed and is summarized in Table 5.4. Medium strength spatial autocorrelation is found on all of the variables.

Table 5.4. Moran's I for Use component variables.

	Moran I statistic	Expectation	Variance
Dry Irrigation	0.579	0.000	0.000
Agr. Area Per Capita	0.468	0.000	0.000
Population Depending on Water	0.572	0.000	0.000
Wet Irrigation	0.501	0.000	0.000

5.1.5 Environment

Environment component consist of four variables; threatened amphibians, disaster occurrence, soil degradation and human footprint. Boxplot (Figure 5.12a) drawn from the variables clearly indicate that the first three variables are categorical ones. The scores for all of the variables are in the top half of the range, except Threatened Amphibians which occupies the entire score range. The only variable that changes between seasons is disaster occurrence; the difference is shown in Figure 5.12b. The change is not dramatic, however, the range of scores in wet season is wider and the overall score is on average better. This suggests that drought disasters are somewhat more frequent than floods and landslides. In fact, 67% of villages in the data experienced drought disasters when the same figure for floods and landslides are 32% and 14%. The vast majority of these villages also report frequent disasters, and in this statistic droughts have a higher representation than the wet season disasters as well.

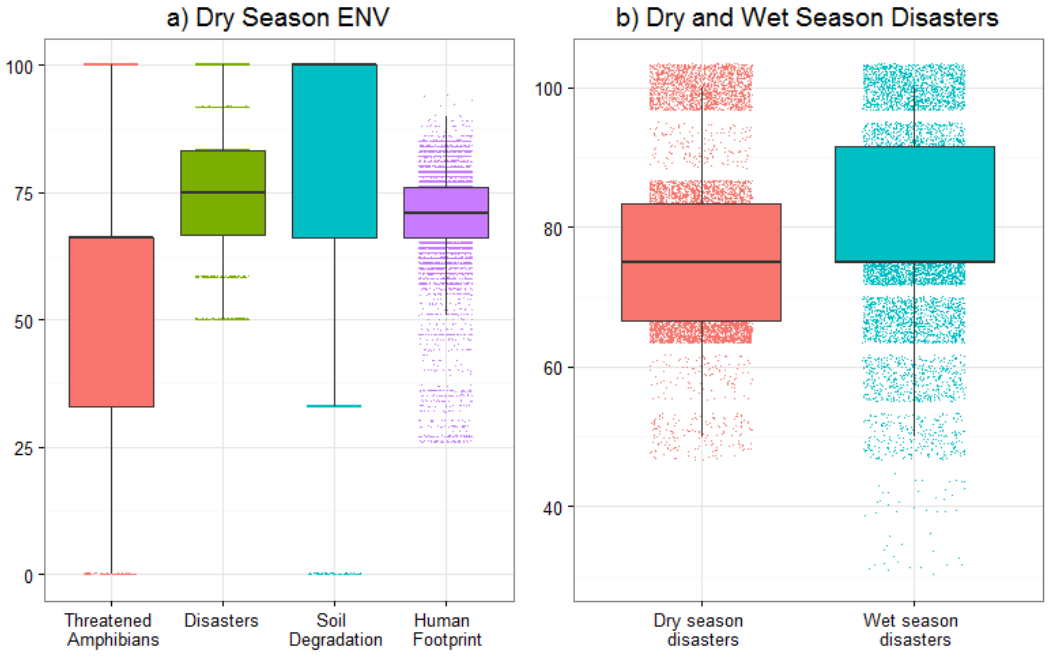


Figure 5.12. a) Environment component variability for the dry season and b) disaster scoring difference between the seasons.

Villages score fairly high in the Environment variables with the exception of threatened amphibians. Scores lower than 50 are found in the northwest (Phongsaly, Oudomxai, Louangnamtha, Bokeo) and in the southeast (Xekong, Attapeu). A small region in Louangnamtha is assigned with 0 score. Disaster scores are on the better half of 50, with

the highest scores found in the northwest and in Vientiane Capital and Vientiane Province. Southern part of the country score more uniformly at approximately 60-70, suggesting that a disasters occur at a higher rate than in the north. Soil degradation is random seems random where it exists, except for two distinct areas; the areas around mountainous Xekong-Saravane and southern Phongsaly-northern Oudomxai. Human Footprint clearly shows where the population live in Laos, with a highlight on the low score of the capital city. However, the score is fairly high due to low amount of population leaving large areas relatively wild.

The variables are not correlated with each other, with the maximum coefficient being 0.26 between human footprint and threatened amphibians. As with the previous components, Moran's I was calculated for each variable and is presented in Table 5.5. Human Footprint and Threatened Amphibians are extremely strongly clustered in the data. The extremely high value for Threatened Amphibians is due to the data being aggregated to river basins, which create large areas of similar values while neighbouring river basins likewise are assigned by a single value. The weakest spatial autocorrelation for all variables in all components, are found in the disaster occurrences and soil degradation. Regardless of showing the lowest degree of spatial relationship in the data used, Moran's I for these variables is approximately 0.3 – commonly interpreted as a strong spatial autocorrelation (Getis, 2010). This means that all variables, despite being described as weak earlier in the text (they are weak only in the relative context of this study), show strong spatial dependence. This partly answers to the first research question on spatial differences in water poverty. At this point we know for certain that the indicators *do* vary spatially beyond doubt. The next section explores the actual Water Poverty Index to find a definite answer to the question.

Table 5.5. Moran's I for Environment component variables.

	Moran I statistic	Expectation	Variance
Threatened Amphibians	0.970	0.000	0.000
(Dry) Disasters	0.364	0.000	0.000
(Wet) Disasters	0.312	0.000	0.000
Soil Degradation	0.265	0.000	0.000
Human Footprint	0.912	0.000	0.000

5.2 Spatial Dimensions of Water Poverty

WPI was calculated for dry and wet seasons explicitly, using variables described in Section 4.2. The result is a distinct index number for each village in the data used for the analysis. The following sections first describe the spatial dimensions of the dry and wet season WPI separately, answering to the first research question: "*Are there distinct differences between areas in their water poverty*?"

5.2.1 Dry Season

Scores of individual components vary quite substantially with lowest scores found in USE component and the highest in CAP and ENV. Maps of geographically weighted mean for each component as well as overall WPI are shown in Figure 5.13 (village level plots are provided in Appendix 3). Clear regions of higher and lower score can be found for each

of the components. Northern part of Laos score higher in Resources than the rest of the country. This is due to the monsoon not having as big effect in the mountainous north as in the rest of the country: Rainfall is more stable throughout the year. Access component shows surprising pattern with lower score near Mekong and Thailand (the less water poor regions), and with capital area scoring medium high. The highest Access score is found in the north in a few small ”islands”. The variables underlying ACC are fairly randomly distributed across the country with the exception of Toilet type, which scores high near the capital area. Capacity on the other hand seems highly autocorrelated with the highest scores found near the Thai border. The further away from the border, the lower the Capacity. USE component shows three distinct ”stripes” of high values with lower values in between and around them. The score more or less visually corresponds to the variables used to calculate the component. The last component, ENV exhibits two regions of high scores; in the south between Bolaven Plateau and around the capital area. Notably, the capital itself scores low due to high impact of Human Footprint. Figure 5.13f presents the computed WPI for Laos using a PCA derived ”objective” weighting and computed using multiplicative (geometric mean) function as explained in Section 4.2.6. Dry season water poverty seems high (low index values) in the northern mountainous areas as well as in the southeast corner of the country. Low water poverty region follows the Mekong river in the Thai border with an intrusion to northern Laos through Vientiane and Xiengkhuang provinces.

Spatial autocorrelation was analysed for the component scores and WPI using the ”spdep” package in R. Resources, Use and Environment component along with the overall WPI all have medium strength spatial autocorrelation with Moran’s I between 0.48 to 0.63. Capacity on the other hand has a stronger spatial dependence with the index value reaching 0.78. Access, on the other hand has a weaker spatial relationship with Moran’s I of 0.32. The semi-strong index value of WPI tells, along with the map in Figure 5.13f, that there are distinct areas of low and high water poverty in the country during dry season. This partly answers to the first research question of whether there are distinct differences between areas in their water poverty.

In order to explore the provincial dimension, a density plot for dry season WPI is given in Figure 5.14 for each province. For better view on the plots, interactive version is available in the authors website at http://markokallio.fi/waterpoverty/. In addition, a table of summary statistics is provided in Appendix 3. The density plots clearly indicate a left-tailed normal distribution for the country level WPI values, with a mean of 54.60. Looking at the individual densities, Xekong appears to be the poorest of all provinces with majority of the poorest villages located in the province. In addition, Houaphan and Phongsaly also include some of the villages with lowest dry season scores. In the other end of the scale, Vientiane Capital and Bolikhamxai are the most water rich provinces in the dry season, with a clear margin. The same provinces inhabit the bottom and top when looking at the mean and median of village WPI rankings. Three categories can be identified; Xekong, Oudomxai, Houaphan, Xayabouly, Phongsaly and Louangnamtha form the water poor provinces, in the order of mention. The median WPI rank of these provinces fall between 1700 and 2890 (out of 8215). The water-rich (or at least relatively water-rich) provinces are Bolikhamxai and Vientiane Capital by a clear margin with median WPI ranks of 6922 and 6653 respectively. The mid-range group consists of the remaining provinces and their median WPI ranks fall between 4081 and 4968. It should be noted that each province contains villages through the entire range, meaning that the provinces are far from being universally water-poor or water-rich. A map of the village ranking is given in Figure 5.15.

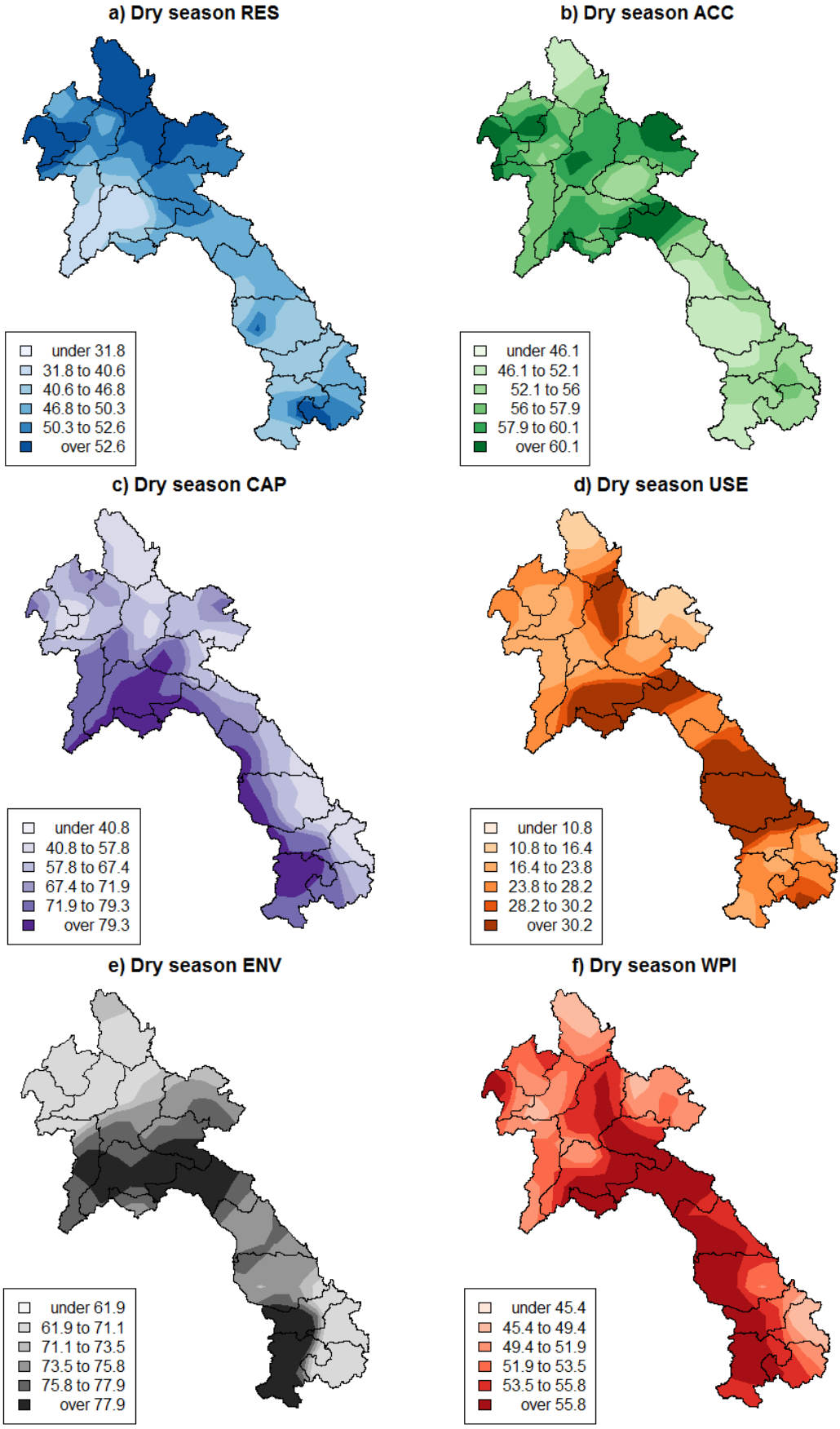


Figure 5.13. WPI components and WPI for dry season. The WPI score is calculated using PCA derived weighting scheme from the components and combined using a multiplicative function. Note that the colouring is relative to the component, not over the entire range 0-100.

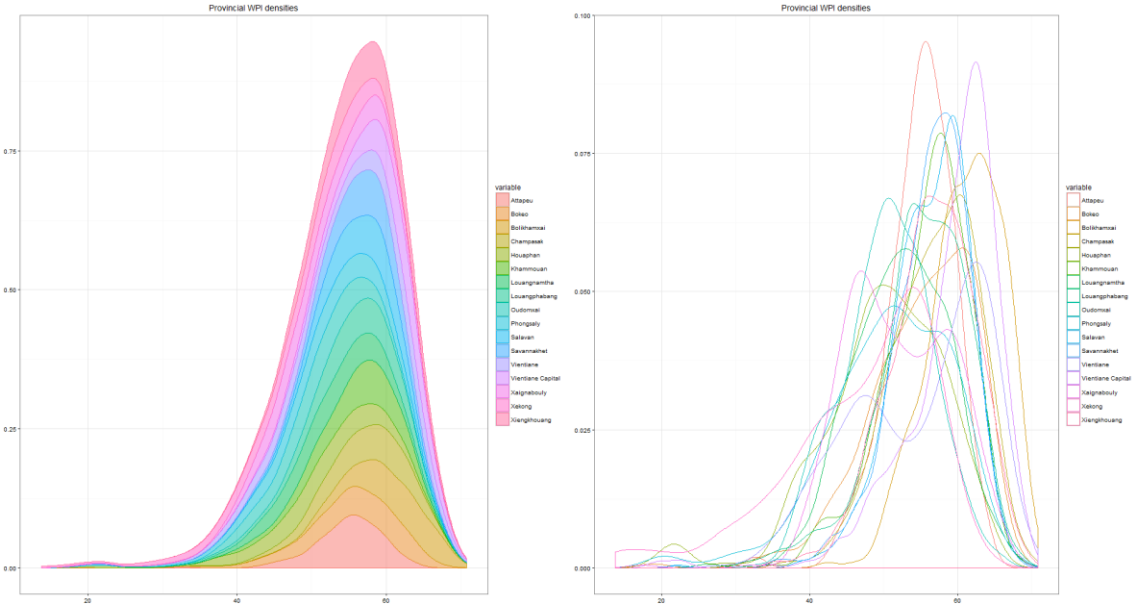


Figure 5.14. Stacked density (left) and normal density plots for dry season WPI in each province.

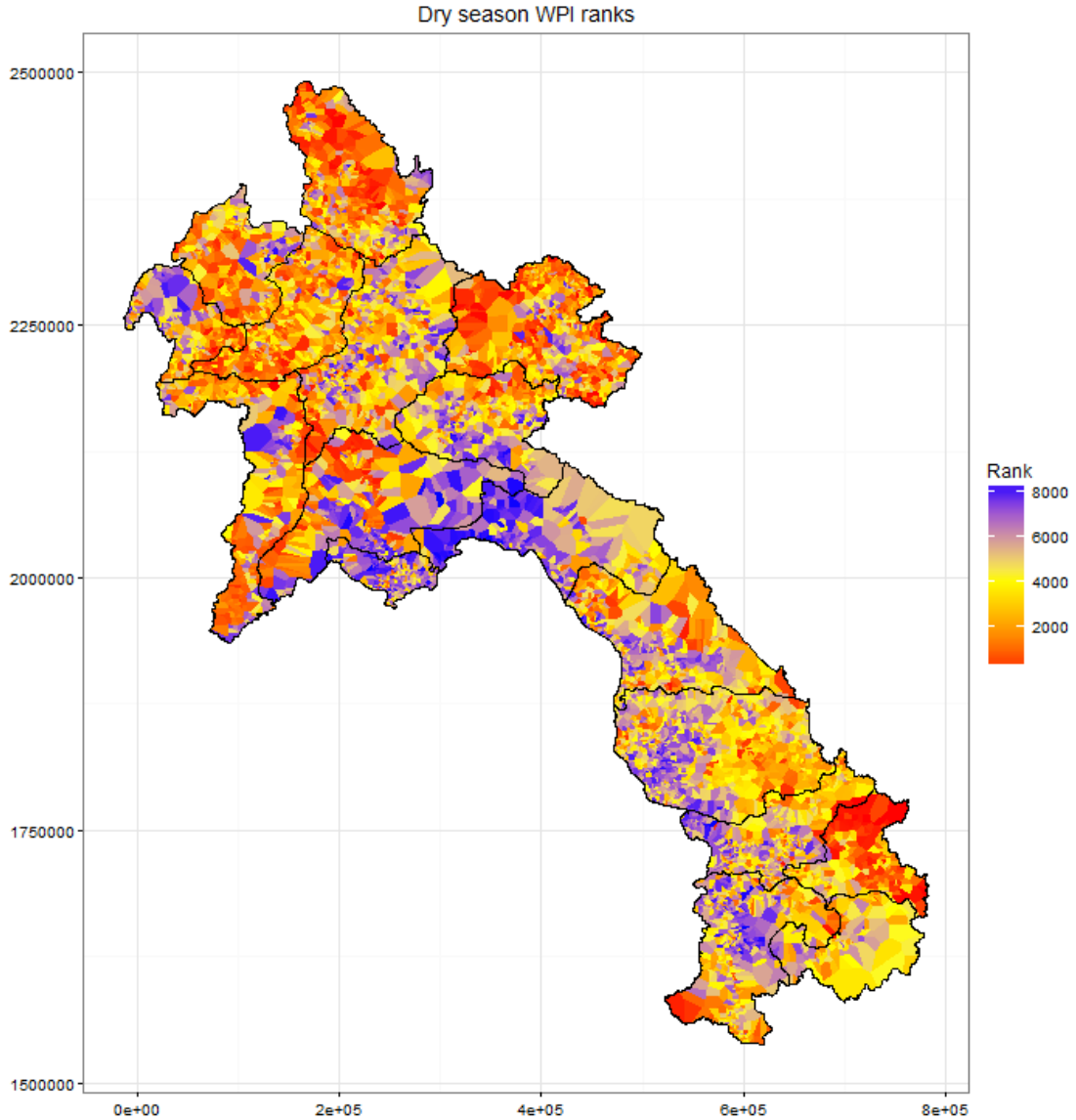


Figure 5.15. Rank plot of dry season WPI. Ranks are ordered so that the highest ranks are given to villages with the highest WPI.

A pairwise t-test was run for the WPI values to find out whether the provincial differences in mean WPI is statistically significant. As a result, null hypothesis of there being no differences is rejected with extremely high confidence level. Naturally, some provinces are similar in their distributions, but overall the results support the interpretation of significant differences. This and a visual evaluation of the figures presented gives a confirming answer to the first research question. Significant spatial differences in the water poverty level across Laos do exist in the dry season.

5.2.2 Wet Season

Localized statistics were also calculated for wet season components using GWSS with an adaptive bandwidth of 400 nearest neighbours, presented in Figure 5.16. The resulting maps are close to the ones shown for dry season (Figure 5.13). The main difference between the seasons is, as may be expected, in the Resources component due to the Monsoon. The mean value for RES rises from 48.6 in the dry season to 83.4 in the wet season. Access component is identical between the seasons, and the other components Capacity, Use and Environment show only minor differences. However, the direction of the change is interesting. Capacity is, in fact, lower in the wet season than in the dry season. The reason for this is road access; a large number of villages are cut off from the road network during the wet season. In Use component, the change is mostly positive, however, in the south of the country (Savannakhet, Attapeu, Champasak and Xekong) some villages show changes are towards the negative. Finally, ENV scores generally higher in the wet season, however, there are occasional differences as the distribution of scores increases in range (see Figure 5.12). The distributions of seasonal WPI looks similar, however the values are approximately 10 points higher than in the counterpart. Villages located close to the Mekong/Thai border again, score higher in than villages near the borders to Viet Nam and China. The border area between provinces of Xiengkhuang, Houaphan and Louang Prabang, which scored high in dry season, fall to mid-range score in the wet season.

The components show a somewhat higher degree of spatial autocorrelation in the wet season than in the dry season. Two of the components, Resources and Capacity have very strong spatial relationships with Moran's I of 0.81 and 0.75 respectably. Access component does not have variables that change between seasons, and therefore there is no change in Moran's I either. The value for Use is slightly lower (0.44) than in the dry season, and for overall WPI score, it is slightly higher (0.66).

Density plot for the province WPI values is shown in Figure 5.17, as was shown for dry season in the previous section. The same provinces stand out as poor as in the dry season. Xekong is as a clear outlier and scores by fat the lowest. Phongsaly and Houaphan are joined by Oudomxai to form the group of poor provinces. Bolikhamxai, on the other hand, is a clear outlier in the rich part of the wet season WPI range. Again, it is the richest province with a clear margin, followed by a group formed by Vientiane, Vientiane Capital and Salavan. Three of these four are neighbours in central Laos to the north and east of the capital city, and Salavan located in southern part of the country, just north of Bolaven Plateau.

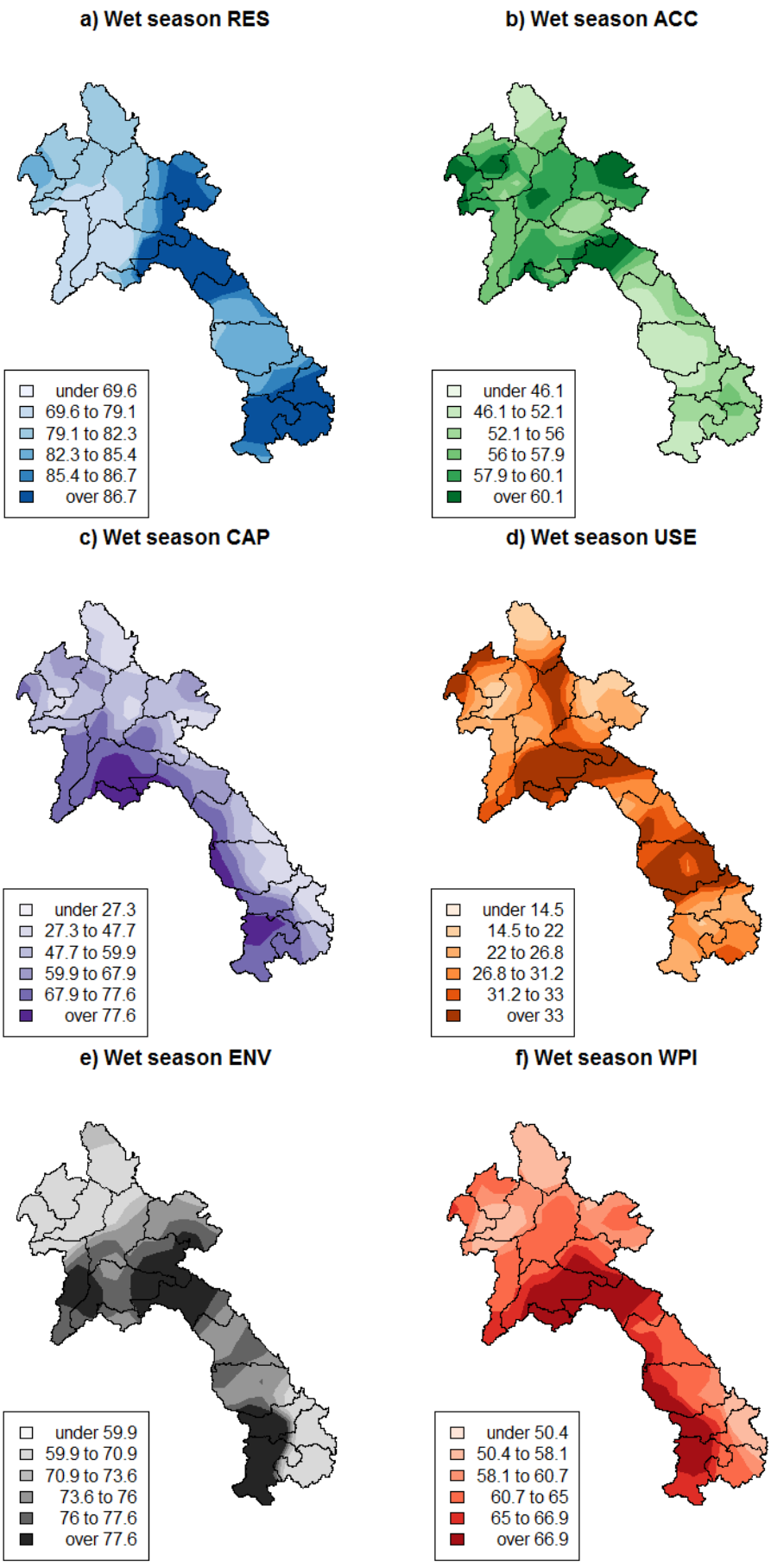


Figure 5.16. WPI components and WPI for wet season. The WPI score is calculated using PCA derived weighting scheme from the components and combined using a multiplicative function. Note that the colouring is relative to the component, not over the entire range 0-100.

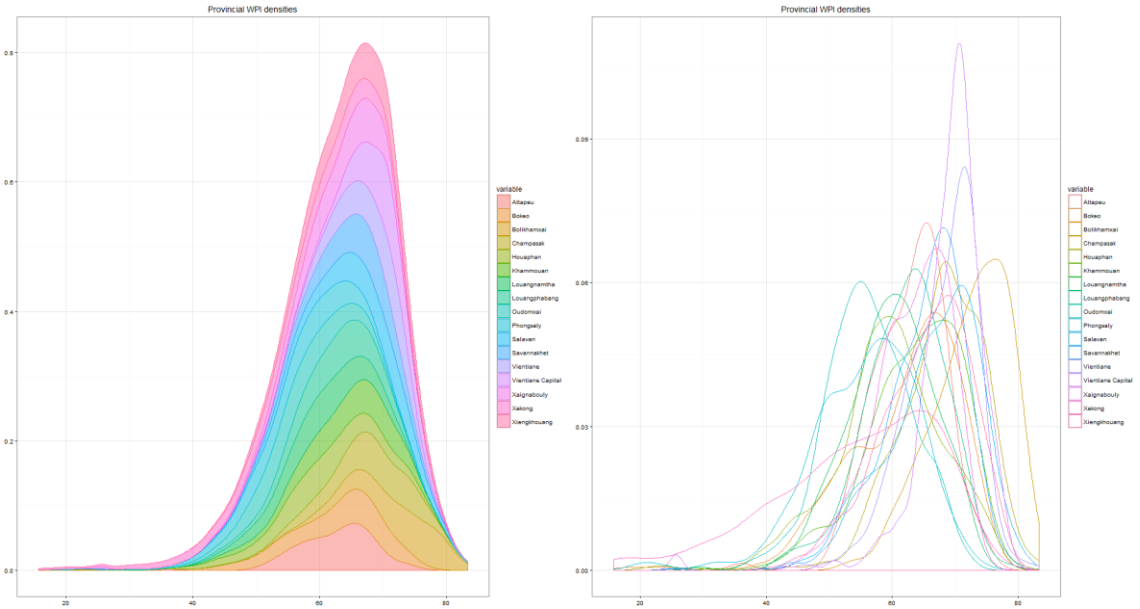


Figure 5.17. Stacked density and normal density plots for wet season WPI in each province.

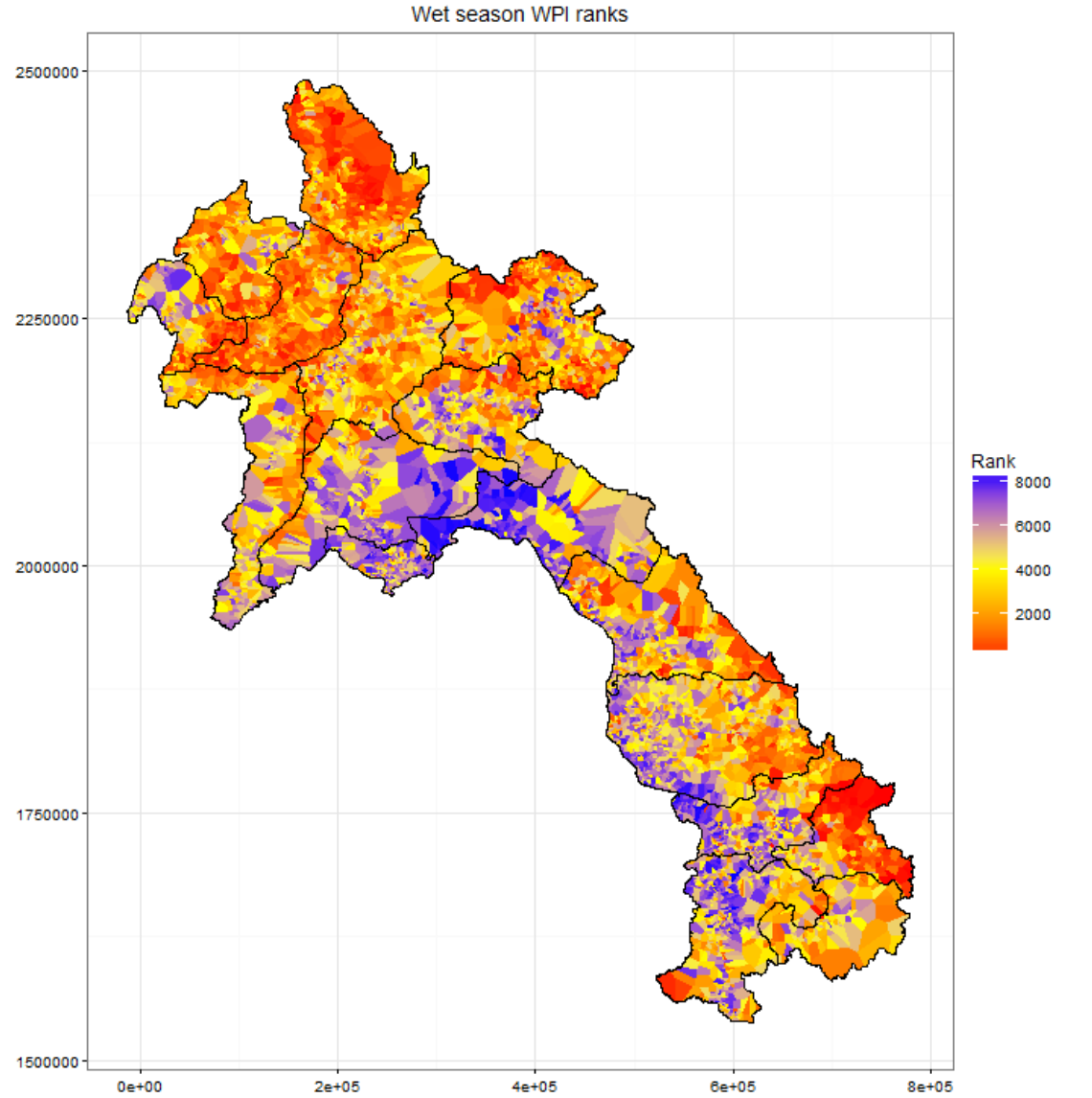


Figure 5.18. Wet season WPI ranks, ordered so that higher rank is given to the villages with higher WPI.

The wet season rank plot for villages in Figure 5.18 again confirms what was found for wet season above. However, as opposed to the dry season ranks (Figure 5.15), the poor regions are much more clustered. Phongsaly, Oudomxai, Xekong, Louangnamtha and Houaphan contain most of the poorest villages both visually and when inspecting the median provincial WPI rank. Bolikhamxai is the most water-rich province by a large margin in this statistic; the median WPI rank is 7529 out of 8215 villages. Vientiane Capital is trailing with median rank of 6351.

A pairwise t-test confirms that the mean WPI for the provinces are different, giving the final confirmation that water poverty indeed is behaving spatially varying phenomenon. This applies to the variables used to build the index, to all of its components and to the two seasonal indices.

5.3 Seasonal Water Poverty

The third section of Results chapter looks at the question whether there are significant inter-annual differences in water poverty by first examining weighting schemes that could be derived from a PCA. The two seasonal WPI's are then compared to determine the answer to the second research question: *"Are there distinct spatio-temporal differences in water poverty?"*

5.3.1 Weighting Schemes

Principal Component Analysis was used to derive an objective weighting scheme, as suggested by Jemmali and Matoussi (2013). PCA was applied to the wet and dry season components separately as well as using data from both seasons combined. The reason for this is that performing PCA on dry and wet season separately will not yield a weighting scheme that can be compared. Instead, they only measure the variability of the data *within* the season. The weighting schemes derived using all 5 Principal Components (PC) are shown in Table 5.6.

Table 5.6. Objective weights for the components derived using Principal Component Analysis.

Component	Dry	Wet	Both
RES	20%	12%	27%
ACC	16%	24%	7%
CAP	7%	14%	14%
USE	11%	16%	12%
ENV	46%	34%	40%

The single-season weighting schemes are close with minor differences. Resources component is assigned a higher weight in the dry season than in the wet season, suggesting a higher overall variability in the dry season (which is logical, looking at Figure 5.3 on page 36). Environment is the heaviest component for both, dry and wet seasons by a large margin. In the dry season, Access, Capacity and Use are given very low weights. In the wet season, importance of Environment and Resources are reduced in favour of the "hu-

man" components of WPI. Resources component (the physical availability of water) becomes is the *least* important component in the wet season. This is attributed to the fact that, for the low population, there is ample of water present even during the dry season. The picture changes when both seasons are used to derive the weighting scheme: Environment and Resources together weight nearly 70% of the overall weights. Access, which does not vary between seasons, is assigned a very low weight of only 7% and Capacity and Use 14% and 12%, respectively. Looking at all of these together, differences in water poverty in the dry season are due to environmental factors, and in the wet season due to the human factors.

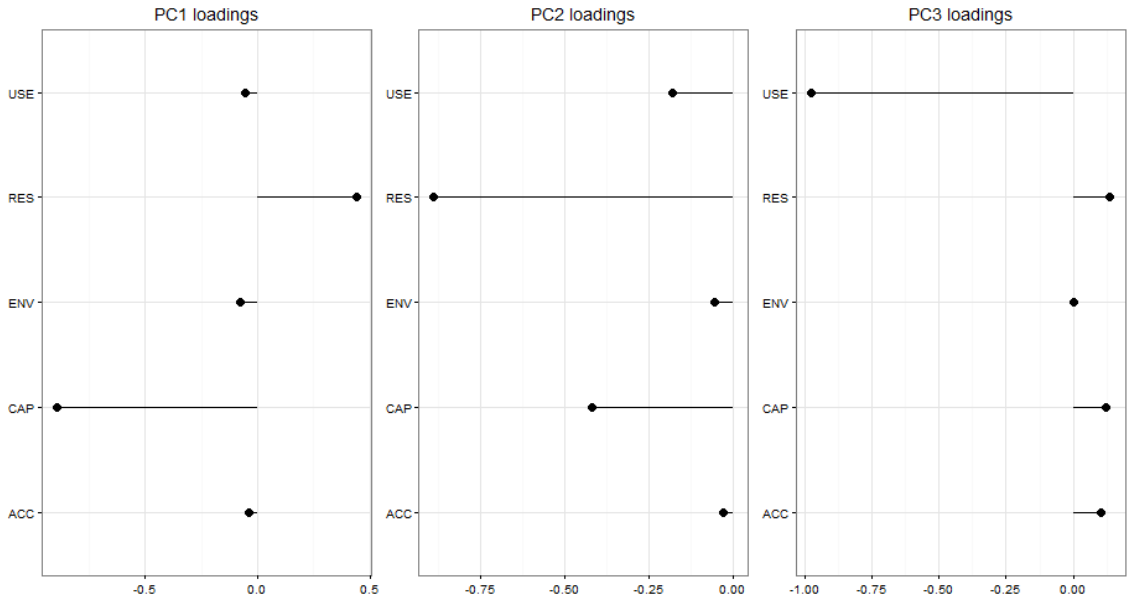


Figure 5.19. Loadings of the first three principal components from scheme using data from both seasons.

Figure 5.19 presents the loadings of the first three principal components from the scheme including both seasons. From the first PC we can see that the largest variability in the dataset is represented by Resources gaining a medium strength positive loading and Capacity given a very strong negative loading, representing areas with high resources availability and low capacity. In the second PC, Resources is assigned with a very strong negative loading in the second PC with Capacity again assigned a medium strength negative loading. As a contrast to the first one, the second PC represents poor areas with low resources, low capacity and with high number of people dependent on the resource. All of the loadings are negative, suggesting that the second PC could be characterized as the "poor PC". The third component represents nearly entirely Use component, which is given maximum negative loading.

WPI calculated with the objective schemes is presented in Figure 5.20. There is only a small difference between the single-season PCA weightings. This is largely due to the smaller weight of RES in the wet season weights and a higher weight in ACC, which is identical to both of the seasons. However, major differences are evident when both seasons are used together to determine the objective weights. The distribution of areas of high and low water poverty are similar between wet and dry season, regardless of weighting scheme. Areas near the Mekong and Thai border, in general, score high on WPI while the remote mountainous areas near Vietnamese border and in the north score

low. The distribution of index values has a wide peak at 8-10 Index points in favour of wet season with mean and median of approximately 8.8. There is a long tail on the higher-difference side.

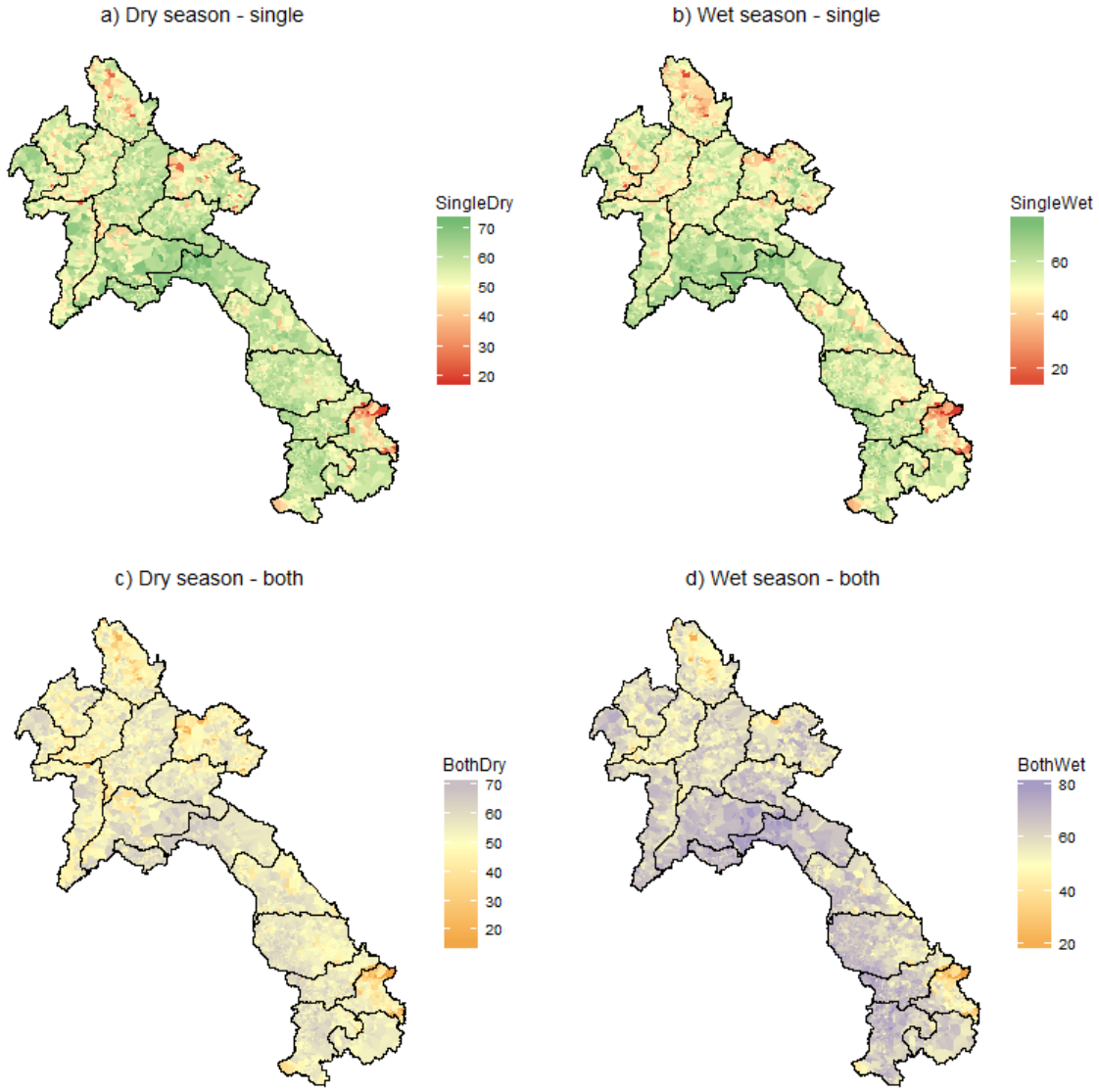


Figure 5.20. WPI calculated from the three objective weighting schemes: a) dry season WPI with weights derived from dry season only, b) wet season WPI from weights derived from wet season only, c) dry season WPI from weights derived using both seasons, and d) wet season WPI from weights derived using both seasons.

The difference between single-season and both-seasons weighting schemes is presented in Figure 5.21. During the dry season, both-season weighting gives lower WPI score to the villages than using a single-season scheme. The effect is opposite in the wet season where both-season scheme results in higher WPI values. Both of these apply to the entire country, however, the amount of difference varies location by location. This effect is due to the weighting differences shown in Table 5.6. To summarize, comparing the weighting schemes suggest that the relationship between components differ according to the season, providing the first evidence that there is a difference between season. This is also the first evidence towards the research question about the drivers of water poverty; environmental

conditions and resource availability being more important in the dry season and the human components in the wet season.

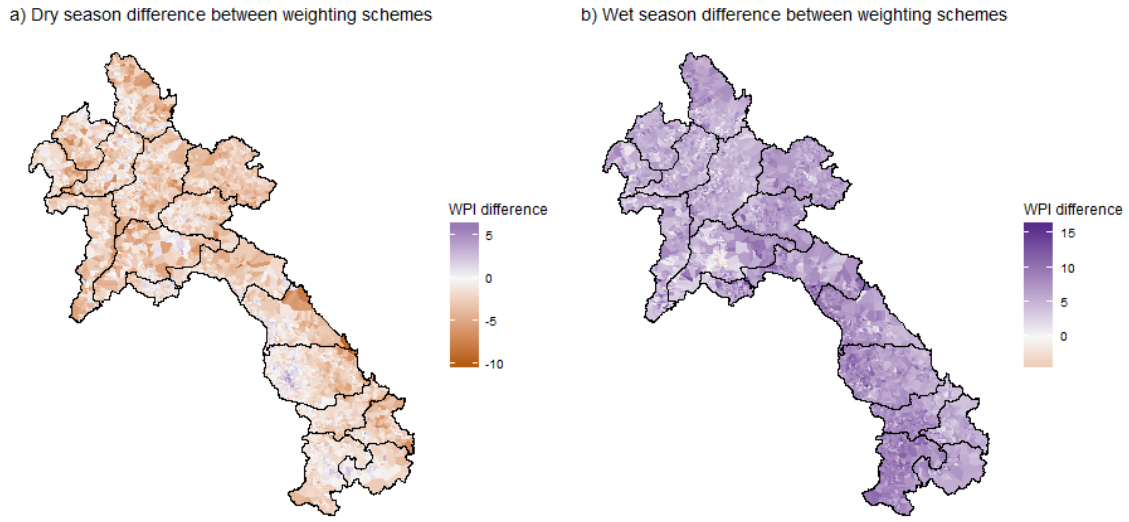


Figure 5.21. WPI difference between single-season and both-seasons weighting schemes. Calculated by subtracting single-season weighted WPI from the both-season weighted WPI.

5.3.2 Difference in Seasonal Index Scores

The degree in which dry and wet season WPI differs varies substantially across the country. The difference is presented in Figure 5.22. As a rule of thumb, areas where overall wet season WPI is high, the difference between seasons is also high. To put it in other words, low WPI score in the dry season generally coincide with low increase in WPI towards the wet season. What this means is that wet season does not substantially improve the water poverty situation in water poor areas, but instead, the difference between the better- and worse-off areas is growing with the arrival of rain. Figure 5.23a presents a dumbell plot which aggregates the seasonal differences in to provincial averages. The plot seems to confirm that, the four provinces at the bottom are the poorest in both seasons. In addition, the three provinces in the bottom, Xekong, Phongsaly and Oudomxai are the provinces in which WPI score increases the least in the wet season. The plot also shows that two provinces, Vientiane Province and Xayabouly, stand out as the increase in WPI in wet season is markedly higher than in the other provinces. This difference suggests that there is a more extreme seasonality in these provinces compared to all other provinces in Laos.

Figure 5.23. Mean provincial a) WPI and b) WPI rank for dry and wet seasons presents a dumbell plot from the perspective of relative ranks. The difference in the relative rank is not as dramatic as using index value, however, it provides insight in the season differences. 9 out of 17 provinces fall in their relative rank towards the wet season. Notable is that some of the poorest provinces, namely Phongsaly, Oudomxai and Louangnamtha drop in rank by a significant amount in the wet season. These provinces are all located in the northwestern corner of Laos. This provides additional evidence that, in the North, the inter-annual difference in precipitation and water resource availability is much more stable than in the south where Xekong is located.

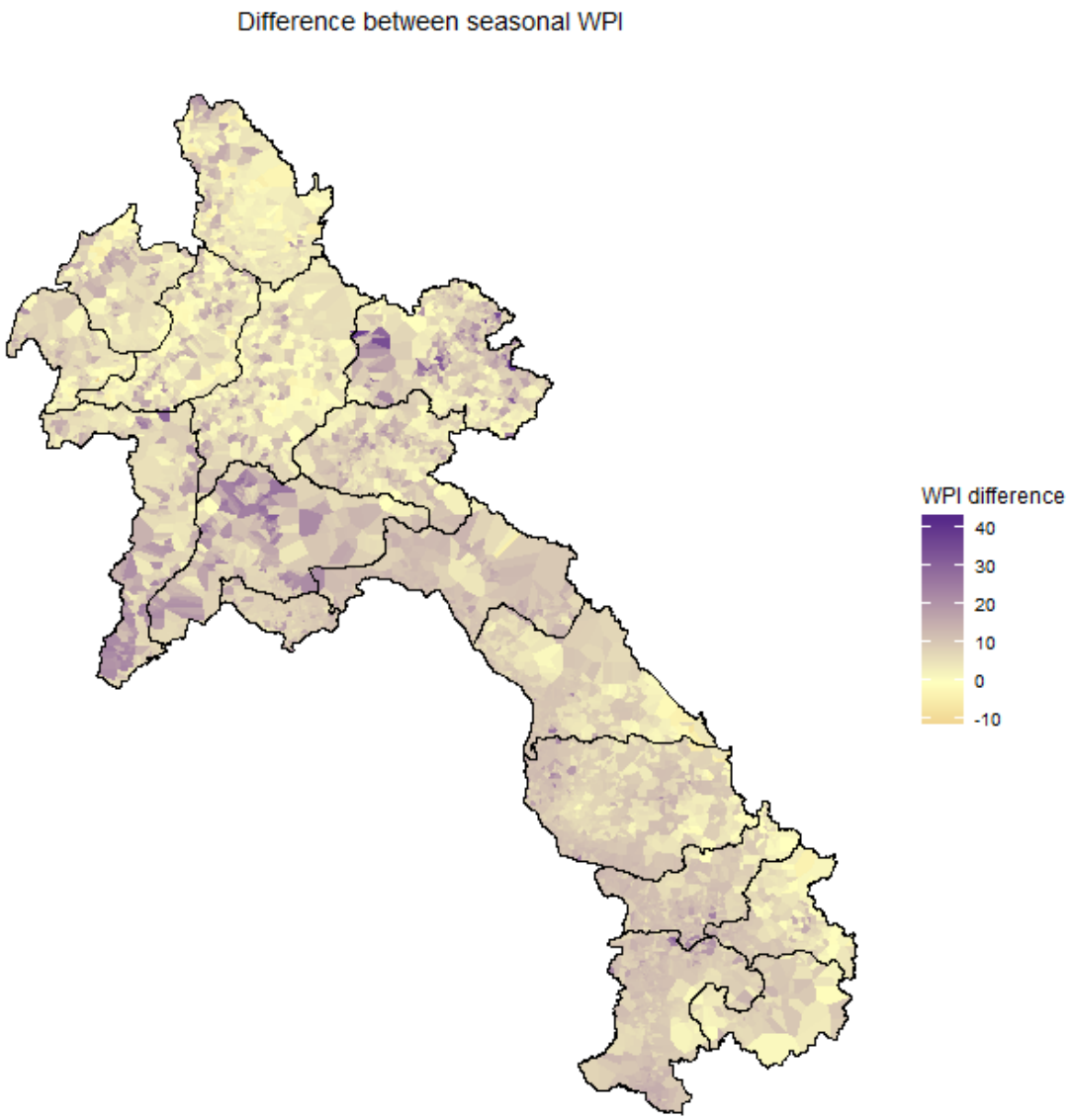


Figure 5.22. Difference between wet and dry season WPI. Calculated by subtracting dry season from wet season WPI (both seasons weighting).

The extreme increase in WPI in the provinces of Xayabouly and Vientiane (and Houaphan) is further confirmed by plotting the localized mean WPI for the entire country (see Figure 5.24a and Figure 5.24b). Areas within Xayabouly and Vientiane have low dry season WPI score while having a high wet season WPI. The same applies to Houaphan to a slightly lower degree. We can also visually confirm the low increase in WPI in the provinces of Xekong and Phongsaly. There are villages in Xekong (in the Southeast) which remain in the lowest class even during the wet season, however, the area is extremely remote with very little population. Phongsaly is located in the other extreme of the country, however the situation is very similar there as it similar to the Xekong. Southern part of Oudomxai near the bordering provinces of Xayabouly and Bokeo also stands out with high water poverty during both seasons.

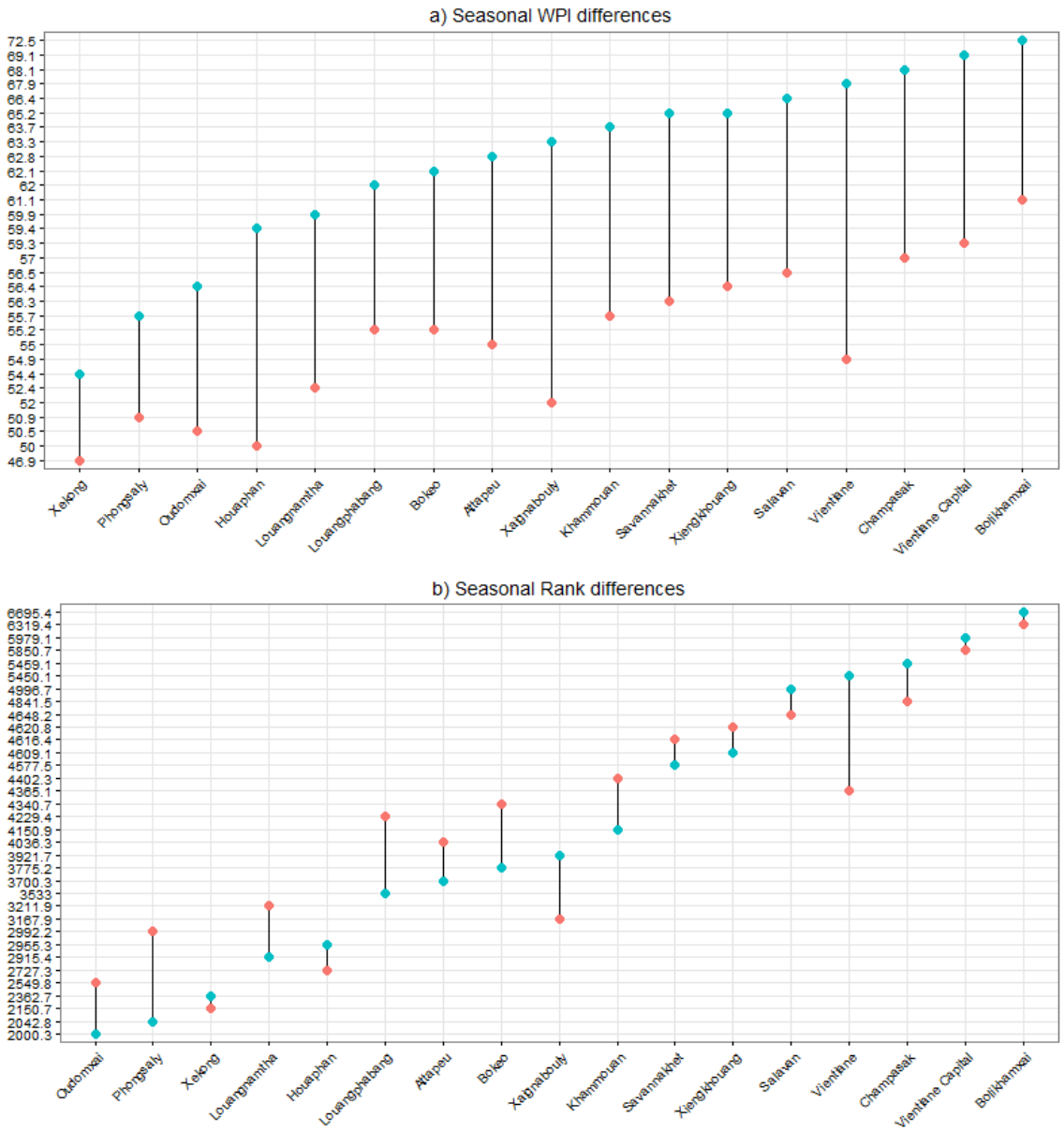


Figure 5.23. Mean provincial a) WPI and b) WPI rank for dry and wet seasons. The provinces are ordered according to the wet season.

In addition to the local mean, Figure 5.24 presents the local standard deviation for WPI. Some very interesting observations can be made from the maps. For instance, it becomes clear that in the areas that, consistently, where water poverty is low (WPI value high), standard deviation is low. In practise, although there are large differences between individual villages, the areas with high WPI are fairly uniform in their situation. However, areas with low WPI exhibit large variation in the local water poverty. This can be illustrated with the example of the two provinces with the largest difference between seasons; Vientiane Province and Xayabouly. In the dry season, overall WPI is low (between 52 and 55), and standard deviation between 7 and going in places above 13. This means that there are villages in the region ranging from extremely water poor to villages with very low water poverty rates. In Xekong and Phongsaly, the poorest of provinces in both seasons, variation is likewise extreme. Oudomxai shows up as an anomaly in this sense, since there is an area with low WPI and low standard deviation. For the wet season, the same rule of high WPI and low standard deviation applies, however there is somewhat larger

variation in it. Houaphan sticks out as having a high WPI in the wet season accompanied with a high variation.

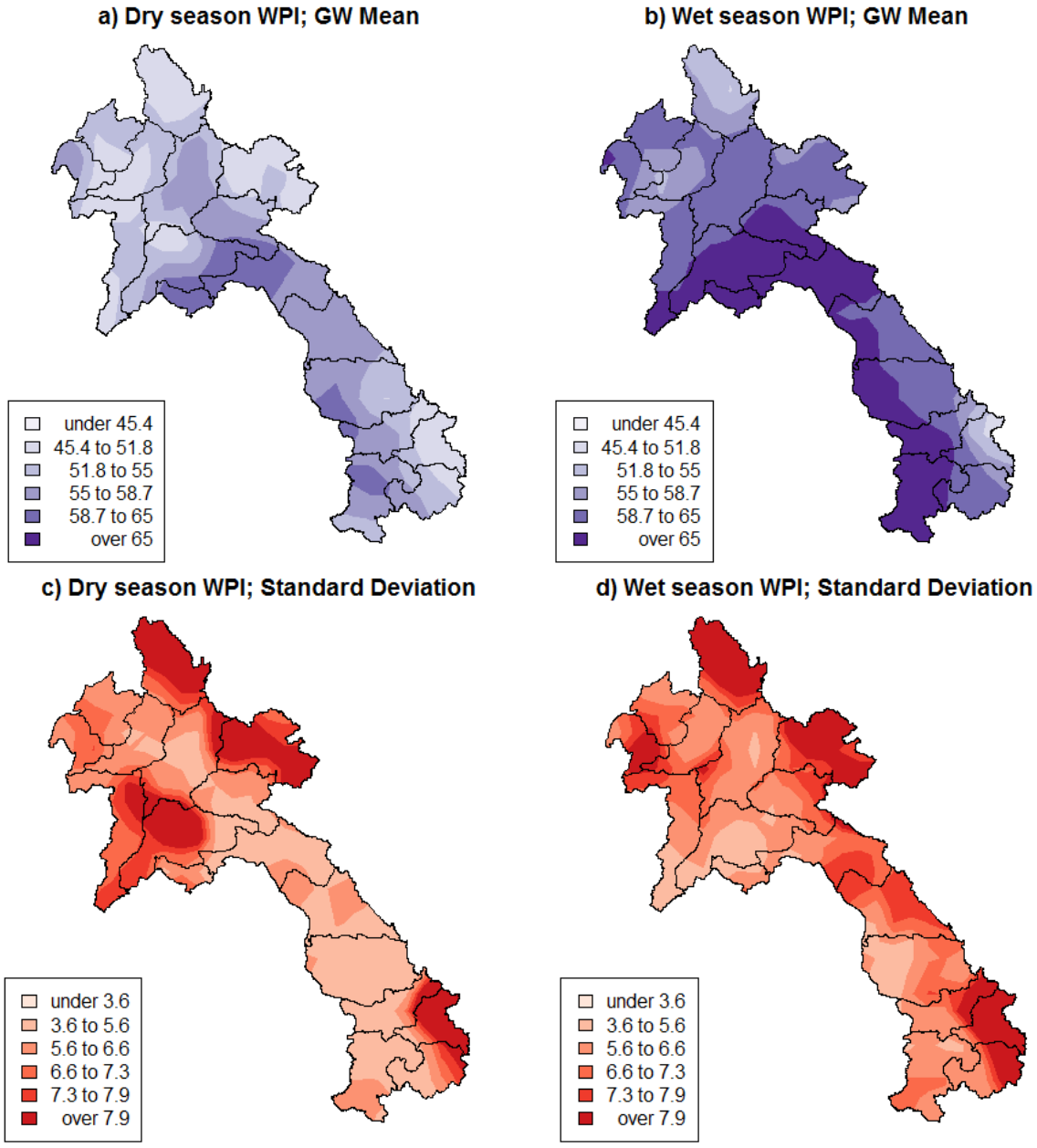


Figure 5.24. Local mean WPI and local standard deviation for dry and wet season. Bandwidth for the calculation is 400 nearest neighbours using Gaussian weighting scheme.

The relationship of village level dry and wet season WPI is plotted in a scatterplot in Figure 5.25. From the plot it is evident that majority of the villages in Laos fall in the better half of the plot range and that extremely poor villages are in the minority. Interestingly the plot tells us that there are a number of villages with dry season WPI higher than in the wet season. These are mostly located in the northwest with 71 villages in Oudomxai, 69 in Phongsaly, 41 in Louangphabang and 26 in Louangnamtha. This can be explained by a lower degree of variability in the Resources component and a high number of villages with only dry season road access. For Oudomxai this means that 15% of the villages in the province are better off, water poverty wise, in the dry season. For Phongsaly the same figure is 13%. This is potentially a significant find.

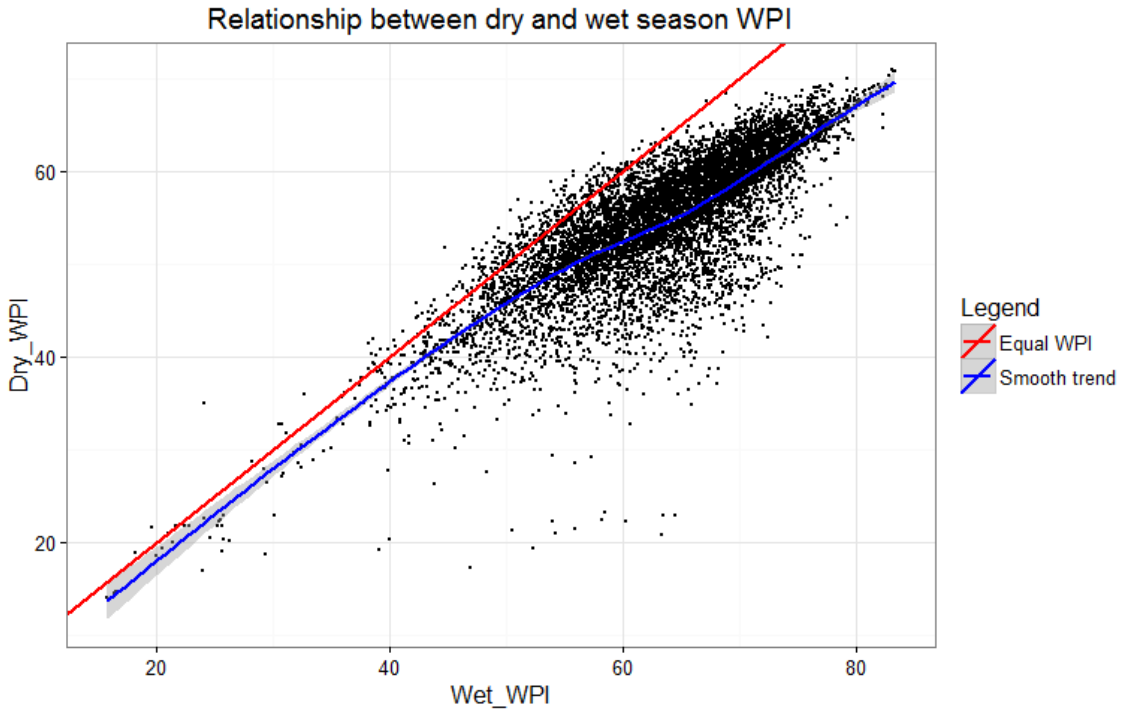


Figure 5.25. Scatterplot of the WPI between wet and dry seasons.

Local correlations show that, despite USE component has a low weight in computing WPI, it correlates highly with overall WPI score. The strength of correlation in both seasons are approximately 0.65. Weak correlation can also be found in Resources and Capacity in the dry season with coefficients of 0.36 and 0.31 respectively. In the wet season correlation between CAP and WPI strengthens into 0.51. Resources and WPI no longer correlate, but it is replaced by Environment with a coefficient of 0.31. None of the correlations between the components of WPI are not significant. All of the distributions are approximately normal. The correlations between components are summarized in Table 5.7.

Table 5.7. Mean local correlations between seasonal WPI and their corresponding components.

	DryRES	DryACC	DryCAP	DryUSE	DryENV
DryWPI	0.36	0.15	0.31	0.64	0.21
DryRES	*NA*	0.00	-0.02	-0.02	-0.08
DryACC	*NA*	*NA*	0.17	0.03	-0.03
DryCAP	*NA*	*NA*	*NA*	0.03	-0.10
DryUSE	*NA*	*NA*	*NA*	*NA*	0.02
	WetRES	**WetACC**	**WetCAP**	**WetUSE**	**WetENV**
WetWPI	0.16	0.21	0.51	0.67	0.31
WetRES	*NA*	0.05	-0.02	0.02	0.00
WetACC	*NA*	*NA*	0.16	0.01	-0.03
WetCAP	*NA*	*NA*	*NA*	0.08	-0.07
WetUSE	*NA*	*NA*	*NA*	*NA*	0.01

In addition, selected components' local correlations are plotted in Figure 5.26. These show interesting patterns; for the dry season (Figure 5.26a, Figure 5.26b and Figure 5.26c), Resources strongly correlates with WPI in the north and a few places in the south. Capacity strongly correlates with some of the poorest regions (Xekong, Phongsaly), but some not (Oudomxai, Louangnamtha). Weaker correlation is found in the more water rich regions. Use correlates very strongly in the south and in Houaphan Province, albeit the correlation is significant across the entire country. There is an interesting interplay between correlations of RES and USE in the north; a clear line is drawn southwards along the border of Houaphan and Louangprabang. Houaphan side correlates strongly with USE and the west side correlates with RES.

For the wet season the picture looks rather different. Capacity seems to correlate with the water-poor areas, while in the water-rich regions correlation is low. Use, on the other hand behaves in the opposite way, although some poor areas are also correlating. The last significant component, Environment, behaves in a more random way with clear regions of high and low correlations among both, water-poor and water-rich areas.

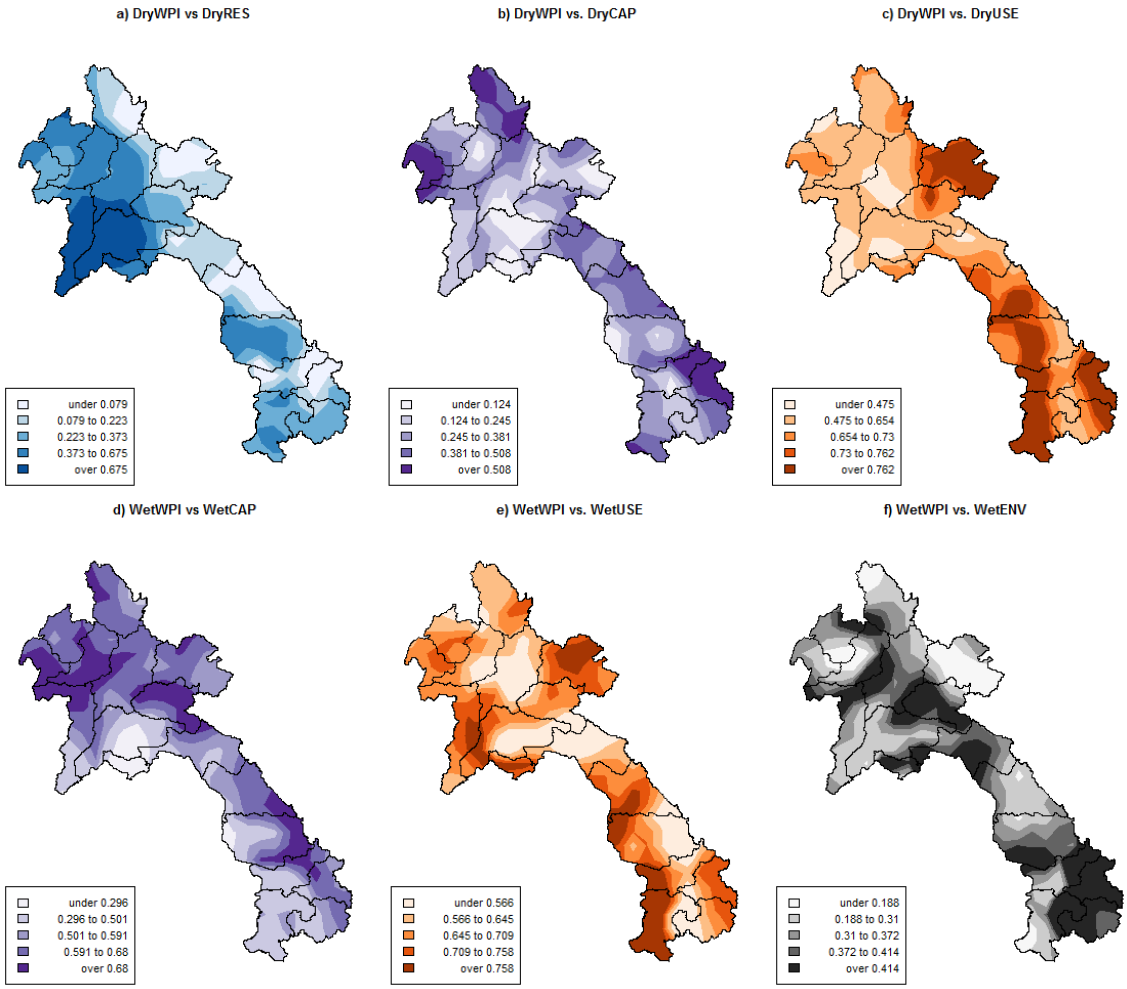


Figure 5.26. Local correlations for the components which (on average) significantly correlate with seasonal WPI's. Dry season WPI correlates with a) RES, b) CAP and c) USE. Wet season correlates with d) CAP, e) USE and f) ENV.

Pairwise t-test result in high confidence in the conclusion that there is a statistical difference between wet and dry season WPI. The investigations have provided ample evidence that there is a significant difference in dry and wet season water poverty, giving a confirming answer to the second research question: ”*Are there distinct spatio-temporal differences in water poverty?”* The investigation now proceeds to determining what are the causes that drive water poverty.

5.4 Mining the Causes of Water Poverty

The causes of water poverty are explored in this section through three SDM methods; namely cluster analysis, GWPCA and GWR. Reporting on these three methods are divided in their own sub-sections, starting with clustering and ending with the regression model.

5.4.1 Cluster Analysis

As a first step in cluster analysis, the determination of the number of clusters was attempted analytically. For this, R package ”NbClust” (Charrad, et al., 2014) provides a useful tool which includes 30 of tests on the dataset. The package *”proposes the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods.”* NbClust analyses were done using a random sample of 2000 villages (approximately 25% of all villages) due to the heavy computational requirements of the functions.

Running the diagnostic with method *k*-means for dry season WPI components with spatial variables (i.e., running the diagnostic with a collection of variables RES, ACC, CAP, USE, ENV and scaled Latitude and Longitude), yields in a proposition of the best number of clusters to two (when number of clusters the algorithm considered was from two to fifteen). The same diagnostic for wet season components and spatial variables proposes four clusters as the best option. Summary for the results is shown in Table 5.8.

Table 5.8. Number of clusters proposed by the 24 indices used by NbClust package.

Best number of clusters	Number of indices	
	Dry season	Wet season
2	11	6
3	7	6
4	0	7
7	0	1
8	3	0
9	1	1
11	0	1
14	1	0
15	1	1

The number of clusters was also experimented with and visually inspected to validate the analytical solution from NbClust package. It was found that aspatial clustering resulted

in dissatisfactory results where the villages are assigned to clusters in a seemingly random fashion (in a geographical sense), and therefore aspatial clustering is ruled out. However, the best number of clusters for the spatial clustering, visually, is challenging. In the dry season, four clusters seem to be most useful, as it splits the country along two axes; north-south and rich-poor. For wet season, the challenge is selecting between four and five clusters. Four clusters split the country along the same axes as in the dry season and the two seem very similar to each other. Five clusters add another class, ”the capital area”, which may be useful in determining the local causes of water poverty. However, four clusters are selected for further exploration due to the analytical solution suggesting four, which is also the number selected for dry season clustering. Figure shows the selected clustering schemes for dry and wet season. Maps for clustering with $k = 3$ to 6 are given in Appendix 4.

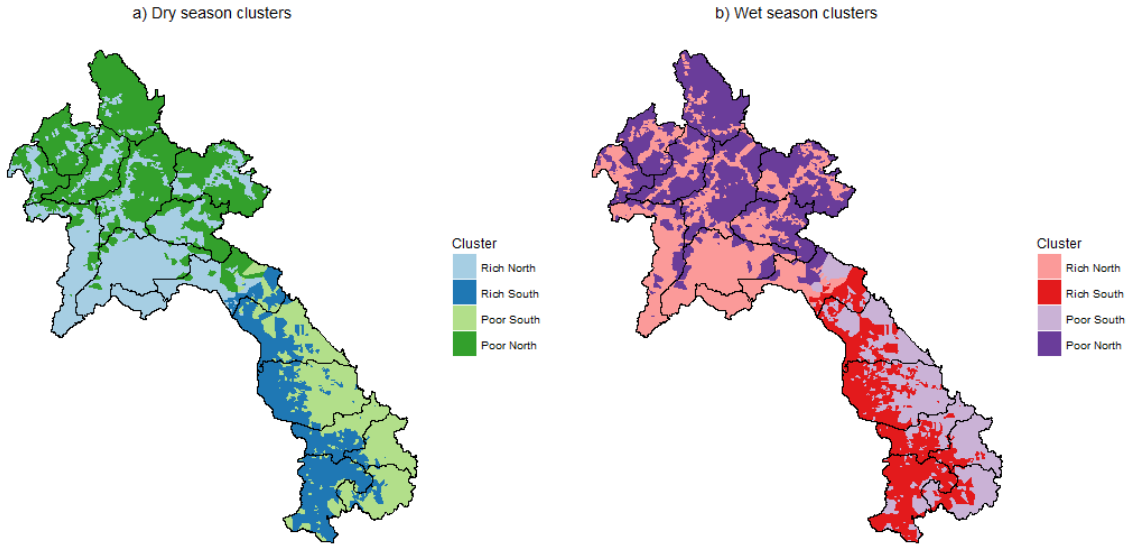


Figure 5.27. Selected spatial *k*-means clustering schemes for a) dry and b) wet season.

The clusters for dry and wet season are very similar to each other. Clusters were given indicative names according to the divisive axes, however it should be noted that not every village in these clusters are poor or rich. The rich cluster follows Mekong River for almost the entire length of it. In addition, major national roads can be clearly seen as long strings of rich villages in the Northern part of the country. Looking at the components (in Figure 5.28) for each of the clusters, we can see that largest difference between the rich and poor clusters is in Capacity component. The difference is approximately 20-30 index points in the dry season and 40-50 in the wet season. The rich clusters also score better than poor in the Environment component. For others, the picture is more varied; Poor North scores the best in Resources in the dry season and poorly in the wet season. Poor North also has the lowest score in Use and Capacity for both seasons.

The rich clusters can be divided mainly by differences in Environment and Access, where the North cluster scores considerably better in Access and slightly worse in Environment, and vice versa for the South. Interestingly, although the southern rich cluster gets a higher average score in the two most important components (RES and ENV) in the wet season, the overall WPI is still less than in the northern counterpart.

Looking at the variables inside the components, following characterizations can be made:

- Dry season water availability is the lowest in the Rich North.
- Southern clusters score significantly lower in the average length of dry period.
- Only Rich North score better than zero (apart from a number of outliers) in Toilet Type.
- All Capacity indicators are considerably higher in the rich clusters.
- Dry season irrigation is higher in the rich clusters.
- Soil degradation is significantly worse in Poor clusters.
- Human footprint is higher (lower score) in rich clusters.
- North clusters receive less rainfall than the southern ones in the wet season.
- Wet season road access is very low in poor clusters.

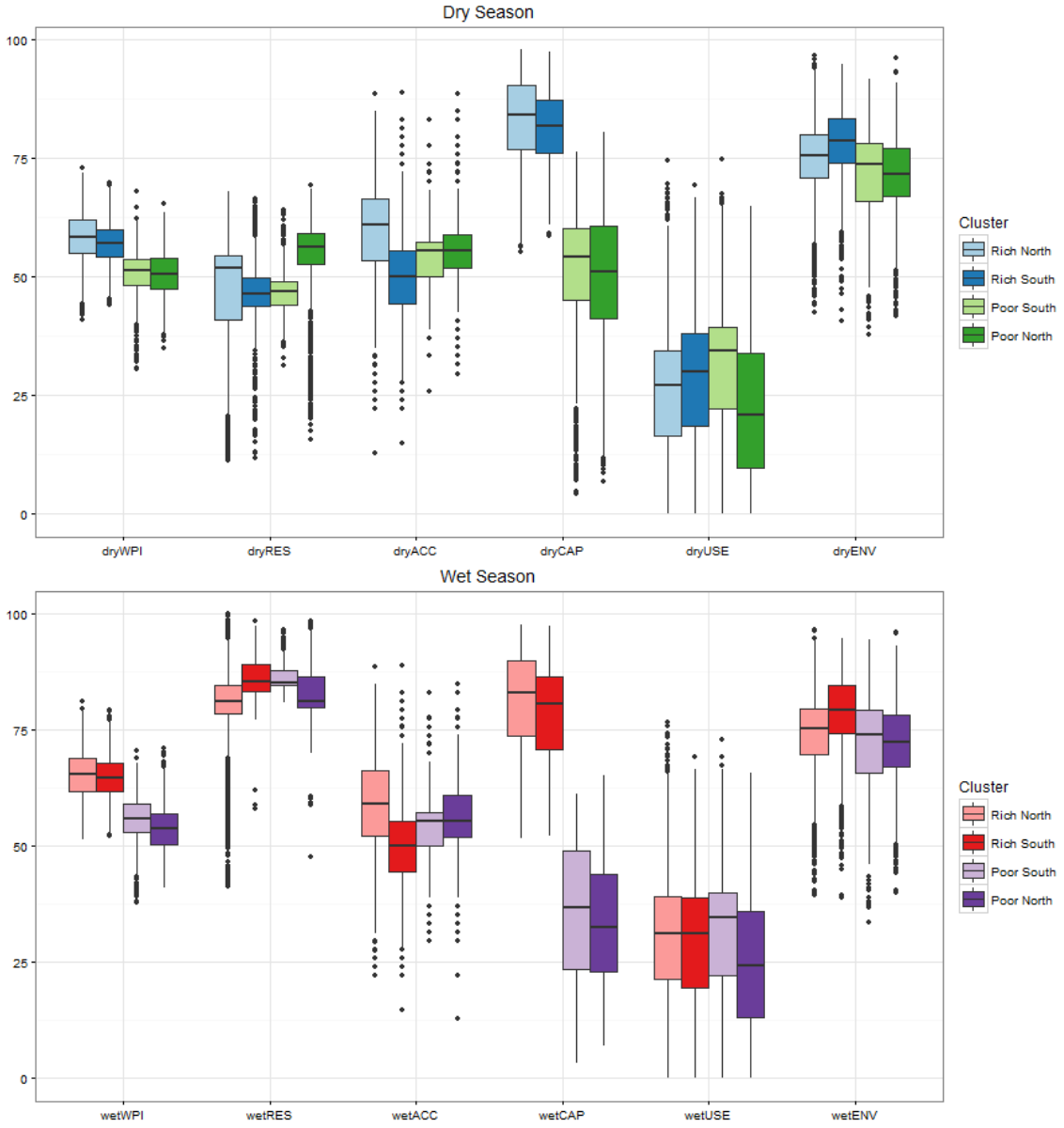


Figure 5.28. Boxplot of WPI components for each cluster in dry and wet season.

From the list above, road access is an interesting find. In the dry season, 25% of the Poor North villages do not have road access, and 16% in the Poor South. In the wet season

however, these figures grow to 86% in the North and 79% in the South. Road access increases mean WPI by 7-8 index points, which is approximately as nearly high improvement in the situation as is the difference between wet and dry seasons.

Table 5.9 provides the percentages of villages belonging to rich and poor clusters per province. The same provinces can be classified poor using clustering as could be identified in Section 5.2. Looking at the share of villages in the provinces, Phongsaly, Xekong, Oudomxai and Houaphan all have more than half of villages classified as poor in both seasons. Louangnamtha and Attapeu have more than half of villages in the poor cluster in the dry season. Of these, Phongsaly is identified as overwhelmingly poor, with only 16% and 25% of villages in rich clusters in dry and wet season, respectively. In the other end of the spectrum, Vientiane Capital clusters 100% and Vientiane Province with more than 90% share in the rich side. In addition, Bolikhamxai, Xayabouly and Champasak all are rich with more than 80% share.

Table 5.9. Share of villages assigned in rich clusters for each province.

Province	% of dry season rich cluster	% of wet season rich cluster	Difference
Phongsaly	16%	25%	9%
Xekong	31%	43%	12%
Oudomxai	38%	43%	5%
Louangnamtha	39%	55%	16%
Houaphan	42%	46%	4%
Attapeu	43%	55%	12%
Louangphabang	52%	52%	0%
Bokeo	53%	61%	8%
Savannakhet	62%	64%	3%
Xiengkhouang	66%	59%	-7%
Salavan	72%	72%	0%
Khammouan	74%	64%	-10%
Champasak	86%	90%	4%
Xayabouly	87%	87%	0%
Bolikhamxai	87%	86%	-1%
Vientiane	95%	91%	-4%
Vientiane Capital	100%	100%	0%

Khammouan and Xiengkhouang, Bolikhamxai and Vientiane Province are interesting provinces due to being the only provinces in which less villages are clustered to rich classes in the wet season than in the dry season.

In addition to clustering using the WPI values, clustering was also done for ranks. Analytical solution suggests that, for the dry season three clusters is optimal with 9 out of 23 indices suggesting this (an 8/23 suggesting two clusters). The remaining indices suggest high numbers of over 10 clusters. For the wet season, the highest number of indices suggest two clusters (8/23 indices). The rest is more spread out with indices suggesting from three to five clusters. However, visual inspection of the clusters does not support 2-3 cluster scheme. Instead, bearing in mind what has been found earlier in the exploration of WPI, in the dry season six clusters provide a cluster division that supports earlier findings. For wet season, five clusters seem to give clusters that approximately follows the

earlier findings. Lower numbers of clusters break the country in two or three uniform regions that seem to be driven by location only, not by components. The higher number of clusters was selected in order to break the area in smaller pieces to create a more detailed view on the causes of water poverty. The selected cluster schemes are presented in Figure 5.29.

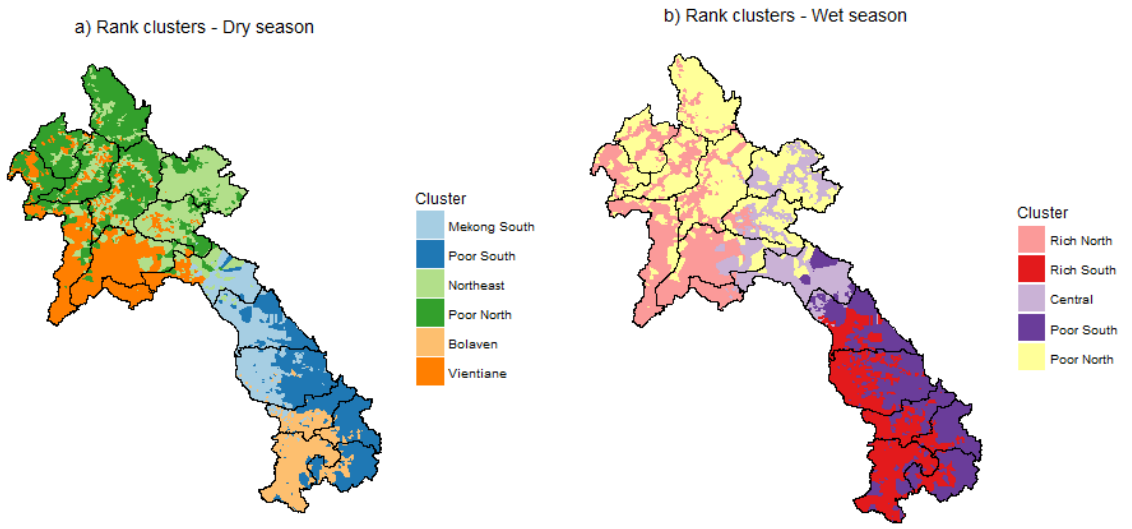


Figure 5.29. Selected Rank based clusters for a) dry and b) wet season.

Rank clustering provides little extra information over the clustering with the WPI values. The area around the Capital (Vientiane cluster in the dry season and Rich North cluster in the wet season) scores the highest in both seasons, followed by the clusters Bolaven plateau (dry season) and Rich South (wet season). In addition, it is clear that the cluster named Poor South (both seasons) are the most problematic of all the clusters. The major difference is in Capacity, as was found in the WPI clustering.

As a summary, cluster analysis suggests that the major difference between rich and poor areas comes from the Capacity component. Specifically, the water-rich and water-poor very strongly correlates with wet season road access.

5.4.2 Geographically Weighted Principal Component Analysis

GWPCA was performed using an adaptive bandwidth of 400 nearest neighbours, which is equal to the bandwidth used for GWSS and very close to the optimal bandwidth for GWR. Following the suggestions in Charlton et al (2010) and Demsar et al (2013), the ”winning variable”, meaning the variable with the highest loading, for the three first principal components for dry season were plotted in Figure 5.30. In the first PC, Capacity is the most loaded variable for most of the country, and mostly in the areas that were identified as poor in the cluster analysis. Rich areas are loaded highest with Use, with the exception of Bolikhamxai. Interestingly, provinces of Vientiane and Xayabouli are the only major areas loaded with Resources. The second PC is more varied. Majority of the areas that were loaded with Capacity in the first PC, are loaded with Use in the second PC. The areas determined as rich in the cluster analysis are mostly divided between Access and Capacity and some areas in the Poor North are loaded with Resources. In the third PC, Access dominates while the rest of the components are more or less evenly

loaded. Environment is the winning variable in some of the most water poor areas (Xekong, Phongsaly). Use is the most loaded in Bolikhamxai (the province with least water poverty), and in the northwestern provinces of Oudomxai and Louangnamtha. In addition, it is the most important variable in the southwestern tip of Xayabouly.

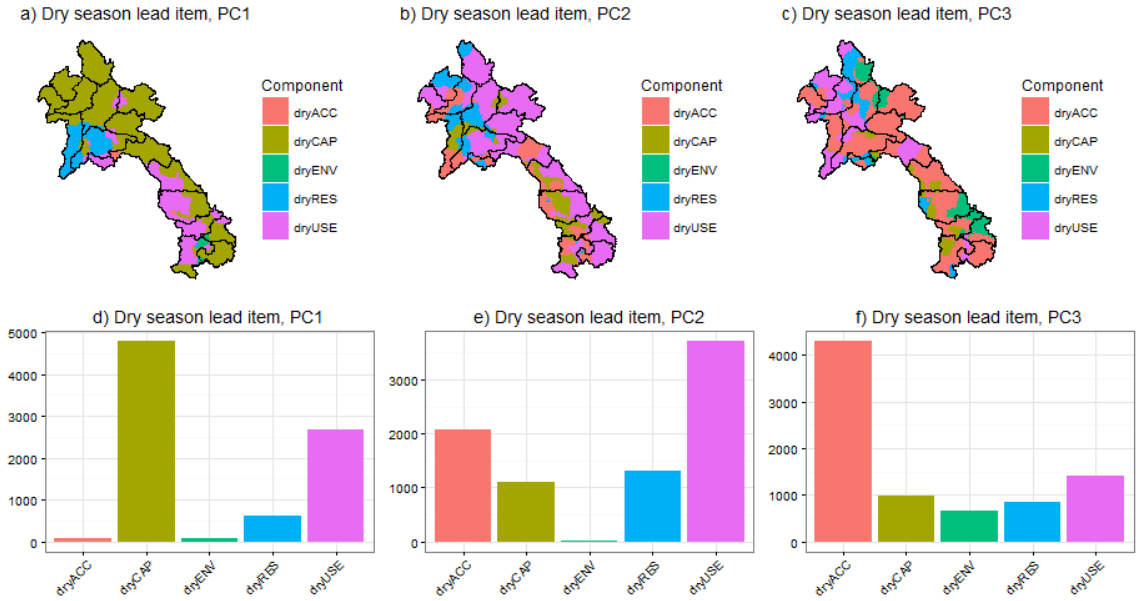


Figure 5.30. The highest loading ("winning") dry season WPI components for the first three Principal Components and bar plots of their frequencies.

The wet season winning variables are far easier to interpret. In each of the first three PC's there's a single component that accounts for nearly the entire country. In the first PC, Capacity is the most loaded component for almost the entire country. The only exceptions are a small number of villages in the south along the Mekong, loaded with Use. In addition, the southwestern tip of Xayabouly is loaded with Access, and a small area in Vientiane Province is loaded with Resources. The second PC is nearly entirely loaded with Use, however again, the same areas stand out as different. Capacity is the leading variable along the Mekong as well as in the tip of Xayabouly. In addition, Access is the most important in Vientiane Capital, and parts of Xayabouly and Oudomxai provinces. The third PC is likewise dominated by a single component, Use. Environment is the most loaded in Xekong and mountainous Saravane and Savannakhet as well as in the north in Phongsaly. Xayabouly again stands out being most loaded with Use and Resources. It appears that Xayabouly stands out from all the other provinces.

Since looking at the components provides information only on collections on variables, GWPCA was also performed on the individual variables that make out the components (these plots are provided in Appendix 5). The results confirm the findings from the GWPCA on components. In Xayabouly and Vientiane, dry season water availability is the most important driver in the first component, exactly as is seen in Figure 5.30a for Resources component. Elsewhere in northern Laos, road access and travel time to the provincial capital are found important. In the south, percentage of population depending on agri- or aquaculture is loaded the highest. In the second PC, dry availability remains important in the north west. Agri- and aquaculture dependence is again the highest loaded variable, only this time it is so throughout the country. The final third PC covered is very mixed, with five variables similarly loaded. These are agri- and aquaculture dependence, agricultural area per person, road access, soil degradation and toilet type.

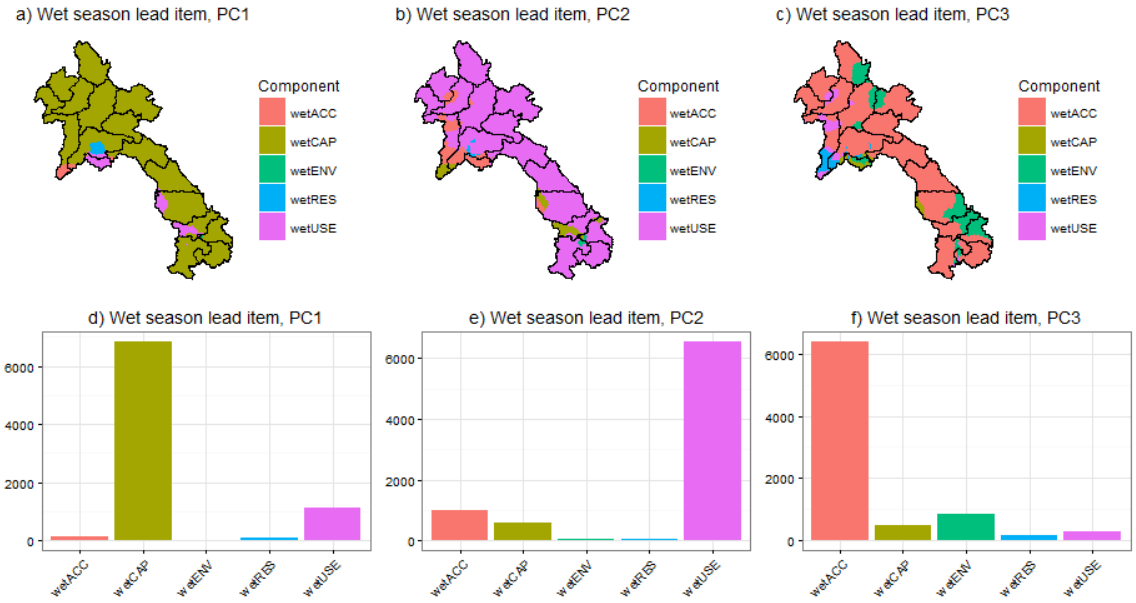


Figure 5.31. The highest loading ("winning") wet season WPI components for the first three Principal Components and bar plots of their frequencies.

Winning variables for wet season variable GWPCA also confirm the earlier findings. Wet season road access is the most important variable in the entire country apart from surroundings of Vientiane Capital. In the second PC, agri- and aquaculture dependence is similarly dominating the entire country, however there are small areas in which other variables become more important. The third PC is, like in dry season, much more mixed, however a few characterizations can be made. Again, agri- and aquaculture dependence is a major factor, along with road access. In addition, toilet type is important across the country.

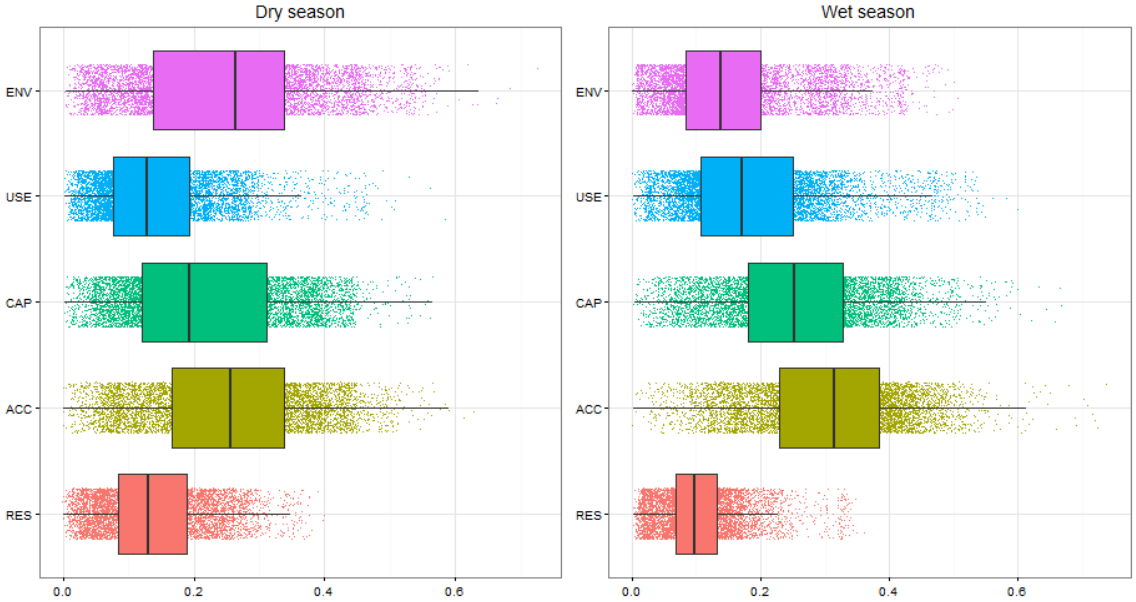


Figure 5.32. Boxplot of GWPCA derived weights for individual villages.

In addition to looking at the winning components and variables, localized weights were calculated for each village and are presented in the boxplot in Figure 5.32. As can be seen from the graph, the weights have very large ranges, which may be expected due to the

short bandwidth of 400 nearest neighbours. However, interestingly, a similar relationship exists between dry and wet season weights as is in the ones derived by global PCA (Table 5.6). As in the global counterpart, Resources and Environment get generally higher weights in dry season than in wet season, and the social components of Use, Capacity and Access all have higher weight in the wet season. However, a major difference between the global and local weights is that in the local ones, Environment is assigned much lower weights. In addition, the social components are assigned a considerably higher weight in the local version of the weights.

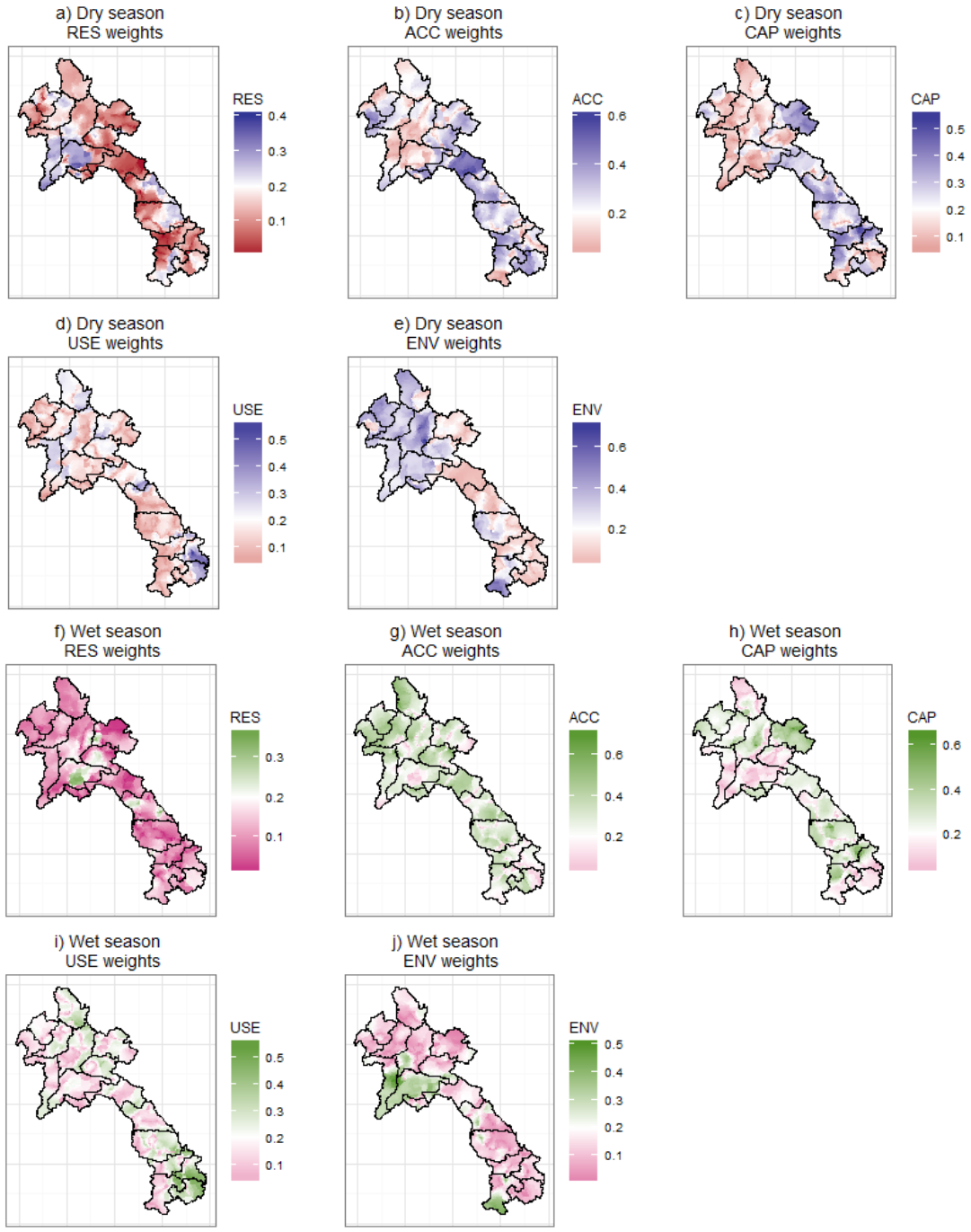


Figure 5.33. Spatial variation of the locally derived component weights. Subplots a-e show the dry season weighting scheme while f-j present the wet season weighting. Neutral colour signifies an equal weighting scheme where all components are weighted at 0.2.

Such a big difference in weights can be explained with Tobler's famous First Law of Geography: *"Everything is related to everything else, but near things are more related than distant things"* (Tobler, 1970). In other words, it is natural that the environmental components (Environment, Resources) are objectively more important in the scale of the entire country (there are big differences between regions) than in the local scale, where the environment is likely to be similar village-to-village. Hence, in the case of Laos, global and local weights emphasize different aspects of water poverty. Global scale emphasizes Resources and Environment, to which humans have only limited ability to control. Local scale on the other hand emphasizes the social components of Use, Capacity and Access, which we, as a society, can significantly affect.

The weights also show major spatial variation, especially in the dry season. The dry season "barrier" of correlations (Figure 5.26) can be identified in the local weights with Capacity getting lower weights in the northwest while Environment is assigned a higher weight in the same region. The stark border is visible in different components; RES and USE divided the regions in the correlation plots, while here, in the local weights, is seen in CAP and ENV. For the other dry season components, weights vary across the country creating several hot- and cold spots. In Resources, weights are generally low, however there is a major hotspot of higher weights in Xayabouly and southwestern Vientiane Province. This is the area that contains many of the villages under water scarcity, as was shown in Figure 5.4. Weights for Access component are also low in this area, as opposed to generally high scores in the rest of the country. In the wet season weights, the area that mostly point out in the local weights is formed by Xayabouly, Vientiane Province and Xiengkhouang. These areas contain low CAP and high ENV weights, which is opposite from the rest of the country (apart from smaller hot and cold regions).

Spatial autocorrelation of the weights was analysed and found to autocorrelate to an extreme degree, despite the seemingly random patterns visible in Figure 5.33. Moran's I, for every weight on both seasons is above 0.9, whereas conventional interpretation for strongly autocorrelated spatial phenomenon is Moran's I of above 0.3 (Getis, 2010).

WPI calculated with the local weights is shown in Figure 5.34 for both seasons. Curiously, the distribution of WPI in both seasons are almost identical with mean and median WPI between 53 and 55 for both of the seasons. In the wet season the distribution is slightly wider. Similar spatial pattern between high and low WPI can be found in the maps as is evident in the global case (correlation between global and local WPI is 0.81 for dry season and 0.82 for wet season). However, it should be noted that the maps calculated with global weights are not directly comparable to the maps in Figure 5.34 due to the global weights used to calculate WPI are derived using both seasons while the local ones utilize only a single-season. Therefore, the seasonal calculated WPI values are not directly comparable to each other either. To make the interpretation a little bit tougher still, the WPI values are *local*, meaning that one can only directly compare a village WPI value to the immediate neighbours only (the WPI is calculated with different weights for each village – Tobler's First Law of Geography applies here too).

Bearing in mind the limitations mentioned in the previous paragraph, the maps a) and b) in Figure 5.34 add evidence to some earlier findings. First, the maps confirm that the difference between water poor and water rich areas increase in the wet season. This can be seen when looking at the intensity of colours in the maps: in the dry season we can see more mid-range colours whereas in the wet season we see much more distinct borders

between the poor and the rich. Second, the same areas appear poor regardless of the weighting scheme (global vs. local).

In addition to the locally weighted WPI, Figure 5.34c and Figure 5.34d show the difference between single-season (globally derived) WPI and the locally weighted counterpart. WPI derived from local weights on average results in a lower index value than the locally weighted ones. For both seasons, the difference is slightly over 3 index points with a near-identical distribution. The spatial pattern reveal that the local weighting assigns a higher WPI in the water-rich areas where as in water-poor areas local weights results in a lower WPI.

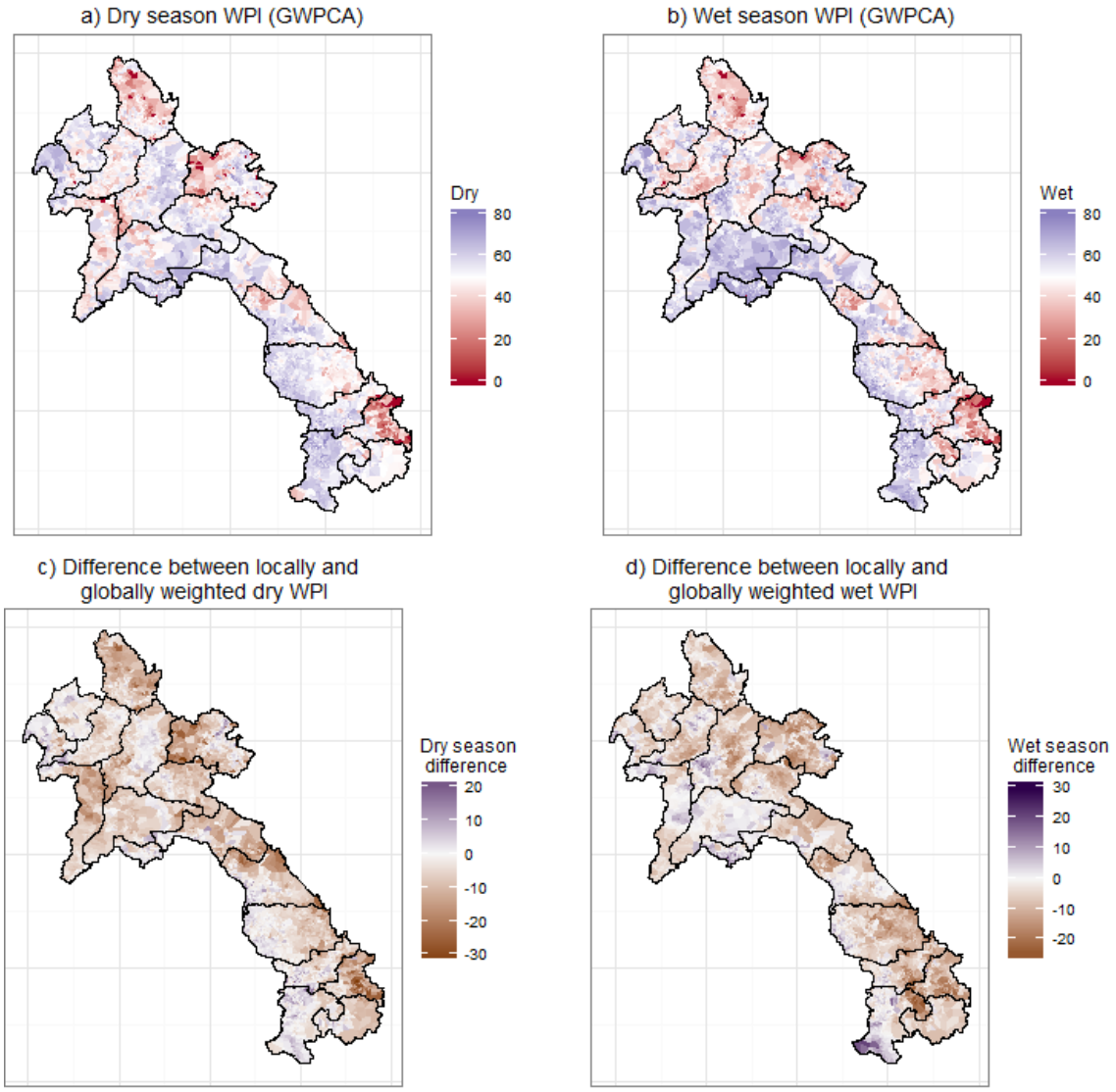


Figure 5.34. WPI calculated using locally derived weights for a) dry and b) wet season. The plots on the second row show the difference between WPI calculated from locally and globally weighted WPI for c) dry and d) wet season. The difference is calculated by subtracting globally weighted WPI from the locally weighted one.

In addition, the difference between the GWPCA-weighted seasonal WPI was plotted and shown in Figure 5.35. This map represents the relative water poverty between seasons. In other words, the areas with a negative value (WPI higher in dry season) are relatively

better off during the dry season than in the wet season when comparing to their immediate neighbours (400 nearest neighbours), and areas with a positive value are relatively better off in the wet season. However, this interpretation does not explain the large patches of uniform colouring. It was mentioned above that the distribution of WPI scores between the season are, in practise, identical. The colouring can therefore, despite the fact that weighting was derived locally, be used to compare the overall position of WPI ranks (in fact, it is nearly identical to a rank map, provided in Appendix 5) between seasons. In this interpretation, the blue areas have a much better rank in the wet season than in the dry season, and red ones vice versa. There are big regional differences on how this ranking relates to water poverty. Generally, the north (excluding Vientiane Capital, Vientiane Province and Xayabouly) get a higher rank in the dry season than in the wet season. The most likely reason for this is the lower Resource seasonal variation (physical availability of water increases less toward the wet season) and that majority of villages without road access are located in the north (thus, lowering the wet season CAP score compared to dry season CAP). Similar effect can be seen the southern mountains.

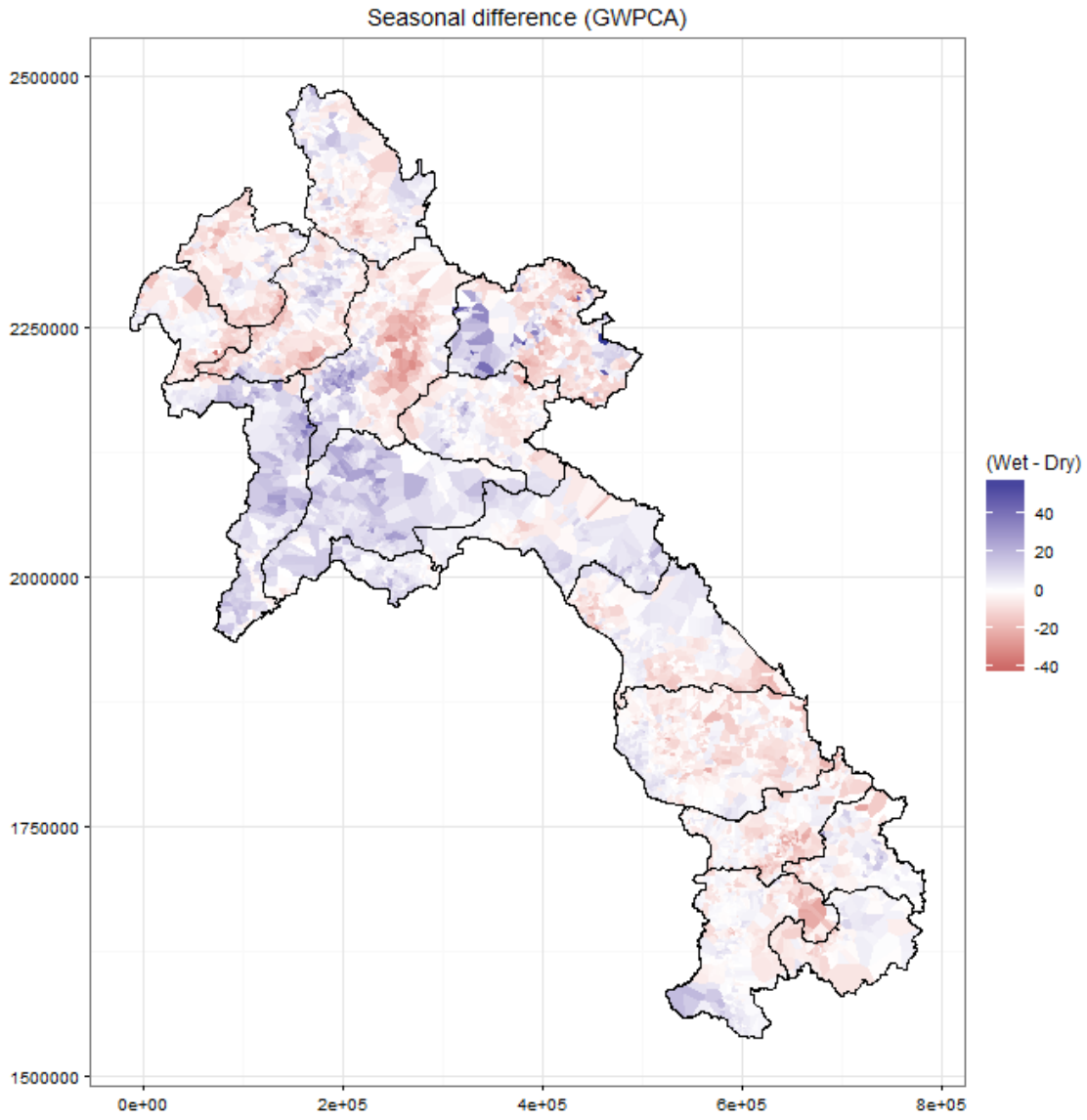


Figure 5.35. Dry season WPI subtracted from wet season WPI, both calculated using GWPCA weighting scheme.

Concluding the section, the winning components and variables and the local component weights show that there are important spatial differences in the components that explain majority of the variability in the attribute space. However, major similarities also exist: In GWPCA analysis for both seasons, Capacity is the highest loading in the first PC, Use in the second and Access in the third. The dominance of these variables is much higher in the wet season than in the dry season. This is interpreted here so that in the wet season, Resource or the Environment does not play a major role in the water poverty differences between areas. In the dry season, Resource availability plays a bigger role. The relative causes of water poverty, when looking at the variables, is much higher. Agri- and aquaculture dependence in the villages is the most important, or one of the three most important variables in the entire country across seasons. These findings provide some answers towards the third research question: *"What are the causes of water poverty in Laos? Do the causes differ across space and seasons?"* There are important spatial and seasonal differences between the causes. The most important causes, based on the GWPCA analysis are agri- and aquaculture dependence, road access and toilet type. In addition, water availability and travel time to province capital are important in the dry season. To a smaller degree

5.4.3 Geographically Weighted Regression

GWR analysis was done starting with selecting significant variables to be used in the model. The variable selection was done through a step-wise selection with an algorithm supplied in R's "GWmodel" package. The algorithm starts by calibrating a GWR model with a single independent variable. Cross Validation (CV) score is recorded for each variable, and the variable which produces the smallest CV is selected. Then, the algorithm introduces the remaining independent variables in addition to the already selected one. These steps are repeated until CV does not significantly improve. The input for the algorithm were the original values (i.e. not values processed to the scores) and consisted of all variables used to calculate WPI components, added with a number of additional, relevant variables (all of these are listed in Appendix 6). After the independent models were selected with the above algorithm, bandwidth was optimized using a function in GWmodel package. The optimized bandwidths for both dry and wet season are identical; 368 nearest villages for adaptive and 56.5 km for Euclidean bandwidth (however, only adaptive bandwidth was used due to edge effects and to ensure sufficient sample size for each regression point).

The geographically weighted model fare significantly better in modelling WPI for both of the seasons. R^2 is significantly better, as is residual sum of squares, CV and AICc. This gives us additional strong evidence for the first research question. A summary of the model goodness statistics is given in Table 5.10.

The variables chosen can explain a high proportion of the wet season water poverty, however, dry season phenomena are captured to a lower degree. In addition, the goodness of the model has a strong spatial variability (see Figure 5.36). While the selected models can explain a large share of WPI, in the dry season the country is split into an eastern part which can be explained to a high degree and (north)western part in which R^2 drops to less than 0.70. In addition, the southern Bolaven Plateau stands out as a region where the dry season model does not fare well. This means is that there is a spatial process in play which cannot be captured by the list of variables introduced to the model selection algorithm (the entire list is provided in Appendix 6). Curiously, this effect only applies to the dry

season; the wet season model performance is good throughout the country (with the exception of the area surrounding the capital city and to a lower degree, Bolaven Plateau). Another implication of this find is that water poverty is driven by different causes in different combinations of geographical and seasonal dimensions.

Table 5.10. Model goodness statistics for dry and wet season global and local models.

Season	Statistic	Global model	Local model
Dry	R^2	0.60	0.73
	RSS	175 384	120 744
	AICc	48 544	46 321
Wet	R^2	0.76	0.84
	RSS	139 752	93 563
	AICc	46 666	44 091

The same areal division can be seen in the local correlation maps between WPI and RES and USE in Figure 5.26. The northwestern part where local R^2 is exhibits a high correlation between WPI and Resources, while the eastern part in the division correlates with Use component. This seems to suggest that, in the northeast, there is an environmental phenomenon driving water poverty that is not represented by the variables introduced to the step-wise selection algorithm. This division can also be seen in the coefficient maps (provided in Appendix 6). A number of variables are significant mostly in the northwest, or the sign of coefficient estimates changes between the northwest and the rest of the country. Interesting examples are dry and wet season surface water availability; these are significant predictors only in the northwest and in the poor southeast. In addition, they change sings; in the capital area and southeast, higher dry season water availability is a negative predictor of dry season WPI, while wet season water availability is a positive one. In the northwest, these two variables behave in the opposite manner. Interesting here is that the water availability is not a significant predictor for water poverty in central Laos.

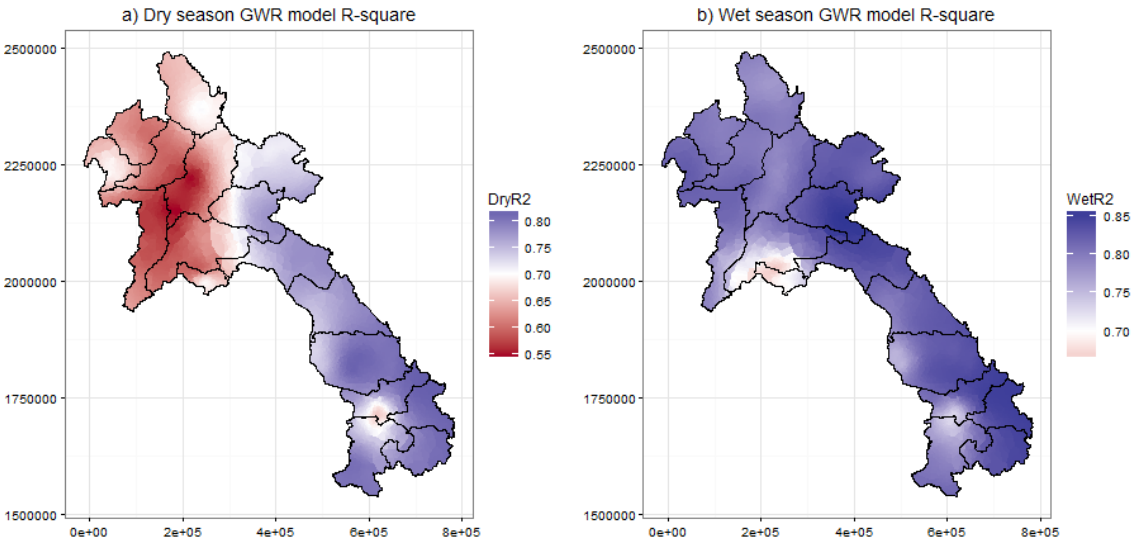


Figure 5.36. GWR model R2 for a) dry and b) wet seasons.

Table 5.11. (Step-wise) selected model variables in the order of selection and the p-values of Monte Carlo test for spatial homogeneity of the coefficients.

	Dry Season			Wet season		
	Variable	Name	p-value	Variable	Name	p-value
1	(Intercept)		0	(Intercept)		0
2	PopDepCrop	% of population depending on agriculture	0	RoadAcc	Road access	0
3	TimeProCap	Travel time to province capital	0	PopDepCrop	% of population depending on agriculture	0
4	SoilDeg	Soil degradation	0.01	LitPopSh	Literacy rate	0
5	RoadAcc	Road access	0	SoilDeg	Soil degradation	0.09
6	TotalPop	Total population	0	TimeDisCap	Travel time to district capital	0
7	TotAgrArea	Total agricultural area	0	TotAgrArea	Total agricultural area	0
8	LitPopSh	Literacy rate	0	TotalPop	Total population	0
9	ShDryIrr	% of agricultural area under dry season irrigation	0	IrrAreaSh	% of agricultural area under wet season irrigation	0
10	WetPrec	Wet season precipitation	0	WetPrec	Wet season precipitation	0
11	IncPov	Incidence of Poverty	0.04	IncPov	Incidence of Poverty	0
12	IrrAreaSh	% of agricultural area under wet season irrigation	0	PestsFre	Frequent disasters, type "Pests"	0.44
13	DryPrec	Dry season precipitation	0	LandslFre	Frequent disasters, type "Landslide"	0
14	PestsFre	Frequent pest disasters	0.34	DryPrec	Dry season precipitation	0
15	AvMaxDDay	Average length of longest yearly dry period	0	Flood	Frequent disasters, type "Flood"	0.12
16	HumanFP	Human footprint	0	DrinkRainW	Drinking water source: Rain	0.11
17	DrySurf	Dry season surface water availability	0	HumanFP	Human footprint	0
18	WetSurf	Wet season surface water availability	0	TotalCons	Total water consumption	0
19	DrinkRainW	Drinking water source: Rain	0.05	TimeProCap	Travel time to province capital	0
20	Drought	Drought disaster occurrence	0.38	AvMaxDDay	Average length of longest yearly dry period	0
21	DrinkOther	Drinking water source: Other	0.13	OtherDisFr	Frequent disasters, type "Other"	0.88
22	TotalCons	Total water consumption	0	TotIrrArea	Total agricultural area under irrigation	0
23	LandType	Land type	0.05	IrrReservo	Irrigation type: Reservoir	0
24	ConsIrr	Irrigation water consumption	0	DrinkSurfW	Drinking water source: Surface water	0.23
25	Irrigation	Irrigation present in the village	0.03	DrinkOther	Drinking water source: Other	0.05
26	Elevation	Median elevation in a 5kmx5km grid cell	0	HealthCent	Travel time to health center over 2 hours?	0.58
27	TimeDisCap	Travel time to district capital	0	ToiletType	Toilet type	0
28	IrrTempWei	Irrigation method: Temporary weir	0	PopDepAqua	% of population depending on aquaculture	0.85
29	TotIrrArea	Total agricultural area under irrigation	0	Slopeclass	Slope classified to 1-7 (1 low slopes, 7 extreme slopes) in a 5kmx5km grid cell	0
30	AmountRF3a	Amount of rain changed recently in the village?	0.38	DroughtFre	Frequent disasters, type "Drought"	0.63
31	OtherDisFr	Frequent disasters, type "Other"	0.8	Elevation	Median elevation in a 5kmx5km grid cell	0
32	DrinkSurfW	Drinking water source: Surface water	0.05	LandType	Land type	0
33	DisFreNS	Frequent disasters, type not specified	0.99	ConsIrr	Irrigation water consumption	0
34	DisFreNo	No frequent disasters occurring	0	StartRF3a	Onset of rainfall (wet season) changed recently?	0.67
35	FloodFre	Frequent disasters, type "Flood"	0.84	WatSupp	Type of the main drinking water supply	0
36	PopDepAqua	% of population depending on aquaculture	0.17	-	-	-
37	IrrGabion	Irrigation type: Gabion	0	-	-	-
38	Disaster	Disasters occurring?	0.18	-	-	-
39	IrrReservo	Irrigation type: Reservoir	0	-	-	-
40	Slopeclass	Slope classified to 1-7 (1 low slopes, 7 extreme slopes) in a 5kmx5km grid cell	0	-	-	-
41	ToiletType	Toilet type	0	-	-	-

The first variables picked in the step-wise selection process are uniform through the entire country: higher total population, lack of road access, higher soil degradation, high incidence of poverty and higher share of population depending on agriculture predict lower WPI. On the other hand, high agricultural area, high literacy rate, and high rate of dry season irrigation predict higher WPI. Majority of the other selected variables change sign in different locations and/or the variable is not statistically significant in some parts of the country.

Wet season coefficient estimates differ from the dry season estimates in three ways: first, less variables are picked by the step-wise algorithm. Second, there is no similar division visible in the estimates as there is in the dry season. Third, the variables change (see Table 5.11). Nine of the first 11 variables are the same, they only change order. Percentage of crops irrigated changed from dry season to wet season irrigation, and travel time to prov-

ince capital changed to district capital. This find suggests that, in the wet season, administrative (and other) services should be available in the district capital rather than the provincial capital. Interestingly, in the very southeast corner of the country, higher precipitation predicts lower WPI while the sign is positive for the rest of the country. This coincides with a very strong negative coefficient for slope class, which is a much weaker (or not significant at all) predictor in the rest of the country. Additionally, it is noteworthy that surface water availability is *not* a significant variable in the wet season model.

The very strong negative coefficient for slope class is the most probably cause for the extremely low prediction (and high residuals) of the water poorest villages in Xekong province. Maps of predicted values and residuals for both seasons are shown in Figure 5.37. The residuals are normally distributed, but both show a long, narrow tail on the negative side due to the inability of the models to predict WPI for the poorest villages (Xekong, Phongsaly). In addition, it is seen that dry season residuals are higher than the wet season ones. The general patterns of low and high WPI areas, however, follow well the distribution seen earlier in Figure 5.24. This further confirms the good model fit for wet season.

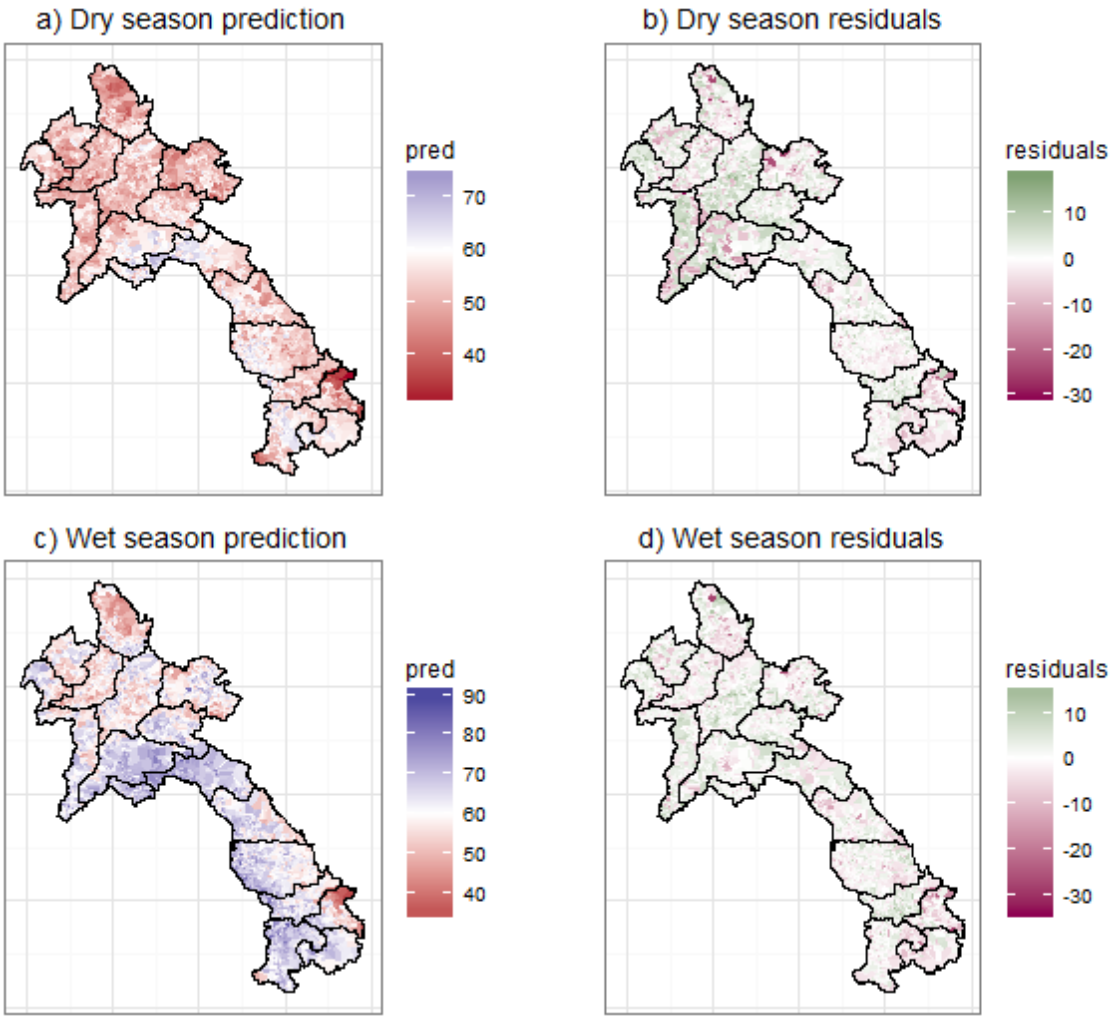


Figure 5.37. Dry season model prediction (a) and residuals (b) and wet season model prediction (c) and (d) residuals.

The spatial heterogeneity of coefficient estimates from both of the models were tested under Monte Carlo randomization. The p-values signifying the probability of the variable being a global one (showing no spatial differences) are shown in Table 5.11. Only a handful of variables are global with a high certainty. In the dry season these are frequent non-specified disasters (99%) and frequent flood disasters (84%). In the wet season: frequent disasters of the type 'Other' (88%), share of population depending on aquaculture (85%) and to a lower degree the change of onset of wet season (67%), frequent droughts (63%) and travel time to a health centre (58%). The majority of all variables exhibit spatial behaviour with a very high confidence, adding to the mounting evidence towards research questions 1 and 3a. The models were also tested with Leung's F1 and F2 tests (F3 test was running for more than 120 hours until the process crashed and therefore could not be finished). The tests measure whether the GWR model fare significantly better than an ordinary least squares regression using residual sum of squares (F1) and analysis of variance (F2). As a result, both tests signify, with an extremely high confidence (p-value of $2.2 * 10^{-16}$ on null hypothesis that there is no significant difference) that the local model is better at explaining WPI. This applies for both, the dry and wet season models.

Table 5.12. Variables with collinearity problems according to VIF and VDP diagnostics.

Season	Variable	VIF >10	VDP >0.5	Comment
Dry Season	Intercept		x	Problematic in some regions
	TimeProCap	x		Problematic mainly in Houaphan
	TimeDisCap	x		Problematic mainly in Houaphan
	WetPrec	x	x	Collinearity in Vientiane Capital, northwest (Bokeo/Louangnamtha) and small areas in Champasak, Khammouane and mountainous Saravane.
	DryPrec	x	x	Collinearity in Bokeo, Louangnamtha, Vientiane Capital and Houaphan
	AvMaxDDay	x	x	Collinear only in Champasak
	DrySurf	x	x	Very high VIF, medium in VDP
	WetSurf	x	x	Very high VIF, medium in VDP
	TotalCons	x	x	Very high VIF, small areas in VDP
	ConsIrr	x	x	Very high VIF, small areas in VDP
Wet Season	TimeDisCap	x		Problematic mainly in Houaphan
	WetPrec	x	x	Collinearity in Vientiane Capital, northwest (Bokeo/Louangnamtha) and small areas in Champasak, Khammouane and mountainous Saravane.
	DryPrec	x	x	Collinearity in Bokeo, Louangnamtha, Vientiane Capital and Houaphan
	TotalCons	x	x	Very high VIF, medium in VDP
	ConsIrr	x	x	Very high VIF, medium in VDP
	DrySurf	x		Very high VIF
	WetSurf	x		Very high VIF
	AvMaxDDay		x	VDP high in northwest, northeast, mountainous south and Khammouane Province

In addition, local collinearity was analysed using Variance Inflation Factor (VIF), Variance Decomposition Proportion (VDP) and Condition Number. The analysis reveals that there is significant collinearity present in some variables (or variable pairs). The collinearity problems are summarized in Table 5.12. Condition number is very high all across the country, suggesting that the models may be unstable. This is most likely due to extreme collinearity between the two surface water availability and the two water consumption variables. Dry and wet season GWR model without some of the collinear variables (dry/wet precipitation, irrigation water consumption, dry/wet surface water availability) was run resulting in very small changes in R^2, which suggests that these variables were not relevant in the model. In addition, the collinearity diagnostics suggest that care must be taken when drawing conclusions from coefficient estimates especially in Houaphan, the northwest corner where Bokeo and Louangnamtha are located, and Khammouane Province.

Summarizing the GWR analysis, a number of causes for water poverty has been identified. The most important predictors for WPI are nearly identical between dry and wet season: Higher share of population depending on aquaculture, travel time to district or province capital, worse soil degradation, lack of road access, higher total village population and higher incidence of poverty all predict lower water poorness across the entire country. Positive variables are total agricultural area, high literacy rate, high share of irrigated crops and high rainfall (except in the very south where higher rainfall predicts lower WPI). These are mutual factors for both of the seasons. The remainder of variables slightly differ: Dry season model contains more environmental variables than the wet season model. This is interpreted as further evidence of earlier conclusions; water poverty is driven by humanistic drivers in the wet season while in the dry season actual water availability is a meaningful factor. In addition to the variables determining WPI, the research question of whether the causes change according to location and season has been addressed. GWR analysis found evidence that there is a significant difference between the variables explaining seasonal WPI, and it was found that the coefficient estimates for these variables also vary to a high degree. GWR analysis therefore supports a confirming answer to research question 3a.

6 Discussion

Water poverty *per se* has not been studied in Laos in this detail – in fact the only (to the authors knowledge) specifically water poverty-related study that involves Lao PDR is an international comparison by Lawrence et al (2002). The results of this study is on similar lines, with average dry season WPI being comparable to the study by Lawrence et al, with a two notable differences. In this study Access scores were twice as high as in the international study and Use is scored several times higher in the international comparison. This is attributed to the variables used; the components in these two studies measure different things.

However, despite there is not a large number of studies to compare the results to, some indications can be used. Poverty and water are well known to have deep rooted links (Sullivan, 2002; International Fund for Agricultural Development, 2014; Ward & Kaczan, 2014), and this study found that WPI is low (meaning that water poverty is high) in the mountainous areas – where IFAD (2014) places the poorest population. This study does not include recommendations on how to alleviate water poverty mainly because the one of the main data sources (Population Census) is already more than 10 years old, and it is known that significant improvement has already occurred since the main data sets were collected in 2005 and 2011. Based on the results it seems that measures taken to tackle poverty in general are the exact measures that possibly would address water poverty issues – it was found that the human components of Access, Capacity and Use are the most problematic ones.

Some significant difficulties were also encountered while exploring water poverty in Laos. This mainly attributed to the high computational requirements of many of the methods employed. Despite running analyses on a relatively powerful PC, some tests and analyses ran for several days or crashing with problems of available memory. These problems mostly manifested in the GWR analyses – Leung's F3 tests could not be finished due to R crashing after 5 days (more than 120 hours) in on the test. In addition, mixed GWR with a few stationary variables crashed due to memory problems, as did robust GWR for the dry season model. Robust model for the wet season, however, could be finished probably due to fewer significant variables (the result of robust wet season model was worse than basic GWR by R^2, AICc and RSS). This may be due to the choice of using R for the analyses – it is known that R may suffer from memory problems when running analyses on large databases. Using alternative environments, e.g. Python, may have had resolved this issue, however it could not be done under the time limit for this study.

The third research question about the causes of water poverty is answered thoroughly in this study, however, the causes found in this study are not explicitly *proven* in this thesis. More precisely, it has been shown that the human components are the most important. Links between poverty (e.g. Sullivan (2002), Ward and Kaczan (2014) among many others) economic growth and jobs (UNESCO, 2016) and water has been shown in many studies. In practise, poverty as a multidimensional phenomenon is directly linked to all of the human components – Access, Capacity and Use. Despite this general knowledge, the reader should note that the causes presented are actually correlations that have been found by different techniques and assumed as the causes in the light of knowledge from literature.

The following two sub-sections deal with the weaknesses and strengths identified with the chosen methodology.

6.1 Weaknesses

The major potential weakness with this study relates to the tool of assessment – WPI – and how it is employed here. WPI is a relative index – traditionally (most of) the chosen variables are standardized through minimum and maximum of the sample used. This makes it very difficult to compare results with other studies – unless the variables and their ranges are comparable. This can be addressed e.g. by setting standard upper and lower limits for variables, as was done in this study for some variables. However, not all variables were standardized in this way. In addition, the variables used for the components are slightly different in this study than what are normally used to calculate WPI. I used Access to measure presence of water infrastructure – normally it includes variables that measure penetration of safe water and sanitation in the communities. In addition, Use is generally standardized to the available water resource – I used it without this standardization.

When it comes to research on WPI itself, Fenwick (2010) concluded in her doctoral dissertation that in addition to scale (Sullivan, et al., 2006), entirely different geographical areas may require different indicators to capture the water poverty situation. In the course of this research, it was found that the seasonal water poverty is very different in a same way as a rural and urban area are different to each other. It should be considered whether selecting different indicators for dry and wet season should be done when applying WPI in areas with stark seasonal differences.

Another major issue when developing WPI for this particular study was data. Unfortunately, many aspects of the environment could not be taken into account in a satisfactory matter and therefore they were dropped out. Water quality data is available only in a handful of data points. To use them meaningfully, they would have needed to be aggregated to very large areas which would have not made sense, as the aim was to calculate village-level WPI. In addition, no reliable data about existing dams were found to have them included in the water resource modelling. However, water scarcity occurs in a limited area in Xayabouly and Vientiane Provinces, and the main impact on component scores would have occurred here – if at all. Final major disadvantage of the variable choices is in using road access alone to measure the access to the villages. In reality, rivers are often used as pathways in areas which road network does not cover – this dimension of transport is not included in this study. River access could have a large impact in the Capacity component scores.

Another weakness is that, despite the dataset having a high coverage of all villages in Laos, there are areas where no villages are included. The reason of their exclusion was that they were not represented in either Population or Agricultural Censuses. Had there been data, some province level conclusions would have likely been different. Primarily this would have influenced Bolikhamxai, which was found to be the least water-poor province of all. However, all of the mountainous villages in this province were dropped from the final dataset – had they been present, the final numbers would have dropped considerably.

Finally, Fenwick (2010), along with a number of other researches, argue that, at community level, the local population should be included in the selection of variables relevant to their perception of water poverty. This unfortunately could not be done for this thesis. The work is entirely depending on datasets that are openly available, which however were not collected specifically for the needs of water poverty research.

6.2 Strengths

In addition to weaknesses, this study does have some distinct strengths. To my knowledge, this is the first study that explicitly applied WPI to dry and wet seasons. Their differences are often mentioned in literature, but there is a lack of studies which compare the impacts of season to water-related socio-economic indicators. This study found that there are significant differences present in other components than just actual physical water availability.

WPI is a composite index and as such, suffers from the disadvantages of composite indices. One of the drawbacks is that often the components strongly correlate with themselves, reducing their usefulness. However, in this study none of the components show strong correlation, and only one (Access and Capacity) correlate with a medium strength. This is advantageous since it means that each component measures a different water poverty linked phenomenon.

In addition, some weaknesses described by Fenwick (2010) in her doctoral dissertation are evaded in this study. It has been found in many studies that components correlate despite carefully selecting indicators, and often it is argued whether a component is a useful addition or not. As mentioned, the components do not correlate to a high degree, and this is therefore interpreted as one of the strengths of this study. In addition, a problem often raised in literature is weighting (e.g. Molle and Mollinga (2003) and Garriga and Foguet (2010)) being a political choice. It is countered here by using statistical methods for ”objective” weighting. However, Fenwick (2010) argues that since WPI is intended as a policy tool, purely statistical weighting may not result in an optimum from policy maker’s point of view. Additionally, it is argued in many papers that one should not concentrate on WPI, but direct analysis on the components – which is what was mainly done in this research.

Finally, the data and R code used for analysis are opened for anyone to check and replicate. This increases the transparency of the study, something that is unnecessarily often lacking in research, and allows for easy modification of the analysis if one deems it necessary.

7 Conclusion

This study started with an aim to investigate water poverty in Lao PDR in a time before the current, very fast dam building commenced in the country. The primary goal was to establish the spatial and seasonal differences of water poverty using WPI, and to provide useful information to the work fulfilling the new Sustainable Development Goals introduced in 2015. Secondarily, the usefulness of inclusion a temporal (or more precisely, seasonal) dimension to WPI has been addressed (although, not explicitly reported in this paper). The actual study followed three key research questions:

1. Are there distinct differences between areas in their water poverty?
2. Are there distinct spatio-temporal differences in water poverty?
3. What are the causes of water poverty in Laos? Do the causes differ across space and seasons?

A meaningful answer to all of the research questions was found during the research project. The following sections concludes the answer for each of these questions.

7.1 Spatial Variation in Water Poverty

The spatial variation in water poverty was explored in three levels: The variables that make out the components, the components individually, and finally the WPI itself. Significant spatial variation was found in all three levels: The smallest Moran's Index value was found to be 0.28, which is commonly considered as a strong spatial relationship, while the highest value was more than 0.9 – variable nearly entirely determined by location. While the variables and components spatially autocorrelate, however, they show very little correlation with each other. There seem to be different spatial processes that drive each of the components. However, correlation can be found when the components are compared against the final index value: Use component strongly correlates with WPI (correlation coefficient of 0.64-0.67) in both of the seasons. It appears that a higher livelihood dependency on water can predict a high portion of the overall water poverty. This is especially surprising, since the Use component scores are low accompanied with low weight in the used PCA-derived weighting scheme. Additionally, Use component does not correlate with any other component. Relevant to the first research question, the correlations vary across the country from locally weak to locally strong correlations.

WPI was also analysed through provinces, and it was found that there are statistically significant differences in the water poverty in different provinces. Three groups of provinces were identified: Water-Poor (Xekong, Oudomxai, Phongsaly and to a lower degree Louangnamtha and Houaphan), Average, and Water-Rich (Vientiane Capital, Vientiane and Bolikhamxai). In fact, cluster analysis revealed that Laos can be divided into two main groups: the water-poor and the water-rich.

As a conclusion, there is a large amount of evidence to support a claim that there *are* distinct differences between regions in their water poverty. This includes both, evidence from visual exploration as well as statistical analysis.

7.2 Spatio-Temporal Variation in Water Poverty

Spatio-temporal dimension of water poverty was analysed through WPI calculated for the dry and wet season. All components, except Access, included variables that change between the seasons. WPI was found to change, with fairly small differences in other WPI components except Resources, which doubles its value in when season changes from dry to wet. This dramatic increase in Resources also results in a significant increase in the overall WPI value. However, despite small differences, visual examination of component maps reveal that spatial changes do occur. In addition, in some areas, the change between dry and wet WPI is dramatic; increase in the index value can be as high as 40 index points (from a maximum value of 100).

Changing component and WPI values were not the only evidence found in the study. Correlations between the components and WPI also change according to seasons. Dry season WPI correlates (correlation coefficient >0.30) with RES, CAP and USE while wet season correlates with CAP, USE and ENV. In addition, weighting scheme derived from PCA result in significantly different schemes; wet season weights emphasize the human components (ACC, CAP, USE) while dry season emphasize the environment components (RES, ENV). Additionally, the clusters found to be either water-poor or water-rich slightly change between seasons, with a higher share of villages assigned to a rich cluster in the wet season than in the dry season. Finally, the relative rank among the provinces changes between the seasons. Some provinces improve dramatically their relative as well as absolute water poverty as wet season starts. A small number of villages, however, behave in the opposite way with wet season WPI proving to be lower than dry season WPI.

The visual and statistical evidence clearly supports the conclusion that there *are* significant spatio-temporal differences in the water poverty across the two seasons.

7.3 Causes of Water Poverty

The causes of water poverty vary somewhat according to the explorative and data mining method employed. As a generalization, it can be said that most of the water-related problems in Laos are in the human components – Access, Capacity and Use – with some spatial differences. Clustering analysis reveals that the biggest difference between water-rich and water-poor is in the Capacity component, and that the clusters correlate with road access to a very high degree. In addition, Environment scores higher in the rich clusters than in the poorer ones. Capacity, as a component, is a common theme across the analyses. It has a large range, spanning near the entire range of Index, and proving to correlate with WPI. In GWPCA, Capacity is the most loaded component in the first PC on both season. Finally, in GWR, variables that make Capacity are all among the 11 first picked by the step-wise model selection algorithm. Therefore, it is concluded that Capacity is the single most important component driving water poverty in Laos.

However, the picture is not as simple as singling Capacity out as the cause. The other human components (Access, Use) are important throughout the country, and in the dry season, Resources is a significant driver in the northwest of the country. Especially Use is an important indicator, as it correlates strongly with dry and wet season water poverty despite scoring low and having a low weight. This is despite Use being the lowest scoring component. In fact, in absolute terms, water poverty could be addressed by improving the infrastructure and with more efficient use of water.

According to GWPCA, share population that depends on agri- or aquaculture (belonging to Use component) as the main source of their income is the most important variable in the dry season, and second most important in the wet season. Road access (Capacity component) is the most loaded variable in the wet season, and an important variable in the dry season as well. In addition, surface water availability (Resources) and the travel time to capitals (Capacity) are significant predictors in the dry season. Toilet type (Access) is the most loaded in the third PC in the wet season. A similar picture is painted by GWR models. The most important predictors are shared by the seasons; they are mostly from Capacity component (all four are included in the top 11 predictors from both seasons), Use (all three included), precipitation from the Resources component and soil degradation from ENV. The rest of the components are largely the same, however, an important difference is that dry season predictors include more environmental and infrastructure variables. The causes of water poverty, however, were not completely discovered as is evident from the low R^2 of the GWR model in the northwest and around the capital city.

The answer to the last research question is therefore that Capacity and other human components are important (in relative terms) everywhere in Laos with Resources being a major factor in the northwest during dry season. Use and Access are the least scoring, and biggest *absolute* improvements can be found in these two components. Dependence on agriculture and road access are the two most important predictors of water-related poverty, suggesting that rural population is more vulnerable. Additionally, it was found that significant spatial variation in variables correlating with WPI.

7.4 The Way Forward

This study created a foundation for exploring water poverty further in Laos (and in Southeast Asia). It has also highlighted additional research questions that need to be answered in order to create a comprehensive picture in the dimensions that cause and drive water poverty.

The study showed that the causes of water poverty change according to space and region. WPI is reliant on geospatial data (Sullivan, 2002) and thus does take the spatial dimension in to account. However, conventionally seasonal variability is only taken into account in Resources component as a variable that describes the seasonal changes of the resource. This study has made it clear that, at least in a highly seasonal environment, single index is not enough to describe the differences between seasons. This is due to the nature of chosen variables which are changing according to season. WPI aims to be a holistic tool (Sullivan, 2002) to assess the complex phenomenon of water-related poverty. Further research is therefore needed whether the difference between seasons warrants development of WPI in a direction that better represents the changing drivers across seasons.

In addition, continued assessment of water poverty is needed to describe the *current* situation in the Mekong – this study was based on data mainly collected prior the accelerated building of large scale water infrastructure. It is likely that the dam building currently occurring in Laos changes the picture painted by WPI. A follow-up study using more recent data (e.g. Population Census conducted in 2015, which was unavailable at the time of this research) should therefore be done to support impact assessments on the infrastructure projects and to better direct efforts to achieve SDG's laid out by the United Nations. It is also known that significant improvement in poverty has occurred since the data collection for the 2005 Population Census (World Bank Group, 2016; Coulombe, et al., 2016).

References

Anselin, L., 1998. Exploratory Spatial Data Analysis in a Geocomputational Environment. In: P. Longley, S. Brooks, R. McDonnell & W. Macmillian, eds. *Geocomputation: A Primer.* New York: Wiley and Sons, pp. 77-94.

Babel, M. & Wahid, S., 2009. *Freshwater Under Threat: Southeast Asia. Vulnerability Assessment of Freshwater Resources to Environmental Change. Mekong River Basin.,* Nairobi, Kenya: United Nations Environment Programme.

Behrens, J., 1997. Principles and Procedures of Exploratory Data Analysis. *Psychological Methods,* 2(2), pp. 131-160.

Bivand, R., 2010. Exploratory Spatial Data Analysis. In: M. Fischer & A. Getis, eds. *Handbook of Applied Spatial Analysis.* Berlin: Springer-Verlag, pp. 219-254.

Bivand, R., Pebesma, E. & Gomez-Rubio, V., 2013. *Applied Spatial Data Analysis with R.*. 2nd ed. New York: Springer.

Bivand, R. & Piras, G., 2015. Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software,* 63(18), pp. 1-36.

Bivand, R. & Yu, D., 2015. *spgwr: Geographically Weighted Regression. R package version 0.6-28.* [Online] Available at: https://CRAN.R-project.org/package=spgwr

Brunsdon, C. & Chen, H., 2014. *GISTools: Some further GIS capabilities for R. R package version 0.7-4.*. [Online] Available at: https://CRAN.R-project.org/package=GISTools

Brunsdon, C., Fotheringham, A. & Charlton, M., 2002. Geographically weighted summary statistics — a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems 26,* pp. 501-524.

Brunsdon, C., Fotheringham, S. & Charlton, M., 1996. Geographically Weighted Regression - A Method for Exploring Spatial Nonstationarity. *Geographical Analysis,* pp. Vol. 28, No. 4, p. 281-298.

Brunsdon, C., Fotheringham, S. & Charlton, M., 1998. Geographically Weighted Regression-Modelling Spatial Non-Stationarity. *Journal of the Royal Statistical Society. Series D (The Statistician),* pp. Vol. 47, No. 3 , p. 431-443.

Charlton, M. et al., 2010. Principal Components Analysis: from Global to Local. *13th AGILE International Conference on Geographic Information Science 2010 ,* pp. Available at: https://agile-online.org/Conference_Paper/CDs/agile_2010/ShortPapers_PDF/114_DOC.pdf.

Charrad, M., Ghazzali, N., Bolteau, V. & Niknafs, A., 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software,* 61(6), pp. 1-36.

Coulombe, H., Epprecht, M., Pimhidzai, O. & Sisoulath, V., 2016. *Where Are the Poor? Lao PDR 2015 Census-Based Poverty Map: Province and District Level Results,* Vientiane: Lao Statistics Bureau.

Cressie, N. & Wikle, C., 2011. *Statistics for Spatio-Temporal Data.* Hoboken, New Jersey: Wiley & Sons.

Demsar, U., Fotheringham, A. & Charlton, M., 2008. Combining Geovisual Analytics with Spatial Statistics: the Example of Geographically Weighted Regression. *The Cartographic Journal,* pp. Vol.45, No.3, p. 182-192.

Demsar, U. et al., 2013. Principal Component Analysis on Spatial Data: An Overview. *Annals of the Association of American Geographers,* pp. Vol. 103, No. 1, p. 106-128.

Epprecht, M. et al., 2008. *The Geography of Poverty and Inequality in the Lao PDR,* Bern: Geographica Bernensia: Swiss National Center of Competence in Research (NCCR) North-South, University of Bern, and International Food Policy Research Institute (IFPRI).

Erwig, M., Güting, R., Schneider, M. & Vazirgiannis, M., 1999. Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. *Geoinformatica ,* 3(3), pp. 269-296.

Falkenmark, M., Lundqvist, J. & Widstrand, C., 1989. Macro-scale Water Scarcity Requires Micro-scale Approaches. *Natural Resources Forum,* 13(4), pp. 258-267.

FAO/IIASA/ISRIC/ISSCAS/JRC, 2012. *Harmonized World Soil Database (version 1.2),* Rome: FAO.

Fenwick, C., 2010. *Identifying the Water Poor: an Indicator Approach to Assessing Water Poverty in Rural Mexico. Doctoral Disseration.,* London: University College London.

Garriga, R. & Foguet, A., 2010. Improved Method to Calculate a Water Poverty Index at Local Scale. *Journal of Environmental Engineering,* 136(11), pp. 1287-1298.

Gassert, F. et al., 2014. *Aqueduct Global Maps 2.1: Constructing Decision-Relevant Global Water Risk Indicators." Working Paper.,* Washington, DC: World Resources Institute.

Getis, A., 2010. Spatial Autocorrelation. In: M. Fischer & A. Getis, eds. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications.* Berlin: Springer-Verlag, pp. 255-278.

Gohel, D., 2016. *ggiraph: Make 'ggplot2' Graphics Interactive Using 'htmlwidgets'.* [Online] Available at: https://github.com/davidgohel/ggiraph

Gollini, I. et al., 2015. GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. *Journal of Statistical Software,* 63(17), pp. 1-50.

Guo, D., 2009. Multivariate Spatial Clustering and Geovisualization. In: H. Miller & J. Han, eds. *Geographic Data Mining and Knowledge Discovery. 2nd Edition.*. Boca Raton, Florida: Chapman & Hall, pp. 325-347.

Guo, D. & Mennis, J., 2009. Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems,* Volume 33, pp. 403-408.

GWSP Digital Water Atlas, 2008a. *Map 13: Water Consumption for Domestic Sector (Dataset) (V1.0).* [Online] Available at: http://atlas.gwsp.org [Accessed 15 02 2016].

GWSP Digital Water Atlas, 2008b. *Map 15: Water Consumption for Irrigation (Dataset) (V1.0).* [Online] Available at: http://atlas.gwsp.org [Accessed 15 02 2016].

GWSP Digital Water Atlas, 2008c. *Map 16: Water Consumption (Total) (v1.0).* [Online] Available at: http://atlas.gwsp.org [Accessed 15 02 2016].

Haining, R., Wise, S. & Ma, J., 1998. Exploratory Spatial Data Analsis in a Geographic Information System Environment. *The Statistician,* 47(3), pp. 457-469.

Hand, D., Mannila, H. & Smyth, P., 2001. *Principles of Data Mining.* s.l.:The MIT Press.
Han, J., Lee, J. & Kamber, M., 2009. An Overview of Clustering Methods in Geographic Data Analysis. In: H. Miller & J. Han, eds. *Geographic Data Mining and Knowledge Discovery. Second Edition.*. Boca Raton, USA: CRC Press, pp. -189.

Harris, P., Brunsdon, C. & Charlton, M., 2011. Geographically weighted principal components analysis. *International Journal of Geographical Information Science,* pp. Vol. 25, No.10, p. 1717-1736.

Heer, J., Bostock, M. & Ogievetsky, V., 2010. A Tour Through the Visualization Zoo. *Communications of the ACM,* 53(6), pp. 59-67.

International Fund for Agricultural Development, 2014. *nvesting in Rural People in the Lao People's Democratic Republic,* Rome, Italy: International Fund for Agricultural Development.

International Rivers, 2015. *Laos.* [Online] Available at: https://www.internationalrivers.org/campaigns/laos [Accessed 23 June 2016].
Isard, W., 1970. On Notions of Models of Time. *Papers of the Regional Science Association,* Volume 25, pp. 7-32.

Jemmali, H. & Matoussi, M., 2013. A multidimensional analysis of water poverty at local scale: application of improved water poverty index for Tunisia. *Water Policy,* Volume 15, pp. 98-115.

Kazar, B. & Celik, M., 2012. *Spatial Autoregression (SAR) Model: Parameter Estimation Techniques.* New York: Springer.

Koponen, J., Lauri, H., Veijalainen, N. & Sarkkula, J., 2010. *HBV and IWRM Watershed Modelling User Guide,* Phnom Penh: MRC Information and Knowledge Management Programme.

Laffan, S. & Bickford, S., 2005. *Using spatial randomisations to improve the utility of Geographically Weighted Regression model results.* s.l., Modelling and Simulation Society of Australia and New Zealand.

Lao Statistics Bureau, 2005. *Census of Population and Housing 2005.* [Online] Available at: Dataset downloaded from http://www.decide.la/ on 10 Feb 2016

Lao Statistics Bureau, 2011. *Lao Agriculture Census 2010/2011.* [Online] Available at: Dataset downloaded from http://www.decide.la/ on 10 Feb 2016

Larose, D., 2005. *Discovering Knowledge in Data: An Intoduction to Data Mining.* Hoboken, New Jersey: John Wiley & Sons Ltd..

Lauri, H., Räsänen, T. & Kummu, M., 2014. Using Reanalysis and Remotely Sensed Temperature and Precipitation Data for Hydrological Modeling in Monsoon Climate: Mekong River Case Study. *Journal of Hydrometeorology,* Volume 15, pp. 1532-1545.

Lawrence, P., Meigh, J. & Sullivan, C., 2002. *The Water Poverty Index: an International Comparison,* Keele, UK: Keele University Department of Economics.

Leung, Y., Mei, C. & Zhang, W., 2000. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A,* Volume 32, pp. 9-32.

Lloyd, C., 2010. Analysing population characteristics using geographically weighted principal components analysis: A case study of Northern Ireland in 2001. *Computers, Environment and Urban Systems ,* pp. Vol. 34, p. 389-399.

Luan, J., 2002. Data Mining and Its Applications in Higher Education. *New Directions for Institutional Research,* 2002(113), pp. 17-36.

Maechler, M. et al., 2016. *cluster: Cluster Analysis. R package version 2.0.4.,* s.l.: s.n.

Mei, C., Xu, M. & Wang, N., 2016. A bootstrap test for constant coefficients in geographically weighted regression models. *International Journal of Geographical Information Science,* p. DOI: 10.1080/13658816.2016.1149181.

Mekong River Commission, 2007. *Diagnostic Study of Water Quality in Lower Mekong Basin. MRC Technical Paper No. 15.,* Vientiane: Mekong River Commission.

Mekong River Commission, 2011. *Planning Atlas of the Lower Mekong River Basin,* Vientiane: Mekong River Commission.

Mekong River Commission, 2012. *Flood Management and Mitigation Programme. Working Paper 2011-2015.,* Phnom Penh: Mekong River Commission.

Miller, H. & Han, J., 2009. *Geographic Data Mining and Knowledge Discovery.* 2nd ed. Baton Rouge: CRC Press.

Molle, F. & Mollinga, P., 2003. Water poverty indicators: conceptual problems and policy issues. *Water Policy,* Volume 5, pp. 529-544.

NIST/SEMATECH, 2013. *e-Handbook of Statistical Methods.* [Online] Available at: http://www.itl.nist.gov/div898/handbook/ [Accessed 29 May 2016].

Openshaw, S., 1983. *The Modifiable Areal Unit Problem,* Norwich, UK: Geo Books.
O'Sullivan, D. & Unwin, D., 2010. *Geographic Information Analysis.* 2nd ed. New Jersey, USA: John Wiley & Sons Ltd.

Perez-Foguet, A. & Garriga, R., 2011. Analyzing Water Poverty in Basins. *Water Resources Management,* Volume 25, pp. 3595-3612.

QGIS Development Team, 2016. *QGIS Geographic Information System (Version 2.14 Essen).* s.l.:Open Source Geospatial Foundation Project.

R Core Team, 2016. *R: A language and environment for statistical computing,* Vienna, Austria: R Foundation for Statistical Computing.

Räsänen, T., Koponen, J., Lauri, H. & Kummu, M., 2012. Downstream hydrological impacts of hydropower development in the Upper Mekong Basin. *Water Resources Management,* 26(12), pp. 3495-3513.

Salmivaara, A. et al., 2015. Exploring the Modifiable Areal Unit Problem in Spatial Water Assessments: A Case of Water Shortage in Monsoon Asia. *Water,* 7(3), pp. 898-917.

Shekhar, S. & Chawla, S., 2003. *Spatial Databases: A Tour.* Upper Saddle River, New Jersey: Prentice-Hall.

Sievert, C. et al., 2016. *plotly: Create Interactive Web Graphics via 'plotly.js'. R package version 3.6.0.*. [Online] Available at: https://CRAN.R-project.org/package=plotly

Steltman, H., 2015. *Experimental Design and Analysis.* [Online] Available at: http://www.stat.cmu.edu/~hseltman/309/Book/ [Accessed 14 July 2016].

Sullivan, C., 2002. Calculating a Water Poverty Index. *World Development ,* pp. Vol. 30, No. 7, pp. 1195–1210.

Sullivan, C., Meigh, J. & Lawrence, P., 2006. Application of Water Poverty Index at Different Scales: A Cautionary Tale. *Water International,* 31(3), pp. 412-426.

The International Water Association, 2014. *An Avoidable Crisis: WASH Human Resource Capacity Gaps in 15 Developing Economies,* Seacourt, UK: The International Water Association.

The United Nations Department of Economic and Social Affairs, 2016. *Sustainable Development Knowledge Platform.* [Online] Available at: https://sustainabledevelopment.un.org [Accessed 18 July 2016].

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography,* pp. Vol. 46, No. 2,a p. 234-240.

Ty, T., Sunada, K., Ichikawa, Y. & Oishi, S., 2010. Evaluation of the state of water resources using Modified Water Poverty Index: a case study in the Srepok River basin,

Vietnam – Cambodia. *International Journal of River Basin Management,* 8(3-4), pp. 305-317.

UNESCO, 2016. *The United Nations World Water Development Report 2016. Water and Jobs.,* Paris, France: UNESCO.

United Nations in Lao PDR, 2015. *Country Analysis Report: Lao PDR,* Vientiane: United Nations.

van der Vywer, C., 2013. Water Poverty Index Calculation: Additive or Multiplicative Function?. *Journal of South African Business Research,* Volume 2013, p. Article ID 615770.

Ward, J. & Kaczan, D., 2014. Challenging Hydrological Panaceas: Water poverty governance accounting for spatial scale in the Niger River Basin. *Journal of Hydrology,* 519(C), pp. 2501-2514.

Wei, C. & Qi, F., 2012. On the estimation and testing of mixed geographically weighted regression models. *Economic Modelling,* pp. Vol. 29, No. 6, p. 2615-2620.

Wheeler, D., 2006. *Diagnostic Tools and a Remedial Method for Collinearity in Geographically Weighted Regression. Dissertation.,* Columbus, Ohio: Ohio State University.

Wheeler, D. & Tiefelsdorf, M., 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems,* Volume 7, pp. 161-187.

Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Wildlife Conservation Society - WCS; Center for International Earth Science Information Network - CIESIN - Columbia University, 2005. *Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Footprint Dataset (IGHP).* s.l.:s.n.

World Bank Group, 2016. *Lao Economic Monitor. Challenges in Promoting More Inclusive Growth and Shared Prosperity,* Washington DC.: World Bank Group.

World Economic Forum, 2016. *The Global Risks Report 2016. 11th Edition.,* Geneva: World Economic Forum.

Zhang, J. & Goodchild, M., 2003. *Uncertainty in Geographical Information.* 3rd ed. London: Taylor & Francis.

Ziv, G. et al., 2011. Trading-off fish biodiversity, food security, and hydropower in the Mekong River Basin. *PNAS,* 109(15), p. 5609–5614.

Appendix 1. Vmod Model Description

The Vmod IWRM (Integrated Water Resources Management) modelling software used to model water availability in this study is a distributed physically based/conceptual hydrological model. The modelled catchment is represented by a grid (raster) with processes simulated using simplified, physically based formulations. State of water balance, runoff and water quality are calculated separately for each grid cell, commonly ranging from 0.01 to 1 km^2 (or even up to 5 km^2), as in the case of the model application used in this study. Each grid cell is assigned with an outflow point which determines the destination cell (downstream) for computed runoff. The model simulates lakes in addition to rivers. A detailed description of the model equations can be found in the model manual. (Koponen, et al., 2010)

Required input data for the model are an elevation model, land cover and soil type, shown in Figure A1.2 for the Mekong model employed in this study, in addition to meteorological and hydrological timeseries. For simulation, each grid cell is divided into four layers; vegetation, surface and two soil layers (Figure A1.1. Model layers and computation processes.).

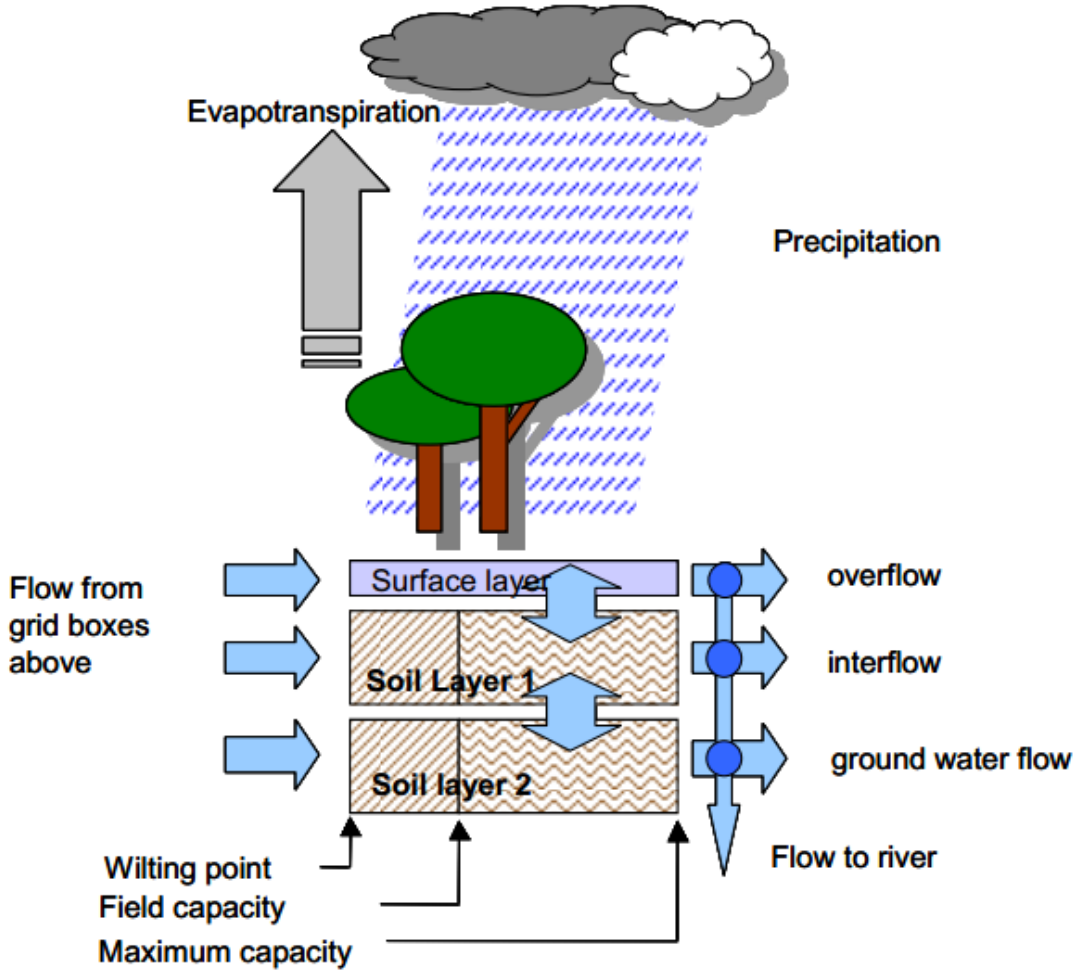


Figure A1.1. Model layers and computation processes. (Koponen, et al., 2010)

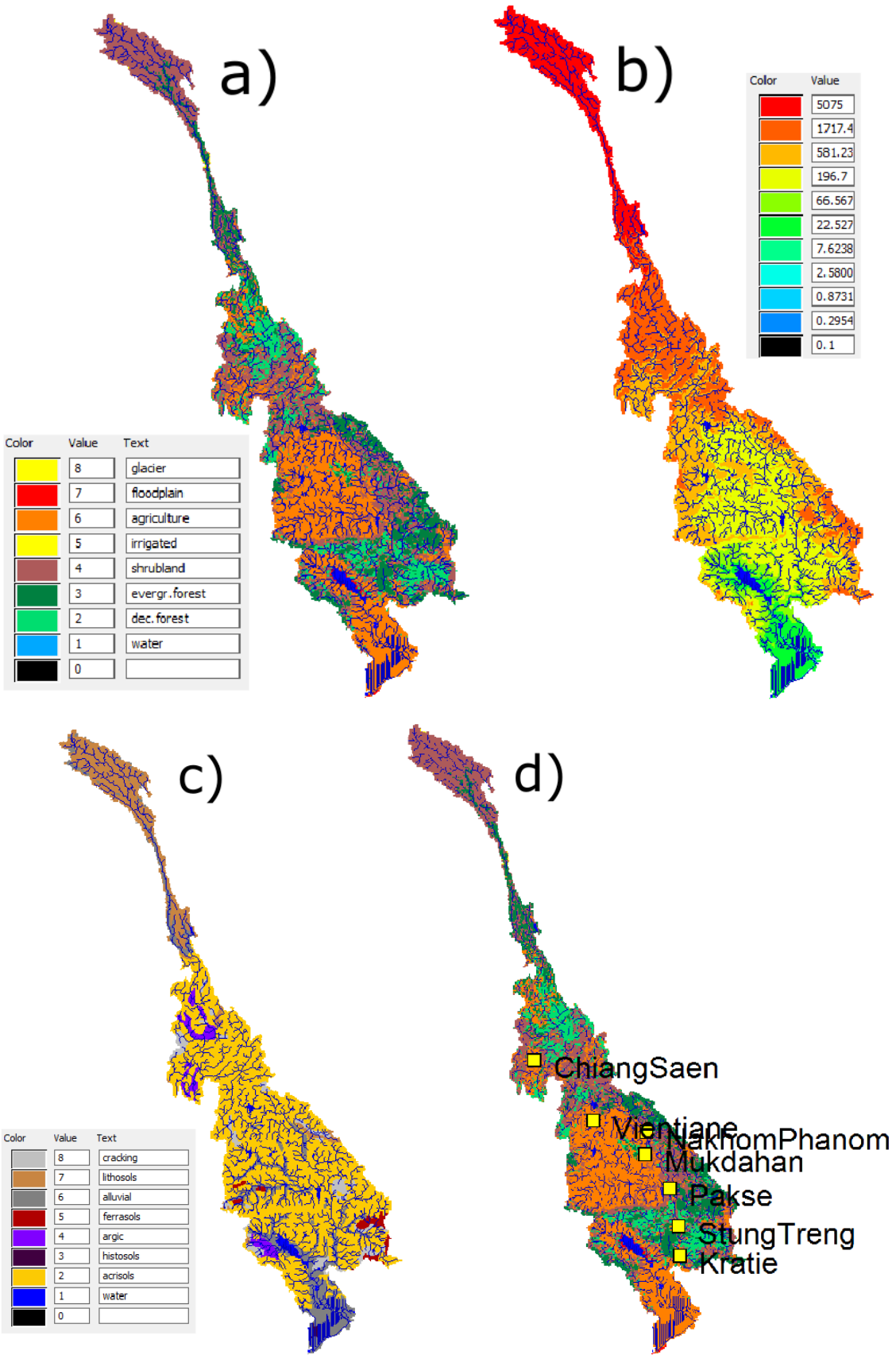


Figure A1.2. a) Landuse layer, b) Elevation model and c) Soil layer used in the Mekong model. Calibration points are shown in d). The blue lines indicate the river network.

Computation process first interpolates and corrects (e.g. effects of elevation) temperature and precipitation data to cover the entire model area. Then, interception by vegetation is calculated before processing infiltration to the ground and accumulation to pond storage and surface runoff. Evaporation from interception storage, ground surface and transpiration of vegetation are taken into account. Plant growth is simulated, as is crop water demand in case the in-built FAO56 agricultural model is used. Water movement takes place between soil layers, from grid cell to the next and from grid cell to river or lake. Accumulation and melting of snow, soil freezing and glacier melting are also included.

The model was calibrated in five points along the Mekong River; Chiang Saeng, Vientiane, Nakhom Phanom, Pakse and Stung Treng. The model overestimates discharge at the measurement stations by approximately 5%. Nash-Sutcliffe Model Efficiency (NSME) is 0.87-0.91 in the south of Laos, up from a worse figure of 0.75 in the North. The poorer performance in the North can be explained by the exclusion of the large dams, which have a major impact on discharge and water levels (Räsänen, et al., 2012) especially in the North and thus, to the NSME. In addition, the validation period includes years when the large Chinese dams were in fill-up phase, affecting discharge in the Mekong. These findings are summarized in Table A1.1.

Table A1.1. Model efficiency and validation period of calibration.

	Average discharge m3		Ratio		Data period	
Point	Measured	Modelled	modelled/observed	NSME	Start	End
Chiang Saeng	2662	2630	0.99	0.75	01/04/1981	30/12/2002
Vientiane	4208	4689	1.11	0.77	01/04/1981	31/12/2001
Nakhom Phanom	7050	7384	1.05	0.88	01/04/1981	31/12/2000
Pakse	9759	10165	1.04	0.91	01/04/1981	31/12/2001
Stung Treng	13159	13895	1.06	0.87	01/04/1981	30/12/2002

In addition to the Mekong model, three smaller catchments in the province of Houaphan were modelled, because they did not fall into the Mekong model. Modelling of these three catchments used the same calibration as in the Mekong model due to lack of discharge data in these locations. The positions of the additional models are shown in Figure A1.3.

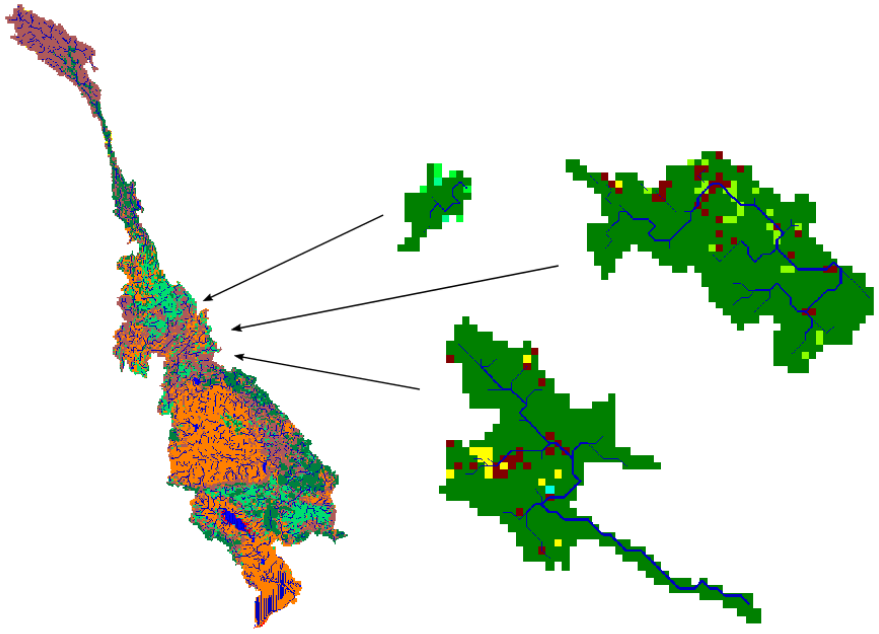
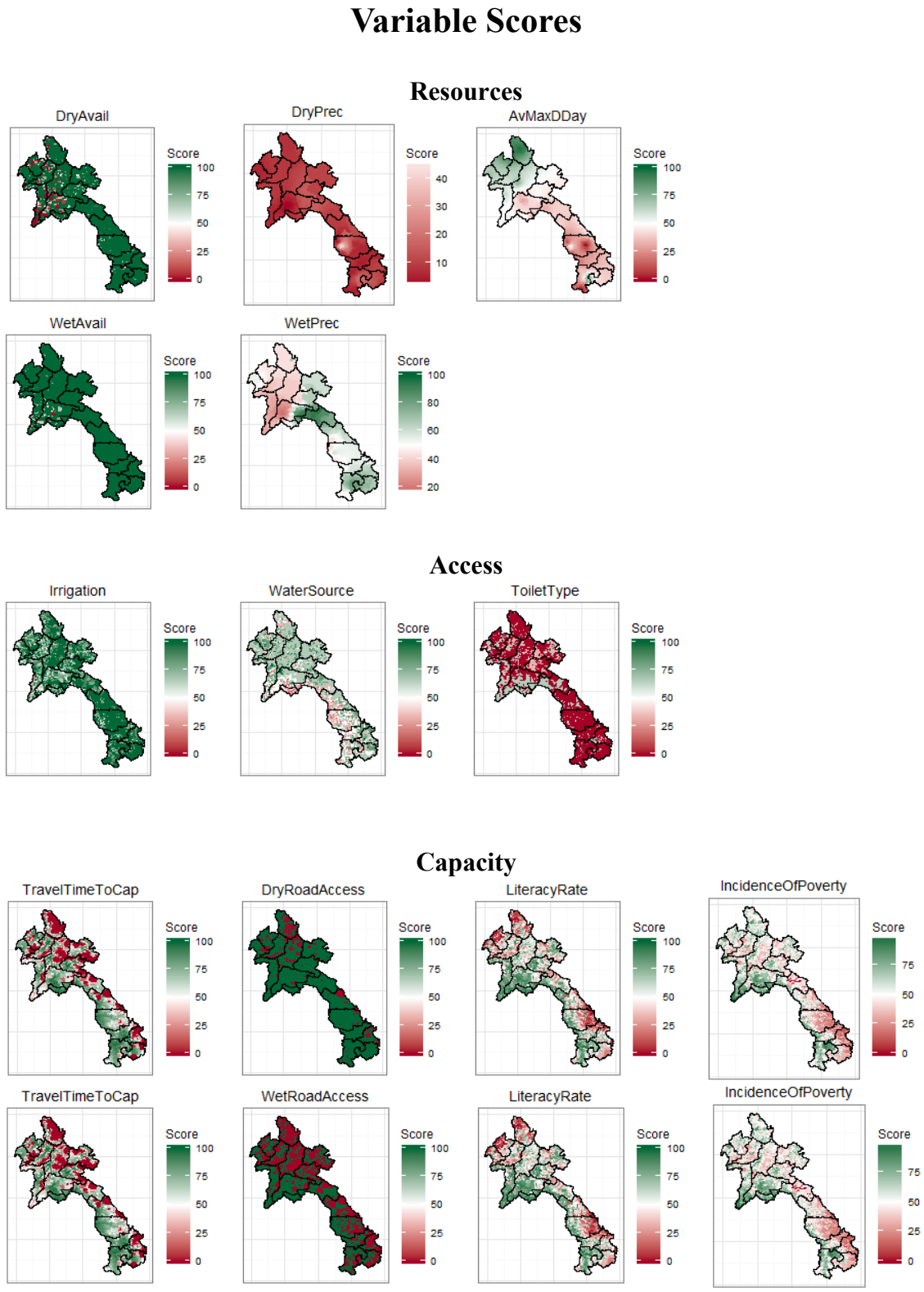


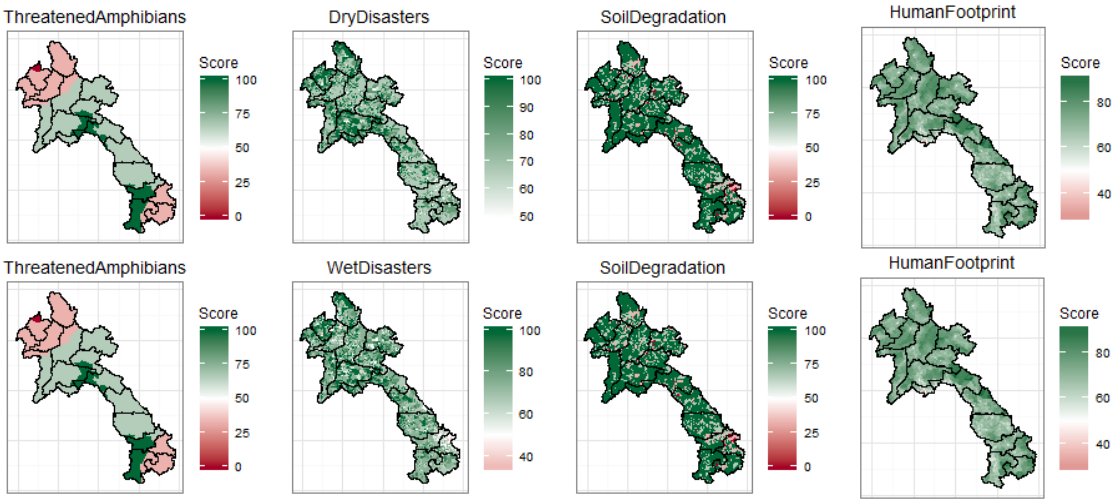
Figure A1.3. Positions of the three small modelled catchments.

Appendix 2. Additional Data for Initial Dataset Exploration

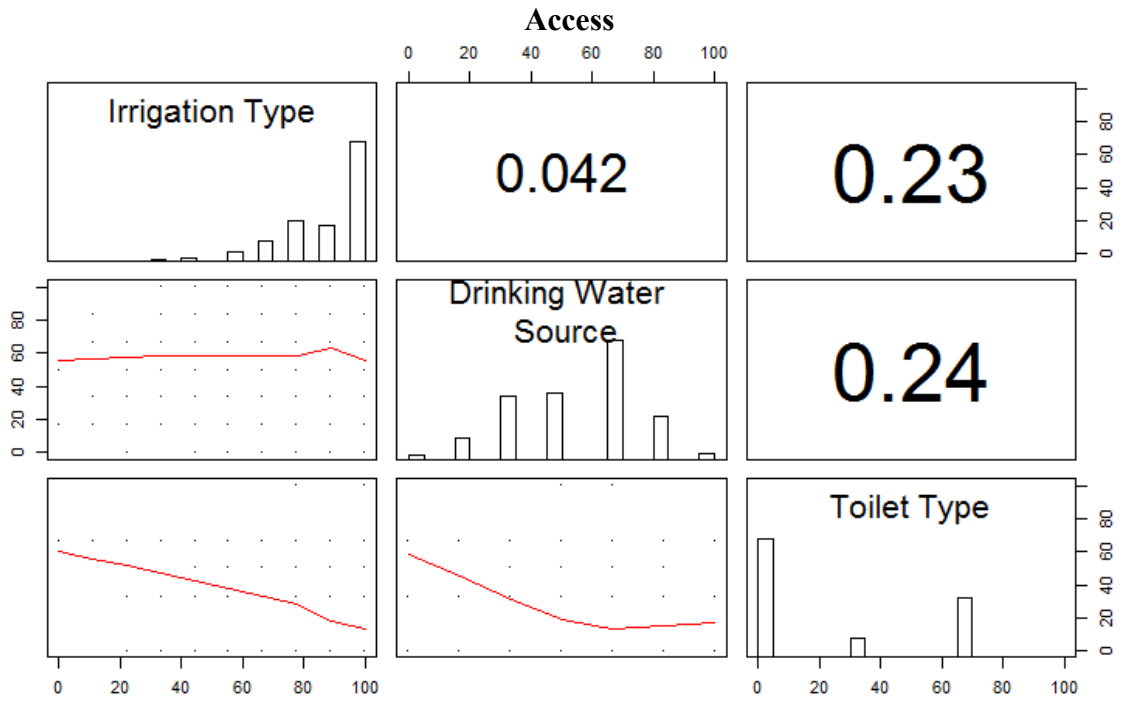


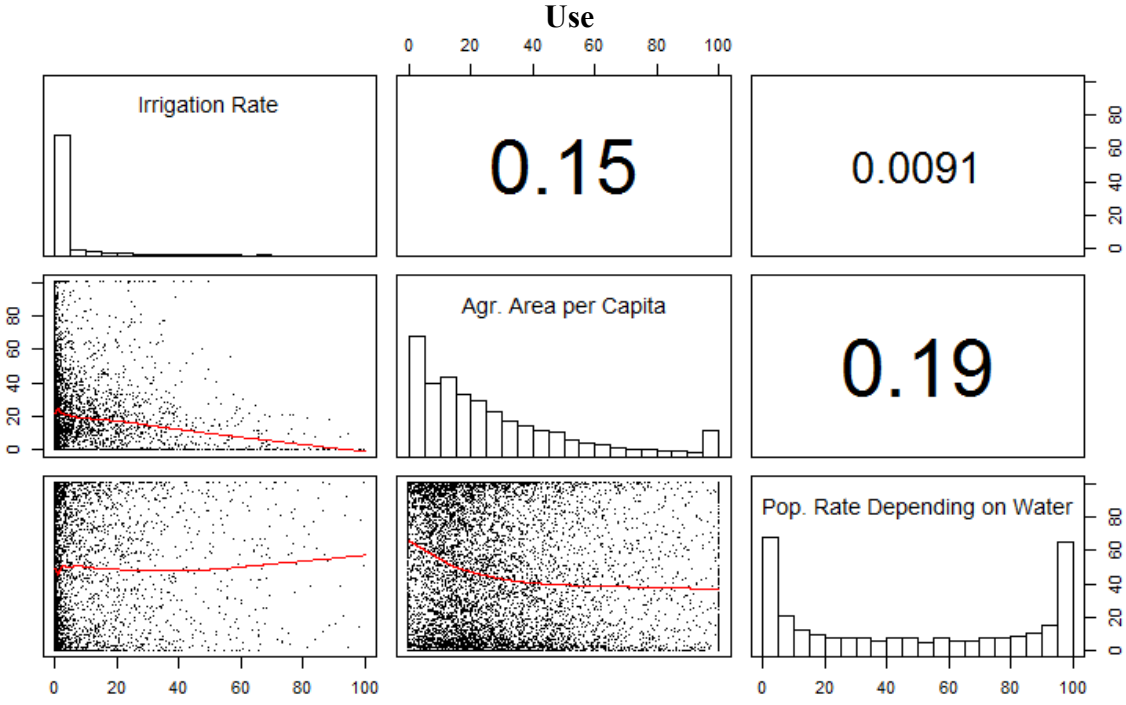
Use

Environment

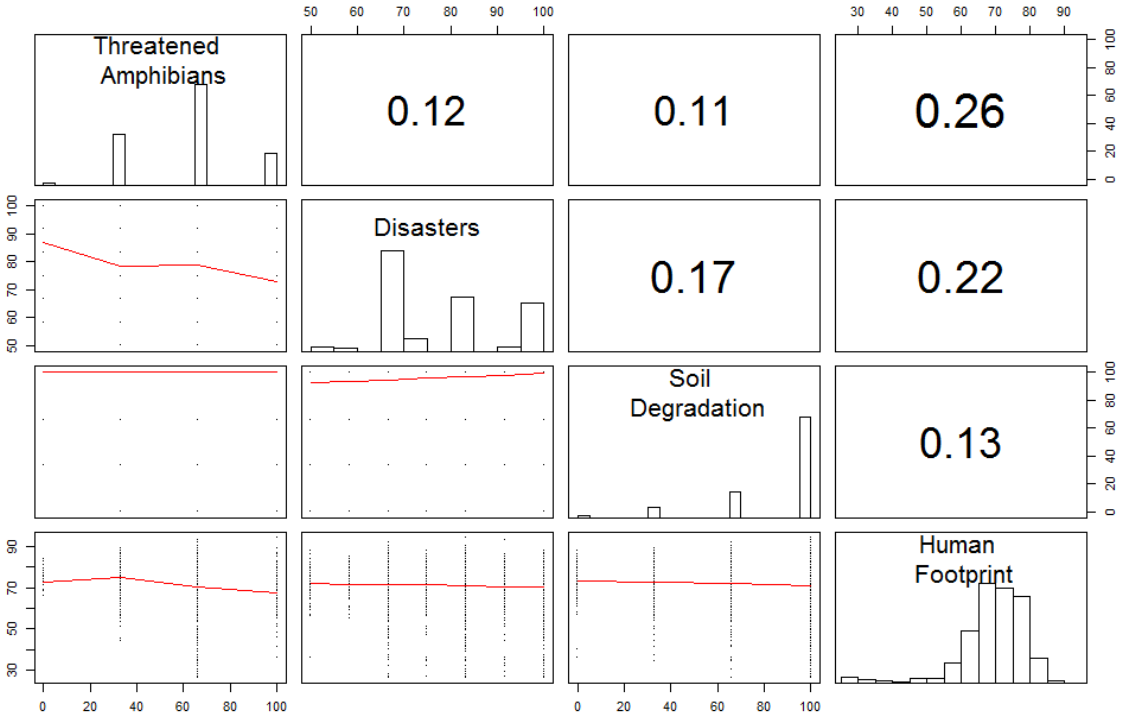


Scatterplot Matrices





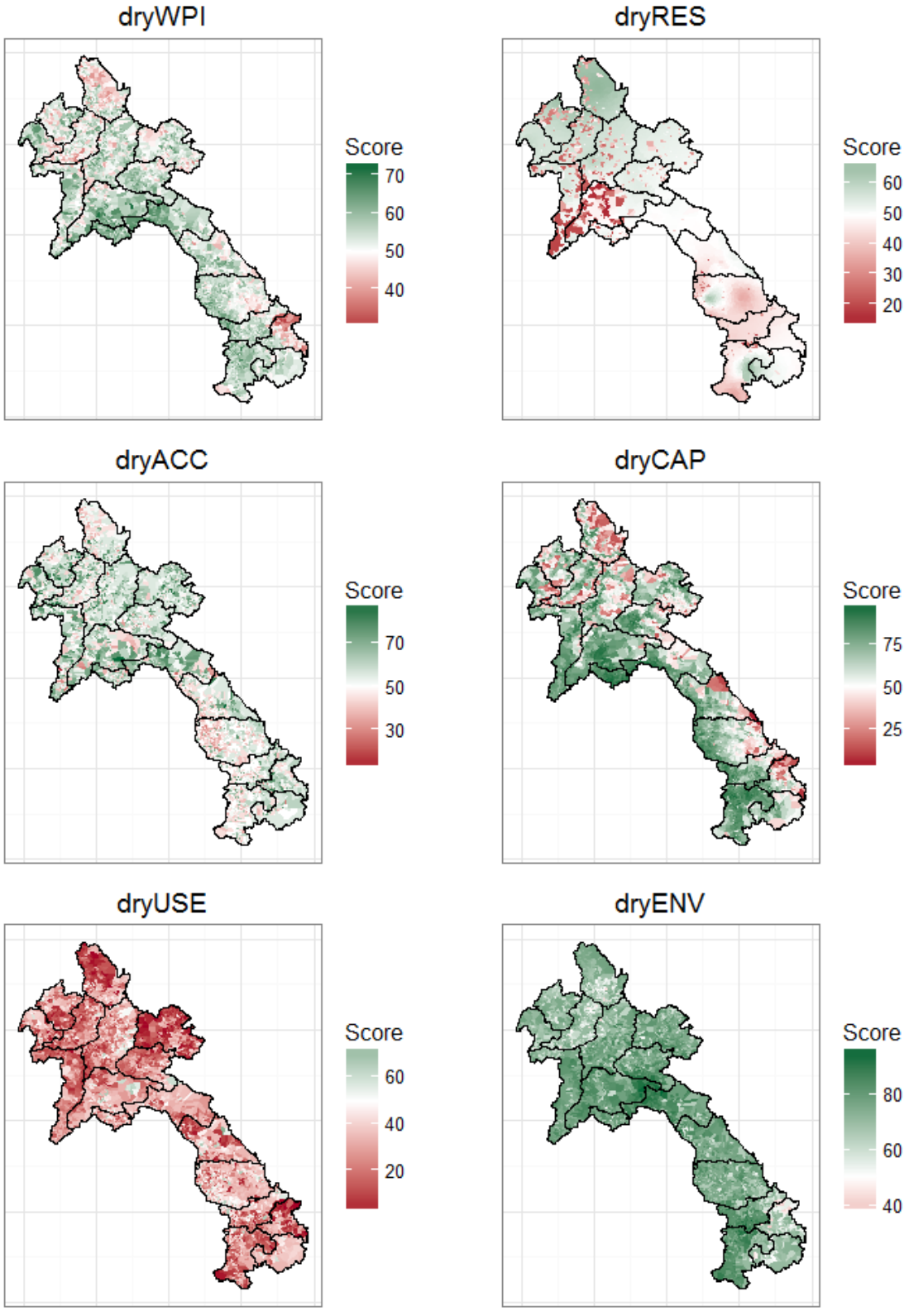
Environment



Appendix 3. Additional Data for Water Poverty Index Exploration

Dry Season

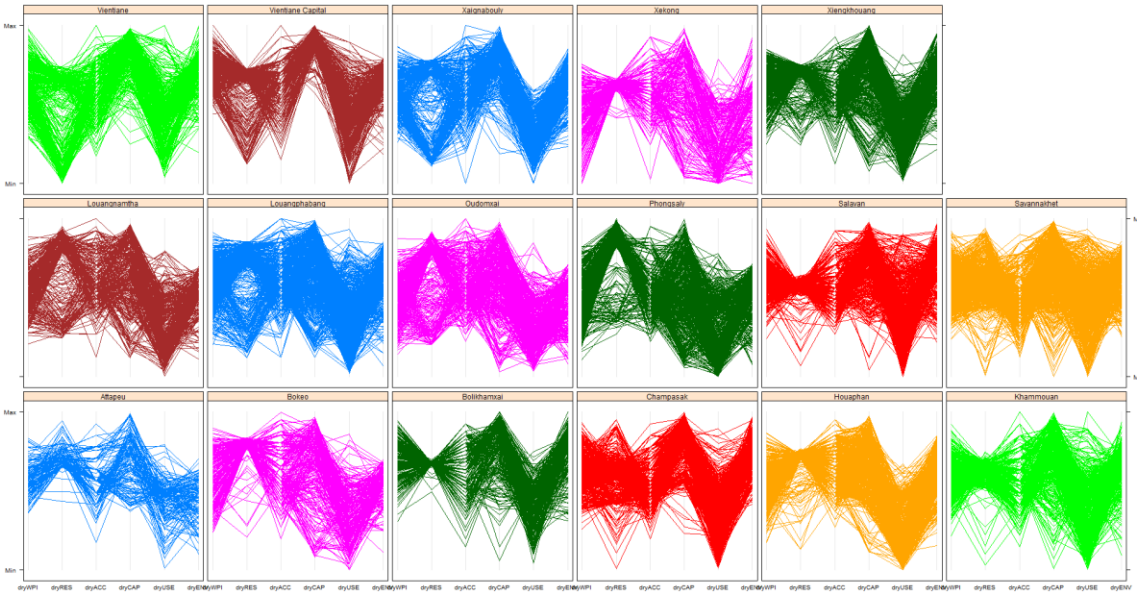
Dry Season WPI Component Scores per Village



Province Summary Statistics of Dry Season WPI score

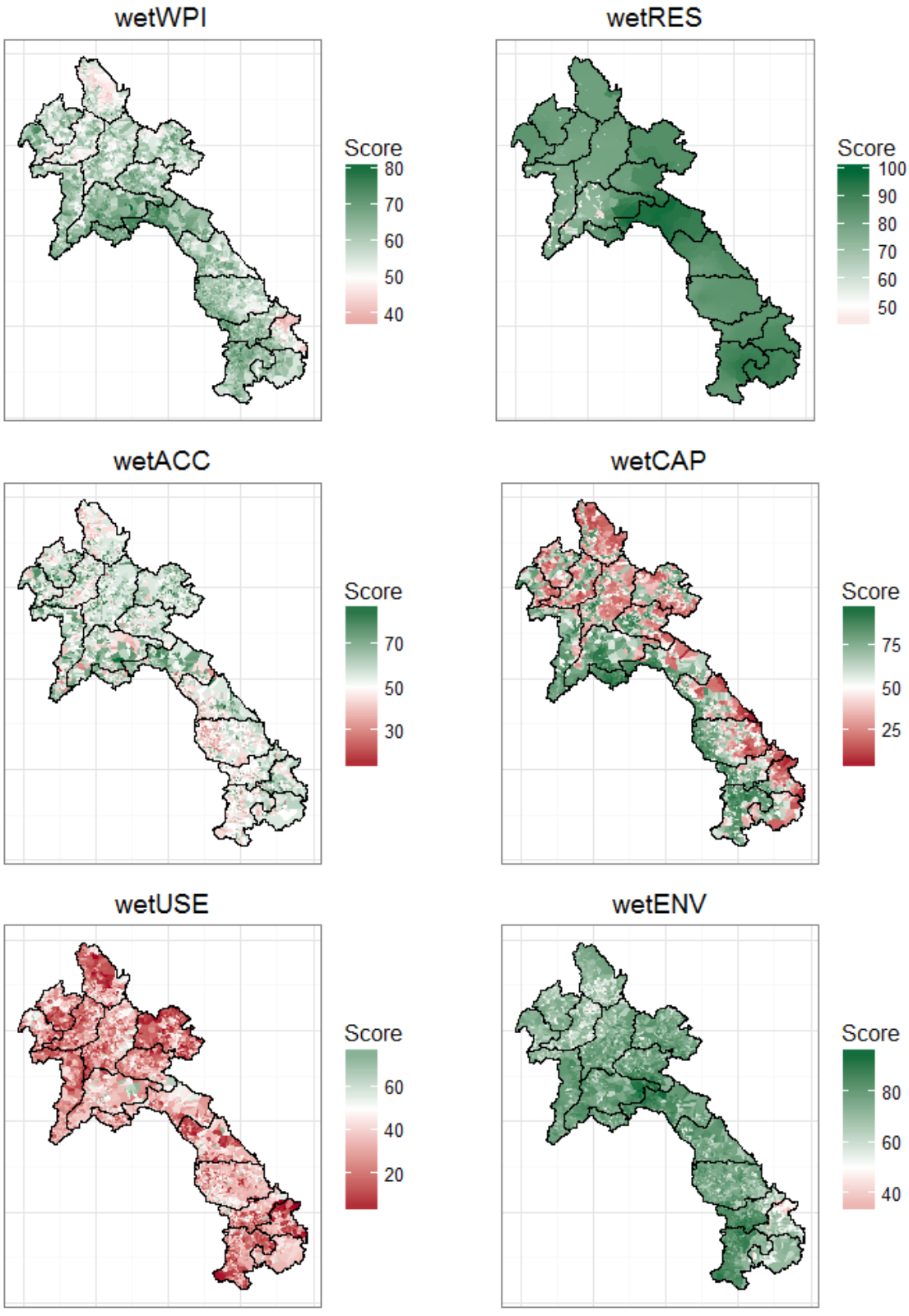
Province	Min.	1st Qu.	Median	3rd Qu.	Max.	Mean
Attapeu	32.7	52.5	55.6	58.2	64.1	55.0
Bokeo	18.9	50.8	56.4	60.9	66.5	55.2
Bolikhamxai	42.4	58.0	61.7	64.9	70.7	61.1
Champasak	36.2	53.1	57.5	61.3	70.7	57.0
Houaphan	18.8	45.7	50.7	56.1	65.4	50.0
Khammouan	29.7	52.3	56.7	59.7	68.5	55.7
Louangnamtha	26.4	47.9	52.7	57.2	67.6	52.4
Louangphabang	30.9	52.0	55.7	59.7	68.7	55.2
Oudomxai	33.2	46.7	50.6	54.7	65.8	50.5
Phongsaly	18.5	45.8	51.8	57.3	66.8	50.9
Salavan	22.2	53.3	57.2	60.1	69.2	56.5
Savannakhet	21.8	53.3	57.0	60.0	68.4	56.3
Vientiane	28.5	47.7	57.4	62.6	70.8	54.9
Vientiane Capital	19.0	56.8	61.0	63.4	69.9	59.3
Xayabouly	17.3	46.5	51.5	58.2	66.5	52.0
Xekong	13.7	41.3	49.5	54.5	61.7	46.9
Xiengkhouang	35.3	53.0	56.7	60.5	69.0	56.4

Parallel Coordinate Plot of Dry Season WPI Components for Each Province



Wet Season

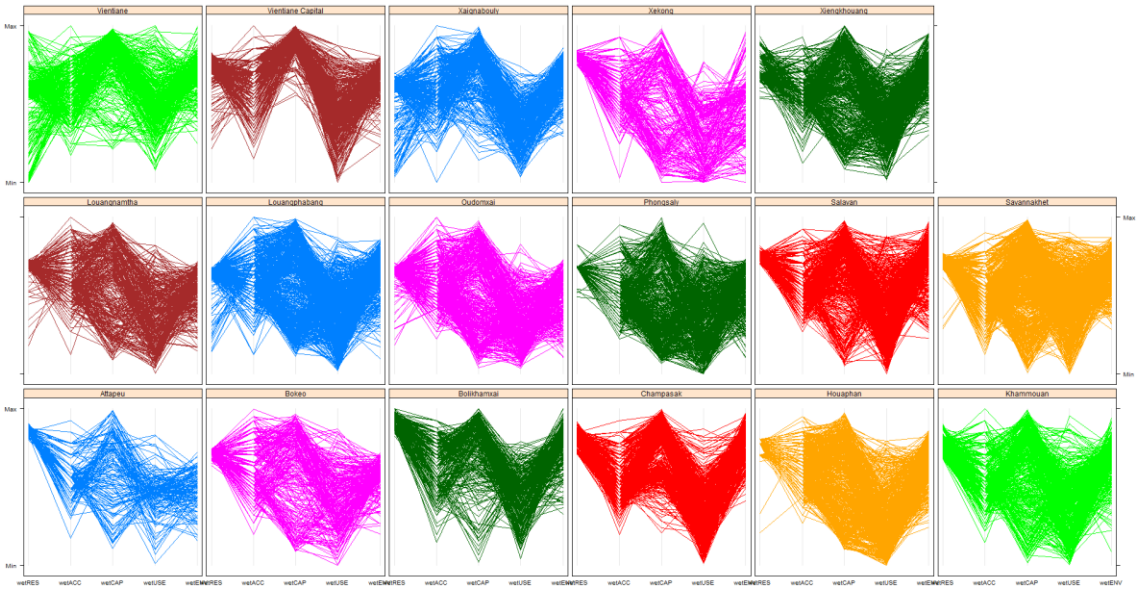
Wet Season WPI Component Scores per Village



Province Summary Statistics of Wet Season WPI score

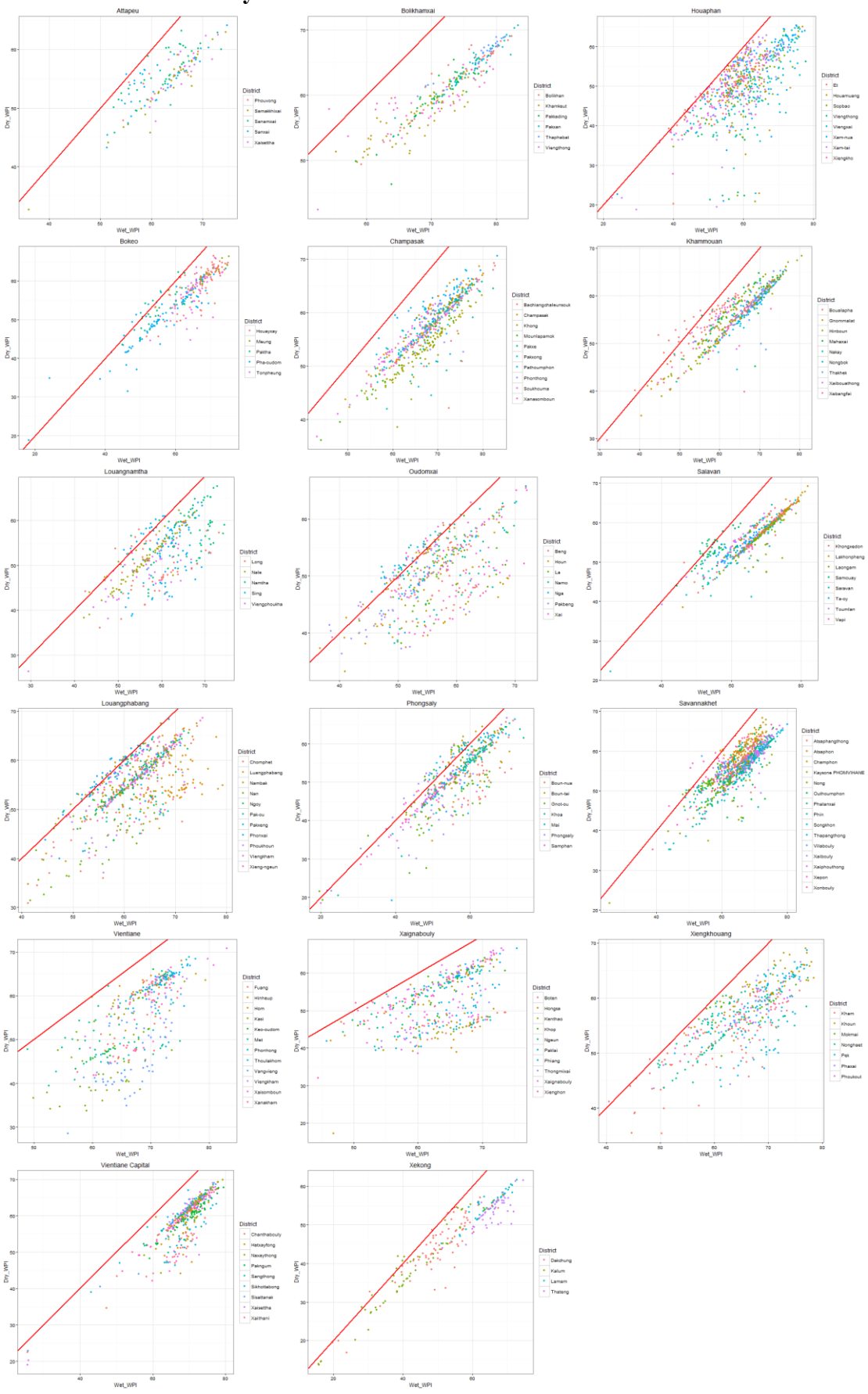
Province	Min.	1st Qu.	Median	3rd Qu.	Max.	Mean
Attapeu	36.04	58.74	63.53	66.57	74.75	62.75
Bokeo	18.17	55.72	64.19	68.23	75.25	62.05
Bolikhamxai	52.45	68.61	73.58	77.27	83.37	72.51
Champasak	43.15	64.13	68.5	72.8	83.22	68.07
Houaphan	20.99	54.97	59.78	64.82	77.79	59.41
Khammouan	31.81	59.19	64.56	69.4	80.6	63.68
Louangnamtha	29.4	55.83	60.08	64.47	74.4	59.86
Louangphabang	41.27	57.92	62.53	66.49	79.23	62.04
Oudomxai	36.63	51.95	56.15	61.07	72.04	56.41
Phongsaly	19.63	50.38	56.53	61.94	74.45	55.72
Salavan	25.35	62.05	67.48	71.86	81.93	66.41
Savannakhet	25.55	61.32	66.33	69.82	79.96	65.19
Vientiane	48.9	64.28	69.05	71.97	83.08	67.87
Vientiane Capital	25.54	67.12	70.06	72.23	79.54	69.1
Xayabouly	44.46	59.48	64.33	68.04	75.34	63.32
Xekong	15.68	47.6	56.65	64.95	74.52	54.42
Xiengkhouang	40.49	60.51	66.18	70.46	78.35	65.24

Parallel Coordinate Plot of Wet Season WPI Components for Each Province

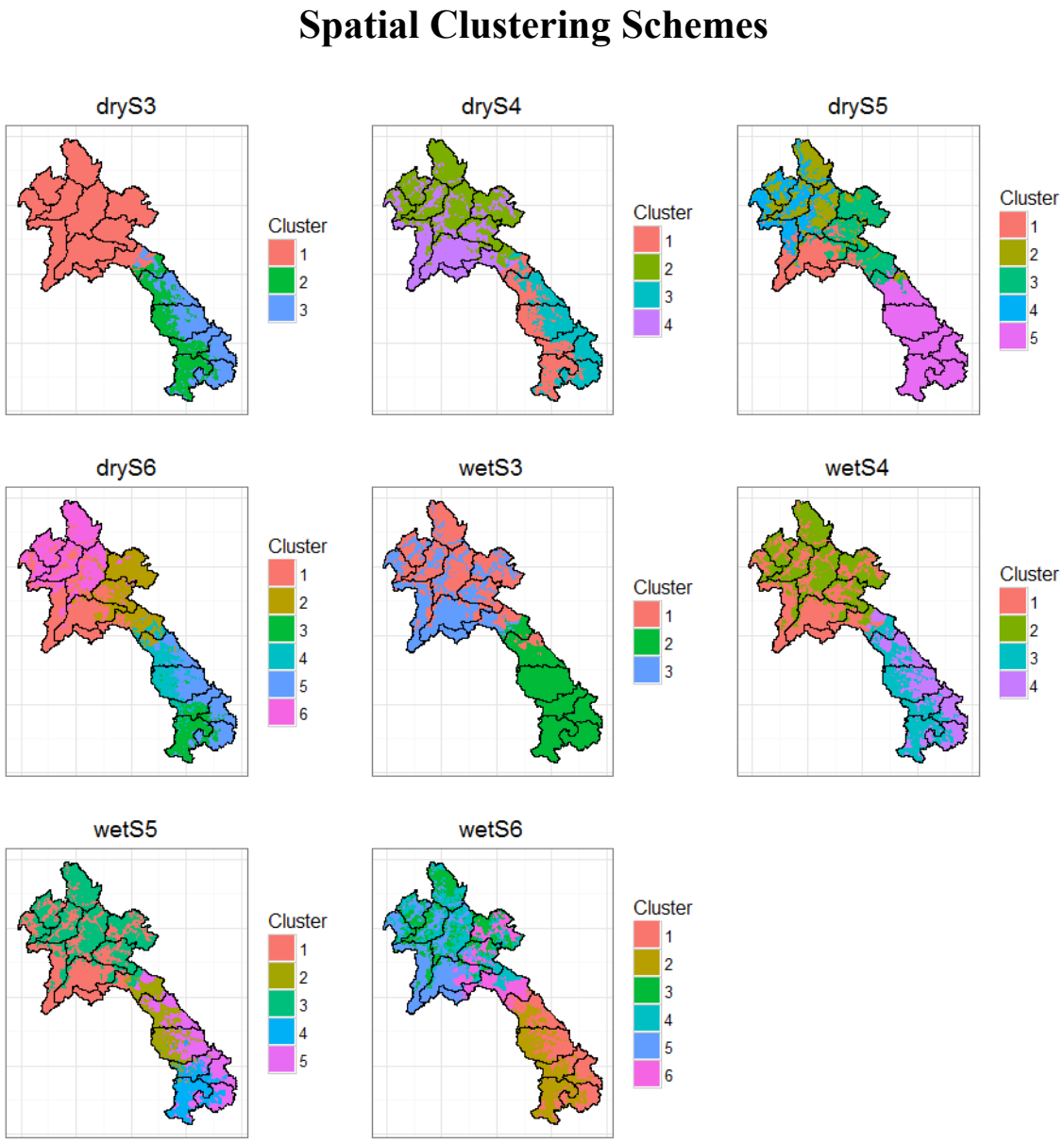


Comparison

Dry vs. Wet Season WPI for Each Province

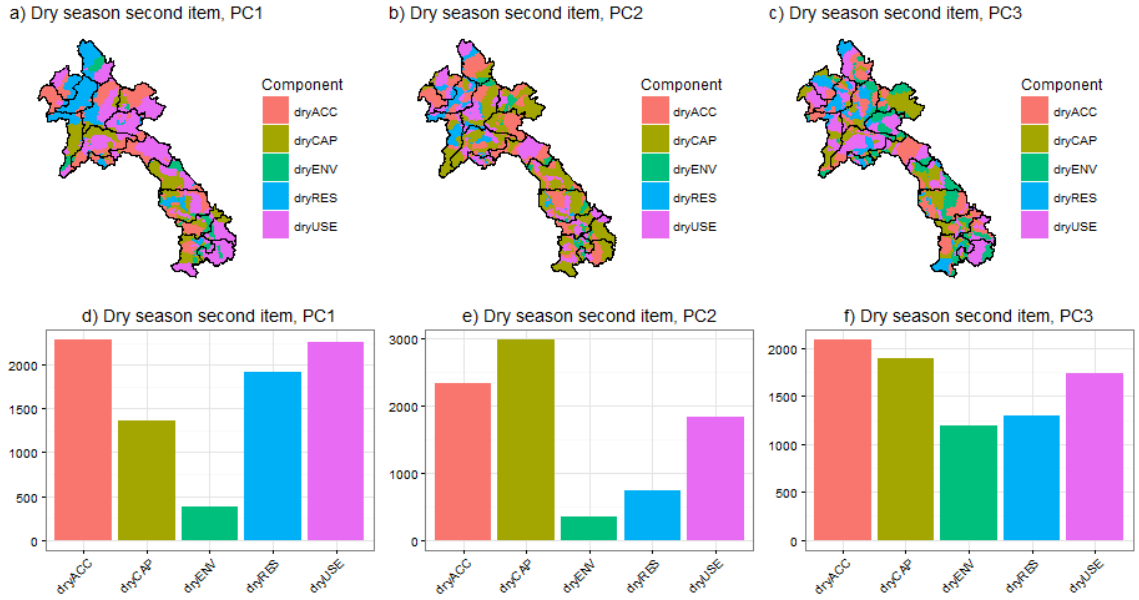


Appendix 4. Additional Data for Cluster Analysis

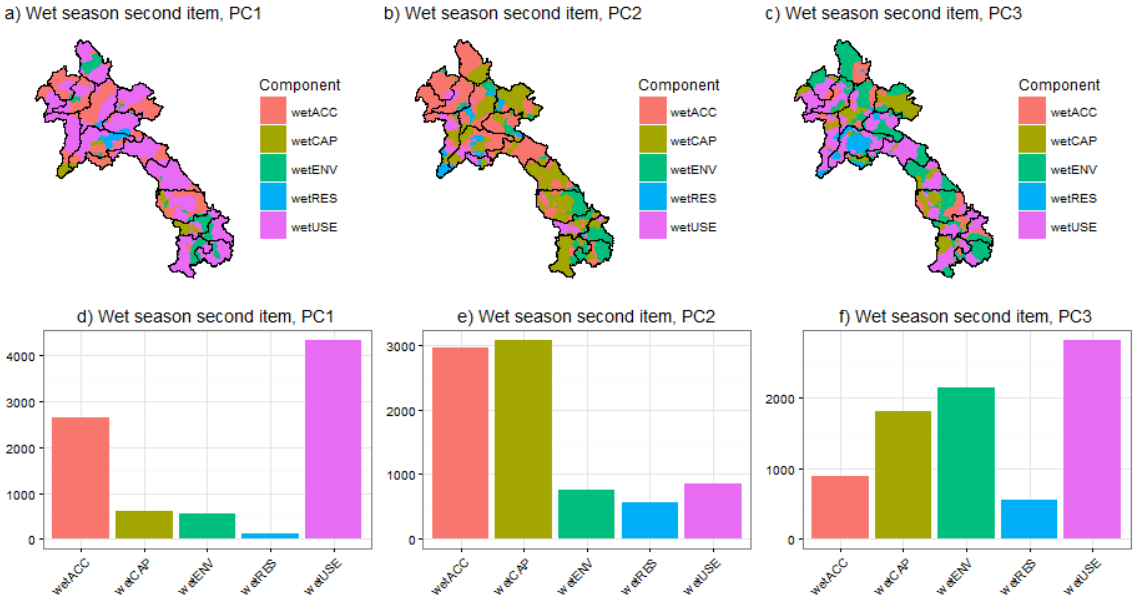


Appendix 5. Additional Data for Geographically Weighted Principal Component Analysis

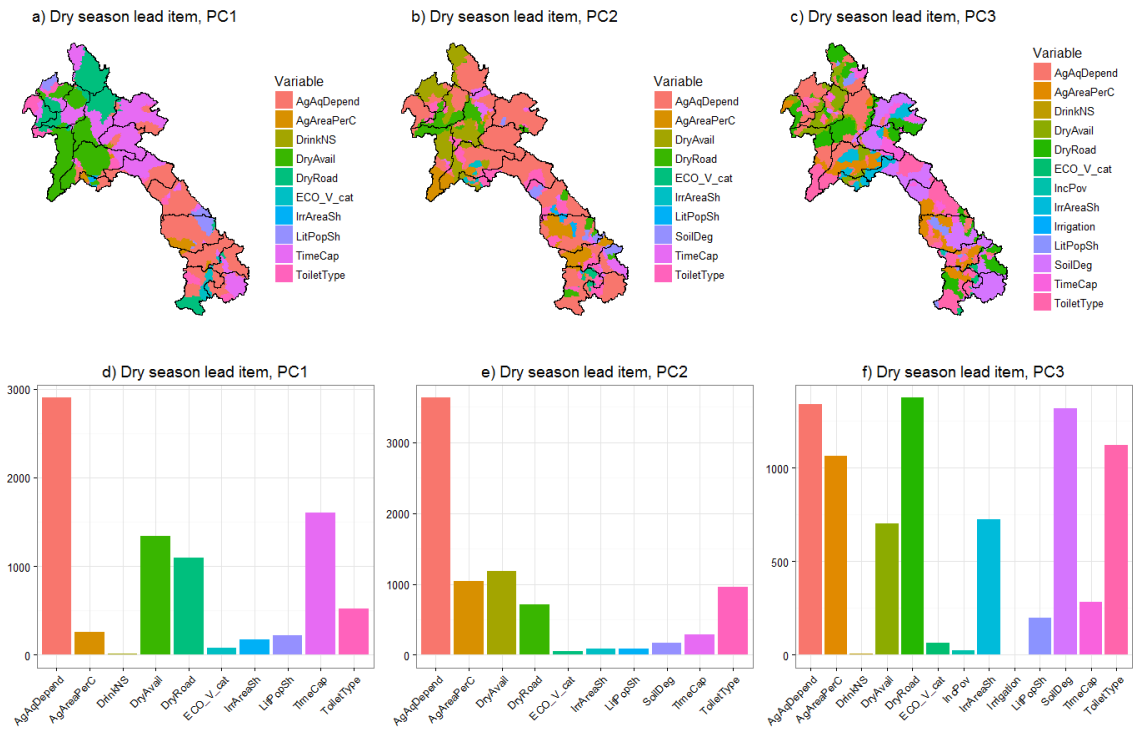
Dry season GWPCA second highest loaded *components*



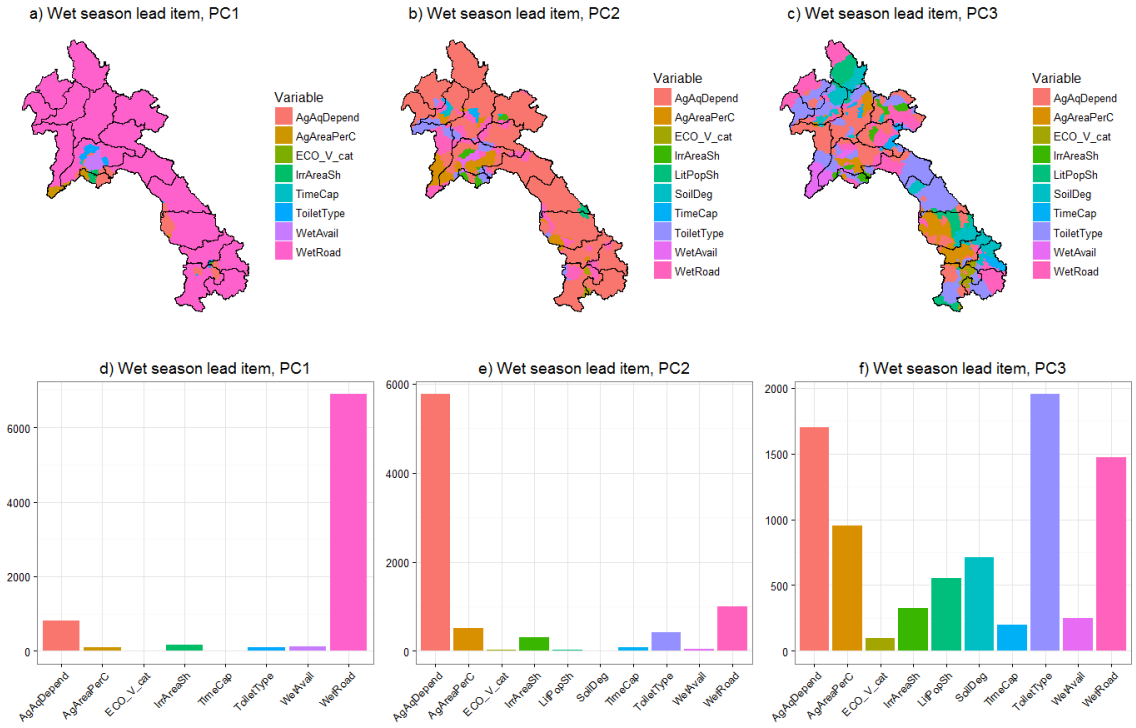
Wet season GWPCA second highest loaded *components*



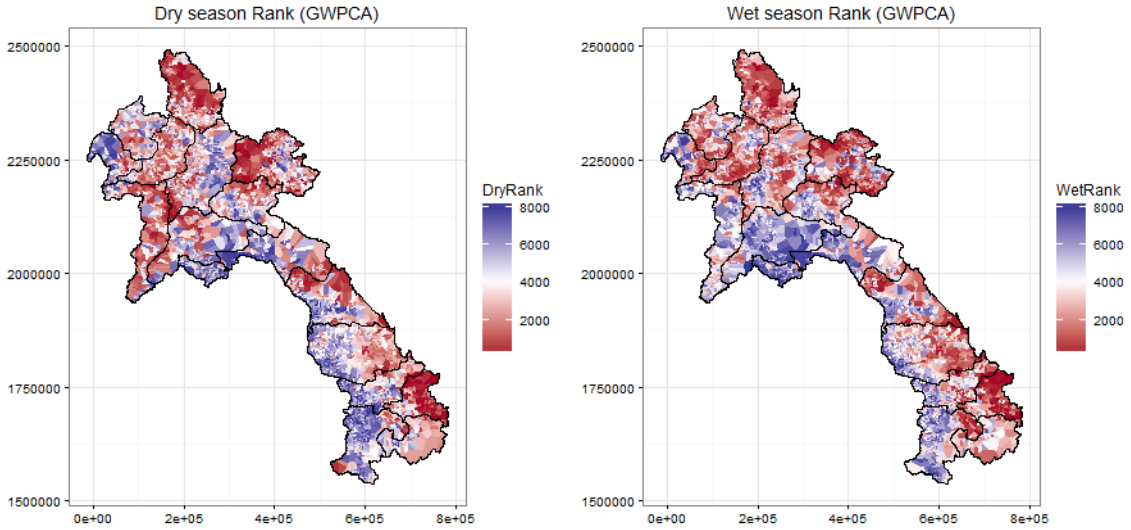
Dry season GWPCA highest loaded *variables*



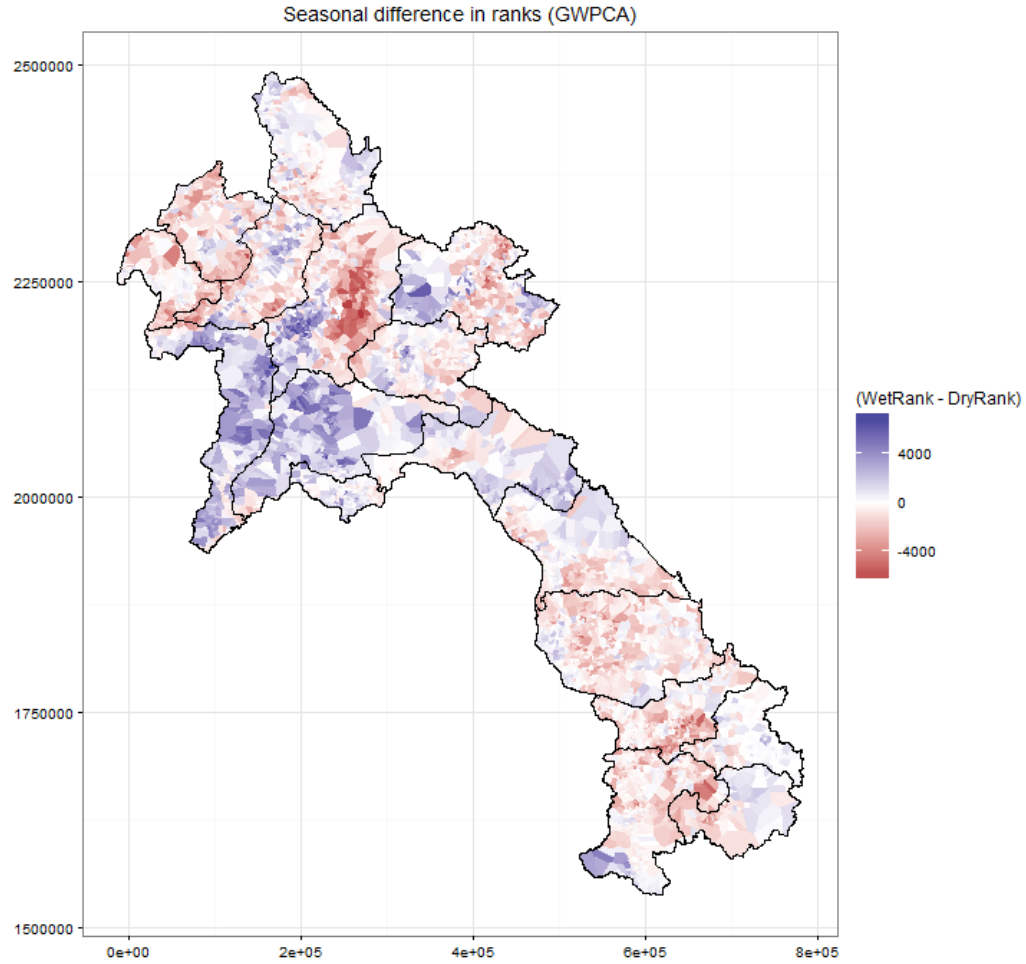
Wet season GWPCA highest loaded *variables*



GWPCA Weighted WPI Ranks

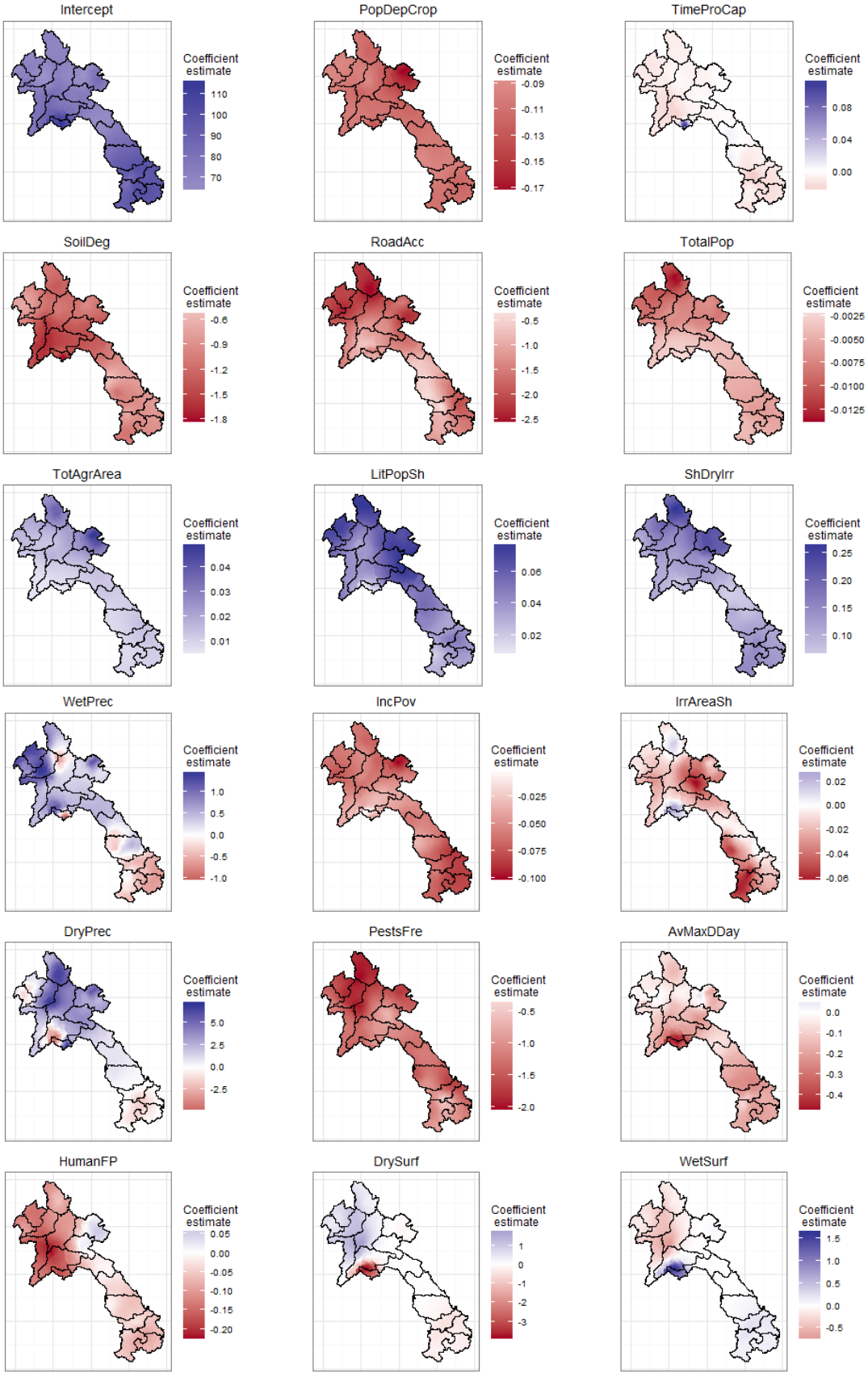


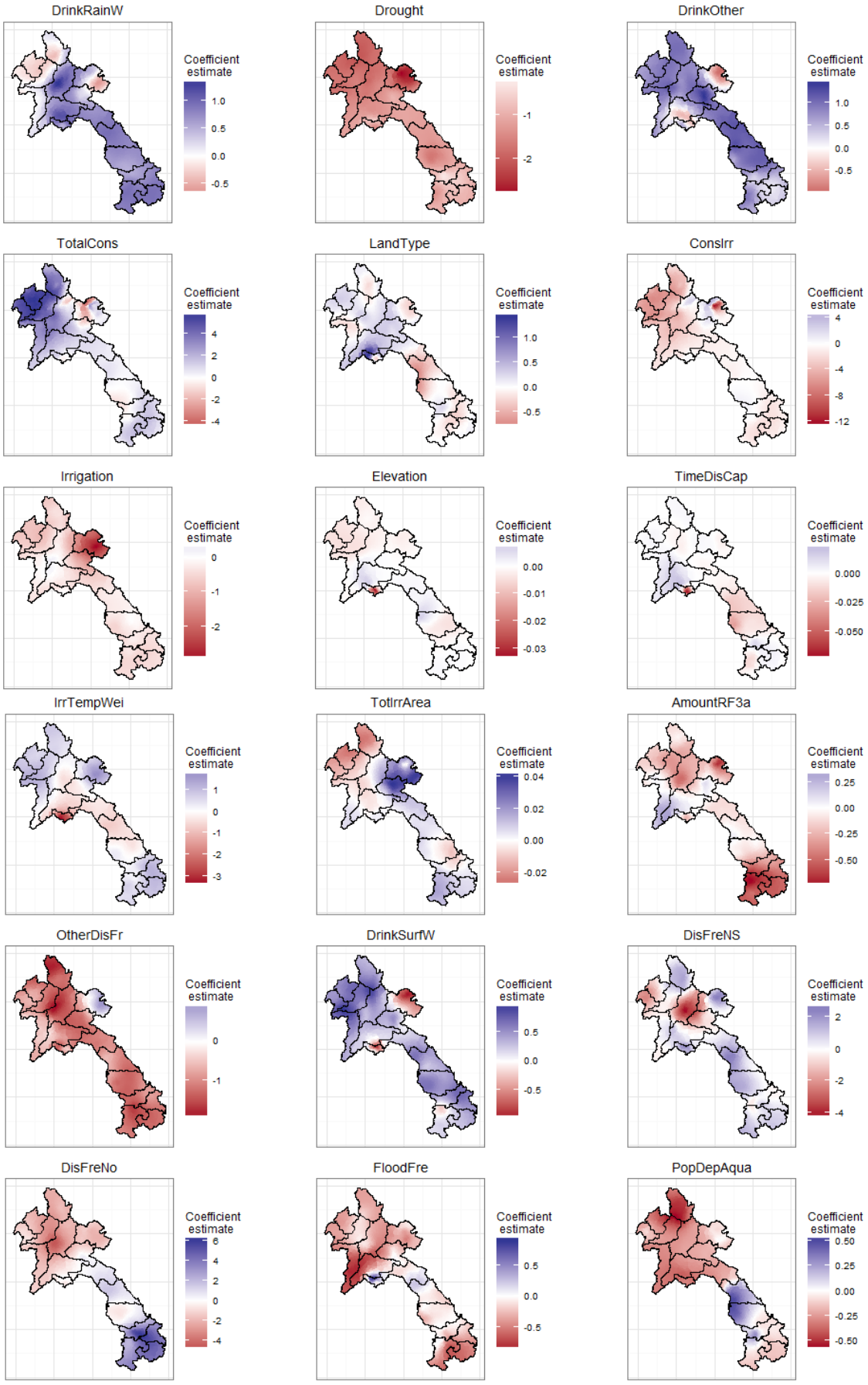
Seasonal Difference Between GWPCA Derived WPI Ranks

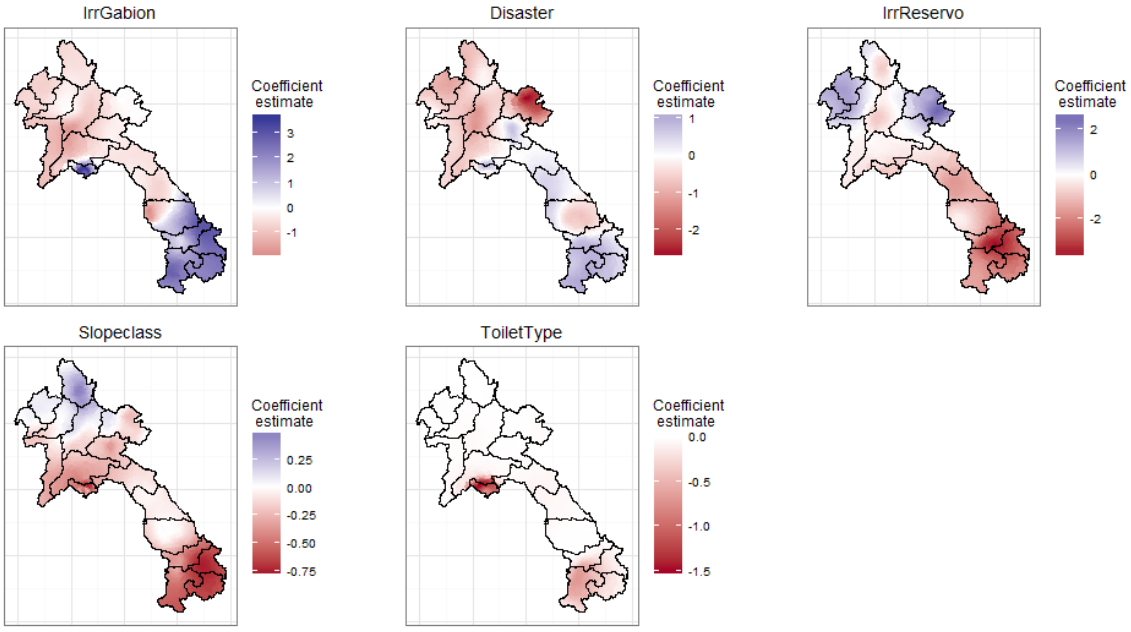


Appendix 6. Additional Data for Geographically Weighted Regression

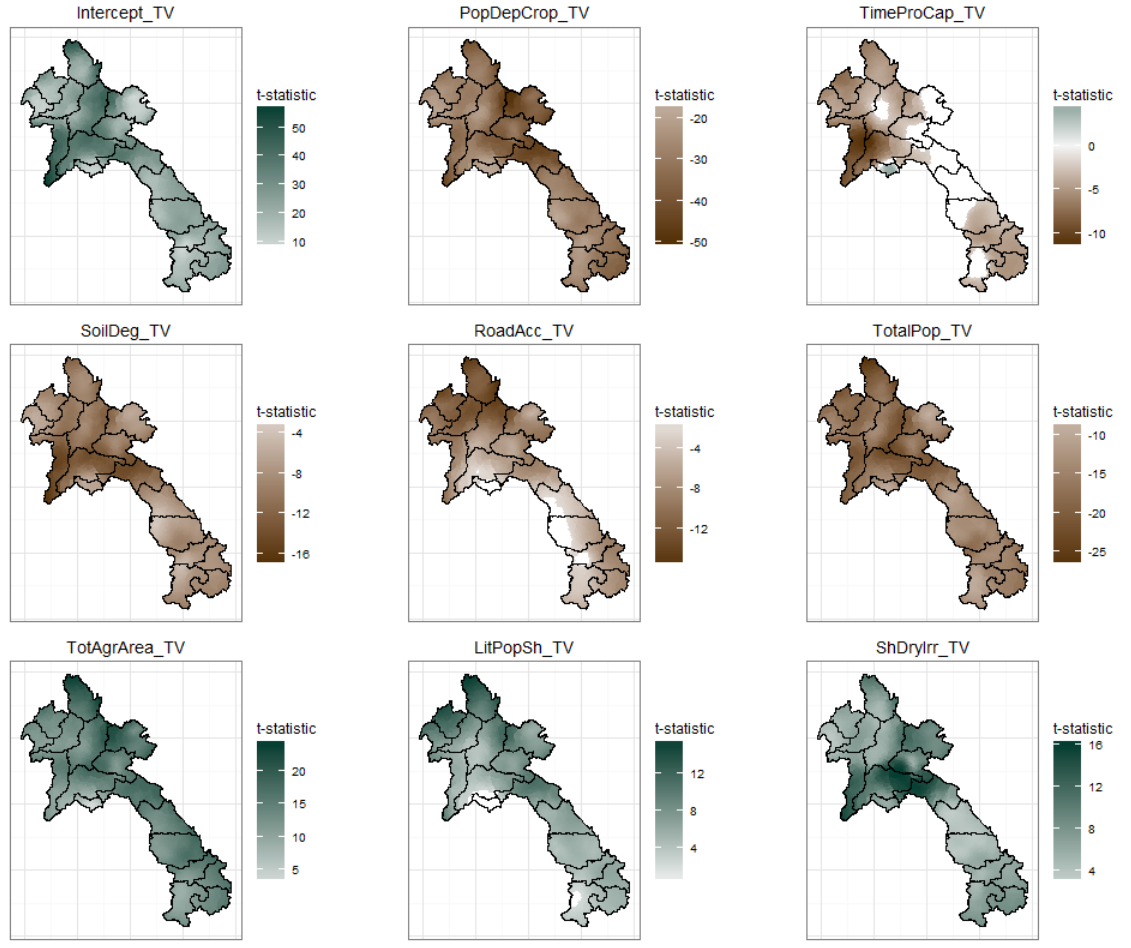
Dry Season GWR Coefficient Estimates

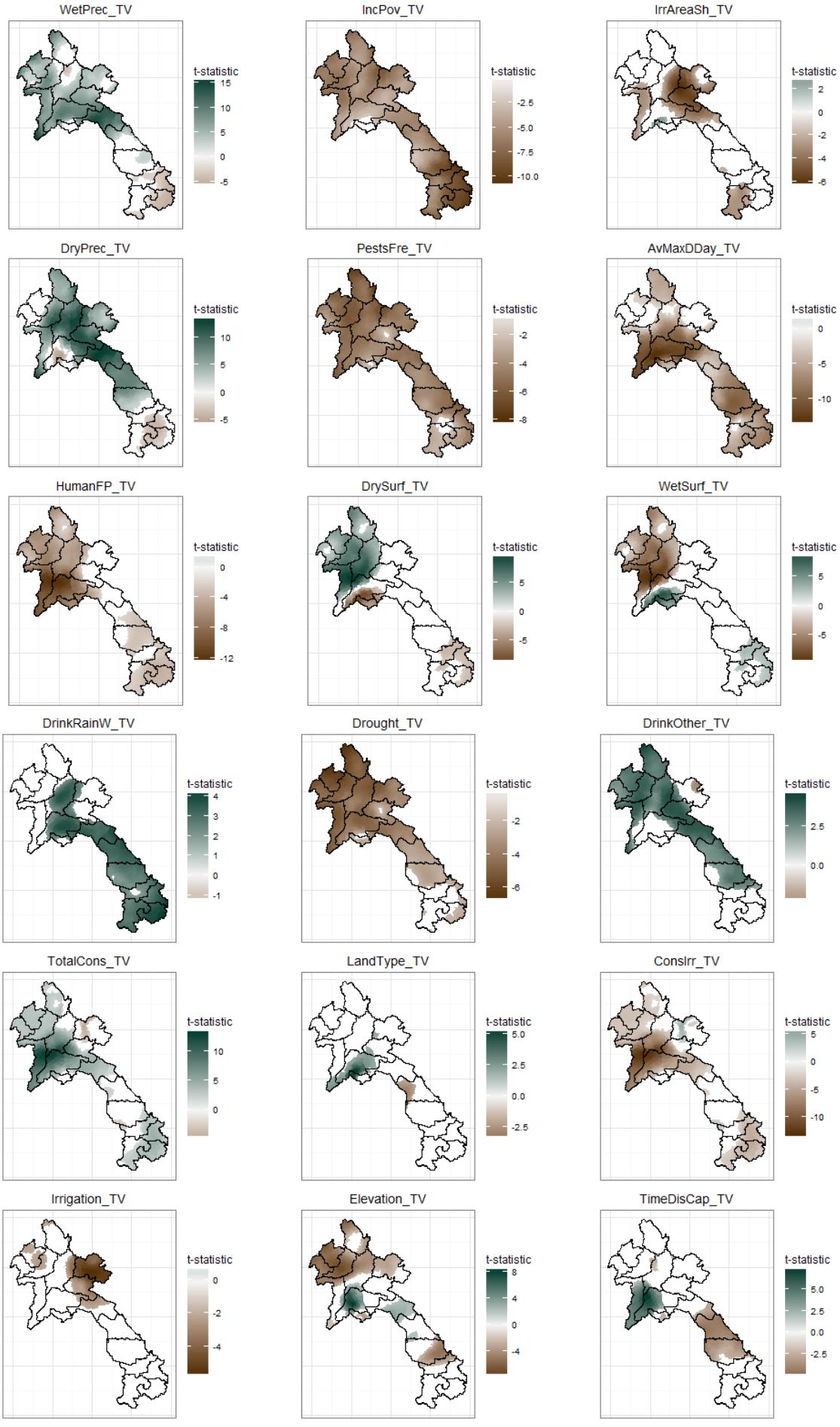


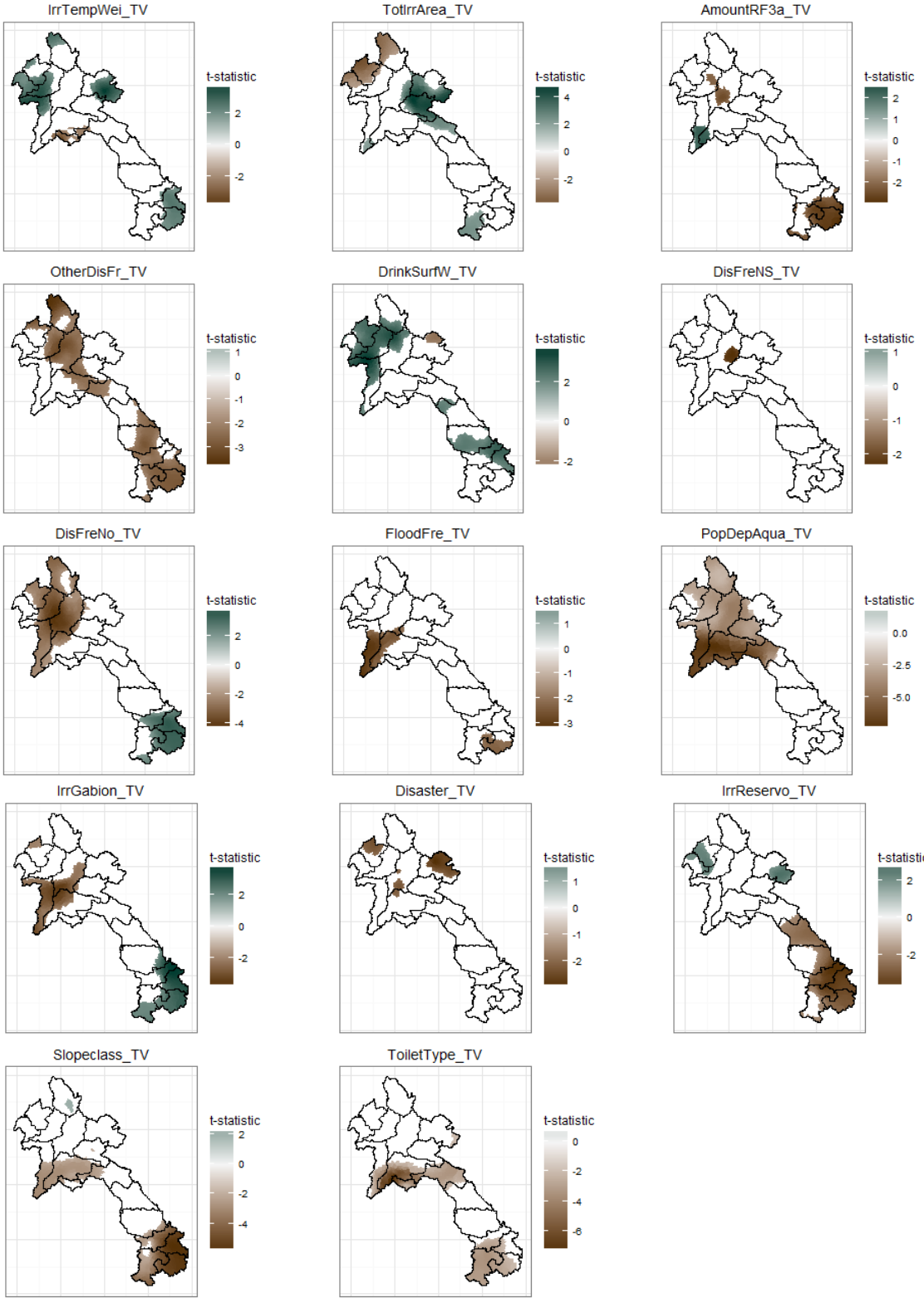

DrinkRainW
Coefficient estimate
1.0
0.5
0.0
-0.5
Drought
Coefficient estimate
-1
-2
DrinkOther
Coefficient estimate
1.0
0.5
0.0
-0.5
TotalCons
Coefficient estimate
4
2
0
-2
-4
LandType
Coefficient estimate
1.0
0.5
0.0
-0.5
Conslrr
Coefficient estimate
4
0
-4
-8
-12
Irrigation
Coefficient estimate
0
-1
-2
Elevation
Coefficient estimate
0.00
-0.01
-0.02
-0.03
TimeDisCap
Coefficient estimate
0.000
-0.025
-0.050
IrrTempWei
Coefficient estimate
1
0
-1
-2
-3
TotIrrArea
Coefficient estimate
0.04
0.02
0.00
-0.02
AmountRF3a
Coefficient estimate
0.25
0.00
-0.25
-0.50
OtherDisFr
Coefficient estimate
0
-1
DrinkSurfW
Coefficient estimate
0.5
0.0
-0.5
DisFreNS
Coefficient estimate
2
0
-2
-4
DisFreNo
Coefficient estimate
6
4
2
0
-2
-4
FloodFre
Coefficient estimate
0.5
0.0
-0.5
PopDepAqua
Coefficient estimate
0.50
0.25
0.00
-0.25
-0.50



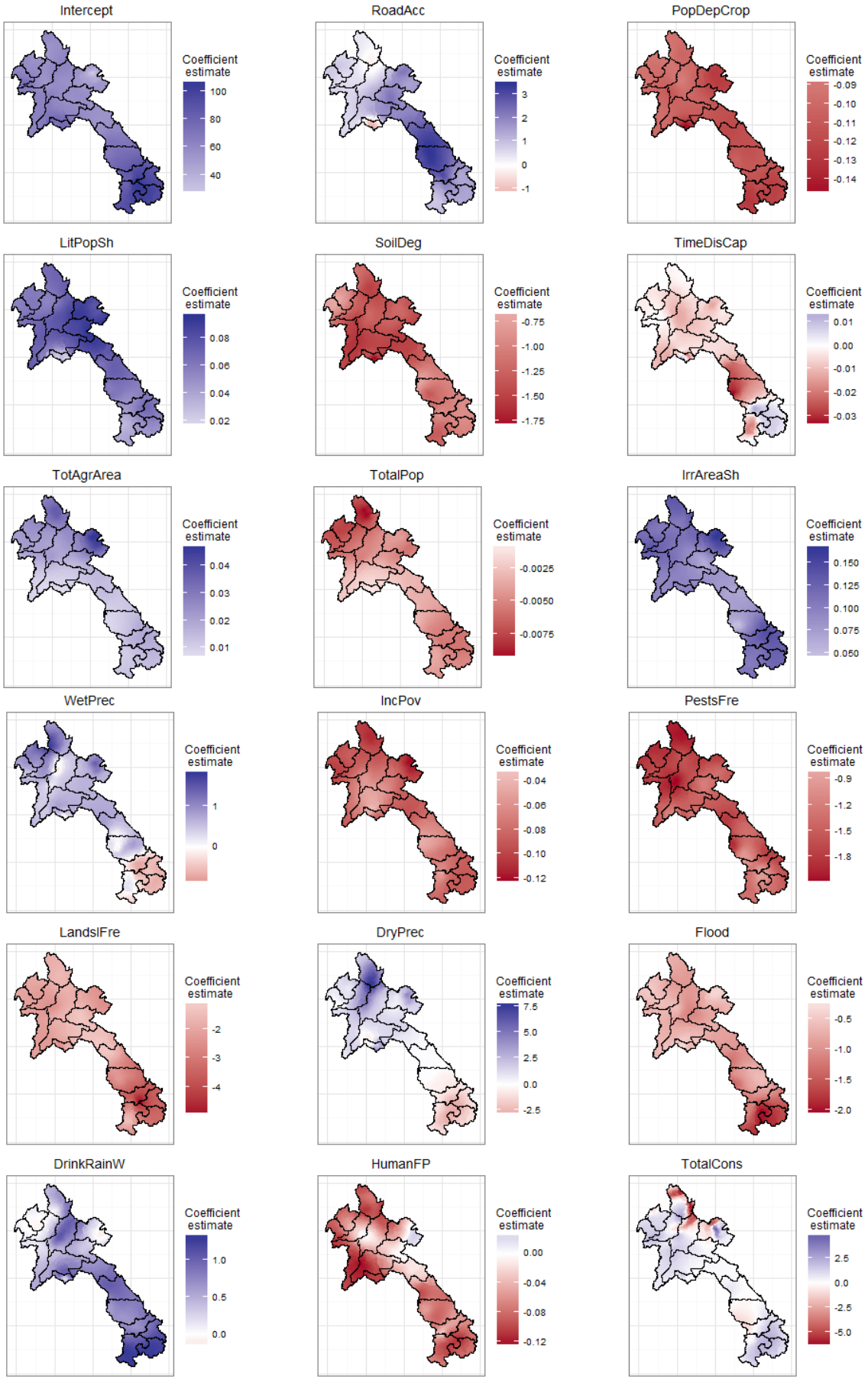
Dry Season Coefficient t-values

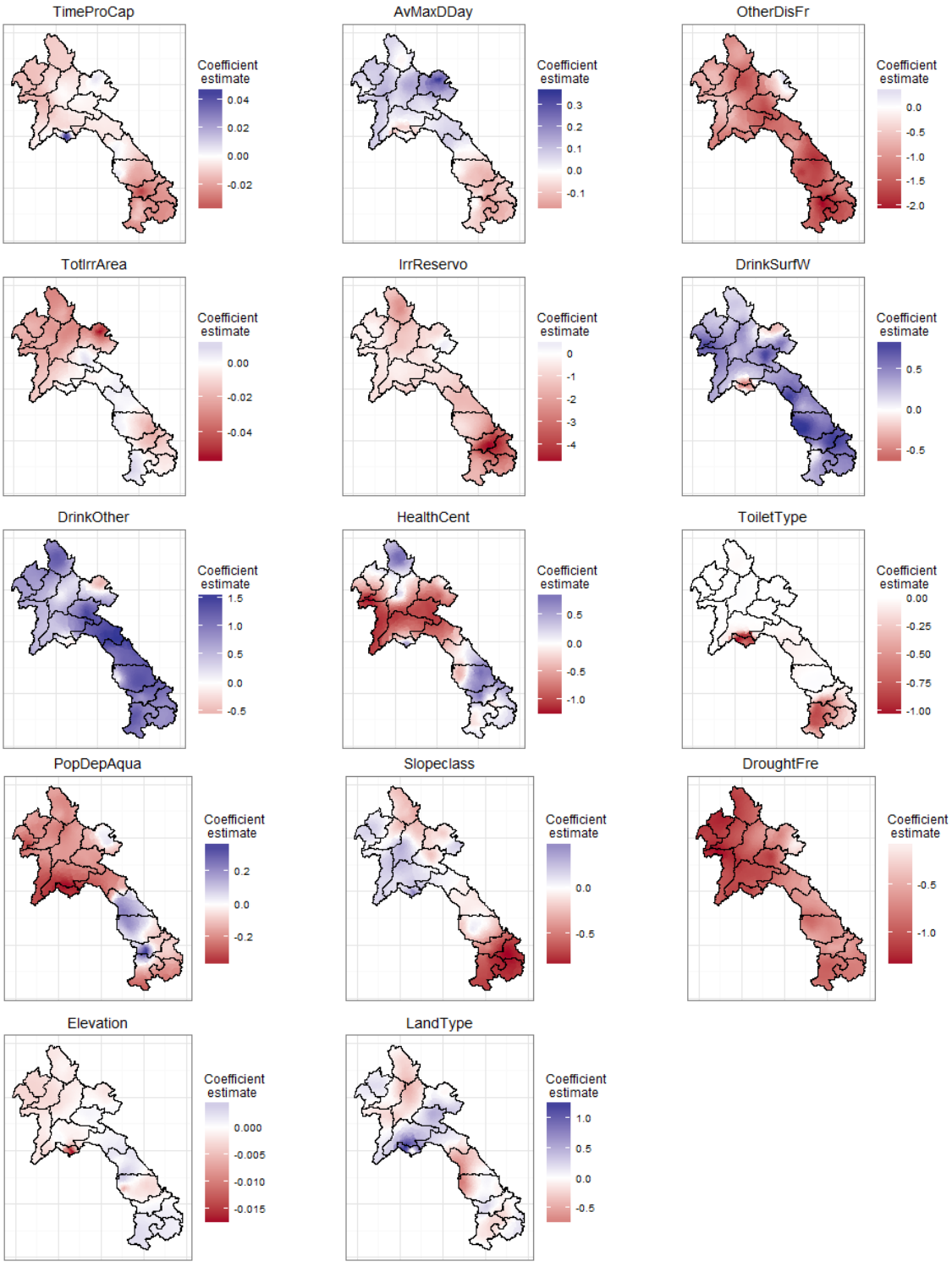



WetPrec_TV
t-statistic
15
10
5
0
-5
IncPov_TV
t-statistic
-2.5
-5.0
-7.5
-10.0
IrrAreaSh_TV
t-statistic
2
0
-2
-4
-6
DryPrec_TV
t-statistic
10
5
0
-5
PestsFre_TV
t-statistic
-2
-4
-6
-8
AvMaxDDay_TV
t-statistic
0
-5
-10
HumanFP_TV
t-statistic
0
-4
-8
-12
DrySurf_TV
t-statistic
5
0
-5
WetSurf_TV
t-statistic
5
0
-5
DrinkRainW_TV
t-statistic
4
3
2
1
0
-1
Drought_TV
t-statistic
-2
-4
-6
DrinkOther_TV
t-statistic
2.5
0.0
TotalCons_TV
t-statistic
10
5
0
LandType_TV
t-statistic
5.0
2.5
0.0
-2.5
ConsIrr_TV
t-statistic
5
0
-5
-10
Irrigation_TV
t-statistic
0
-2
-4
Elevation_TV
t-statistic
8
4
0
-4
TimeDisCap_TV
t-statistic
5.0
2.5
0.0
-2.5

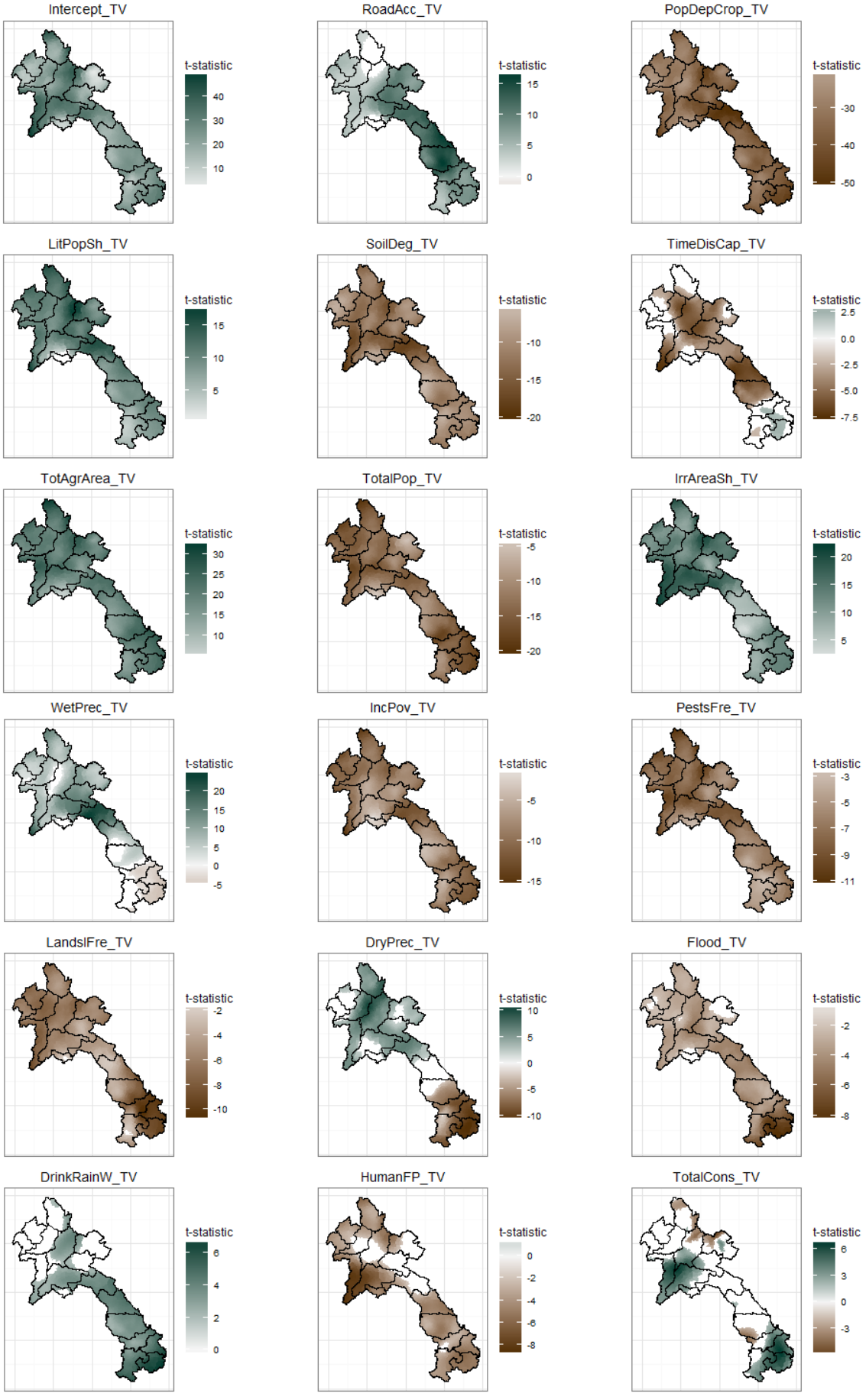

IrrTempWei_TV
t-statistic
2
0
-2
TotIrrArea_TV
t-statistic
4
2
0
-2
AmountRF3a_TV
t-statistic
2
1
0
-1
-2
OtherDisFr_TV
t-statistic
1
0
-1
-2
-3
DrinkSurfW_TV
t-statistic
2
0
-2
DisFreNS_TV
t-statistic
1
0
-1
-2
DisFreNo_TV
t-statistic
2
0
-2
-4
FloodFre_TV
t-statistic
1
0
-1
-2
-3
PopDepAqua_TV
t-statistic
0.0
-2.5
-5.0
IrrGabion_TV
t-statistic
2
0
-2
Disaster_TV
t-statistic
1
0
-1
-2
IrrReservo_TV
t-statistic
2
0
-2
Slopeclass_TV
t-statistic
2
0
-2
-4
ToiletType_TV
t-statistic
0
-2
-4
-6

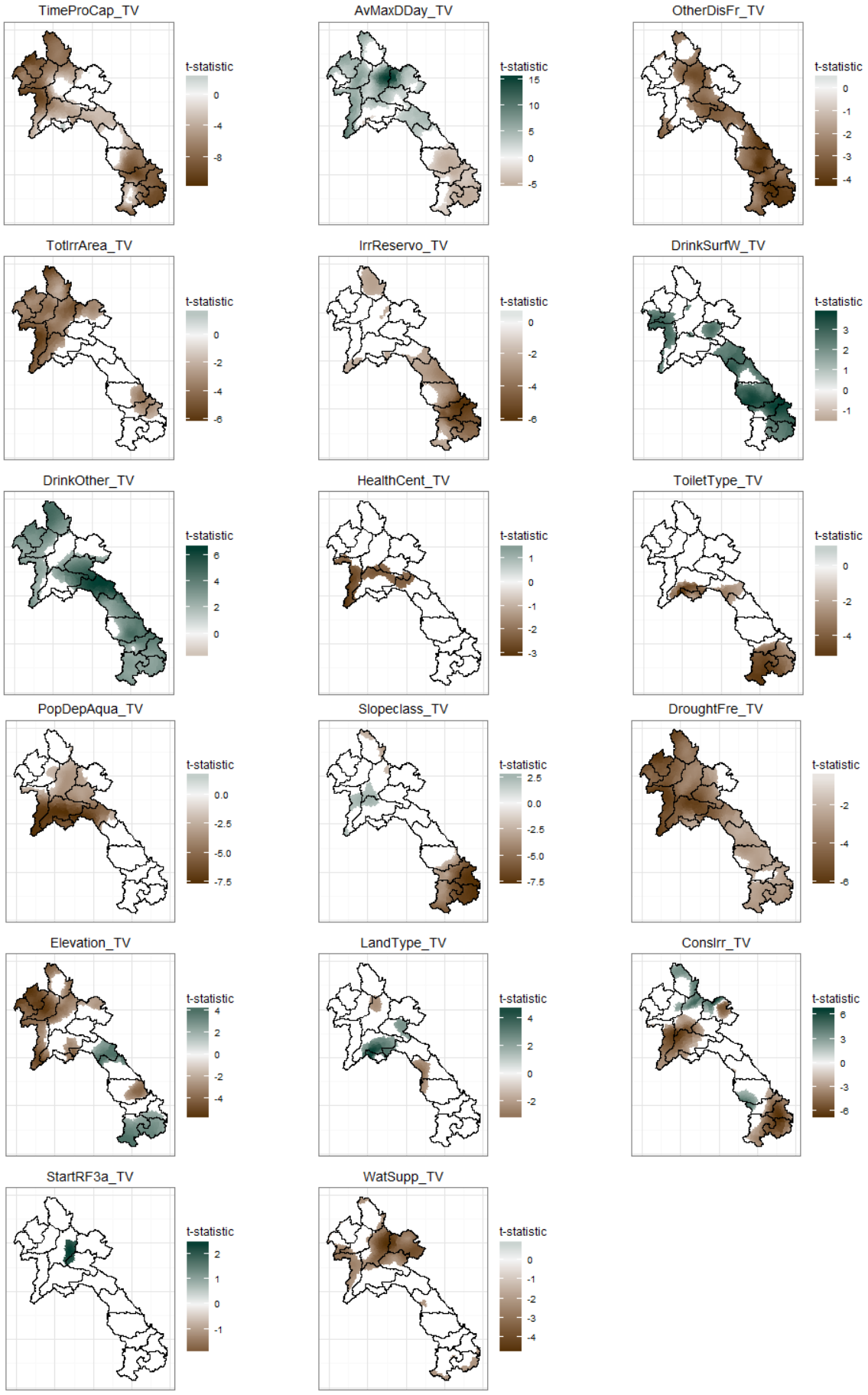
Wet season GWR Coefficient Estimates




TimeProCap
Coefficient estimate
0.04
0.02
0.00
-0.02
AvMaxDDay
Coefficient estimate
0.3
0.2
0.1
0.0
-0.1
OtherDisFr
Coefficient estimate
0.0
-0.5
-1.0
-1.5
-2.0
TotIrrArea
Coefficient estimate
0.00
-0.02
-0.04
IrrReservo
Coefficient estimate
0
-1
-2
-3
-4
DrinkSurfW
Coefficient estimate
0.5
0.0
-0.5
DrinkOther
Coefficient estimate
1.5
1.0
0.5
0.0
-0.5
HealthCent
Coefficient estimate
0.5
0.0
-0.5
-1.0
ToiletType
Coefficient estimate
0.00
-0.25
-0.50
-0.75
-1.00
PopDepAqua
Coefficient estimate
0.2
0.0
-0.2
Slopeclass
Coefficient estimate
0.0
-0.5
DroughtFre
Coefficient estimate
-0.5
-1.0
Elevation
Coefficient estimate
0.000
-0.005
-0.010
-0.015
LandType
Coefficient estimate
1.0
0.5
0.0
-0.5

Wet Season Coefficient t-values




TimeProCap_TV
t-statistic
0
-4
-8
AvMaxDDay_TV
t-statistic
15
10
5
0
-5
OtherDisFr_TV
t-statistic
0
-1
-2
-3
-4
TotIrrArea_TV
t-statistic
0
-2
-4
-6
IrrReservo_TV
t-statistic
0
-2
-4
-6
DrinkSurfW_TV
t-statistic
3
2
1
0
-1
DrinkOther_TV
t-statistic
6
4
2
0
HealthCent_TV
t-statistic
1
0
-1
-2
-3
ToiletType_TV
t-statistic
0
-2
-4
PopDepAqua_TV
t-statistic
0.0
-2.5
-5.0
-7.5
Slopeclass_TV
t-statistic
2.5
0.0
-2.5
-5.0
-7.5
DroughtFre_TV
t-statistic
-2
-4
-6
Elevation_TV
t-statistic
4
2
0
-2
-4
LandType_TV
t-statistic
4
2
0
-2
ConsIrr_TV
t-statistic
6
3
0
-3
-6
StartRF3a_TV
t-statistic
2
1
0
-1
WatSupp_TV
t-statistic
0
-1
-2
-3
-4

Variables Used in Step-wise Model Selection

	GWR Variables		
	Variable	Name	Source
1	LandType	Land type (lowland, upland, plateau, or other)	Agricultural Census 2010/2011
2	NormRF3a	Normal rainfall last 3 years (< 10 years ago, the same, > 10 y ago)	Agricultural Census 2010/2011
3	StartRF3a	Onset of rainfall last 3 years compared to 10 years ago (earlier, same, later)	Agricultural Census 2010/2011
4	AmountRF3a	Rainfall in 2010 compared to 10 years ago (less, same, more)	Agricultural Census 2010/2011
5	SoilDeg	Degree of soil degradation (none, light, moderate, severe, don't know)	Agricultural Census 2010/2011
6	Disaster	Natural disaster occur	Agricultural Census 2010/2011
7	Drought	Droughts occur	Agricultural Census 2010/2011
8	Flood	Floods occur	Agricultural Census 2010/2011
9	Landslide	Landslides occur	Agricultural Census 2010/2011
10	Pests	Pests occur	Agricultural Census 2010/2011
11	OtherDis	Other disasters occur	Agricultural Census 2010/2011
12	DisastNS	Disaster type not specified	Agricultural Census 2010/2011
13	FloodFre	Floods occur every 1-2 years	Agricultural Census 2010/2011
14	DroughtFre	Droughts occur every 1-2 years	Agricultural Census 2010/2011
15	LandslFre	Landslides occur every 1-2 years	Agricultural Census 2010/2011
16	OtherDisFr	Other natural disasters occur every 1-2 years	Agricultural Census 2010/2011
17	DisFreNo	No frequent natural disasters occur	Agricultural Census 2010/2011
18	Irrigation	Irrigation facilities present in the village	Agricultural Census 2010/2011
19	IrrPermWei	Type of irrigation facility: permanent weir	Agricultural Census 2010/2011
20	IrrReservo	Type of irrigation facility: reservoir	Agricultural Census 2010/2011
21	IrrPump	Type of irrigation facility: pump scheme	Agricultural Census 2010/2011
22	IrrTempWei	Type of irrigation facility: temporary weir	Agricultural Census 2010/2011
23	IrrGabion	Type of irrigation facility: gabions	Agricultural Census 2010/2011
24	IrrOther	Type of irrigation facility: other irrigation facilities	Agricultural Census 2010/2011
25	IrrNS	Irrigation type not specified	Agricultural Census 2010/2011
26	HealthCent	Health center present in the village	Population Census 2005
27	TWalkHC	Hours walk to nearest dispensary or hospital (<2h, >2h)	Agricultural Census 2010/2011
	RoadAcc	Seasons that the road is accessible (dry, all year, no road)	Population Census 2005
28	DrinkPipe	Drinking water source: piped	Agricultural Census 2010/2011
29	DrinkPrWel	Drinking water source: protected well/bore	Agricultural Census 2010/2011
30	DrinkUnprW	Drinking water source: unprotected well/bore	Agricultural Census 2010/2011

31	DrinkSurfW	Drinking water source: river/stream/dam	Agricultural Census 2010/2011
32	DrinkRainW	Drinking water source: rainwater from tank	Agricultural Census 2010/2011
33	DrinkOther	Drinking water source: other source of drinking water	Agricultural Census 2010/2011
34	DrinkNS	Drinking water source not specified	Agricultural Census 2010/2011
35	Drink70Pip	Over 70% of households have piped water	Agricultural Census 2010/2011
36	DrinkSourc	Main source of drinking water in the village	Agricultural Census 2010/2011
37	TimeProCap	Mean travel time (min) to province capital	Population Census 2005
38	TimeDisCap	Mean travel time (min) to district capital	Population Census 2005
39	WatSupp	Village(s) with water supply	Population Census 2005
40	LitPopSh	Percentage of literate population	Population Census 2005
41	IncPov	Incidence of poverty	Population Census 2005
42	ToiletType	Main type of toilet	Population Census 2005
43	TotAgrArea	Total area of agriculturally used land	Agricultural Census 2010/2011
44	TotIrrArea	Total irrigated area	Agricultural Census 2010/2011
45	IrrAreaSh	Share of agricultural land irrigated	Agricultural Census 2010/2011
46	AgrPop	Agricultural Population	Agricultural Census 2010/2011
47	PopDepCrop	Population with main income from Crops	Agricultural Census 2010/2011
48	PopDepAqua	Population with main income from aquaculture	Agricultural Census 2010/2011
49	TotalPop	Total Village Population	Population Census 2005
50	PopElec	Share of population with Electricity	Population Census 2005
52	TotalCons	Total water consumption	GWSP Atlas
53	ShDryIrr	Share of agricultural area irrigated in the dry season	Agricultural Census 2010/2011
54	ConsIrr	Irrigation water consumption	Agricultural Census 2010/2011
55	AvMaxDDay	Average length of the longest no-rain sequence in the dry season	Calculated and interpolated from historical records
56	Elevation	Mean elevation in 5kmx5km grid cell	Harmonized World Soil Database
57	SlopeClass	Slope class in a 5kmxkm grid cell	Harmonized World Soil Database
58	HumanFP	Human footprint aggregated to 5kmx5km grid cell	SEDAC Last of the Wild v2