# EVALUATION OF SPATIAL PREDICTION METHODS FOR SELECTED SOIL PROPERTIES TO SUPPORT LAND SUITABILITY MAPPING FOR RICE IN ZAMBIA

By

MIRRIAM MAKUNGWE

A thesis submitted to the University of Zambia in fulfilment of the requirements for the degree of Doctor of Philosophy in Integrated Soil Fertility Management

THE UNIVERSITY OF ZAMBIA

LUSAKA

2021

## DECLARATION

I, **Mirriam Makungwe**, hereby declare that all the work presented in this dissertation is my own and has never been submitted for a degree at this or any other University.

Some parts of this work have been published as journal articles (Makungwe et al., 2021a) in Geoderma titled "*Performance of linear mixed models and random forests for spatial interpolation of soil pH*", https://doi.org/10.1016/j.geoderma. 2021.115079 and (Makungwe et al., 2021b) in Geoderma Regional titled "*Assessing land suitability for rainfed paddy rice production in Zambia*", https://doi.org/10.1016/j.geodrs.2021.e00438

**……..** **……**

**Mirriam Makungwe** **Date**

## APPROVAL PAGE

This Thesis of **Mirriam Makungwe** is approved as fulfilling part of the requirement for the award of the degree of Doctor of Philosophy in Integrated Soil Fertility Management by the University of Zambia.

| **Examiner's Name** | **Signature** | **Date** |
|---|---|---|
| 1. …… | ……. | …… |
| 2. …… | ……. | ……. |
| 3. …….. | ……. | ……. |
| **Principal Supervisor** | | |
| 4. …… | ……. | ……. |
| **Board of Examiners' Chairperson** | | |
| 5. …….. | ……. | ……. |

## ABSTRACT

Rice is one of the staple food crops and is a profitable smallholder cash crop in Zambia. It has the potential to contribute to increased incomes and employment among rural producers. However, rice is the only staple crop for which domestic production does not meet domestic demand due to low productivity, among other factors. One step towards addressing this problem is the identification of land with greatest potential for production. This can be done through a land suitability evaluation. This study focuses on how to map the spatial variation of selected soil properties across Zambia to support evaluation of land suitability for rice production. When mapping the spatial variation of selected soil properties, legacy data on the target variable were available along with additional environmental covariates as predictor variables. The options were to undertake spatial prediction by geostatistical or machine learning methods. Also addressed was how to robustly validate models from legacy data when these have, as is often the case, a strongly clustered spatial distribution. The validation statistics results showed that the empirical best linear unbiased predictor (EBLUP) with the only fixed effect a constant mean performed better than the other methods used for predicting soil pH and the EBLUP with fixed effect performed better than other methods used for predicting soil organic carbon and soil Phosphorus. Random forests had the largest model-based estimates of the expected squared errors in all predictions. It was observed that the random forest algorithm was prone to select as "important" spatially correlated simulated random variables. The maps produced using the best performing methods were used as factors in land suitability assessment of paddy and upland rice under rainfed and irrigation conditions. Land suitability was evaluated while accounting for important multiple factors, and which considers their joint effect of a hierarchical model of constraints. The suitability classes were ranked according to the FAO land suitability classification. Four land suitability maps where produced. Results showed that potential land classified as highly and moderately suitable was 27 percent for rainfed paddy rice, 29 percent for rainfed upland rice, 25 percent for irrigated paddy rice and 54 percent for irrigated upland rice. The results show limited potential for production of rainfed paddy, rainfed upland as well as irrigated paddy rice production but great potential for irrigated upland rice production. Therefore, irrigated upland rice production in Zambia would help expand the potential production area of rice.

## DEDICATION

To my husband Moses Tolopu, and Children Shamendi and Kuwunda Tolopu.

To my father, Joe Mwandawamufu Makungwe, and Mother, Lydia Mwangala Mwitumwa.

To the memory of my late grandmother, Florence N. Sitamulaho, and young sister, Walusiku Makungwe.

## ACKNOWLEDGEMENTS

I wish to express my gratitude to all those who made completion of this study possible.

My special thanks go to my husband, Moses Tolopu, my daughter, Shamendi Tolopu, and Son, Kuwunda Tolopu, for their support and patience while I spent many months away from home working on this study. This study could not have been easier without them.

Gratitude is due to my academic supervisors: Dr. Benson H. Chishala; Dr. Lydia Mumbi Chabala; and, Prof. R. Murray Lark, for their patience, encouragement, expert advice and for guiding me through the entire process.

This study was made possible with support from the Zambian Ministry of Higher Education through their 2016/2017 Science and Technology Female Postgraduate Scholarships.

Sincere appreciation to the International Institute for Applied Systems Analysis (IIASA) for allowing me to spend time at their institution in Austria and use their resources to work on some sections of this study through funding from the Global Environment Facility (GEF) and support of the United Nations Industrial Development Organization (UNIDO) as a part of the Integrated Solutions for Water, Energy, and Land (ISWEL) project (GEF Contract Agreement: 6993) that they implemented. Thank you to Dr. Michiel van Dijk for his patience, encouragement, expert advice while at IIASA.

Many thanks to the Commonwealth Scholarships Commission (CSC) through their split-site scholarship for enabling me to spend one year of my PhD study in UK at the University of Nottingham under the supervision of Professor R. Murray Lark.

Thank you to Indaba Agricultural Policy Research Institute (IAPRI) for allowing me to use their RALS 2012 survey crop and soils data.

The assistance of all other persons not mentioned by name is gratefully acknowledged.

Thank You.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS AND ACRONYMS

| Abbreviation | Definition |
|---|---|
| AIC | Akaike Information Criterion |
| CNBL | Channel Network Base Level |
| CSO | Central Statistical Office |
| DEM | Digital Elevation Model |
| EBLUP | Empirical Best Linear Unbiased Predictor |
| ESA | European Space Agency |
| ESE | Expected Square Error |
| EVI | Enhanced Vegetation Indexed |
| FAO | Food and Agriculture Organization |
| FDR | False Discovery Rate |
| IAPRI | Indaba Agricultural Policy Research Institute |
| IIASA | International Institute for Applied Systems Analysis |
| ILWIS | Integrated Land and Water Information System |
| JICA | Japan International Cooperation Agency |
| LMM | Linear Mixed Model |
| MCE | Multi-Criteria Evaluation |
| NDVI | Normalized Difference Vegetation Index |
| NERICA | New Rice for Africa |
| NRDS | National Rice Development Strategy |
| OOB | Out-Of-Bag |
| PSU | Primary Sampling Unit |
| RALS | Rural Agricultural Livelihoods Survey |
| REML | Residual Maximum Likelihood |
| RF | Random Forests |
| RI | Random Index |
| RSP | Relative Slope Position |
| SEA | Standard Enumeration Area |
| SSF | Soil and Site Factors |
| VIV | Variable Importance Value |
| ZARI | Zambia Agricultural Research Institute |

# CHAPTER ONE

## 1.0. INTRODUCTION

Singha and Swain (2016) described land suitability evaluation as a process of determining the appropriateness of land in a specific location for a particular use. The suitability of a particular area of land for a crop depends on various factors, some of which cannot feasibly be modified by management practices and so are absolute constraints. If we are to make effective use of land, then we need to analyse these requirements, and then identify which land uses are appropriate at some location of interest, or where land is suitable for particular uses of interest (Suheri et al., 2018; Agidew, 2015). Land suitability evaluation can identify constraints, opportunities and potential of the land resource for a given use (Mohammed, 2011; Mokarram and Aminzadeh, 1996). It also plays an important role in sustainable agricultural practice and management (Tanasă et al., 2010) as it provides information to farmers, extension staff, policy makers and other stakeholders on how suitable the land is in terms of soil limitations (Olaleye et al., 2002).

When carrying out land suitability evaluation, one of the major functional factors is soil information which is represented in soil maps (Costantini, 2009). These soil maps describe the spatial variation of soil types and provide important information on spatial variation of soil properties (Kempen et al., 2010). Mapping of soil properties is important as it provides policy makers with a synoptic view of the state of the soil, and agricultural stakeholders with information about where soil problems might occur (Lark et al., 2019). Soil maps are generated using various soil mapping methods which can be divided into conventional and pedometric approaches (Kienast-Brown et al., 2010; Hengl, 2003).

Conventional soil survey represents soil variation in terms of profile classes and corresponding map legend units. It can provide a basis for spatial prediction of soil properties and may also serve as a structure for recording substantial information on soil management and for systematizing knowledge of the distribution of soils in the landscape. Conventional approaches were based largely on manual processes which are costly and time consuming (Kienast-Brown et al., 2010) mainly because of long fieldwork periods (Moonjun et al., 2010). Pedometric approaches are based on the application of mathematical and statistical methods for the primary purpose of

predicting the values of soil properties where these have not been observed directly (McBratney et al., 2000). A well-established statistical approach to doing this is the application of model-based geostatistics (Stein, 1999; Diggle and Ribeiro, 2007). In this approach, the variation of the soil is represented in a linear mixed model (LMM) as a combination of fixed effects (which may be a constant unknown mean, or a function of predictive covariates such as remote sensor data), and random effects, including Gaussian random fields which exhibit spatial correlation. The parameters of the LMM model can be estimated by Residual Maximum Likelihood (REML) method developed by Patterson and Thompson (1971), which allows parameters of the random effects to be estimated with small bias arising from uncertainty in the fixed effects (Kitanidis, 1987; Swallow and Monahan 1984; Zimmerman and Zimmerman, 1991; Lark and Cullis, 2004). When the model is fixed, values of the soil property at unsampled sites can be obtained by the empirical best linear unbiased predictor (EBLUP) (Stein, 1999; Lark et al., 2006; Lark and Webster, 2006; Minsay and McBratney, 2007).

There has been a growing interest in the potential of machine learning methods (e.g. Breiman, 2001) as an alternative to statistical modelling for spatial prediction of soil properties (Hengl et al., 2015; Behrens and Scholten, 2007). The main difference between geostatistical approaches and random forest is that geostatistics is based on a statistical model. This provides a basis for formal inference about the validity of the model (including the task of selecting which covariates to use in prediction), and for producing a prediction distribution at unsampled sites of interest. One may then derive point predictions from this distribution (typically the mean), and measures of uncertainty. On the other hand, machine learning methods such as random forests, are predictive tools applied to identify empirical relationships between the target variable in a training data set and associated predictive covariates and to extrapolate these to unsampled sites. With no model, there can be no formal inference, but empirical approaches, based on internal cross-validation are used. For example, to evaluate the evidence that a particular variable is predictive. One particular strength of the geostatistical approach is that the estimation of coefficients for predictor variables, and inferences about them, are based on a model of the spatial dependence of the random variation. This accounts for the fact that data which are strongly spatially clustered are

likely to be correlated, and so do not provide independent evidence to support the fitted model.

## 1.1 Statement of the Problem

Rice together with maize, cassava, sorghum, millet, wheat, sweet and Irish potato, has been identified to be one of the strategic food crops and a profitable smallholder cash crop in Zambia. However, the suitability of Zambian soils for rice production has not been studied. Therefore, in order to ensure the optimum production of rice, there is need to grow the crop, where it is suitable. When carrying out land suitability evaluation, one of the major functional factors is soil information which is represented in soil maps. There has been a growing interest in machine learning methods such as random forest as an alternative to statistical models for digital mapping of soil properties leaving one with the challenge of choosing whether to use machine learning or statistical models. Soil sampling and analysis are time consuming and expensive. However, when legacy data is available from previous surveys which can be used for digital soil mapping, one is faced with the problem of how to robustly validate models from legacy data when these have, as is often the case, a strongly clustered spatial distribution.

## 1.2 Rationale

When assessing the suitability of land, accurate predictions of the variation of soil properties is very important. This is because such predictions allow stakeholders to understand the current state of soils, how they are changing and the pressure placed upon their quality.

Rice has potential to contribute significantly to increased incomes and employment among rural producers in Zambia but land suitability assessment has not yet been done. The result of this study provides information for all value chain players including farmers wanting to include production of rice on their farms. The maps produced in this study can be used as a guide in selecting suitable sites for rice production. This study will also help identify the main limiting factors for rice production and enable decision makers to develop crop management practices.

Ministry of Agriculture (2016) observed the need to map the different rice ecologies in Zambia in order to have a picture of the total land that would potentially support

rice production in each ecology. This study contributes to this need and help government to put in place more specific interventions for rice production in Zambia.

## 1.3 Main Objective

The main objective of this study was to assess approaches for digital mapping of soil pH, soil phosphorus and soil organic carbon at national scale across Zambia to support evaluation of land suitability for rice production.

## 1.4 Specific Objectives

(1) To predict soil variation of P, SOC and pH using random forest (RF) and Linear Mixed Models.

(2) To evaluate the performance of Linear Mixed Models and Random Forests in mapping the spatial distribution of P, SOC and pH using legacy data with strongly clustered spatial distribution.

(3) To generate land suitability maps for rice production using the spatial predicted soil properties in a geographical information system (GIS) decision support process model.

## 1.5 Research Hypothesis

(1) There are statistically significant variations in the spatial distribution of P, SOC and pH in the soils of the study area.

(2) There are statistically significant differences in the performance of the random forest and linear mixed models in mapping spatial variability of soil properties when legacy data with strongly clustered spatial distribution is used.

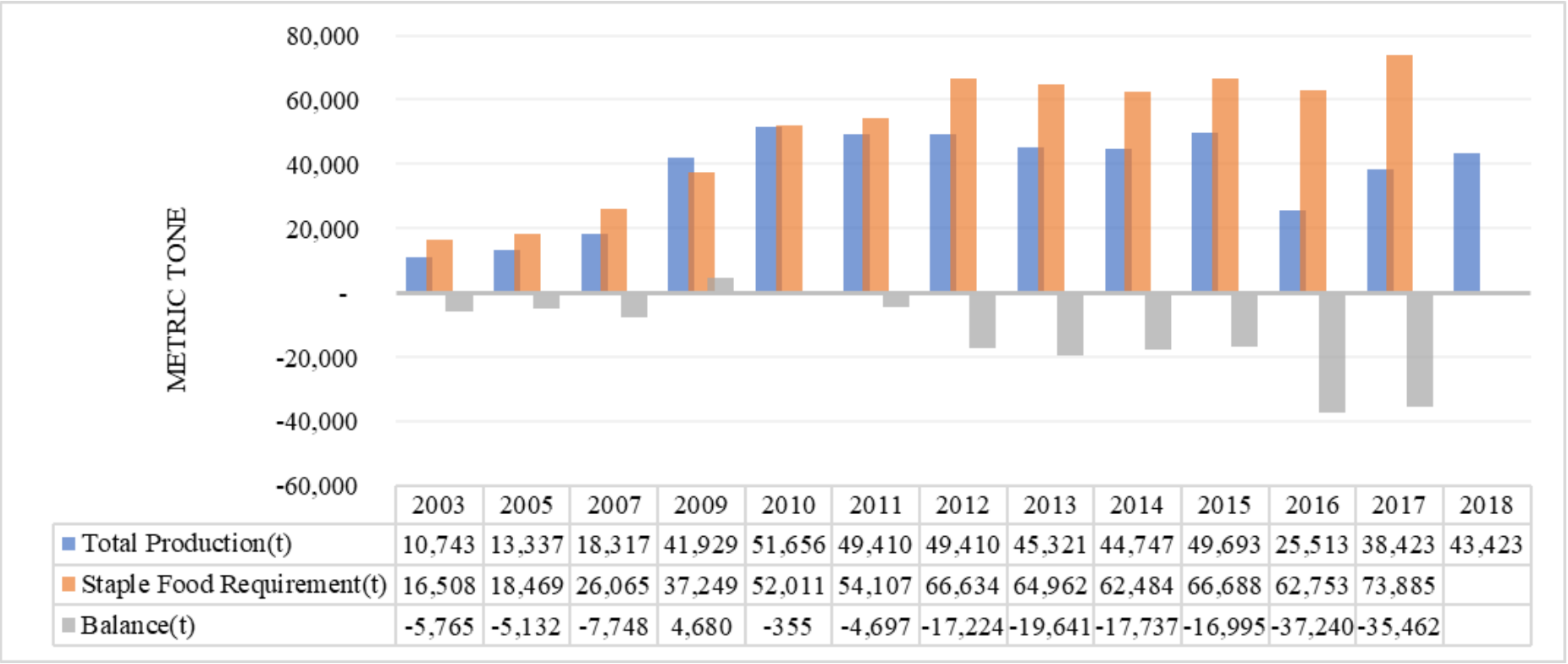
# CHAPTER TWO

# 2.0. LITERATURE REVIEW

## 2.1. Rice Production in Zambia

Rice in addition to maize, cassava, sorghum, millet, wheat, sweet and Irish potato, has been identified to be one of the strategic food crops (Mwila et al., 2008; Styger, 2014) and a profitable smallholder cash crop in Zambia (Chizhuka, 2009). The current status of rice is evidence of its growing importance. The annual demand for rice rose steadily from below 20,000 tones to almost 70,000 tones during the period 2003 to 2017 as illustrated in Figure 1 (CSO, 2018).

However, the annual demand for rice exceeds production. To meet this deficit, the country imports between 5,000 and 20,000 tons of milled rice annually (Ministry of Agriculture, 2016). In response, the government through the Ministry of Agriculture, developed the National Rice Development Strategy (NRDS) in 2016, whose overall objective was to increase local rice production by at least 50 percent and to enhance its competitiveness on the market by the year 2020. However, to date the national average yield of rice has not increased, neither has the area planted, although the staple requirement continues to increase (Table 1).




| | 2003 | 2005 | 2007 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total Production(t) | 10,743 | 13,337 | 18,317 | 41,929 | 51,656 | 49,410 | 49,410 | 45,321 | 44,747 | 49,693 | 25,513 | 38,423 | 43,423 |
| Staple Food Requirement(t) | 16,508 | 18,469 | 26,065 | 37,249 | 52,011 | 54,107 | 66,634 | 64,962 | 62,484 | 66,688 | 62,753 | 73,885 | |
| Balance(t) | -5,765 | -5,132 | -7,748 | 4,680 | -355 | -4,697 | -17,224 | -19,641 | -17,737 | -16,995 | -37,240 | -35,462 | |

*Figure 1: Rice production, consumption and net-trade balance in Zambia*

*Source: (Ministry of Agriculture/Central Statistical Office Crop Forecast Survey 2002/03-2017/18, Ministry of Agriculture Food Balance Sheets 2002/03-2017/18 https://www.zamstats.gov.zm/; https://zambia.opendataforafrica.org/etqmqgf/agriculture-statistics-2017)*

*Table 1: Area planted and yield of paddy rice in Zambia.*

| **Year** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** |
|---|---|---|---|---|---|---|---|---|
| Total Area Planted (ha) | 33,995 | 31,388 | 38,528 | 40,974 | 42,983 | 25,594 | 33,303 | 34,217 |
| Yield t/ha | 1.45 | 1.44 | 1.16 | 1.21 | 0.59 | 1.04 | 1.15 | 1.26 |

*Source: Ministry of Agriculture (2016); Ministry of Agriculture/Central Statistical Office Crop Forecast Survey 2010/11-2017/18*

Poor yield is one of the factors that has contributed to Zambia's inability to meet the increasing demand for rice through local production. The average rice yield is 1.3 t/ha (CSO/MAL/RALS, 2015) which is quite low when compared to other Eastern and Southern African countries such as South Africa, Kenya, Uganda and Zimbabwe where national average yields were 2.61, 5.24, 2.30 and 2.26 t/ha respectively for the year 2013 (Ministry of Agriculture, 2016).

Apart from soil constraints (Aune et al., 2014), poor water management is also one of the factors that limits rice yields (Styger and Uphoff, 2014). Most of the rice grown in Zambia is rainfed paddy rice and this limits its cultivation to flooded or semi-flooded lowland environments (Mutale et al., 2010). With frequent occurrence of droughts, floods and other extreme weather conditions due to climate change, farmers generally find it difficult to improve production and productivity (Ministry of Agriculture, 2016). The introduction of upland rice types such as New Rice for Africa (NERICA) varieties, has given the country the option of bringing additional arable land under cultivation (Ministry of Agriculture, 2016).

## 2.2. Land Suitability Evaluation in Zambia

Land suitability evaluation may account for a range of factors that are potential constraints on the land use of interest. The Food and Agriculture Organization (FAO) approach, developed in collaboration with International Institute for Applied Systems Analysis (IIASA), is based on climate, soil and terrain conditions (IIASA/FAO, 2012), and so requires that information on different factors, which constrain land use in different ways, are combined in the evaluation (Fischer et al., 2002). This requires a multi-criteria evaluation (MCE), by which information on several factors (soil and land

constraints and requirements) are used to produce a single index which can, for example, be presented as a map (Malczewski, 1999).

In Zambia, land evaluation began in 1948. The first land evaluation method was developed between 1960 and 1970 and it was called Land Use Branch Land Capability System. This system had two weaknesses: one, it was not sufficiently detailed to asses land suitability for many uses as it was based on physical factors easily recognizable in the field. The second weakness was that it was not understood internationally. It was in this regard that in 1987, this system was succeeded by the Zambia Land Evaluation System (ZLES) which was developed based on the FAO framework for land evaluation (Chinene, 1991).

A review of the literature showed that there are few studies on land suitability assessment in Zambia using multi-criteria evaluation. One study was carried out by Munene et al., (2017) to assess land suitability for soybean production in Kabwe District. The other was carried out by Chirwa et al., (2016) who evaluated the soil fertility status and suitability of land for groundnut and maize production by smallholder farmers in Chisamba District.

### 2.3. Linear Mixed Model

The theory of residual maximum likelihood (REML) in combination with the empirical best linear unbiased predictor (EBLUP) for spatial interpolation has been illustrated and described in detail by Lark et al., 2006. The LMM takes the form

$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{z}$ is a set of observations of the random variable at sampled locations, $\mathbf{M}$ is the design matrix of fixed effects, which could include covariates such as topographic attributes, $\boldsymbol{\beta}$ is the vector of the fixed effects parameters or regression coefficients, $\mathbf{S}$ is the design matrix of random effects (which is an identity matrix unless analytical duplicate observations are included), $\boldsymbol{\eta}$ is a random effect, a Gaussian random variable which has a mean of zero and, in the spatial setting, a covariance matrix which expresses spatial dependence, $\boldsymbol{\varepsilon}$ is an independently and identically distributed Gaussian residual of mean zero and variance $\sigma^2$. These two random components have a joint distribution

$$\begin{bmatrix}\boldsymbol{\eta}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2\xi\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \sigma^2\mathbf{I}\end{bmatrix}\right), \tag{2}$$

where **I** is the identity matrix and **G** is the correlation matrix of the random effect $\boldsymbol{\eta}$. Element [*i*,*j*] of **G**, at locations $\mathbf{x}_i$ and $\mathbf{x}_j$ depends only on the interval in space between them under an assumption of second-order stationarity. The lag vector $\mathbf{x}_i - \mathbf{x}_j$, under the assumption of isotropy, depends only on the scalar part of this vector, the lag distance and so:

$$\mathbf{G}[i,j] = \rho\left(\left|\mathbf{x}_i - \mathbf{x}_j\right|; \alpha\right), \tag{3}$$

where $\rho(h;\alpha)$ is a correlation function of lag distance *h* with spatial parameters $\alpha$ which control how the correlation decreases with increasing distance. The term $\xi$ is the ratio of the variance of the random effect $\boldsymbol{\eta}$ to $\sigma^2$, the variance of the residual term.

The residuals depend on the fixed effects parameters $\boldsymbol{\beta}$ in the model, and in ordinary maximum likelihood estimation the uncertainty in the estimates of the fixed effects parameters biases the estimates of the random effects' parameters. To avoid this, we use residual maximum likelihood (REML) which is based on the principle that a new random variable, independent of the fixed effects, is computed by projecting the original data **z** into a residual space where the fixed effects can be filtered out (Chai et al., 2008). The log likelihood of the new random variable which we now call the residual log-likelihood because it is independent of fixed effects can be expressed as:

$$\ell_R(\sigma^2, \xi, \alpha|\mathbf{z}) = -\frac{1}{2}\{\log|\mathbf{H}| + \log|\mathbf{M}^{\mathrm{T}}\mathbf{H}\mathbf{M}| + (n-p)\sigma^2 + \frac{1}{\sigma^2}\mathbf{z}^{\mathrm{T}}(\mathbf{I} - \mathbf{W}\mathbf{C}^{-1}\mathbf{W}^{\mathrm{T}})\mathbf{z}, \tag{4}$$

where $\mathbf{H} = \xi\mathbf{M}\mathbf{G}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}$, $\mathbf{W} = [\mathbf{M}, \mathbf{S}]$ and $\mathbf{C} = \begin{bmatrix} \mathbf{M}^{\mathrm{T}}\mathbf{M} & \mathbf{M}^{\mathrm{T}}\mathbf{S} \\ \mathbf{S}^{\mathrm{T}}\mathbf{M} & \mathbf{S}^{\mathrm{T}}\mathbf{S} + \xi^{-1}\mathbf{G}^{-1} \end{bmatrix}$.

Once the covariance parameters $\sigma^2, \xi, \alpha$ have been estimated by REML, they are used to compute the estimated covariance matrix at sampled points. With the estimated covariance matrix computed, the estimated fixed effects parameter, $\widehat{\boldsymbol{\beta}}$, and predicted random effects, $\widetilde{\boldsymbol{\eta}}$, are then computed by solution of mixed model equation:

$$\mathbf{C}\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widetilde{\boldsymbol{\eta}} \end{bmatrix} = \begin{bmatrix} \mathbf{M}^{\mathrm{T}}\mathbf{z} \\ \mathbf{S}^{\mathrm{T}}\mathbf{z} \end{bmatrix} \tag{5}$$

With the covariance matrix for the error of the estimates being:

$$\mathrm{Cov}\begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{bmatrix} = \sigma^2\mathbf{C}^{-1}. \tag{6}$$

After the covariance matrix and the fixed effects parameters have been estimated, they are used in EBLUP to predict the soil property , $\tilde{z}_p$ , at unsampled locations:

$$\tilde{z}_p = \mathbf{M}_{\mathrm{p}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widetilde{\boldsymbol{\eta}}_{\mathbf{p}} = \mathbf{M}_{\mathrm{p}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \mathbf{g}_{\mathrm{o,p}}^{\mathrm{T}}\mathbf{G}^{-1}\widetilde{\boldsymbol{\eta}}, \tag{7}$$

where $\mathbf{M}_{\mathrm{p}}$ is the design matrix for the prediction sites, $\mathbf{g}_{\mathrm{o,p}}$ is a vector computed from the covariance matrix of $\boldsymbol{\eta}$ with the $\boldsymbol{\eta}_{\mathrm{p}}$ values at the unsampled locations ($\mathrm{Cov}\left[\boldsymbol{\eta}, \boldsymbol{\eta}_{\mathrm{p}}\right] = \xi\sigma^2\mathbf{g}_{\mathrm{o,p}}$). The variance of the prediction errors, $Var\left[\tilde{z}_p - z_p\right]$, which accounts for the uncertainty in predicting the fixed effects and uncertainty in predicting the random effects is expressed as:

$$\mathrm{Var}\left[\tilde{z}_p - z_p\right] = \sigma^2\left\{\left[\mathbf{M}_{\mathrm{p}}, \mathbf{g}_{\mathrm{o.p}}^{\mathrm{T}}\mathbf{G}^{-1}\right]^{\mathrm{T}}\mathbf{C}^{-1}\left[\mathbf{M}_{\mathrm{p}}, \mathbf{g}_{\mathrm{o.p}}^{\mathrm{T}}\mathbf{G}^{-1}\right] + \xi\left(\mathrm{g}_{\mathrm{p,p}} - \mathbf{g}_{\mathrm{o.p}}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{g}_{\mathrm{o,p}}\right) + 1\right\}. \tag{8}$$

There are many variables that researchers can use as fixed effects in linear mixed models for spatial prediction of soil properties. However, it is unwise to include variables without regard for evidence that they are of predictive value, the inclusion of predictors unrelated to the target variable may inflate the prediction error variance. To avoid this, variable selection is an important step in model development. One approach to the problem is to base the inclusion or rejection of a predictor based on a hypothesis test in the LMM framework (e.g., by a log-likelihood ratio test) (Verbeke and Molenberghs, 2000). To reduce the risk of including excess predictors because of multiple hypothesis testing, one may use false discovery rate control (Lark, 2017). The false discovery rate (FDR) is the probability that a null hypothesis is true, given that it has been rejected. False discovery rate control can reduce the power to detect real predictors, and Lark (2017) demonstrated how this problem can be reduced, while maintaining FDR control, by the method of alpha investment (Foster and Stine, 2008). This entails an initial ordering of the predictors starting with the one which, a priori (and not based on inspection of the data) is thought most likely to be of predictive value and adding in predictors in declining order of expected predictive power. In this approach the power to detect a predictor is increased by the rejection of the null hypotheses early in the sequence while maintaining control of FDR. This approach has been used elsewhere for spatial prediction (Gashu et al., 2020).

The disadvantage of the LMM approach is that it assumes that the fixed effects are linear in the parameters. Such a model can represent complex and non-linear relations

between soil properties and predictors, for example through the use of polynomial terms in the predictor variables, or spline basis functions, there has been increasing interest in more flexible methods to predict soil properties from covariates, in particular the machine learning method known as the random forest.

## 2.4. Random Forest

The random forest (RF) is an ensemble tree-based method that combines multiple decision trees (classification or regression) to give a prediction (Breiman, 2001). A decision tree is an algorithm that involves recursive partitioning of data into several simple regions using a series of splitting rules. It is called a decision tree because these series of splitting rules can be summarised into an upside-down tree structure as illustrated in Figure 2. Figure 1 shows a structure made up of predictors $(X_1, X_2, \ldots\ldots, X_k)$ which are split into $J$ distinct and non-overlapping regions $(R_1, R_2, \ldots, R_j)$ at test node *t*, and the mean of the response values for the training observations in each region $R_j$ is calculated and assigned as a prediction for every observation that falls in region $R_j$ (James et al., 2013). When growing a decision tree, the following steps are taken; (1) at each test node *t*, a predictor $X_k$ is randomly sampled from all the predictors, then the best split point $S_k$ among all possible splits for the $k$ th predictor is determined; (2) the best split $S^*$ among the $S_k$ is chosen and this $j$th predictor at its identified cut point $C_{S^*}$ is used for the splitting at test node $t$. (3) The predictor $X_k$ is split into two regions ( observations with $X_k < C_{S^*}$ and $X_k \geq C_{S^*}$) at test node $t$. Steps 1-3 are repeated on all descendant nodes to grow a tree $\hat{f}(x)$ (Archer and Kimes, 2008; James et al., 2013).

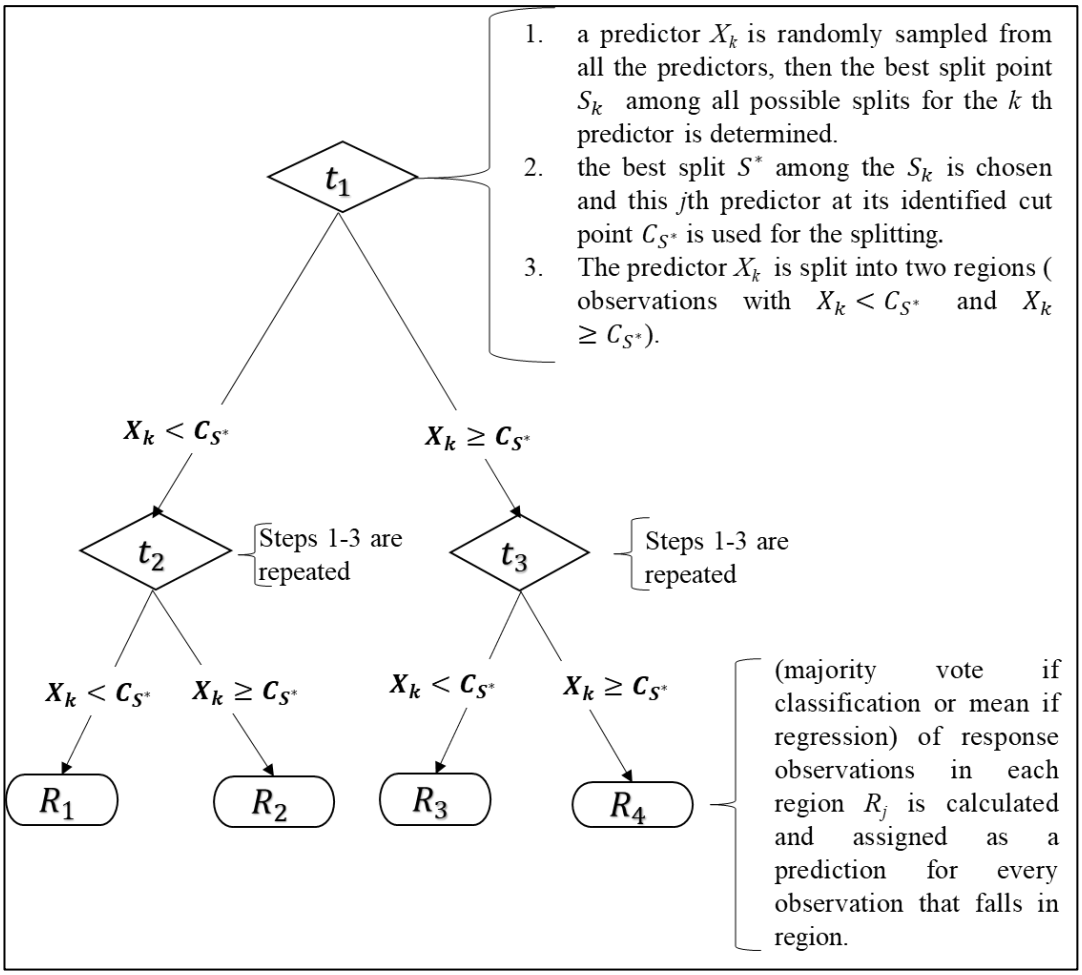



*Figure 2: Illustration of a decision tree*

One major limitation with decision trees is that using only one tree for prediction results in highly unstable predictions. A small change in the data can result into a large change in the final estimated tree. To improve the performance of decision trees, Breiman (1996) introduced an algorithm called Bagging, also known as bootstrap aggregation which takes repeated (bootstrap) samples (where $B$ is the number of bootstrap samples) from training set with replacement and builds a total of $B$ trees $(\hat{f}^1(x), \hat{f}^2(x), \ldots\ldots \hat{f}^b(x))$ which the average of all the prediction trees $\hat{f}_{bag}(x)$ is calculated:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \tag{9}$$

One disadvantage of bagging is that a single predictor may dominate all trees in the bag, meaning that their outputs are strongly correlated. As a result of this, the reduction in variance from the use of multiple trees is very limited (James et al., 2013).

To address this, Breiman (2001) developed a random forest algorithm which is an improvement of bagging. Like bagging, random forest also takes repeated (bootstrap) samples from the training data and builds $B$ decision trees. But in the case of random forest, when building these trees, to avoid using one strong predictor for all bagged trees, at every test node $t$, when splitting, every bagged tree is made to consider only a random subset of predictors by randomly sampling a fresh $m$ predictors from a set of $k$ predictors, and the split is only allowed to use one of these $m$ predictors. For regression trees, the number of $m$ predictors considered at each split is approximately the total number of predictors divided by three ($m \approx {}^{k}/_{3}$ ) and for classification trees, $m \approx \sqrt{k}$. Because of this, random forest results in many uncorrelated trees which give a large reduction in variance when averaged.

The random forest algorithm has three important outputs. These are the out-of-bag Mean Squared Error (OOB Mean Squared Error), the out-of-bag R-squared and the variable importance. The RF model does not use all the data for building the tree. In each bootstrap training set, about one-third of the data are left out. The data that are left out when building the trees is called out-of-bag (OOB) data and after the trees are grown, the OOB data are used as test set to measure the strength (OOB Mean Squared Error) and correlation (OOB R-squared) of the model. In short, random forest has an inbuilt cross-validation. Variable importance is defined as the increase in prediction error when OOB data for that variable is randomly permuted while all others are left unchanged (Liaw, 2002). It analyses the contributions of each predictor to the overall results (Breiman, 2001). The algorithm randomly permutes the predictor $X_m$ several times, breaking its original association with the response variable and assesses the relevance of the predictor by using the permuted predictor together with the other unpermuted predictors to predict the response variable for the out-of-bag observations giving the difference in prediction accuracy before and after permuting. The result is a vector of importance measures for each predictor equivalent to the number of permutations. The algorithm then computes a p-value as a measure of the evidence that a variable is predictive (Strobl et al., 2007; Altmann et al., 2010). This permutation p-value is the probability of observing a permuted model (from the several number of permutations) that is equal to or better than the unpermuted model (Cummings et al., 2004).

Equation (7) presents the E-BLUP from a linear mixed model. The second term on the right-hand side corresponds to the spatial interpolation of the correlated random effect in the model. In this way the E-BLUP combines a regression-type prediction based on the predictor variables with a spatial interpolation component. As described above the random forest predicts a soil property from the predictor variables only, making no use of spatial dependence through interpolation. An attempt was made to include spatially weighted local observations in prediction by random forests by including coordinates as predictors and using weighted buffer distances (Hengl et al., 2018), neighboring observations and their distances to the prediction location were used by Sekulić et al., (2020). Li et al., (2011) and Viscarra Rossel et al., (2014) combined random forest with kriging, just like in regression-kriging, by calculating the random forest residuals and then kriged them to all prediction positions and then added to the results of the prediction positions.

As described above, inferences in the random forest approach are based on an internal cross-validation procedure. This might lead to overoptimistic conclusions about a random forest model, or about the value of a particular predictor if observations from the same clusters appear in the OOB sample and in the data used to develop the trees. That is because the observations within a cluster can be expected to be strongly correlated, and so the validation of a model fitted to data on strongly correlated observations will give an unduly optimistic impression of the model's capacity to predict at an independent location.

# CHAPTER THREE

## 3.0. MATERIALS AND METHODS

### 3.1. Description of Study Area

This study was carried out in Zambia, a landlocked country in Southern Africa with an area of 752, 618 $km^2$ as shown in figure 2 ranging from Acrisols in the northern, Arenosols in the west. It is also made up of four agro-ecological zones

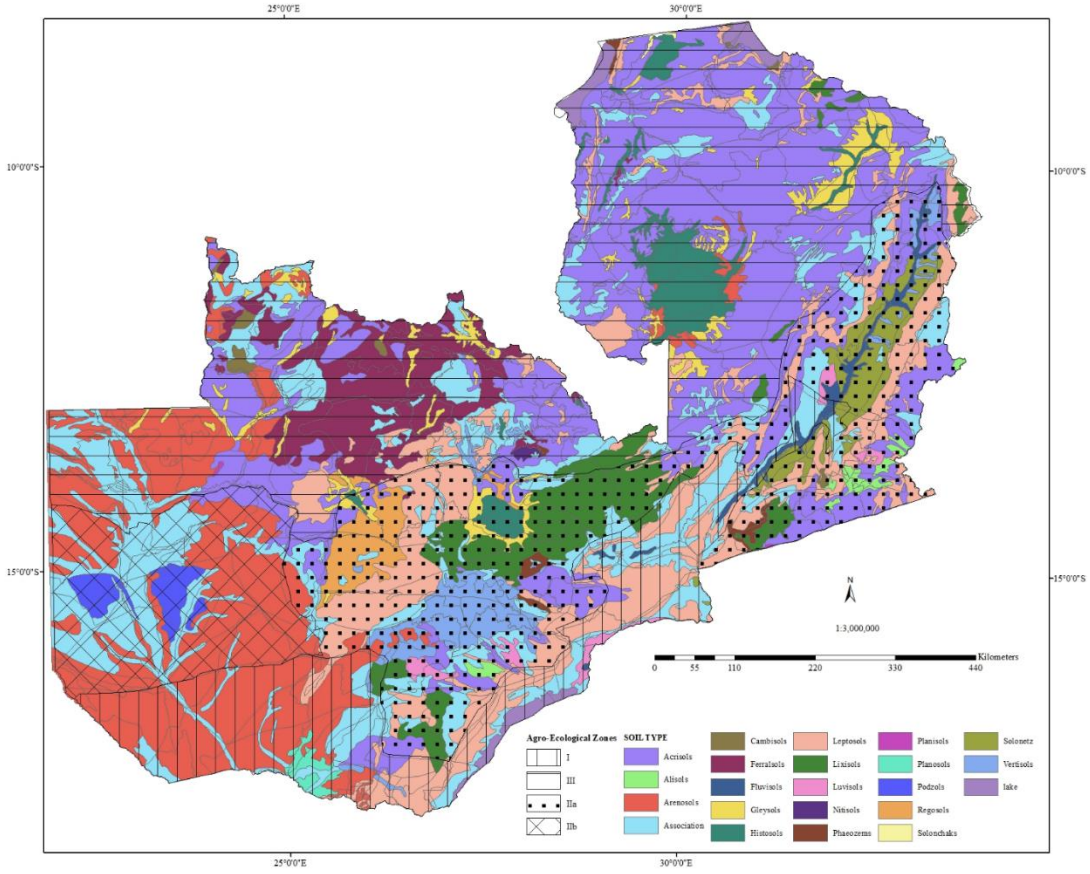



*Figure 3: Agro-ecological zones and Soils map of Zambia Location of Study area. Source: Author`s illustration with data from www.diva-gis.org Author's illustration with data from (GRZ, 1991)*

### 3.2. Land Suitability Evaluation for Rice Production

Figure 4 shows the flowchart of the methodology followed in the study to carry out the suitability assessment for rice production in Zambian under rainfed and irrigated conditions. The first step was literature review to identify the data set that would be required to carry out the land suitability evaluation for rice in Zambia. After review of the literature (FAO, 1976; De Data, 1981; Chisci, 2009) on land suitability for paddy

and upland rice production key soil and site factors (SSF) comprising both constraints and requirements key to evaluation were identified. These are slope, the content percent by volume of coarse fragments (soil particles > 2mm), soil drainage, soil pH, soil organic carbon (SOC), soil cation exchange capacity (CEC), annual rainfall and mean temperature of the growing season. Step 2 involved identifying the sources and type of data that was available to enable land suitability evaluation and available, were two data types, there was climate (annual rainfall and mean temperature), topography (slope, coarse fragment) and some of the soil data (soil drainage and CEC) that was already available in map form all that was need was to just pre-process the maps and then use them in a MCE with GIS for land suitability evaluation. Because some of the soil data (phosphorus, pH and organic carbon) available was legacy data which was in point form, step 3 and Step 4 involved use of models for spatial prediction of these point data to produce maps and then evaluated the performance of these models. Step 5, all maps where now reclassified to a common measurement scale. In step 6, an MCE was performed and then in step 7, the MCE was integrated with GIS to produce the suitability maps for rice in Zambia. Details of how each of these steps were carried out is outlined below.

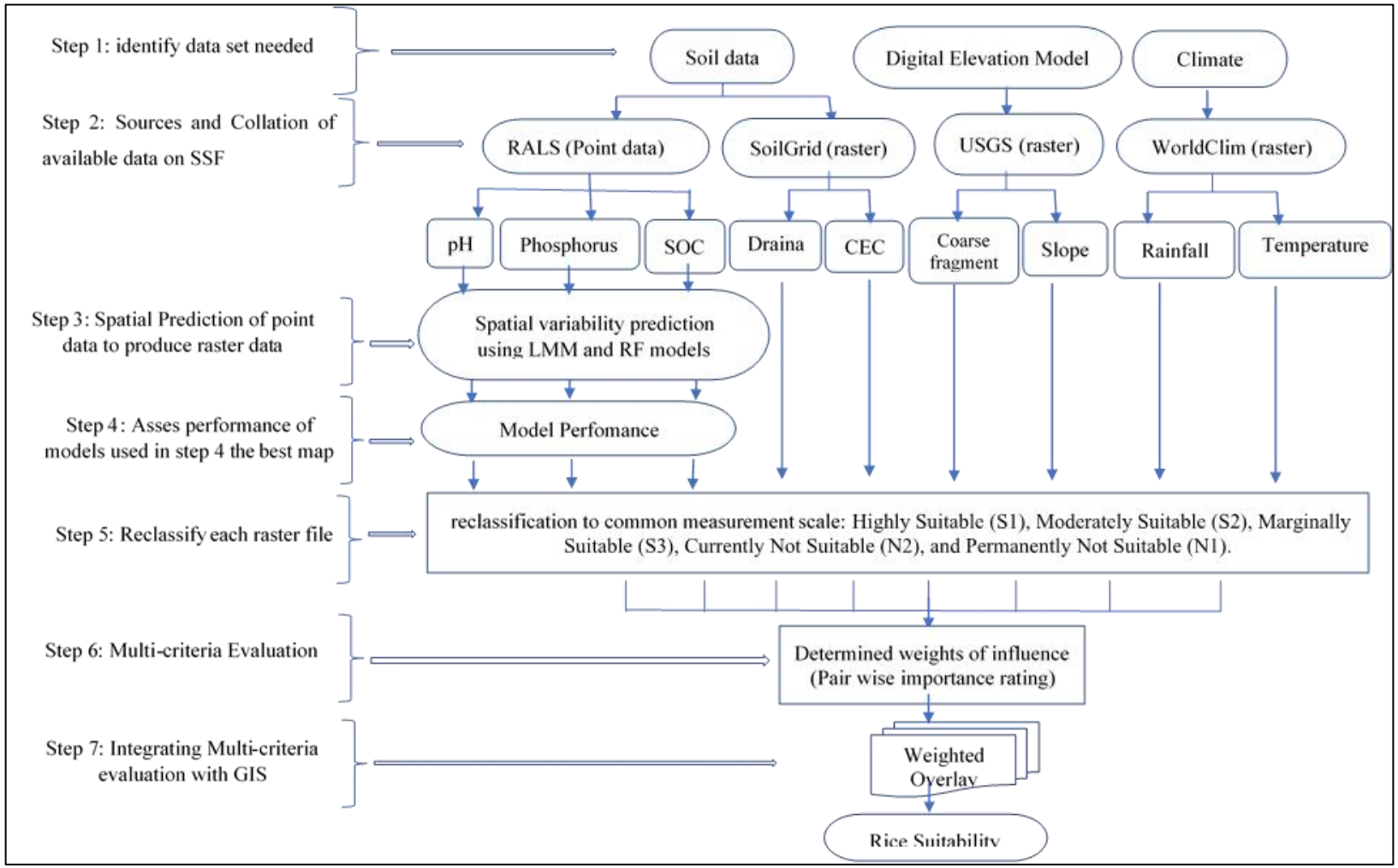



*Figure 4: Flowchart of the methodology followed in the study*

### 3.2.1. Sources and Collation of Information on SSF

Basic information on the soil and land constraints and requirements identified were Slope data which was derived from the NASA Shuttle Radar Topography Mission (SRTM3) global 1-arcsecond (30-m) Digital Elevation Model (DEM) downloaded from USGS (2019); annual mean temperature and annual precipitation which were averages from 1970 to 2000 with spatial resolution of 1km (Fick and Hijmans, 2017). The data on soil properties was downloaded from ISRIC (2017). Hengl et al. (2017) described in detail the analytical and prediction methods that were undertaken to map these soil properties. The SoilGrids system at 250m resolution was updated in June 2016 and provides global predictions for standard numeric soil properties such as CEC, drainage conditions and coarse fragments. Hengl et al. (2017) undertook 10-fold repeated cross-validation and showed that the R-squared of the models for CEC and coarse fragments were 0.64 and 0.56, respectively.

*Table 2: Data, Format and Sources*

| **Data** | **Format** | **Resolution** | **Source** |
|---|---|---|---|
| Topographic Data (DEM) | Raster | 30 x 30m | USGS (https://earthexplorer.usgs.gov) |
| Climatic Data (Rainfall, Temperature) | Raster | 1 x 1km | WorldClim (www.worldclim.org) |
| Soil Data (CEC, Drainage, Course fragment) | Raster | 250 x 250m | Soil Grids (https://soilgrids.org) |
| Soil Data (pH, Phosphorous, Organic Carbon) | Raster mapped from point data | | Indaba Agricultural Policy Research Institute (IAPRI) |

### 3.2.1. Spatial Prediction of Exchangeable Acidity (pH), Soil Phosphorus (P) and Soil Organic Carbon (SOC)

#### 3.2.1.1. Soil data

This study used Rural Agricultural Livelihoods Survey (RALS) of 1713 geo-referenced soil data collected by Indaba Agricultural Policy Research Institute (IAPRI) in collaboration with Central Statistical Office (CSO) and Ministry of Agriculture. The sampling frame for the RALS 2012 survey was based on the 2010 Census of Housing and Population (CSO/MAL/IAPRI, 2015). Full detail of the stratified two-stage sampling design is provided by (CSO 2012). Four households were randomly selected in each Standard Enumeration Area (SEA) and soil samples were collected from the largest maize field. A composite of 10–20 sub-samples of soil collected throughout each field and each sub-sample was a composite of equal parts soil in the 0–10cm and10–20cm depth horizons. Full details on the soil collection and laboratory analysis for soil pH (determined for a soil suspension in $CaCl_2$ with a standard pH meter), soil phosphorus and soil organic matter (measured following the Walkley and Black method) are provided by Burk et al. (2019) and Chapoto et al. (2016). The spatial prediction of soil pH, soil phosphorus and soil organic carbon for Zambia using this data has been studied by Chapoto et al. (2016) who only used ordinary kriging for the prediction.

Data cleaning involved removal of spurious values in the x and y coordinates. The need for this was indicated when the raw data were first plotted, showing points lying outside the borders of Zambia. The mean coordinates of all households were computed in each Standard Enumeration Area (SEA) centroid, and then the households were removed from the data set if the notional distance to the SEA centroid exceeded 10km. This threshold value was decided after discussion with IAPRI staff about plausible values for the distance between a village in the SEA and the centroid. After data cleaning, a total of 1202 soil samples were used for analysis.

The sampling pattern for the RALS survey was not designed for spatial interpolation of soil. Due to the sampling pattern, the data was strongly clustered at the level of SEAs (the SEAs are the clusters) with a total of 362 clusters. For this reason, splitting of the dataset into training set (80 percent) and test set (20 percent) was done at cluster level (the 362 clusters were split into 260 (80 percent) for training and 102 (20 percent)

for validation). Figure 3 shows the training and test clusters with the red solid dots being the training clusters and the blue solid triangles being the test clusters.

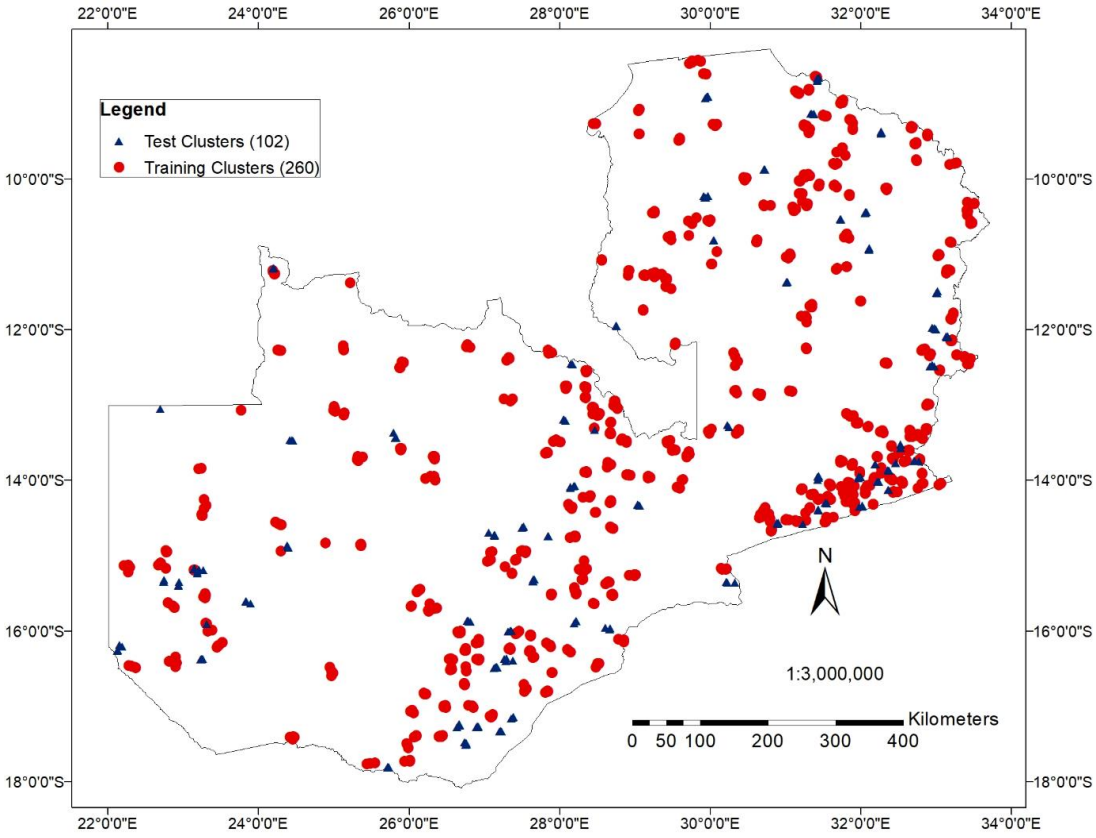



*Figure 5: Cluster locations for the RALS 2012 soil data*

*Note: red solid dots being the training clusters used for spatial prediction of soil pH, Phosphorus and organic matter while the blue solid triangles being the test clusters left out for validation.*

### 3.2.1.2. Environmental Covariates

The environmental predictors available for use in this study were soil class, landcover, mean annual rainfall, elevation, slope, aspect, valley depth, LS-Factor (a combination of slope and slope length, relative slope position (RSP), channel network base level (CNBL) and normalized difference vegetation index (NDVI).

Soil class information was obtained from the 1:1,000,000 scale exploratory soil map of Zambia compiled by the Zambia Agricultural Research Institute (ZARI) - Soil Survey Section in 1991(GRZ, 1991) and then digitized to raster format. Map units are allocated to suborders of the FAO-UNSECO classification as used in the Third Draft

of the legend to the Soil Map of the World (Jahn et al., 2006). A total of 96 soil classes were represented in the data available for model development, but these do not comprise all the classes on the map of Zambia, and so some generalization was required to develop models for spatial prediction. Therefore, the number of classes were reduced, by aggregating the classes from suborder to order level, this reduced the number of classes to 18 and all the classes in the prediction grid were represented in the training set. Land cover data for the years between 2000 and 2015 with spatial resolution of 300m were downloaded from the European Space Agency (ESA), (2017). The data presented a similar situation as that of soil class with landcover classes in the prediction sites not being represented in the training set. The number of landcover classes were also reduced, by aggregating them as shown in Table 3.

*Table3: Aggregated landcover classes based on ESA, (2017)*

| **New Class** | **ESA Class** | **Description** |
|---|---|---|
| | 10 | Rainfed cropland |
| | 20 | Irrigated or post-flooding cropland |
| 1 | 30 | Mosaic cropland (>50%) / natural vegetation (tree |
| | 11 | Herbaceous cover |
| | 40 | Herbaceous cover) (>50%) / cropland (<50%) |
| 2 | 110 | Mosaic herbaceous cover (>50%) / tree and shrub (<50%) |
| | 12 | Tree or shrub cover |
| 3 | 100 | Mosaic tree and shrub (>50%) / herbaceous cover (<50%) |
| | 50 | Hlosed to open (>15%), evergreen, broadleaved, tree cover |
| | 60 | Closed to open (>15%), deciduous, broadleaved, tree cover |
| | 61 | Closed (>40%), deciduous, broadleaved, tree cover |
| 4 | 62 | Open (15-40%), deciduous, broadleaved, tree cover |
| | 120 | Shrubland |
| 5 | 122 | Shrubland deciduous |
| 6 | 130 | Grassland |
| | 160 | Fresh or brackish water, flooded, tree cover |
| | 170 | Saline water, flooded, tree cover |
| 7 | 180 | Fresh/saline/brackish water, flooded Shrub or herbaceous cover |
| 8 | 190 | Urban areas |
| | 200 | Consolidated bare areas |
| 9 | 202 | Unconsolidated bare areas |

Mean annual rainfall data (averages from 1970 to 2000) with a spatial resolution of 1km was downloaded from WorldClim website (Fick and Hijmans, 2017). A 90-m resolution NASA Shuttle Radar Topography Mission (SRTM3) Digital elevation model (DEM) was downloaded from USGS (2019) and projected to WGS 84 UTM 35 S. The DEM was pre-processed by filling sinks using the fill sinks (Planchon/Darboux, 2001) tool in Saga GIS (Conrad et al., 2015), and then elevation, slope, aspect, valley depth, LS-Factor (a combination of slope and slope length), relative slope position (RSP) and channel network base Level (CNBL) data was extracted from the DEM using basic terrain analysis tool in Saga GIS. MODIS land surface reflectance (MOD009GA V6) was downloaded from USGS (2019). After downloading the respective data sets, Quantum GIS was used to project the data sets to WS 84 UTM 35s and then converted to the Integrated Land and Water Information System (ILWIS) format. Then ILWIS was used to harmonise all the raster files to the same extent and cell size of 1km. Normalized difference vegetation index (NDVI) was extracted from the remote sensing images using the image indices of the soilassessment package (Omuto, 2020) for the R platform.

#### 3.2.2.3. Variable Selection by Alpha-Investment

Section 2.1 explained how variable selection for the LMM included false discovery rate control, to avoid over-fitting, with alpha-investment to improve the probability of retaining covariates which are predictive as predictor variables. The alpha-investment approach is most effective if the predictors can be ordered, a priori, with the one thought most likely to be predictive ranked first and so on as shown in Table 15. It must be emphasized that this ranking is based on prior considerations about the underlying process, and not on exploration of the data. The ordering of exhaustive environmental covariates based on how they influence each soil property. The ordering of environmental covariates for the prediction of soil pH, soil Phosphorus and soil organic carbon was therefore done as follows:

***Soil pH***

The ordering of exhaustive environmental covariates was based on how they influence the production of $H^+$ ions and the loss of basic cations from the soil. Rainfall was proposed as the most influential factor at national scale. Soils in environments with large annual rainfall tend to have relatively low pH due to reduced based saturation

resulting from loss of basic cations by leaching (McCauley et al., 2009; Brady & Weil, 2014). For this reason, more acid soils are expected in the northern parts of Zambia (Agroecological Region III) and alkaline soils in the south (Agroecological Region I) where annual rainfall is much smaller (Veldkamp et al., 1984; GRZ and UNDP, 2009). Soil class was ranked second because the soil classes represent variations in soil parent material, weathering and rejuvenation of land surfaces and development of the soils. The old, highly-weathered plateau soils in the northern part of the country have lost most of the basic cations. The sandy soils in the western part are easily leached with little accumulation of basic cations. On the other hand, the Karoo group materials in the valleys are rich in basic cations resulting in high pH values. After soil class, next to be included were topographic variables slope, elevation and valley depth. These should reflect processes such as the movement of water which carries with it dissolved basic cations from steep slope to flat areas, and the rejuvenation of weathered land surfaces which entails the removal of old highly-weathered material to reveal material with a larger content of weatherable minerals. Landcover and the Normalized Difference Vegetation Index (NDVI) were then included. These will reflect effects of land management practices, including agricultural inputs, and decisions on land use which may depend on how local pH limits crop performance. The NDVI will also reflect local vigour of vegetation growth, which may be pH-limited. Finally, some further topographic variables were included which may reflect differences in the soil-forming environment (length-slope factor, channel network base level, relative slope position and aspect).

***Soil phosphorous***

Because the major sources of phosphorus are inorganic (Al, Fe, and Ca phosphates) and organic. It is released from these sources into soil solution through weathering (inorganic P) and mineralization (organic P) process which are mediated by plant roots and soil microorganisms. These processes are highly influenced by pH, temperature, moisture, physical and chemical properties of the soil (Shen et al., 2011; Jones, 2012). It is in this regard that when ordering the environmental covariates, the effect of soil class on spatial variability of phosphorus is put first in the sequence because phosphorus availability in the soil is greatly influenced by the pedological pathway which results into weathering of soils. Because soil classes are classified on the basis of weathering and development, Smeck (1985) proposed that a relationship between P

forms in soil and soil class. The effect of rainfall comes next in the sequence. Both the organic and inorganic pool of P in the soil, are greatly driven by leaching, runoff and changes in soil acidity and these processes are highly influenced by rainfall (Lindsay and Moreno, 1960; Arai and Sparks, 2007; Kleinman et al., 2006). Effect of enhanced vegetation indexed (EVI) on soil phosphorus occurs when photosynthetic carbon from roots, the rhizosphere is characterized by high microbial activity (Jones et al., 2009) which results in vegetation influencing soil microbial activity. As the vegetation dies and decays, it also provides a source or organic carbon for microbial and a source of phosphorus.

***Soil organic carbon***

Just like soil pH and soil phosphorus, soil organic carbon is very important in soil fertility. Some of its functions include improving soil structure and nutrient availability for agriculture production. Accumulation and long-term stability of soil organic carbon (SOC) is highly influenced by soil erosion and deposition processes (Fissore et al., 2017). It is in this regard that soil class was put as first in the sequence because classification of soil is based on a combination of a number of soil properties that highly influence soil erosion and deposition. Slope and rainfall were ranked next in the series as these greatly influence soil erosion.

After ordering the environmental covariates for soil pH, the data points were first projected from WGS 1984 to WSG UTM 35s. A total of 19 observations had duplicate coordinates, which were jittered by adding 100m to each of the coordinates for one site. An exploratory model was fitted to the data with all predictors included, using the likfit function of geoR package (Ribeiro and Diggle, 2001) with residual maximum likelihood (REML) as the likelihood method. The only output from this model which was examined were the residuals, for which summary statistics were calculated, and exploratory plots to check the plausibility of the assumption of normally distributed errors. In addition, the correlation model (defined in equation 3) type (exponential or spherical) was identified for which the residual likelihood was largest, and this model was then used in all further analyses. During the sequence testing of hypothesis, first the null model, $m_0$, (with the only fixed effect a constant mean) was fitted with the likfit function and ML as the likelihood method. Then the next model, $m_1$, with the first predictor in the sequence was fitted in the same way. The likelihood ratio was then calculated:

$$L = 2(L_{\mathrm{m1}} - L_{\mathrm{mo}}), \quad (10)$$

where $L_{\mathrm{m1}}$ is the likelihood for model $m_1$ and $L_{m0}$ is the likelihood for the null model. If the null model is correct, then the asymptotic distribution of $L$ is $\chi^2$ with degrees of freedom equal to the number of additional parameters in model $m_1$ by comparison with $m_0$. If $L$ provided evidence to reject the null model with P<0.05, then the additional predictor in model $m_1$ was retained. The second predictor in the list was then considered. When all predictors had been examined the *P*-values at each step were reassessed in sequence using alpha-investment as described by Lark (2017) and controlling the FDR at 0.05. Details of this approach are provided by Lark (2017), but in summary, successive tests are made against a threshold *P*-value which depends on a quantity, the alpha-wealth, which is either augmented when null hypotheses are rejected or augmented when they are rejected. If the hypotheses are ordered so that the variables which, a priori, are expected to be good predictor variables are considered early, then this alpha investment method increases the probability of selecting predictive covariates while controlling FDR.

#### 3.2.2.4. Spatial Prediction of Soil pH, Phosphorous and Organic Carbon

After variable selection, the likfit function of geoR package (Ribeiro et al., 2007) in R with REML as the likelihood method was used to fit two linear mixed models. One with the selected predictors as fixed effects (Kriging with an external drift) and the other with a constant mean as fixed effect (ordinary kriging). The E-BLUP prediction for both models was then calculated at the validation points.

The ranger function of ranger package (Wright and Ziegler, 2017) was used to fit the random forest model. Because random forest has an inbuilt variable selection that occurs within the model by randomly selecting variables to be used at splitting nodes, two random forest models were fitted, one with all variables and the other with the two variables that were selected during the alpha-investment variable selection procedure. In this study, the ranger package in R was used to compute the permutation variable importance according to Altmann et al. (2010).

When predicting soil properties in space, the random forest algorithm can find apparently predictive relationships between the target variable and arbitrary spatial variables (such as digital images of human faces) when these are presented as candidate predictor variables alongside covariates which pedometricians might

reasonably expect to be predictive of soil properties (Wadoux et al., 2020). This shows that pattern recognition should not be equated to knowledge discovery. It may also suggest that the random forest algorithm is prone to overfitting, as a result of which its predictions at independent locations may be unreliable. To investigate this effect, for soil pH, entirely random spatially autocorrelated candidate predictor variables were generated, independent of the data, which are called null predictors. Six spatially correlated but mutually independent null predictors were used, specifying a spherical variogram with a distance parameter of 100 km, nugget variance of 0 and correlated variance of 1 for each. The function RFsimulate from the RandomFields package for R (Schlather et al., 2015) was used to simulate values of these null predictors at the calibration locations. The ranger package (Wright and Ziegler, 2017) was used to fit a random forest model with all predictors including the null variables as predictors and then computed the permutation variable importance according predictor p-values from the model result.

To examine the possibility of improving RF predictions by an additional kriging step (following Li et al., 2011 and Viscarra Rossel et al., 2014), residuals of the models at training points were derived (subtracting the predicted values from the observed values) and then a variogram (equation 4) was fitted to the residuals using likfit with a constant mean as the only fixed effect. The evidence for spatial dependence in the residuals was assessed by comparing the Akaike Information Criterion (AIC) for the fitted model and for a non-spatial alternative which are reported in likfit output.

### 3.2.2. Performance of RF and LMM in mapping the spatial distribution of Soil pH, Phosphorous and Organic Carbon.

Legacy data on soil pH, soil phosphorus and soil organic carbon were available from 2012 RALS survey (a previous national survey) (CSO/MAL/IAPRI, 2015). As with many such surveys, this followed a two-stage design, and so the observations were spatially clustered. In addition, various exhaustive environmental covariates which could be regarded as potential predictor variables for soil pH, soil phosphorus and soil organic carbon were available. Different forms of linear mixed model, and prediction with the random forest were compared using a validation subset of the data. Prediction errors were evaluated at the validation locations by comparing predictions with observed values. The selection of the validation subset, and the quantification of the uncertainty from the observed prediction errors had to take account of the spatial

clustering of the observations in the legacy data. Because of this clustering, no subset could be regarded as independent random observations.

Validation of each selected model or random forest was done using the validation data set. At each validation location the predicted soil pH, soil phosphorus and soil organic carbon was computed, and the prediction error was calculated as the difference between the predicted and observed (so a positive error is when the predicted value exceeds the observed value). As exploratory summary statistics the mean error, median and mean square error were computed.

The validation data belong to a subset of SEA from the original survey, and as such are strongly clustered. Because of this the sample average of the squared errors may not be a good estimate of the mean square error, because the observations are not independent. A model-based approach was therefore taken to compute the expected squared error of prediction. A LMM was fitted to the prediction errors at the validation site (with a constant mean the only fixed effect). The expected square error (ESE) for each set of predictions was then computed as the sum of the squared mean error and the variances (nugget and spatially correlated) from the LMM. This is the *a priori* mean square error, i.e., the expected square error at a random location, and as such is likely to exceed the MSE computed directly from the errors of clustered data.

### 3.2.3. Reclassification

The DEM was pre-processed. First pit and sink filling was performed on the DEM using the fill tool in spatial analyst tools of ArcMap 10.7.1 (ESRI 2011). The DEM was then filtered using the filter tool in spatial analyst tools which employs a low pass filter using a 3x3 moving window to smooth the raster dataset. Slope was then calculated from the pre-processed DEM using the slope tool in ArcMap. The average values of soil properties (CEC and coarse fragments), over the depth interval 0–30 cm, were obtained by a weighted average of the predictions using the numerical integration trapezoidal rule explained in detail by Hengl et al., (2017). All the datasets whose cell size was less than 1km were then rescaled to 1km using the resampling tool in ArcGIS using the nearest neighbor function (ESRI 2011).

Once all the data on each SSF were acquired and processed, the suitability levels of each SSF were defined, based on the FAO land suitability classification (FAO, 1976) as: Highly Suitable (S1), Moderately Suitable (S2), Marginally Suitable (S3),

Currently Not Suitable (N). Table 4 gives the interpretation of each FAO land suitability class.

*Table 2: Interpretation of the FAO land suitability classification (FAO, 1976)*

| **FAO land suitability class** | **Interpretation** |
|---|---|
| Class S1 | Land with minor limitations to productivity. Not perfect but is the best that can be hoped for |
| Class S2 | Land that is clearly suitable, but which has limitations that either reduce productivity or increase the inputs needed to sustain productivity compared with those needed on S1 land |
| Class S3 | Land with severe limitations that reduces benefits and/or increase the inputs needed to sustain production so that this cost is only marginally justified |
| Class N | Land is marginally not suitable and has limitations that may be surmountable in time, but which cannot be corrected with the existing knowledge or under present social conditions to give acceptable physical productivity. |

Information from the published literature and crop production guides was used to define, for each SSF, a range of values corresponding to the five FAO suitability classes shown in Tables 5 and 6. Most of the information is from Sys et al. (1993) who categorized requirements for various crops, including paddy and upland rice, grown in tropical and sub-tropical regions into the FAO suitability classes and provided recommendations requirements regarding climate, soil condition and topography.

For purposes of further manipulation and display, the FAO suitability categories were reclassified to numerical scores, assigning values 1 (“Not suitable”), 2 (“Marginally suitable”), 3 (“Moderately suitable”) or 4 (“Highly suitable”).

*Table 3: Land use requirements for rainfed paddy rice*

| **Criterion** | **Highly Suitable (S1)** | **Moderately Suitable (S2)** | **Marginally Suitable (S3)** | **Not Suitable (N)** | **Source** |
|---|---|---|---|---|---|
| Mean annual rainfall (mm) | >1500 | 1500 - 1200 | 1200 – 800 | <800 | Ambarwulan et al., 2016 |
| Annual Mean Temperature ($^{o}$C) | 31-24, 31-36 | 24-18, >36 | 18-10, | <10 | Sys et al., 1993 |
| Slope (%) | 0-1 | 1 - 2 | 2 – 3 | >3 | Masoud et al., 2013; Ojara et al., 2017 |
| Coarse fragment (Volumetric % of soil particles >2mm diameter) | 0-3 | 3-15 | 15-35 | >35 | Sys et al., 1993 |
| Drainage (FAO, 2006) | Imperfect, Poor | Moderate, Well | Somewhat Excess | Very poor, Excessive | Sys et al., 1993 |
| Soil pH ($CaCl_2$) | 4.7-7.2 | 7.2 - 7.7<br>4.2 – 4.7 | 3.7- 4.2 | >7.7, <3.7 | Sys et al., 1993 |
| CEC (cmol/kg)) | >40 | 25-40 | 25 – 15 | <15 | Masoud et al., 2013; Ojara et al., 2017 |
| Phosphorous (ppm) | >25 | 25-10 | 10 – 5 | <5 | Agbeshie and Adjei 2019 |
| Soil Organic Carbon (%) | >1.5 | 1.5-0.8 | <0.8 | | Sys et al; (1993);Ambarwulan et al., 2016 |

*Table 4: Land use requirements for rainfed upland rice*

| **Criterion** | **Highly Suitable (S1)** | **Moderately Suitable (S2)** | **Marginally Suitable (S3)** | **Not Suitable (N)** | **Source** |
|---|---|---|---|---|---|
| Annual Rainfall (mm) | >1500 | 1500 - 1200 | 1200 – 1000 | < 1000 | Jones and Garrity (1986) |
| Annual Mean Temperature ($^oC$) | 31-24, 31-36 | 24-18, >36 | 18-10, | <10 | Sys et al., 1993 |
| Slope (%) | 0-8 | 8 - 16 | 16 – 30 | >30 | Masoud et al., (2013); Ojara et al., 2017 |
| Coarse fragment (Volumetric % of soil particles >2mm diameter) | <15 | 15 - 35 | 35-55 | >55 | Sys et al., 1993 |
| Drainage (FAO, 2006) | Moderate, Well | Somewhat excess, Imperfect | Poor, Very poor | Excessive | Sys et al., 1993; FAO, 2006 |
| Soil pH ($CaCl_2$) | 4.7 – 7.2 | 7.2 -7.7<br>4.2 – 4.7 | 3.7 – 4.2 | >7.7, <3.7 | Sys et al., 1993 |
| CEC (cmol/kg)) | >40 | 25-40 | 15-25 | <15 | Masoud et al., 2013; Ojara et al., 2017 |
| Soil Phosphorous (ppm) | >25 | 25-10 | 10 – 5 | <5 | Agbeshie and Adjei 2019 |
| Soil Organic Carbon (%) | >1.5 | 1.5-0.8 | <0.8 | | Sys et al; 1993;Ambarwulan et al., 2016 |

### 3.2.4. Multicriteria Evaluation

The steps outlined above resulted in nine suitability maps, one for each SSF. These datasets needed to be combined and transformed into a single suitability output map. This is the key challenge of multicriteria evaluation. For simplicity, first considered is an example case where just two factors, annual rainfall and slope, are used (Table 7). In a "dominated" situation (Table 7), it is easy to put together such information because site A is highly suitable with respect to both slope and rainfall and site B is unsuitable by both criteria, therefore one can easily conclude that site A is highly suitable and site B is not suitable. But this is not generally the case. Consider a non-dominated case (Table 8) where A, is highly suitable in so far as this is judged by rainfall but is not suitable with respect to slope and site B the converse applies with respect to both factors. In this case, it becomes difficult to evaluate the suitability of each location for paddy rice production. To solve this challenge, weights of influence were introduced. In this approach an overall suitability score is computed which is a linear combination of the scores for different factors. Each factor has a weight which reflects its overall importance in determining the overall suitability of any site. If the weights are constrained to sum to 1 then the resulting weighted combination of values will lie in the same interval as the constituent scores, 1 to 5. It should be noted that this is not the only way in which different scoring systems could be combined in an overall assessment. The key assumption is that no one factor can be absolutely limiting on rice production, because if two or more factors have similar and appreciable weights then a deficiency in one might be substituted by the other being very suitable.

*Table 5: An example of a dominated case*

| | Annual Rainfall (mm) | Slope (%) | Suitability |
|---|---|---|---|
| Site A | 1400 (highly suitable) | 0-1(highly suitable) | Highly Suitable |
| Site B | <800 (not suitable) | >5 (not suitable) | Not Suitable |

*Table 6: An example of a non-dominated case*

| | Annual Rainfall (mm) | Slope (%) | Suitability |
|---|---|---|---|
| Site A | 1400 (highly suitable) | >5 (not suitable) | ? |
| Site B | <800 (not suitable) | 0-1 (highly suitable) | ? |

### 3.2.5. Weighting of the factors

The calculation of weights was based on expert elicitation. The process of elicitation that was used here was based on the method of Saaty (1988) which requires that the expert considers all pair-wise comparisons of factors, evaluating their relative importance according to a fixed scale. The first step involves creation of a pairwise matrix **A** which is $n \times n$ where $n$ is the number of factors. There is $n(n\text{-}1)/2$ unique comparisons between factors, represented by the elements of the matrix $a_{ij}$ where $i < j$. These values were taken from the scale due to Saaty (1988). These scores range from 1/9 to 9. If $a_{ij}$ is equal to 1 this implies that factors $i$ and $j$ are of equal importance; if $a_{ij}$ is equal to 9 this implies that factor $i$ dominates factor $j$ almost completely in any consideration of suitability of a site for rice. Conversely, if factor $j$ dominates factor $i$ almost completely, then $a_{ij}$ is equal to 1/9. In Saaty's (1988) system intervening values of 3, 5 and 7 are assigned if factor $i$ dominates factor $j$ "moderately", "strongly" or "very strongly" respectively, and even-numbered scores can reflect uncertainty or compromise between experts whose opinions are elicited. As before, if factor $j$ dominates factor $i$ "moderately", "strongly" or "very strongly" then $a_{ij}$ is set to 1/3, 1/5 or 1/7 respectively. Once all values $a_{ij}$ are obtained where $i < j$ the matrix may be completed according to the rule:

$$\begin{aligned} a_{j,i} &= \{a_{i,j}\}^{-1}, \quad i \neq j \\ &= 1, \quad i = j. \end{aligned} \tag{10}$$

Table 9 shows the comparison of factors in the rows ($i$) against those in the columns ($j$). The scores in Table 9 were based either on published values from the application of this approach to land suitability evaluation in other studies, or local expert judgements made in consultation with experts comprising extension staff from Ministry of Agriculture, researchers from Zambia Agricultural Research Institute (ZARI) and rice farmers. Table 10 shows the sources for each score in the pairwise matrix. Scores for the comparison of slope, temperature, pH and OC against each other were obtained from Ayoade (2017) who compared these factors against each by carrying out a quantitative analysis of the relationships between rice yield and environmental variables. Scores for CEC/pH, pH/drainage, CEC/drainage and slope/coarse fragments were based on Moreno *et al.* (2007); Dengiz *et al.* (2015); Yohannes and Soromessa (2018) and Massawe *et al.* (2019) respectively.

*Table 7: Pairwise Comparison Matrix (we compare the factors in the rows (i) against those in the columns (j))*

| Criterion | Pairwise Comparison Matrix | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Coarse Fragment | Slope | Drainage | pH | Soil Organic Carbon | Cation Exchange Capacity | Phosphorous | Mean annual Temperature | Annual Rainfall |
| Coarse Fragment | **1** | 1/2 | 1/9 | 1/5 | 1/6 | 1/9 | 1/5 | 1 | 1/9 |
| Slope | 2 | **1** | 1/5 | 1/3 | 1/5 | 1/6 | 1/3 | 2 | 1/6 |
| Drainage | 9 | 5 | **1** | 7 | 4 | 1 | 7 | 9 | 1 |
| pH | 5 | 3 | 1/7 | **1** | 1/3 | 1/7 | **1** | 5 | 1/7 |
| Soil Organic Carbon | 6 | 5 | 1/4 | 3 | **1** | 1/4 | 3 | 6 | 1/4 |
| Cation Exchange Capacity | 9 | 6 | 1 | 7 | 4 | **1** | 7 | 9 | 1 |
| Phosphorous | 5 | 3 | 1/7 | 1 | 1/3 | **1/7** | 1 | 5 | 1/7 |
| Mean Annual Temperature | 1 | 1/2 | 1/9 | 1/5 | 1/6 | 1/9 | 1/5 | **1** | 1/9 |
| Annual Rainfall | 9 | 6 | 1 | 7 | 4 | 1 | 7 | 9 | **1** |

*Table 8: Sources for the scores of the pairwise matrix in Table 7*

| **Criterion** | **Pairwise Comparison Matrix** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Coarse Fragment** | **Slope** | **Drainage** | **pH** | **Soil Organic Carbon** | **Cation Exchange Capacity** | **Mean annual Temperature** | **Annual Rainfall** |
| Coarse Fragment | **1** | | | | | | | |
| Slope | Massawe et al., 2019 | **1** | | | | | | |
| Drainage | Local expert opinion | Local expert opinion | **1** | | | | | |
| pH | Local expert opinion | Ayoade, 2017; Yohannes & Soromessa., 2018 | Dengiz et al., 2015 | **1** | | | | |
| Soil Organic Carbon | Local expert opinion | Ayoade, 2017 | Local expert opinion | Ayoade, 2017 | **1** | | | |
| Cation Exchange Capacity | Local expert opinion | Local expert opinion | Yohannes and Soromessa 2018 | Moreno et al., 2007 | Local expert opinion | **1** | | |
| Phosphorus | Yohannes & Soromessa., 2018 | Yohannes & Soromessa., 2018 | Yohannes & Soromessa., 2018 | Yohannes & Soromessa., 2018 | Yohannes & Soromessa., 2018 | Yohannes & Soromessa., 2018 | | |
| Mean Annual Temperature | Local expert opinion | Ayoade, 2017 | Local expert opinion | Ayoade, 2017 | Ayoade, 2017 | | **1** | |
| Annual Rainfall | De Data, S.K., 1981 | De Data, S.K., 1981 | De Data, S.K., 1981 | De Data, S.K., 1981 | De Data, S.K., 1981 | De Data, S.K., 1981 | De Data, S.K., 1981 | **1** |

A pairwise matrix produced in this way could be either consistent or inconsistent. For example, if in a set of consistent pair-wise comparisons, x is more important than y and y is more important than z then x must be more important than z. There is no guarantee that a matrix **A** obtained by eliciting individual elements from experts will be consistent, and this must be evaluated before the matrix is used further. Saaty (1988) proved that if a pairwise matrix is consistent, then the maximum eigenvalue should be equal to the order of the matrix. The maximum eigenvalue of the pairwise matrix in Table 9 was then computed with the `eigen` function R platform (R Core Team, 2019).

The computed maximum eigenvalue ($\lambda_{max}$) of the matrix in Table 9 was 9.63 which was larger than the order of the matrix (9). This indicates that there was some level of inconsistency in the pairwise matrix. However, Saaty (1988) recognized that, if one thinks of the elicited matrix **A** as an estimate of an underlying consistent matrix, **Å**, with the estimate obtained with some observation error, then some small degree of inconsistency in **A** is likely and is practically tolerable. Saaty proposed that the consistency of **A** is measured by a consistency index *CI*, which was computed by

$$CI = \frac{\lambda_{max} - n}{n-1}, \tag{11}$$

where $\lambda_{max}$ is the maximum eigenvalue of **A** which is of order *n*. Saaty (1980) conducted computational experiments in which matrices of order 3 to 10 were generated by random selection of index values from 1/9 to 9 for elements $a_{ij}$ where $i < j$ with other elements obtained according to Equation (11). For each matrix he computed CI and repeated these 500 times. Table 11 shows the mean values of CI for matrices of order 3 to 10, which Saaty called the Random Index (RI).

*Table 9: Tabulated for random matrices (RI) (Source: Golden and Wang, 1990).*

| Order Matrix ($n$) | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| Random Index | 0.58 | 0.9 | 0.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.49 |

As a rule of thumb Saaty (1988) proposed a consistency ratio, CR, which is the ratio of CI for an elicited matrix **A** to the tabulated value of RI for random matrices of the same order. Saaty (1988) suggested that the matrix may be used if CR is less than 0.1.

In this case:

$$CI = \frac{\lambda_{max} - n}{n - 1} = \frac{9.63 - 9}{9 - 1} = 0.079, \quad (12)$$

$$CR = \frac{CI}{RI} = \frac{0.079}{1.45} = 0.054 < 0.1 \quad (13)$$

This shows that the comparison matrix presented in Table 9 is acceptable for further use.

The pairwise comparison matrix **A** (Table 9) was then normalized by dividing each element ($a_{ij}$) by the corresponding column sum (Equation 14). The elements of the normalized comparison matrix, **B**, are therefore

$$b_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}} \quad (14)$$

Then to obtain the weight of each criterion ($w_i$) the row sum of the normalized matrix was then divided by the matrix order $n$ (Equation 15) and the sum of the criteria weights must equal to one. Table 12 shows the weights of each criterion.

$$w_i = \left(\frac{1}{n}\right) \sum_{j=1}^{n} b_{ij} \quad (15)$$

*Table 10: Criteria weights for rainfed rice*

| Criterion | Coarse Fragment | Slope | Drainage | pH | Soil Organic Carbon | Cation Exchange Capacity | Phosphorous | Mean Annual Temperature | Annual Rainfall |
|---|---|---|---|---|---|---|---|---|---|
| Weight | 0.019 | 0.033 | 0.235 | 0.058 | 0.101 | 0.239 | 0.058 | 0.019 | 0.239 |

For irrigated rice (paddy and upland), temperature and rainfall were removed as in Sys et al. (1993) and equations 16 and 17 gave us the CI and CR for irrigated rice. Table 13 shows the weights

$$CI = \frac{\lambda_{max} - n}{n - 1} = \frac{7.499 - 7}{7 - 1} = 0.083, \quad (16)$$

$$CR = \frac{CI}{RI} = \frac{0.083}{1.32} = 0.063 < 0.1. \quad (17)$$

*Table 11: Criteria weights for irrigated rice*

| Criterion | Coarse Fragment | Slope | Drainage | pH | Soil Organic Carbon | Cation Exchange Capacity | Phosphorous |
|---|---|---|---|---|---|---|---|
| Weight | 0.023 | 0.040 | 0.325 | 0.071 | 0.137 | 0.332 | 0.071 |

### 3.2.6. Weighted Overlay

Once the raster files had been reclassified to a common measurement scale and the weights of influence for each criterion determined, a weighted overlay was performed in ArcMap for all the reclassified criteria raster files. This overlay tool used, combines several raster files to one by first multiplying cell values in each raster by the raster weight of influence and then adds the results to create a single output map. The final values of the output raster were rounded up to whole numbers because the weighted overlay is integer, therefore giving an output raster with the same common scale as that of the input raster.

### 3.2.7. Statistical Evaluation of the Suitability Map

Validation of suitability maps was only performed for rainfed paddy rice as it is the only data that was available. The other maps where not validated because upland rice is still new and very few farmers irrigate rice to give enough data for validation. The only data that was available from Japan International Cooperation Agency (JICA) on rainfed upland rice was not georeferenced making it difficult to use it for the purpose of validation.

Locations for households growing different crops including rice were obtained from the Rural Agricultural Livelihoods Survey (RALS) of 2012 data collected by Indaba Agricultural Policy Research Institute (IAPRI) in collaboration with Central Statistical Office (CSO) and Ministry of Agriculture (CSO/MAL/IAPRI, 2015). RALS is a nationally representative panel survey designed to obtain a comprehensive picture of Zambia's small and medium-scale farming sector using the 2010 census sampling frame. The data obtained through this survey is unique because it is georeferenced. The sampling frame for the RALS 2012 survey was based on the 2010 Census of Housing and Population, CSO/MAL/IAPRI, (2015).

A stratified two-stage sample design (CSO, 2012) was used. The first stage involved identifying the Primary Sampling Unit (PSU) which was one or more Standard Enumeration Areas (SEAs) each comprising a minimum of 30 agricultural households. The SEA was the smallest area with well-defined boundaries identified on census sketch maps. The second stage involved listing and identification of agricultural households in selected SEAs. The listed agricultural households were then stratified into three categories A, B and C (CSO/MAL/IAPRI, 2015). Category C comprised households with 5 to 19.99 ha of land under crops, grown one or more special crops, raising ≥ 50 cattle, ≥20 pigs, ≥ 30 goats and or ≥ 50 chickens. Category B comprised agricultural households with 2 to 4.99 hectares area under crop and category A comprised households with 0 to 1.99 hectares of land under crop or owing livestock numbers less than those specified in category C.

Systematic sampling from the household list in the SEA was then used to select 20 households distributed across the three strata. Where all the three categories had adequate numbers of households listed, the sample household distribution was C=10, B=5 and A=5. Where there were shortfalls in category C, all households in this category where selected and the difference from 20 was equally allocated to categories B and A. If the difference from 20 could not be equally allocated to the two categories, category B was allocated one more sample household than category A. Where there was no household in category C, 10 sample households were allocated to category B, and 10 to category A. Where there was no household in category C and less than 10 in category B, all were included in the sample and the allocation for category A was increased to make up for the shortfall from the required number of 20 sample households. Where all households fall in category A, all the required 20 sample

households were selected from that category. For each stratum, systematic sampling was done to select the required housed holds. First the sampling interval was calculated by dividing the total number of households in the category by the sample number. Then the random start number was selected by randomly selecting a column from the table of random numbers. Starting from the top of that column, the first random number between 1 and the number of households in the category was selected, inclusive as the first corresponding selected household in the sample. To add the next household number, the sampling interval was added to the chosen random number and this procedure was repeated to add remaining households of the sample (CSO, 2012).

The RALS 2012 covered 442 Standard Enumeration Areas (SEAs) across the 10 provinces and a total of 8,840 households (CSO/MAL/IAPRI, 2015). Figure 5 shows the SEA locations for RALS 2012.

The extent to which the distribution of rice producers from the three categories is related to suitability was examined in contingency tables and data from the RALS 2012 survey was used for this analysis. Data cleaning involved removal of spurious values in the x and y coordinates. The need for this was indicated when the raw data were first plotted, showing points lying outside the borders of Zambia. The mean coordinates of all households were computed in each SEA (SEA centroid), and then the households were removed from the data set if the notional distance to the SEA centroid exceeded 10km. After data cleaning, a total of 7,516 households were used to test the null hypothesis that the presence and absence of rice at a sample site and the rice suitability index are independently distributed. Because this evaluation is for paddy rice, only those households that planted local varieties were considered in the presence category. Those that planted improved varieties were put in absence category as it is very likely that some of the improved varieties might be upland rice.

Contingency tables were obtained which show the distribution of observations between Suitability Class (columns) and Crop Presence (rows: rice present or absent). These were then analyzed with the chisq.test function of the package stats for the R platform (R Core Team, 2019). This was done separately for farms in the three categories. The test statistic, $X^2$, is the sum over all cells of the squared difference between the observed number of households and the expected number under a null hypothesis of random association, the squared difference being divided by the expected value. Under the null hypothesis, which is of random association between

crop presence and suitability, the expected number of households in any cell is equal to the product of the corresponding row and column totals divided by the total number of observations in the table. If the null hypothesis is true, then the $X^2$ statistic is distributed as $\chi^2$ with degrees of freedom equal to $(n_r - 1) \times (n_c - 1)$ where $n_r$ and $n_c$ are respectively, the number of rows and columns in the contingency table. The result of this analysis is interpreted as follows. If the null hypothesis is accepted, then there is no evidence of any association between the suitability of land for rice production on one hand, and the presence or absence of a rice crop in the other. However, if the suitability index is informative, then expected is a larger proportion of sites where rice is grown where the suitability index is large than where it is small. This is despite the fact that rice might be grown, for cultural or economic reasons, at some unsuitable sites, and similarly might not be grown at some sites where it is suitable. Thus, if the null hypothesis can be rejected, and there are more sites with rice grown in the classes with larger suitability than expected under the null hypothesis, then this is evidence that the suitability index is, indeed, informative.

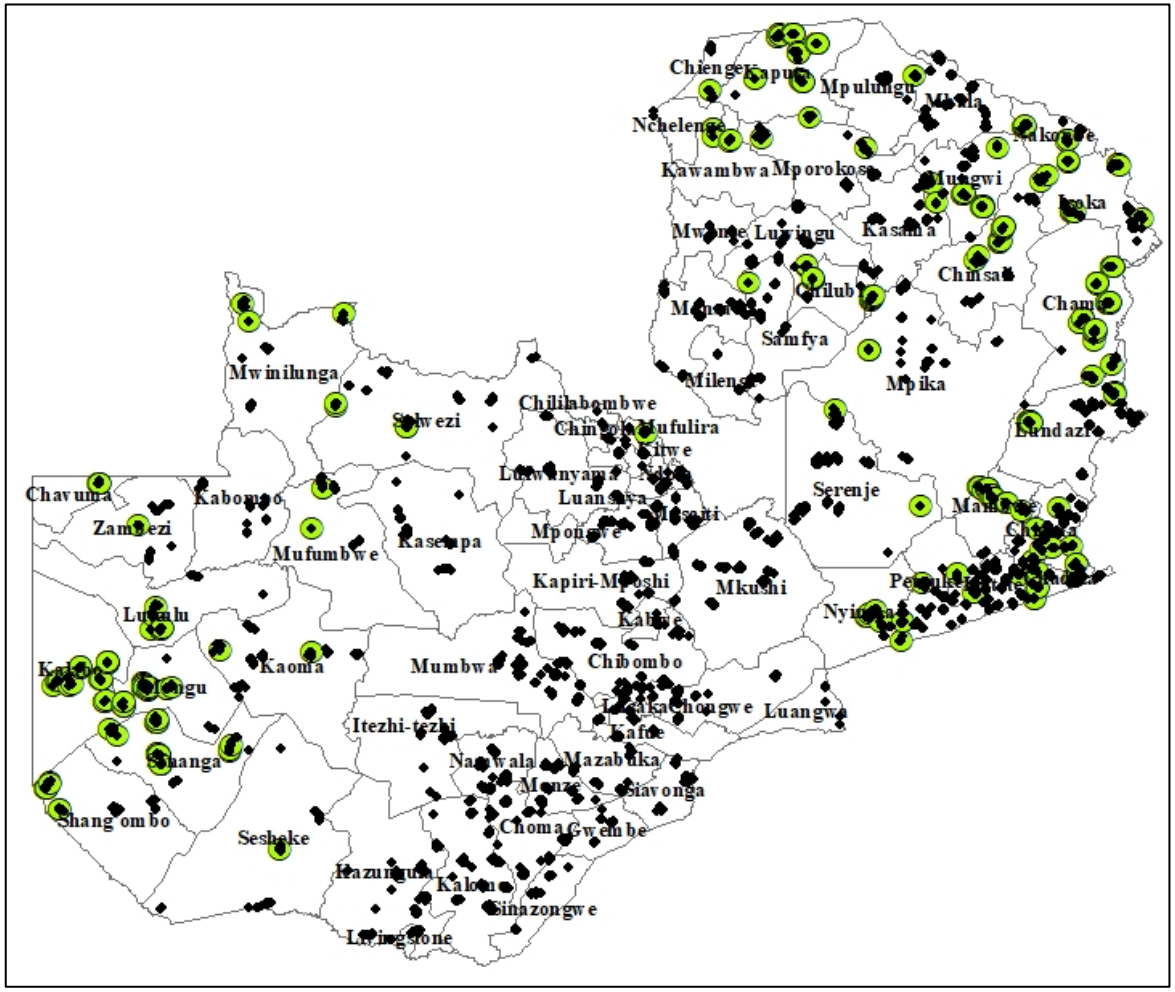



*Figure 6: Standard Enumeration Area (SEA) locations for the RALS 2012 survey*
*Note: The points in black are for all SEAs and the ones with the light green circle are the locations with paddy rice present.*

# CHAPTER FOUR

## 4.0. RESULTS

### 4.1. Spatial Interpolation of Exchangeable Acidity (pH), Soil Phosphorus (P) and Soil Organic Carbon (SOC)

**Variable selection**

Table 14 shows the statistical summary of residuals from the exploratory models for soil pH, soil phosphorus and soil organic carbon and Figures 7, 8 and 9 show the distribution of the residuals from the exploratory models of soil pH, soil phosphorus and soil organic carbon, respectively. The histograms for soil pH and soil organic carbon appear symmetrical and normal. The residuals for soil pH and soil organic carbon have octile skewness inside the range [−0.2,0.2] and skewness inside [−1,1], which would mean that a transformation is not normally considered necessary (Rawlins et al., 2005, Webster and Oliver, 2007). But for soil phosphorus, the histogram was not symmetrical, the residues had octile skewness outside the range [−0.2,0.2] and skewness inside [−1,1], hence transformation was necessary and it can be observed that after transformation (Table 14), the octile skewness inside the range [−0.2,0.2] and skewness inside [−1,1].

*Table 12: Statistical summary of residuals from the exploratory models for soil pH, soil phosphorus and soil organic carbon*

| | **mean** | **Median** | **Variance** | **SD** | **Skewness** | **Octile skewness** | **Kurtosis** |
|---|---|---|---|---|---|---|---|
| Soil pH | 2.366e-16 | -0.040 | 0.167 | 0.408 | 0.487 | 0.159 | 1.044 |
| Soil phosphorus | 1.01e-15 | -0.123 | 2.683 | 1.638 | 0.344 | 0.152 | 0.420 |
| Soil organic carbon | 4.165e-16 | -0.006 | 0.065 | 0.254 | 0.326 | 0.025 | 2.483 |

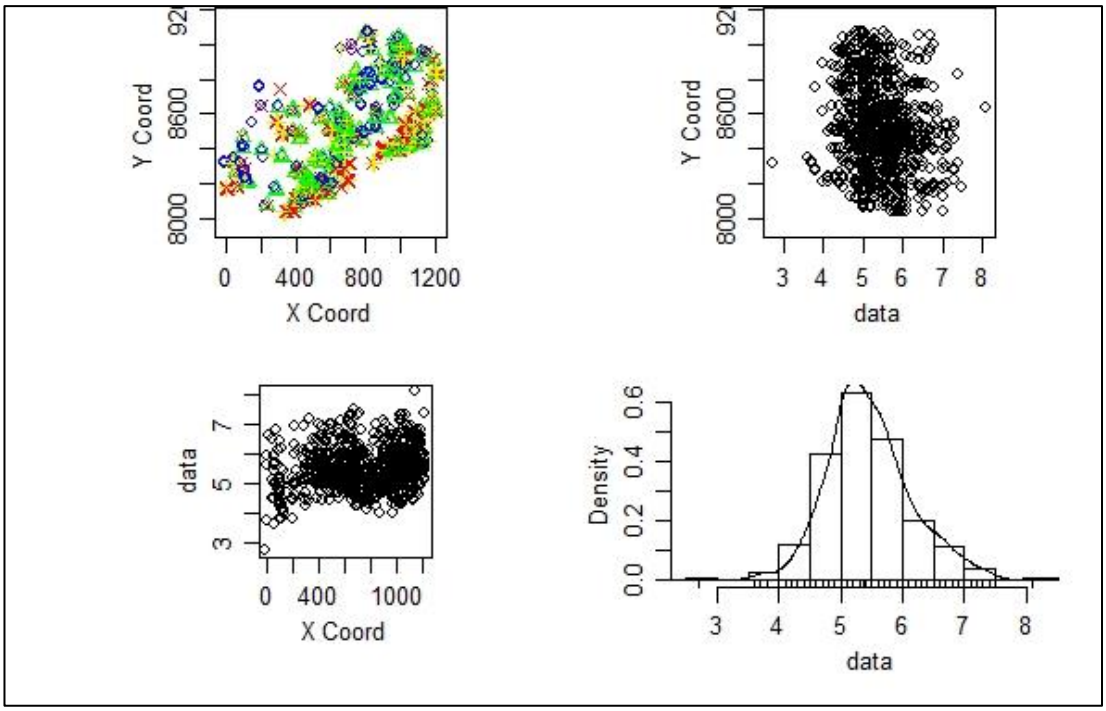



*Figure 7: Top left: point locations, top right and bottom left are the data values against coordinates and the bottom right histogram of the soil pH data*

.

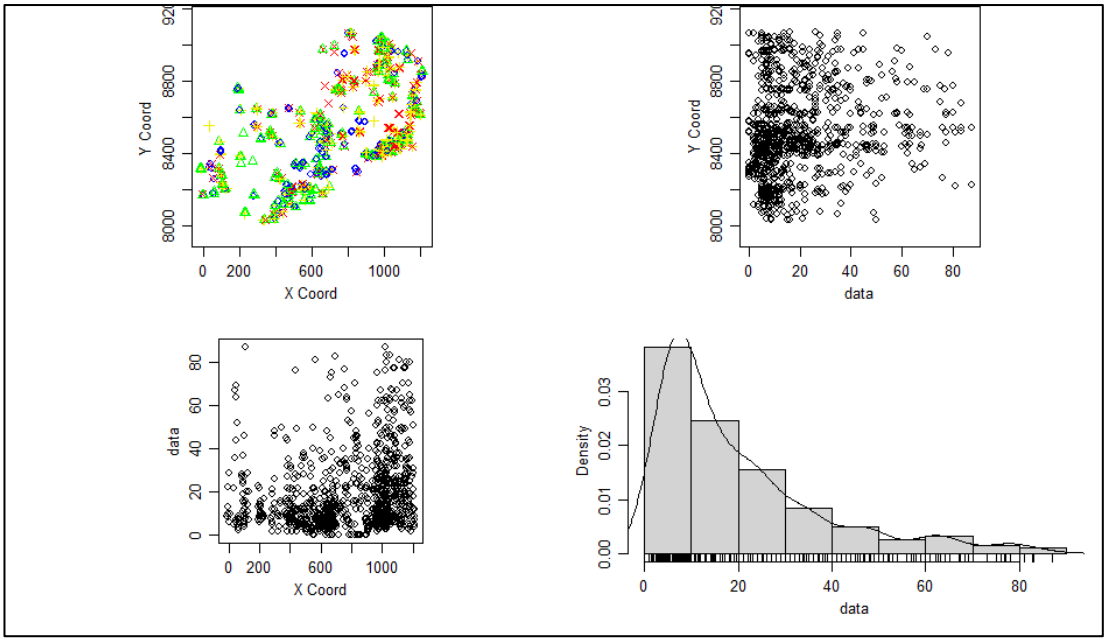



*Figure 8: Top left: point locations, top right and bottom left are the data values against coordinates and the bottom right histogram of the soil phosphorus data*

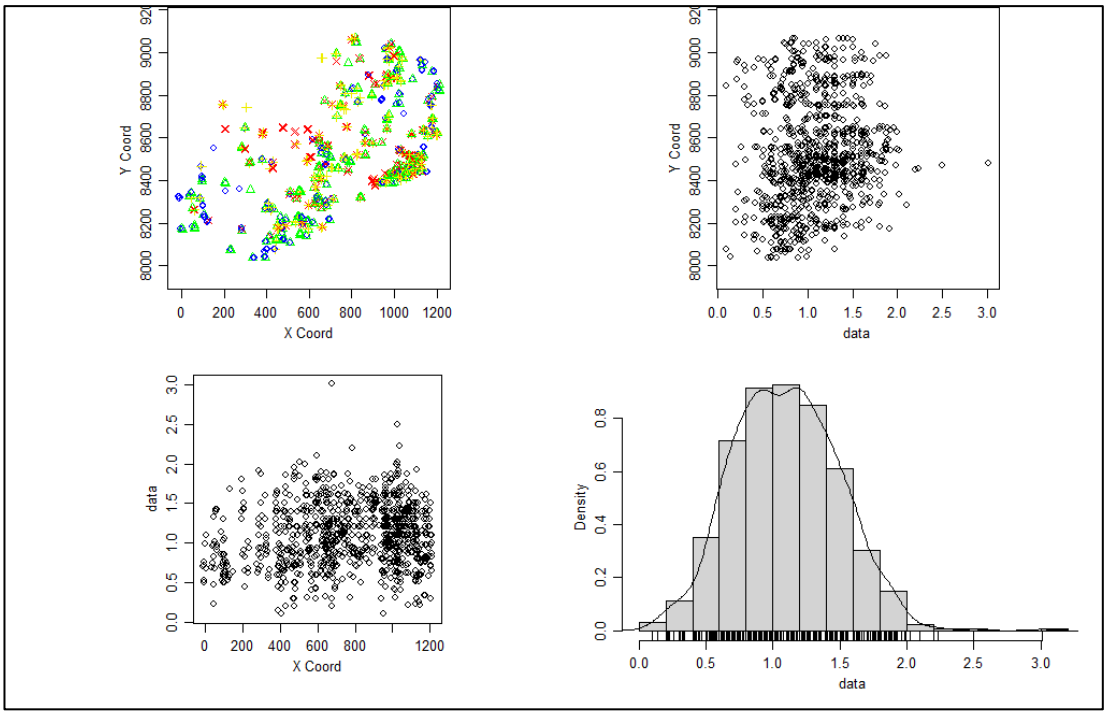



*Figure 9: Top left: point locations, top right and bottom left are the data values against coordinates and the bottom right histogram of the soil organic carbon data*

Table 15 shows the REML estimates of parameters and AIC for the exploratory model, null model and the hypotheses tests for soil pH, soil phosphorus, soil organic carbon. Table 16 shows the log likelihood ratio and p-values of each test at respective degree of freedom (df) and chi-square distribution values for soil pH, soil phosphorus, soil organic carbon. As can be observed for soil pH, the likelihood ratios of tests 1,4 and 8 were greater than the chi-square distribution value values. Therefore, the null hypotheses for these cases were rejected, and the predictors retained during the sequential testing. The rest of the tests had log likelihood ratio less than their respective chi-square distribution values. Hence, the null hypothesis was accepted for these predictors and they were dropped. This same principle was followed for soil phosphorus and soil organic carbon.

*Table 13: REML estimates of parameters and AIC for the exploratory model, null model and the hypothesis tests for soil pH, soil phosphorus, soil organic carbon*

| Test | Predictors | Partial Sill | Range | Nugget | AIC | |
|---|---|---|---|---|---|---|
| | | | | | Max. likelihood | Non-spatial |
| Soil pH | | | | | | |
| | Exploratory (all predictors) | 0.137 | 21.08 | 0.218 | 1570 | 1664 |
| 0 | mean | 0.259 | 68.28 | 0.224 | 1618 | 1946 |
| 1 | Rainfall | 0.217 | 51.87 | 0.222 | 1611 | 1865 |
| 2 | Rainfall + Soil class | 0.197 | 52.71 | 0.221 | 1622 | 1817 |
| 3 | Rainfall +Slope | 0.214 | 51.98 | 0.222 | 1610 | 1856 |
| 4 | Rainfall + Elevation | 0.180 | 39.70 | 0.220 | 1598 | 1774 |
| 5 | Rainfall + Elevation + Valley depth | 0.173 | 33.48 | 0.218 | 1598 | 1770 |
| 6 | Rainfall+ Elevation + Landcover | 0.174 | 40.09 | 0.220 | 1603 | 1769 |
| 7 | Rainfall + Elevation + NDVI | 0.181 | 39.83 | 0.220 | 1600 | 1776 |
| 8 | Rainfall + Elevation + LS | 0.174 | 38.84 | 0.221 | 1596 | 1760 |
| 9 | Rainfall + Elevation + LS + CNBL | 0.171 | 37.40 | 0.221 | 1597 | 1783 |
| 10 | Rainfall + Elevation + LS + RSP | 0.171 | 38.13 | 0.221 | 1598 | 1751 |
| 11 | Rainfall + Elevation + LS + Aspect | 0.173 | 39.18 | 0.221 | 1598 | 1760 |
| Soil Phosphorus | | | | | | |
| | Exploratory (all predictors) | 3.424 | 62.16 | 3.565 | 4063 | 4237 |
| 0 | mean | 3.636 | 65.93 | 3.522 | 4183 | 4441 |
| 1 | Rainfall | 3.480 | 64.61 | 3.526 | 4179 | 4430 |
| 2 | Rainfall + Soil class | 3.253 | 63.90 | 3.527 | 4200 | 4401 |
| 3 | Rainfall +Slope | 3.487 | 64.34 | 3.521 | 4180 | 4432 |
| 4 | Rainfall + Elevation | 3.330 | 61.09 | 3.530 | 4178 | 4401 |
| 5 | Rainfall + Valley depth | 3.480 | 64.6 | 3.526 | 4181 | 4431 |
| 6 | Rainfall+ Landcover | 3.440 | 64.27 | 3.494 | 4196 | 4435 |
| 7 | Rainfall + NDVI | 3.462 | 65.13 | 3.514 | 4176 | 4426 |
| 8 | Rainfall + NDVI+LS | 3.479 | 64.87 | 3.505 | 4178 | 4428 |
| 9 | Rainfall + NDVI + cnbl | 3.299 | 62.48 | 3.522 | 4174 | 4394 |
| 10 | Rainfall + NDVI + cnbl+ RSP | 3.300 | 62.61 | 3.523 | 4176 | 4395 |
| 11 | Rainfall + NDVI + cnbl+ Aspect | 3.282 | 62.37 | 3.526 | 4176 | 4389 |
| Soil Organic Carbon | | | | | | |
| | Exploratory (all predictors) | 0.061 | 40.28 | 0.083 | 667 | 791 |
| 0 | Mean | 0.082 | 69.02 | 0.085 | 658 | 913 |
| 1 | Soil class | 0.063 | 51.03 | 0.082 | 652 | 844 |
| 2 | soillclass + slope | 0.063 | 49.55 | 0.082 | 652 | 845 |
| 3 | sloilclass + rainfall | 0.058 | 40.84 | 0.081 | 649 | 815 |
| 4 | sloilclass + rainfall + NDVI | 0.058 | 41.18 | 0.081 | 651 | 816 |
| 5 | sloilclass + rainfall + cnbl | 0.054 | 32.17 | 0.080 | 648 | 782 |
| 6 | sloilclass + rainfall + valley | 0.058 | 41.21 | 0.081 | 651 | 817 |
| 7 | sloilclass + rainfall + elevation | 0.053 | 31.50 | 0.080 | 646 | 775 |
| 8 | sloilclass + rainfall + elevation+rsp | 0.054 | 34.39 | 0.081 | 648 | 777 |
| 9 | sloilclass + rainfall + elevation + landcover | 0.050 | 32.12 | 0.081 | 664 | 772 |
| 10 | sloilclass + rainfall + elevation + ls | 0.053 | 31.15 | 0.080 | 648 | 776 |
| 11 | sloilclass + rainfall + elevation + aspect | 0.053 | 31.53 | 0.080 | 648 | 777 |

*Table 14: likelihood ratio and p-values of each hypothesis test at respective degree of freedom(df) and chi-square distribution values for soil pH, Soil phosphorus and soil organic carbon*

| **Test** | | **df** | **Chi-square** | **Likelihood ratio** | **p-value** |
|---|---|---|---|---|---|
| Soil pH | | | | | |
| 1 | Rainfall | 1 | 3.841 | 9.533 | 0.002 |
| 2 | Rainfall + Soil class | 17 | 27.587 | 22.506 | 0.166 |
| 3 | Rainfall +Slope | 1 | 3.841 | 2.429 | 0.119 |
| 4 | Rainfall + Elevation | 1 | 3.841 | 14.416 | 0.000 |
| 5 | Rainfall + Elevation + Valley depth | 1 | 3.841 | 1.817 | 0.177 |
| 6 | Rainfall+ Elevation + Landcover | 8 | 15.507 | 10.836 | 0.211 |
| 7 | Rainfall + Elevation + NDVI | 1 | 3.841 | 0.598 | 0.439 |
| 8 | Rainfall + Elevation + LS | 1 | 3.841 | 3.946 | 0.047 |
| 9 | Rainfall + Elevation + LS + CNBL | 1 | 3.841 | 1.047 | 0.306 |
| 10 | Rainfall + Elevation + LS + RSP | 1 | 3.841 | 0.403 | 0.525 |
| 11 | Rainfall + Elevation + LS + Aspect | 1 | 3.841 | 0.213 | 0.644 |
| Soil Phosphorus | | | | | |
| 1 | Rainfall | 1 | 3.841 | 6.131 | 0.013 |
| 2 | Rainfall + Soil class | 17 | 27.587 | 12.491 | 0.770 |
| 3 | Rainfall +Slope | 1 | 3.841 | 0.177 | 0.674 |
| 4 | Rainfall + Elevation | 1 | 3.841 | 3.019 | 0.082 |
| 5 | Rainfall + Valley depth | 1 | 3.841 | 0.000 | 0.999 |
| 6 | Rainfall+ Landcover | 8 | 15.507 | 8.499 | 0.810 |
| 7 | Rainfall + NDVI | 1 | 3.841 | 4.422 | 0.035 |
| 8 | Rainfall + NDVI+LS | 1 | 3.841 | 0.409 | 0.523 |
| 9 | Rainfall + NDVI + cnbl | 1 | 3.841 | 4.075 | 0.044 |
| 10 | Rainfall + NDVI + cnbl+ RSP | 1 | 3.841 | 0.084 | 0.772 |
| 11 | Rainfall + NDVI + cnbl+ Aspect | 1 | 3.841 | 0.222 | 0.637 |
| Soil Organic Carbon | | | | | |
| 1 | Soil class | 17 | 27.587 | 39.714 | 0.001 |
| 2 | Soil class + slope | 1 | 3.841 | 1.388 | 0.239 |
| 3 | Soil class + rainfall | 1 | 3.841 | 4.643 | 0.031 |
| 4 | Soil class + rainfall + NDVI | 1 | 3.841 | 0.133 | 0.716 |
| 5 | Soil class + rainfall + cnbl | 1 | 3.841 | 3.421 | 0.064 |
| 6 | Soil class + rainfall + valley | 1 | 3.841 | 0.289 | 0.591 |
| 7 | Soil class + rainfall + elevation | 1 | 3.841 | 5.023 | 0.025 |
| 8 | Soil class + rainfall + elevation + rsp | 1 | 3.841 | 0.634 | 0.426 |
| 9 | Soil class + rainfall + elevation + landcover | 8 | 15.507 | 8.425 | 0.587 |
| 10 | Soil class + rainfall + elevation + ls | 1 | 3.841 | 0.107 | 0.744 |
| 11 | Soil class + rainfall + elevation + aspect | 1 | 3.841 | 0.003 | 0.955 |

Figures 10a, 11a and 12a shows the alpha wealth (explained in section 3.2.2.3) after each test and it can be observed that the quantity of the wealth is increased when the null hypothesis is rejected and depleted when the null hypothesis retained, and it goes to zero at the end of the sequence. Figures 10b, 11b and 12b shows the p-values (open symbols) for the successive tests of additional predictors, as in Table 15 and 16, and the threshold (solid symbols) against which each successive p-value is tested to achieve FDR. On this basis rainfall and elevation were selected as predictors for soil pH, rainfall alone was selected as predictor for soil phosphorous and soil class alone was selected as predictor for soil organic carbon.

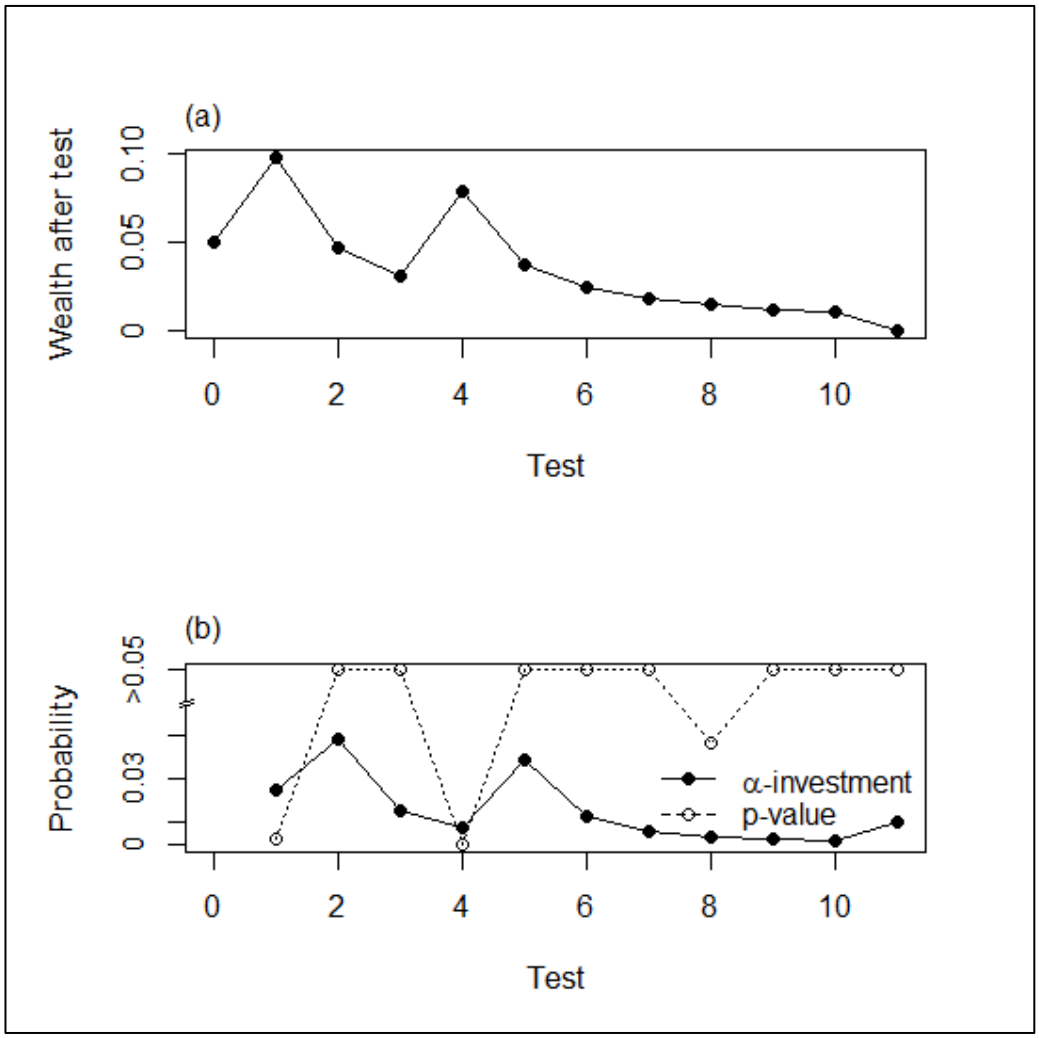



*Figure 10: alpha wealth after each test (a). probability of alpha investment and p-values (b) for soil pH*

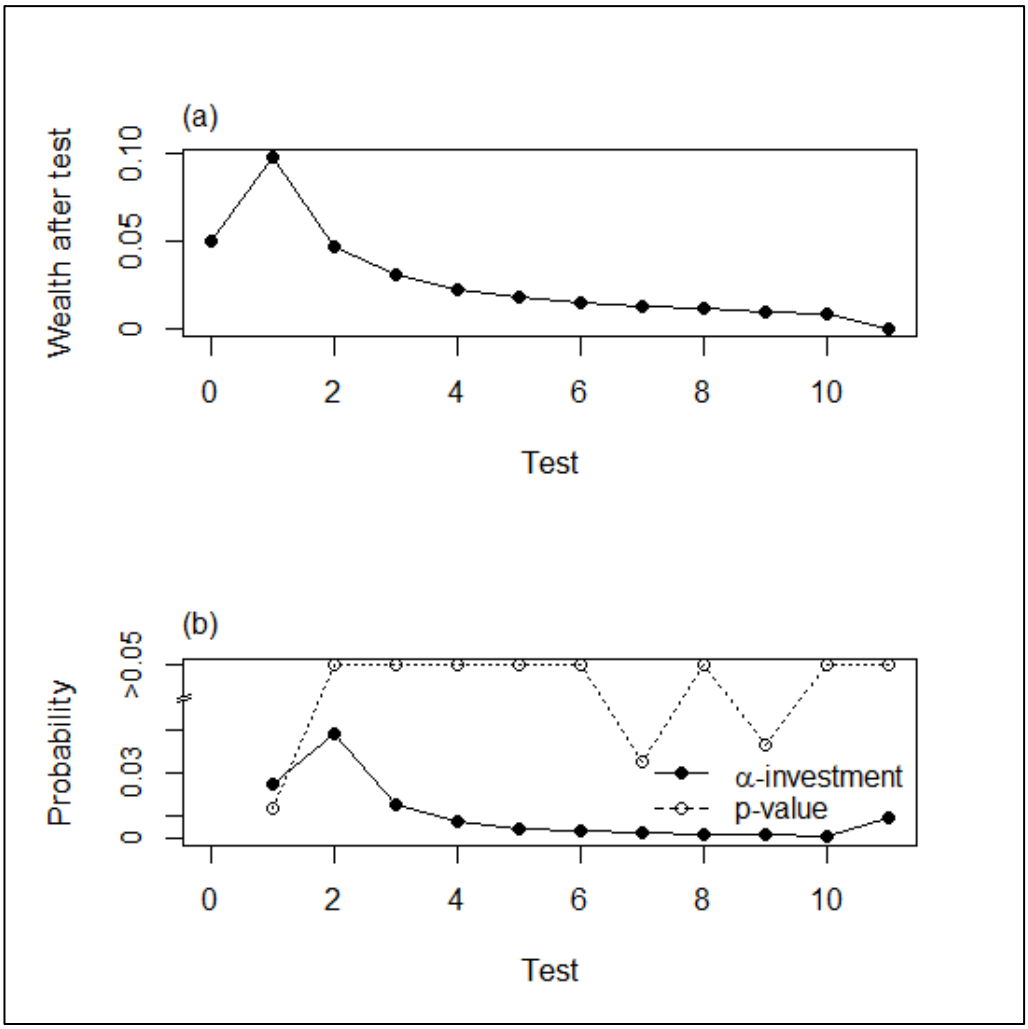



*Figure 11: alpha wealth after each test (a). probability of alpha investment and p-values (b) for soil phosphorus*

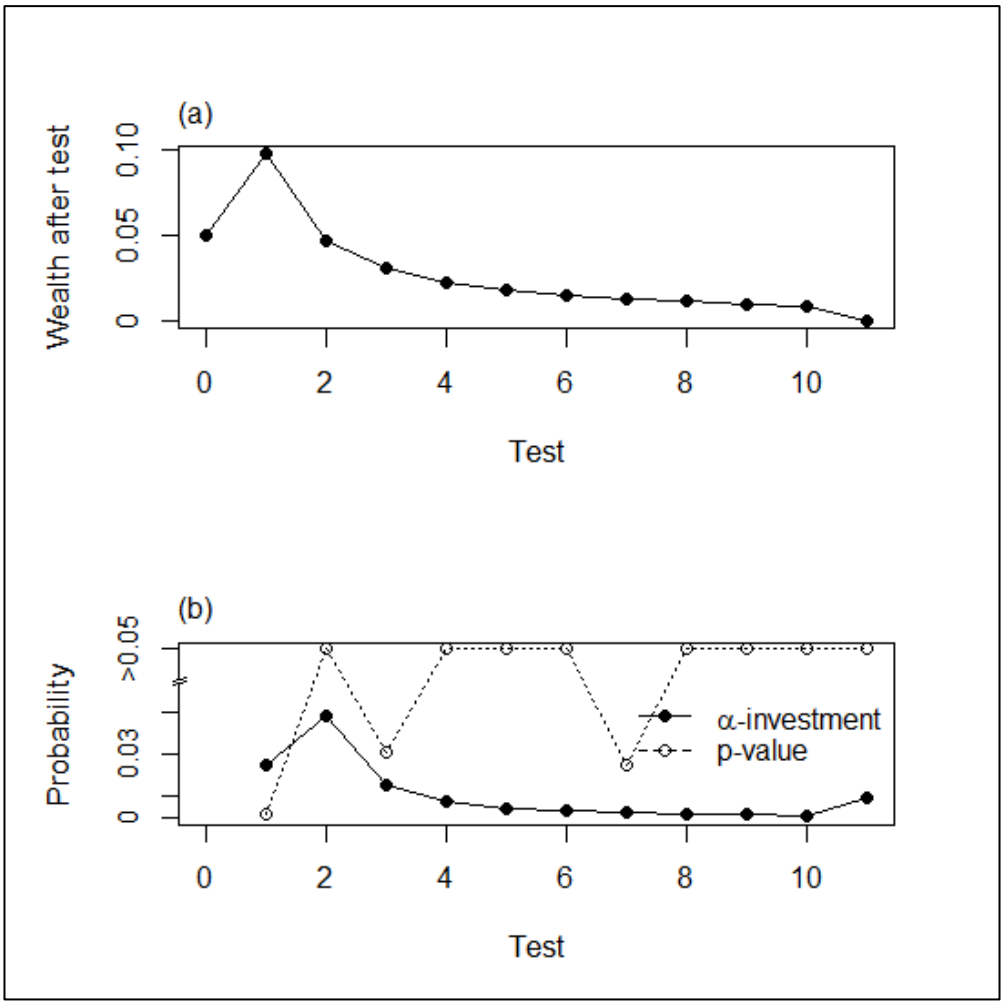



*Figure 12: alpha wealth after each test (a) probability of alpha investment and p-values (b) for soil organic carbon*

**Spatial Prediction of Soil pH, Soil Phosphorus and Soil Organic Carbon**

The estimated covariance parameters for the linear mixed models used for spatial prediction of soil pH by the E-BLUP with elevation and rainfall as fixed effects for prediction (Method A) and a constant mean as the only fixed effect (Method B, equivalent to ordinary kriging) are shown in Table 17. The nugget, partial sill and range for the model with rainfall and elevation as fixed effects are 0.220, 0.195 and 33.95, respectively. These values are smaller than the corresponding parameters for the model with a constant mean as the only fixed effect (0.224, 0.269 and 72.83). AIC values for both models are less than those of respective non-spatial AIC. On this basis concluded that there is evidence for spatial dependence in the random component of the LMM, and so potentially benefits in computing the E-BLUP for spatial prediction at unsampled sites.

*Table 15: Covariance parameters for spatial prediction of soil pH*

| Method | Partial Sill | Range | Nugget | AIC | |
|---|---|---|---|---|---|
| | | | | Max.likelihood | Non-spatial |
| A | 0.195 | 33.95 | 0.220 | 1184 | 1310 |
| B | 0.269 | 72.83 | 0.224 | 1614 | 1945 |

*Note: A= REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), B=REML-EBLUP with the only fixed effect a constant mean (ordinary kriging).*

The estimated covariance parameters for the linear mixed models to be used for spatial prediction of soil phosphorus by the E-BLUP with rainfall as fixed effects for prediction (Method A) and a constant mean as the only fixed effect (Method B, equivalent to ordinary kriging) are shown in Table 18. The nugget, partial sill and range for the model with rainfall as fixed effects are 3.652, 3.742 and 65.11, respectively. These values are smaller than the corresponding parameters for the model with a constant mean as the only fixed effect (3.649, 3.877 and 66.37). AIC values for both models are less than those of respective non-spatial AIC. Just like for soil pH, there is evidence for spatial dependence in the random component of the LMM, and so potentially benefits the computation of the E-BLUP for spatial prediction at unsampled sites.

*Table 16: Soil phosphorous parameters for (A) = REML-EBLUP with rainfall as predictor selected through alpha-investment, (B) =REML-EBLUP (ordinary kriging)*

| Method | Partial Sill | Range | Nugget | AIC | |
|---|---|---|---|---|---|
| | | | | Max.likelihood | Non-spatial |
| A | 3.742 | 65.11 | 3.652 | 4206 | 4463 |
| B | 3.877 | 66.37 | 3.649 | 4215 | 4478 |

*Table 17: Soil organic carbon for A= REML-EBLUP with soil class as predictor selected through alpha-investment, B=REML-EBLUP (ordinary kriging)*

| Method | Partial Sill | Range | Nugget | AIC | |
|---|---|---|---|---|---|
| | | | | Max.likelihood | Non-spatial |
| A | 0.0645 | 55.02 | 0.0826 | 617.3 | 825.3 |
| B | 0.085 | 74.54 | 0.0849 | 654.6 | 913.5 |

Tables 20, 21 and 22 show the number of trees, number of predictors, number of variables considered at each split, target node size and out-of-bag cross validation of the random forest methods for soil pH, soil phosphorus and soil organic carbon

respectively. For soil pH (Table 20), the out-of-bag MSE and R-squared show that there is a slight reduction in performance of the random forest model with rainfall and elevation

*Table 18: Parameters from Random forest model for soil pH*

| **Method** | **ntree** | **predictors** | **mtry** | **Target node size** | **Out-of-Bag MSE** | **Out-of-Bag R-squared** |
|---|---|---|---|---|---|---|
| C | 200 | 11 | 3 | **5** | 0.31 | 0.30 |
| D | 200 | 2 | 1 | 5 | 0.32 | 0.27 |

*Note: ntree = number of trees in the forest; mtry = number of variables considered at each split. (C) random forest with all predictors(D) random forest with rainfall and elevation as predictors selected through alpha-investment.*

*Table 19: Parameters from Random forest model for soil phosphorus*

| **Method** | **ntree** | **predictors** | **mtry** | **Target node size** | **Out-of-Bag MSE** | **Out-of-Bag R-squared** |
|---|---|---|---|---|---|---|
| C | 200 | 11 | 3 | **5** | 263.98 | 0.18 |
| D | 200 | 11 | 3 | 5 | 5.36 | 0.23 |
| E | 200 | 1 | 1 | 5 | 377.06 | -0.17 |
| F | 200 | 1 | 1 | 5 | 8.19 | -0.18 |

*Note: ntree = number of trees in the forest; mtry = number of variables considered at each split. (C) random forest with all predictors data not transformed (D) random forest with all predictors data transformed and then back transformed. (E) random forest with rainfall as predictor selected through alpha-investment data not transformed. (F) random forest with rainfall as predictors selected through alpha-investment data not transformed.*

*Table 20: Parameters from Random forest model for soil organic carbon*

| **Method** | **ntree** | **predictors** | **mtry** | **Target node size** | **Out-of-Bag MSE** | **Out-of-Bag R-squared** |
|---|---|---|---|---|---|---|
| C | 200 | 11 | 3 | **5** | 0.11 | 0.27 |
| D | 200 | 1 | 1 | 5 | 0.14 | 0.07 |

*Note: ntree = number of trees in the forest; mtry = number of variables considered at each split. (C) random forest with all predictors(D) random forest with soil class as predictor selected through alpha-investment.*

Tables 23, 24 and 25 show the permutation variable importance (defined in section 2.4) values for soil pH, Soil Phosphorus and soil organic carbon, respectively when random forest model is fit. For soil pH (Table 23) each predictor when a random forest model is fit with all predictors alone and when we include null predictors (sim1 to sim6) which were generated by simulation to examine how random forest variable importance performs with predictors that have no relation to the data. For the random

forest model, the most important variable is elevation with importance value of 0.166, followed by Channel Network Base Level with value of 0.155. Some variable importance values for some predictors are almost equal or even less, but their p-values are much smaller. Null variables sim1 and sim6 despite have low variable importance values, but very small p-values (P<0.01). The inclusion of these null variables has a substantial effect on the p-values of some predictors such as soil class, slope, landcover.

*Table 21: Permutation variable importance and p-values for soil pH data when a random forest model is fit with all predictors alone and when null predictors (sim1 to sim6) are included.*

| **Predictor** | **No null predictors** | | **Null predictors included** | |
|---|---|---|---|---|
| | **Importance** | **p-value** | **Importance** | **p-value** |
| Rain | 0.0889 | 0.0099 | 0.0912 | 0.0099 |
| Soil class | 0.0231 | 0.0198 | 0.0153 | 0.0099 |
| Slope | 0.0318 | 0.8218 | 0.0268 | 0.2079 |
| elevation | 0.1657 | 0.0099 | 0.1718 | 0.0099 |
| Valley | 0.0761 | 0.0099 | 0.0544 | 0.0099 |
| landcover | 0.0063 | 0.4752 | 0.0074 | 0.0792 |
| NDVI | 0.0499 | 0.0099 | 0.0470 | 0.0099 |
| Ls | 0.0467 | 0.3267 | 0.0286 | 0.1287 |
| Cnbl | 0.1554 | 0.0099 | 0.1477 | 0.0099 |
| Rsp | 0.0554 | 0.0198 | 0.0330 | 0.0495 |
| Aspect | 0.0148 | 0.1188 | 0.0066 | 0.3663 |
| Sim1 | | | 0.0279 | 0.0099 |
| Sim2 | | | 0.0232 | 0.0198 |
| Sim3 | | | 0.0187 | 0.0594 |
| Sim4 | | | 0.0204 | 0.0198 |
| Sim5 | | | 0.0160 | 0.1782 |
| Sim6 | | | 0.0274 | 0.0099 |

*Note: Sim 1 to Sim 6 are null predictors which were generated by simulation to examine how random forest variable importance performs with predictors that have no relation to the data.*

*Table 22: Permutation variable importance and p-values for untransformed and transformed soil phosphorous data when a random forest model is fit with all predictors*

| **Predictor** | **No transformation** | | **Transformed** | |
|---|---|---|---|---|
| | **importance** | **p-value** | **Importance** | **p-value** |
| Slope | 29.4992 | 0.6139 | 0.7824 | 0.1584 |
| Aspect | 22.3515 | 0.0099 | 0.7197 | 0.0099 |
| rainfall | 70.3860 | 0.0099 | 1.4883 | 0.0099 |
| Cnbl | 129.6755 | 0.0099 | 2.2515 | 0.0099 |
| Ls | 28.8579 | 0.6040 | 0.6813 | 0.2871 |
| Valley | 38.2470 | 0.0594 | 0.9548 | 0.0099 |
| Rsp | 47.9913 | 0.0099 | 0.9669 | 0.0099 |
| Soil class | 10.8044 | 0.1485 | 0.2332 | 0.0792 |
| landcover | 6.2560 | 0.3663 | 0.1748 | 0.0792 |
| elevation | 116.5409 | 0.0099 | 1.7054 | 0.0198 |
| NDVI | 39.0748 | 0.0198 | 1.1379 | 0.0099 |

*Note: ls = LS-Factor (a combination of slope and slope length), rsp = relative slope position, cnbl = channel network base Level and NDVI = Normalized difference vegetative index.*

*Table 23: Permutation variable importance and p-values for soil organic carbon data when a random forest model is fit with all predictors.*

| **Variable** | **Importance** | **p-value** |
|---|---|---|
| slope | 0.0131 | 0.2178 |
| aspect | 0.0058 | 0.0495 |
| rainfall | 0.0501 | 0.0099 |
| cnbl | 0.0325 | 0.0099 |
| ls | 0.0149 | 0.1287 |
| valley | 0.0244 | 0.0099 |
| rsp | 0.0170 | 0.0099 |
| Soil class | 0.0160 | 0.0099 |
| landcover | 0.0082 | 0.0099 |
| elevation | 0.0342 | 0.0099 |
| NDVI | 0.0202 | 0.0099 |

*Note: ls = LS-Factor (a combination of slope and slope length), rsp = relative slope position, cnbl = channel network base Level and NDVI = Normalized difference vegetative index.*

Tables 26, 27 and 28 show the estimated parameters of the random forest residuals for the exponential, spherical and pure nugget correlation models for soil pH, soil phosphorus and soil organic carbon respectively. As can be observed from the tables 26 and 27, for soil pH and soil phosphorus, the non-spatial model was preferred because the AIC values for the spatial component was higher than that of the non-spatial component in both random forest predictions. Indeed, the fitted correlated

variance for the spatial covariance function was zero. On this basis, there is no scope to improve the RF predictions by a kriging step. But with soil organic carbon (Table 28) the models for random forest with soil class and soil class means as predictors had AIC values for maximum likelihood less than those of the non-spatial AIC showing some spatial pattern in the residuals. Therefore, the residuals where kriged to the prediction locations and then added to the predicted values.

*Table 24: REML Estimated parameters of the exponential, spherical and pure nugget correlation functions for the residuals of the two random forest predictions for soil pH*

| Method | Parameter | Exponential | Spherical | Pure.nugget |
|---|---|---|---|---|
| **RF (dem + rain)** | Partial Sill | 0 | 0 | 0.188 |
| | range | 0 | 0 | 50 |
| | Nugget | 0.188 | 0.188 | 0 |
| | $AIC_{max.lelihood}$ | 1123 | 1123 | 1123 |
| | $AIC_{non\text{-}spatial}$ | 1119 | 1119 | 1119 |
| **RF (all predictors)** | Partial Sill | 0 | 0 | 0.157 |
| | range | 0 | 0 | 50 |
| | Nugget | 0.157 | 0.157 | 0 |
| | $AIC_{max.likelihood}$ | 951.4 | 951.4 | 951.5 |
| | $AIC_{non\text{-}spatial}$ | 947.4 | 947.4 | 947.4 |

*Table 25: REML Estimated parameters of the exponential, spherical and pure nugget correlation functions for the residuals for the random forest predictions for soil phosphorous*

| Method | Parameter | Exponential | Spherical | Pure.nugget |
|---|---|---|---|---|
| | Partial Sill | 0 | 0 | 0 |
| | range | 0 | 0 | 0 |
| | Nugget | 308.8 | 308.8 | 308.8 |
| | $AIC_{max.likelihood}$ | 7962 | 7962 | 7962 |
| **RF_rain (no transformation)** | $AIC_{non-spatial}$ | 7958 | 7958 | 7958 |
| | Partial Sill | 0 | 0 | 0 |
| | range | 0 | 0 | 0 |
| | Nugget | 307 | 307 | 307 |
| | $AIC_{max.likelihood}$ | 7956 | 7956 | 7956 |
| **RF rain (transformed)** | $AIC_{non-spatial}$ | 7952 | 7952 | 7952 |
| | Partial Sill | 129.1 | 8.791 | 0 |
| | range | 0.106 | 10.62 | 0 |
| **RF (no transformation)** | Nugget | 0 | 120.6 | 129.4 |
| | $AIC_{max.lelihood}$ | 7151 | 7154 | 7154 |
| | $AIC_{non-spatial}$ | 7150 | 7150 | 7150 |
| | Partial Sill | 130.2 | 9.084 | 0 |
| | range | 0.1005 | 9.707 | 0 |
| **RF (transformation)** | Nugget | 0 | 121.2 | 130.3 |
| | $AIC_{max.likelihood}$ | 7159 | 7161 | 7161 |
| | $AIC_{non-spatial}$ | 7157 | 7157 | 7157 |

*Table 26: REML Estimated parameters of the exponential, spherical and pure nugget correlation functions for the residuals for the random forest predictions for soil organic carbon*

| Method | Parameter | Exponential | Spherical | Pure.nugget |
|---|---|---|---|---|
| | Partial Sill | 0.0627 | 0.0616 | 0 |
| | range | 40369 | 78285 | 0 |
| | Nugget | 0.0818 | 0.0833 | 0.1383 |
| | $AIC_{max.lelihood}$ | 630.9 | 640.5 | 828 |
| **RF (soil class)** | $AIC_{non-spatial}$ | 824 | 824 | 824 |
| | Partial Sill | 0.0633 | 0.0619 | 0 |
| | range | 31015 | 66884 | 0 |
| **RF (soil class_means)** | Nugget | 0.0803 | 0.0827 | 0.1407 |
| | $AIC_{max.lelihood}$ | 640.7 | 648.9 | 844.7 |
| | $AIC_{non-spatial}$ | 840.7 | 840.7 | 840.7 |
| | Partial Sill | 0.0072 | 0.0085 | 0 |
| | range | 131.2 | 255.6 | 0 |
| **RF (all predictors)** | Nugget | 0.0458 | 0.0444 | 0.053 |
| | $AIC_{max.likelihood}$ | -88.74 | -89.39 | -87.28 |
| | $AIC_{non-spatial}$ | -91.28 | -91.28 | -91.28 |

Figure 13 shows the predicted spatial variability of soil pH for (a) REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), (b) REML-EBLUP with the only fixed effect a constant mean (ordinary kriging (c) random forest with all predictors (d). random forest with elevation and rainfall as predictors selected through alpha-investment. The spatial pattern is similar for all the models with low pH values (less than 5.5) in the Western and Northern parts and higher values (above 6) in the Southern and Eastern parts. The resolution of the raster maps is 1 km which translates to a scale of 1:2,000,000 according to

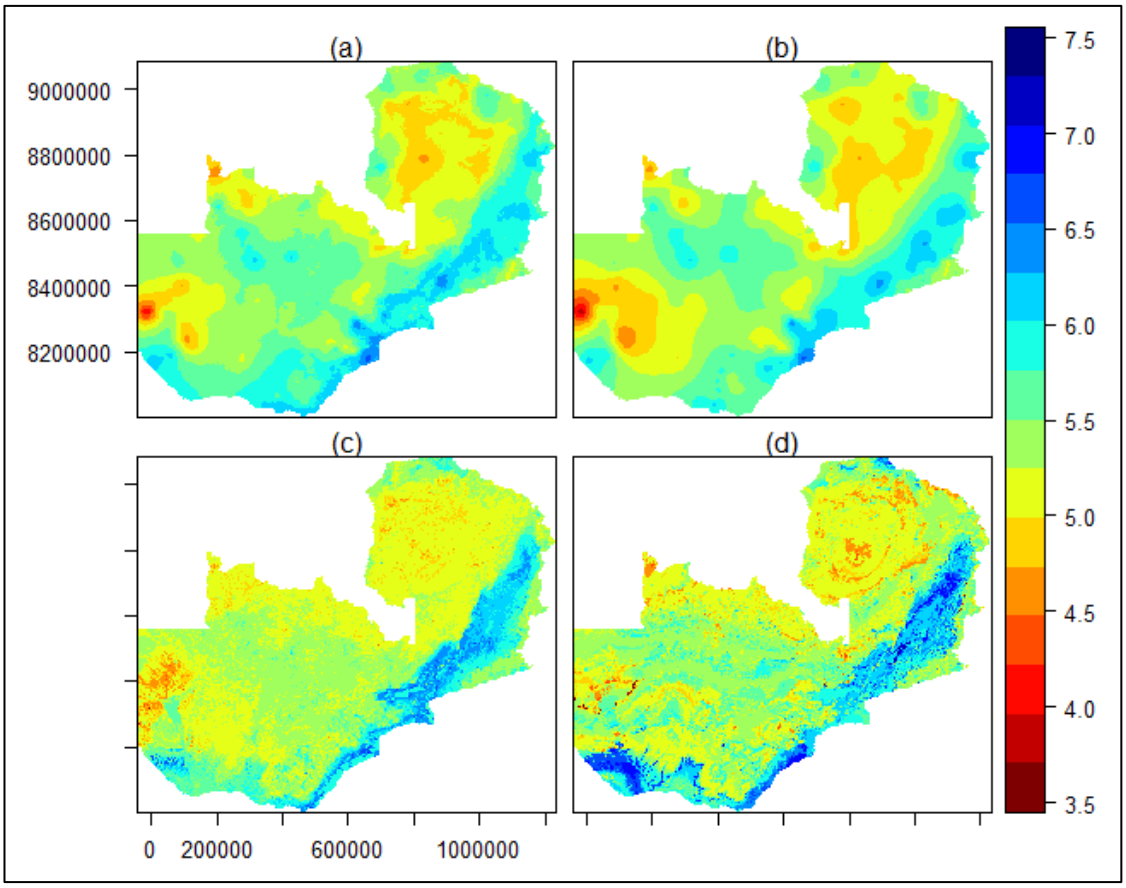



*Figure 13: Prediction maps of soil pH*

*Note: (a) REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), (b)REML-EBLUP with the only fixed effect a constant mean (ordinary kriging (c) random forest with all predictors (d). random forest with elevation and rainfall as predictors selected through alpha-investment. resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*

Figure 14 shows the prediction maps of soil Phosphorous for (a) REML-EBLUP with rainfall as predictors selected through alpha-investment (b) REML-EBLUP with ordinary kriging (c) random forest with all predictors data not transformed (d) random

forest with all predictors data transformed and then back transformed (e) random forest with rainfall as predictors selected through alpha-investment data not transformed (f) random forest with rainfall as predictors selected through alpha-investment data not transformed. As can be observed, from both the maps (Figure 14) and the cross-validation results (Table 30) below comfirm that the results of the transformed and untransformed are the same hence there is no need for data transformation with random forest. Most parts of the country have values below 10 except for a few areas aroud eastern, western and northern provinces.

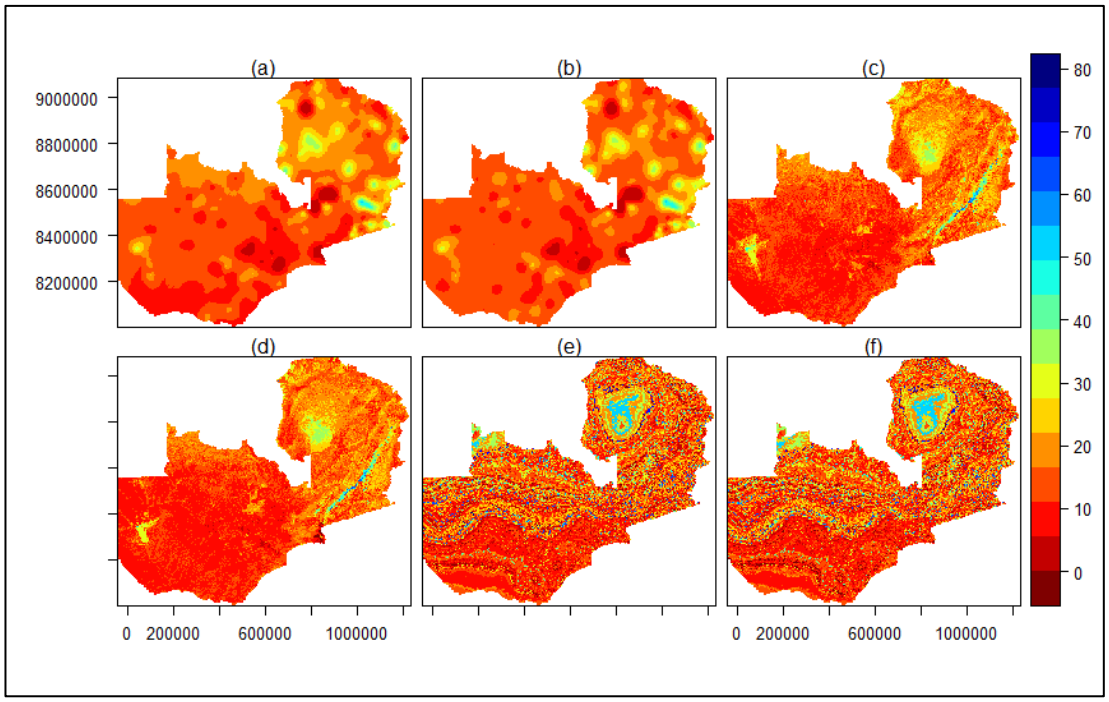



*Figure 14: Prediction maps of soil Phosphorous*

*Note: (a) REML-EBLUP with rainfall as predictors selected through alpha-investment, (b) REML-EBLUP with ordinary kriging (c) random forest with all predictors data not transformed (d) random forest with all predictors data transformed and then back transformed (e) random forest with rainfall as predictors selected through alpha-investment data not transformed (f) random forest with rainfall as predictors selected through alpha-investment data not transformed. Resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*

Figure 15 shows the prediction maps of soil organic carbon for (a) REML-EBLUP with soil class as predictor selected through alpha-investment (b) REML-EBLUP with ordinary kriging (c) random forest with all predictors data (d) random forest with soil class as predictor (e) random forest with soil class plus REML-EBLUP kriged residuals. The western part of the country has very low values, this could be attributed

to the type of soils which are very sandy resulting in leaching. The values are even lower using the random forest models.

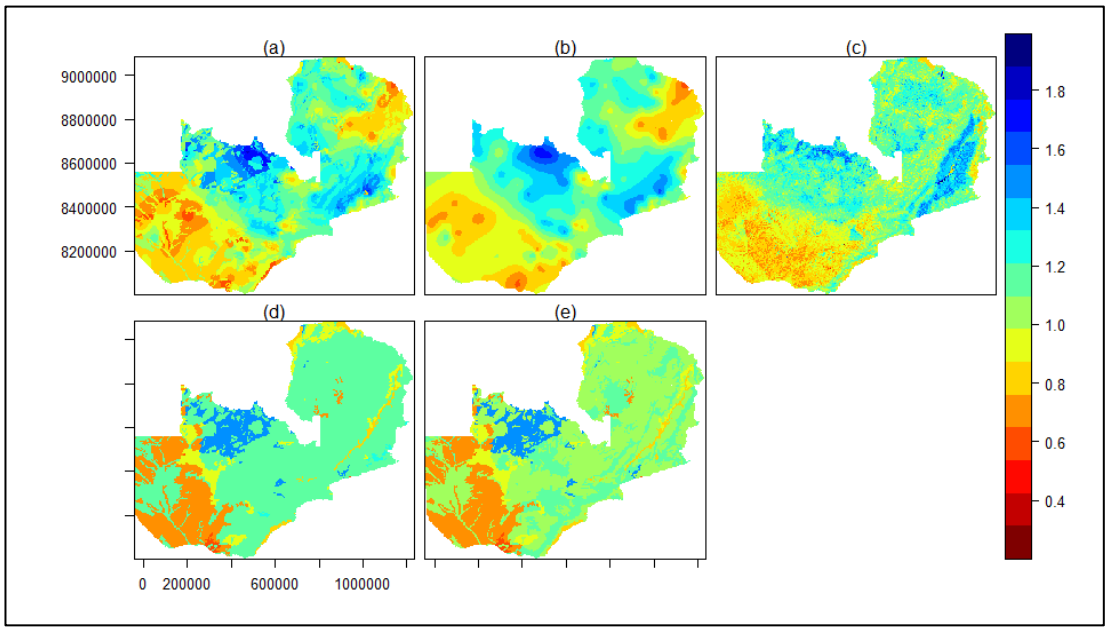



*Figure 15: Prediction maps of soil organic carbon*

*Note: (a) REML-EBLUP with soil class as predictor selected through alpha-investment (b) REML-EBLUP with ordinary kriging (c) random forest with all predictors data (d) random forest with soil class as predictor (e) random forest with soil class plus REML-EBLUP kriged residuals. Resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*

## 4.2. Performance of RF and LMM in mapping the spatial distribution of P, Soil Organic Carbon and pH

Tables 29, 30 and 31 show the summary validation statistics for soil pH, Soil phosphorus and soil organic carbon, respectively. For soil pH (Table 29), the mean and median error values were smallest for the REML-EBLUP (ordinary kriging) method while that of the REML-EBLUP (kriging with external drift) was larger than that of the two random forest methods. But the opposite is true for soil phosphorus (Table 30) and soil organic carbon (Table 31) where the mean and median error values were smallest for the REML-EBLUP (kriging with external drift) method and the random forest methods were larger than those of the two REML-EBLUP methods. For all cases (soil pH, soil phosphorus and soil organic carbon), the MSE and RMSE for the two REML-EBLUP methods were smaller than those of the random forest methods with REML-EBLUP (ordinary kriging) having the smallest values for soil pH and REML-EBLUP (kriging with external drift) having the smallest values for soil

phosphorus and soil organic carbon. There was spatial dependency in the prediction error in all the cases (soil pH, soil phosphorus and soil carbon) with the two REML-EBLUP cases having the smaller partial sill values compared the values for Random forest. The ESE values for all the cases (soil pH, soil phosphorus and soil organic carbon) were larger than the MSE values because the bias (ME) for the models is greater than zero. REML-EBLUP (ordinary kriging) had the smallest ESE value for soil pH while REML-EBLUP (kriging with external drift) had the smallest ESE for soil phosphorus and soil organic carbon.

*Table 27: Soil pH Summary Validation statistics*

| **Variable** | | **A** | **B** | **C** | **D** |
|---|---|---|---|---|---|
| Prediction error | Mean | 0.168 | 0.094 | 0.116 | 0.128 |
| | Median | 0.212 | 0.148 | 0.200 | 0.200 |
| MSE | | 0.417 | 0.388 | 0.463 | 0.551 |
| Corr.Model | | Exponential | Exponential | Exponential | Exponential |
| Partial Sil | | 0.154 | 0.145 | 0.218 | 0.299 |
| Range | | 48.380 | 44.670 | 40.200 | 36.250 |
| Nugget | | 0.240 | 0.240 | 0.237 | 0.238 |
| ESE | | 0.422 | 0.393 | 0.468 | 0.553 |

*Note: (A) REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift) (B)REML-EBLUP with the only fixed effect a constant mean (ordinary kriging) (C) random forest with all predictors (D) random forest with elevation and rainfall as predictors selected through alpha-investment.*

*Table 28: Soil phosphorus Summary Validation statistics*

| **Variable** | | **A** | **B** | **C** | **D** | **E** | **F** |
|---|---|---|---|---|---|---|---|
| Prediction error | Mean | -3.370 | -3.264 | -5.369 | -5.751 | -3.195 | -3.279 |
| | Median | 0.968 | 1.170 | -0.990 | -0.990 | -0.990 | -0.990 |
| MSE | | 238.679 | 244.390 | 286.054 | 285.890 | 366.870 | 369.562 |
| Corr.Model | | Spherical | Spherical | Spherical | Spherical | Spherical | Spherical |
| Partial Sil | | 159.800 | 171.500 | 186.800 | 181.800 | 277.800 | 263.600 |
| range | | 10.820 | 11.590 | 11.580 | 11.440 | 10.110 | 8.679 |
| Nugget | | 92.050 | 91.260 | 92.240 | 92.380 | 110.300 | 116.200 |
| ESE | | 263.207 | 273.413 | 307.866 | 307.254 | 398.308 | 390.551 |

*Note: (A) REML-EBLUP with rainfall as predictors selected through alpha-investment, (B) REML-EBLUP with ordinary kriging (C) random forest with all predictors data not transformed (D) random forest with all predictors data transformed and then back transformed. (E) random forest with rainfall as*

*predictors selected through alpha-investment data not transformed. (F) random forest with rainfall as predictors selected through alpha-investment data not transformed.*

*Table 29: Soil organic carbonic Summary Validation statistics*

| Variable | | A | B | C | D | E |
|---|---|---|---|---|---|---|
| Prediction error | Mean | 0.018 | 0.030 | 0.032 | 0.053 | 0.038 |
| | Median | 0.058 | 0.091 | 0.055 | 0.095 | 0.080 |
| MSE | | 0.138 | 0.139 | 0.168 | 0.157 | 0.156 |
| Corr.Model | | Spherical | Spherical | Spherical | Spherical | Spherical |
| Partial Sil | | 0.066 | 0.067 | 0.092 | 0.082 | 0.082 |
| range | | 38.740 | 55.780 | 60.060 | 41.860 | 41.860 |
| Nugget | | 0.076 | 0.076 | 0.084 | 0.077 | 0.077 |
| ESE | | 0.142 | 0.144 | 0.177 | 0.162 | 0.160 |

*Note: (A) REML-EBLUP with soil class as predictor selected through alpha-investment, (B) REML-EBLUP with ordinary kriging (C) random forest with all predictors data (D) random forest with soil class as predictor (E) random forest with soil class plus REML-EBLUP kriged residuals.*

### 4.3. Land Suitability Evaluation for Rice Production

Figures 16 and 18 show the reclassified maps of suitability levels of each criterion for rainfed paddy rice and rainfed upland rice, respectively. Figures 16 and 18 show the proportions of each suitability classes for each criterion for rainfed paddy rice and rainfed upland rice, respectively. For both rainfed paddy rice, at least 90 percent of the study area has CEC that is not suitable with most of the country having CEC ranging between 5 to 15 cmol/kg which is currently not suitable and part of the western part having CEC less than 5 cmol/kg which is permanently not suitable. Despite having highly suitable temperature and pH, the Eastern and Southern part of the country have slope of greater than 5 percent which is permanently not suitable for paddy rice production. The Southern part of the study area also is affected by low rainfall which is less than 800mm permanently not suitable and the middle part of the country having rainfall between 800 and 1000mm currently not suitable. In areas such as eastern and southern parts with one criterion highly suitable and another not suitable, it becomes difficult to evaluate the suitability levels, hence the introduction of weights of

influence for each criterion (described in section 3.2.7) which were used to produce the final suitability map.

In their study on Soyabean suitability in Kabwe District of Zambia, Munene et al., (2017) also observed some limitations owing to soil pH, low SOC and slope. Chirwa et al., (2016) evaluated the soil fertility status and land suitability for smallholder farmers' groundnut and maize production in Chisamba District of Zambia and concluded that soil pH, low CEC were some of the major soil fertility limiting factors.

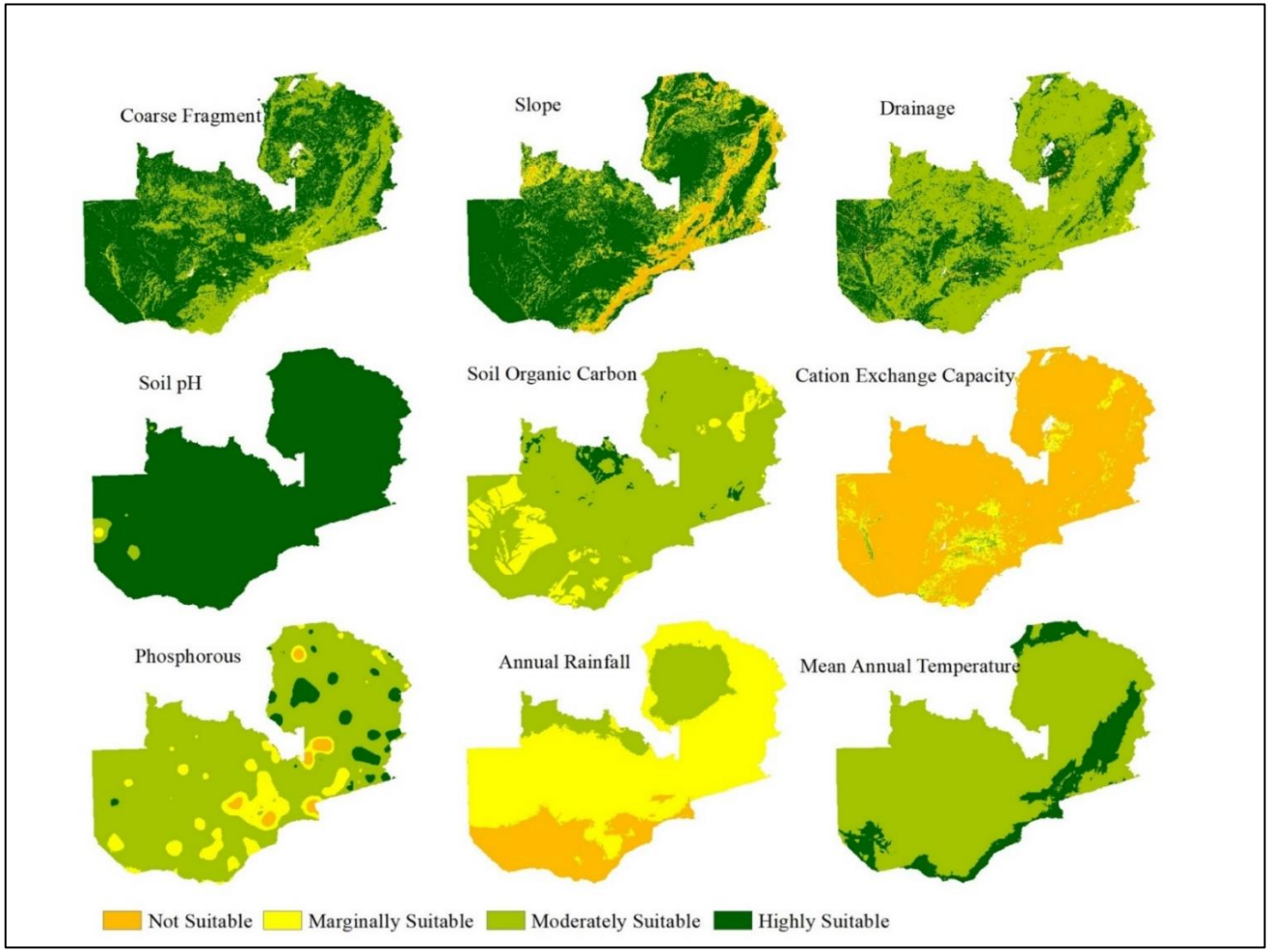



*Figure 16: Rainfed Paddy Rice FAO Suitability ratings for each criterion*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988)*

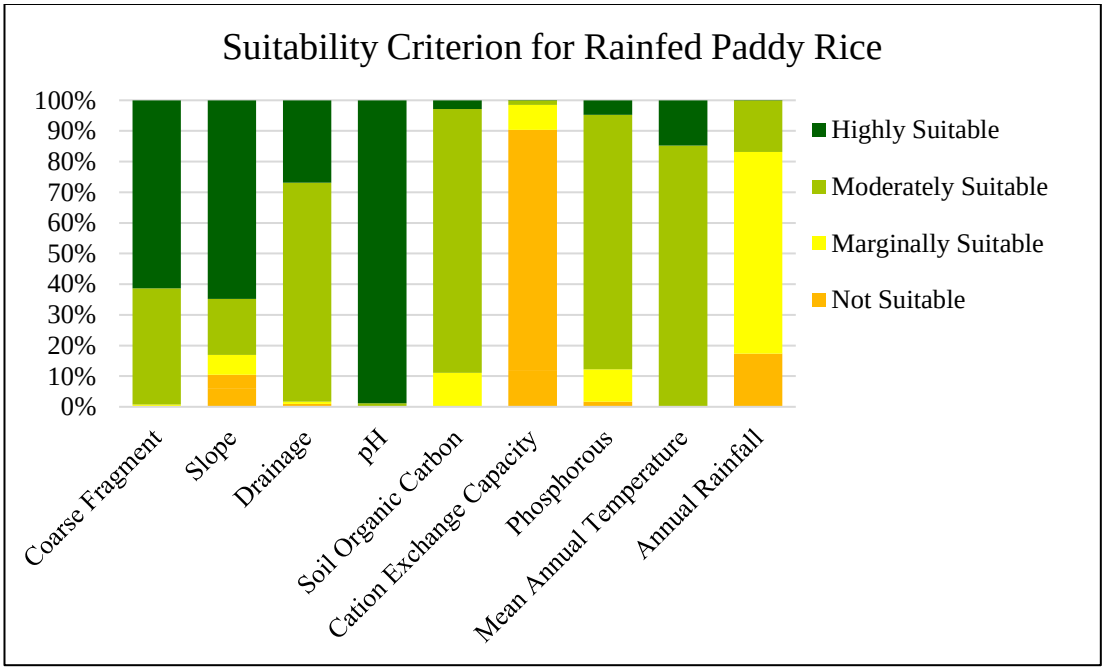



*Figure 17: Proportions of suitability ratings in each suitability criterion for rainfed paddy rice*

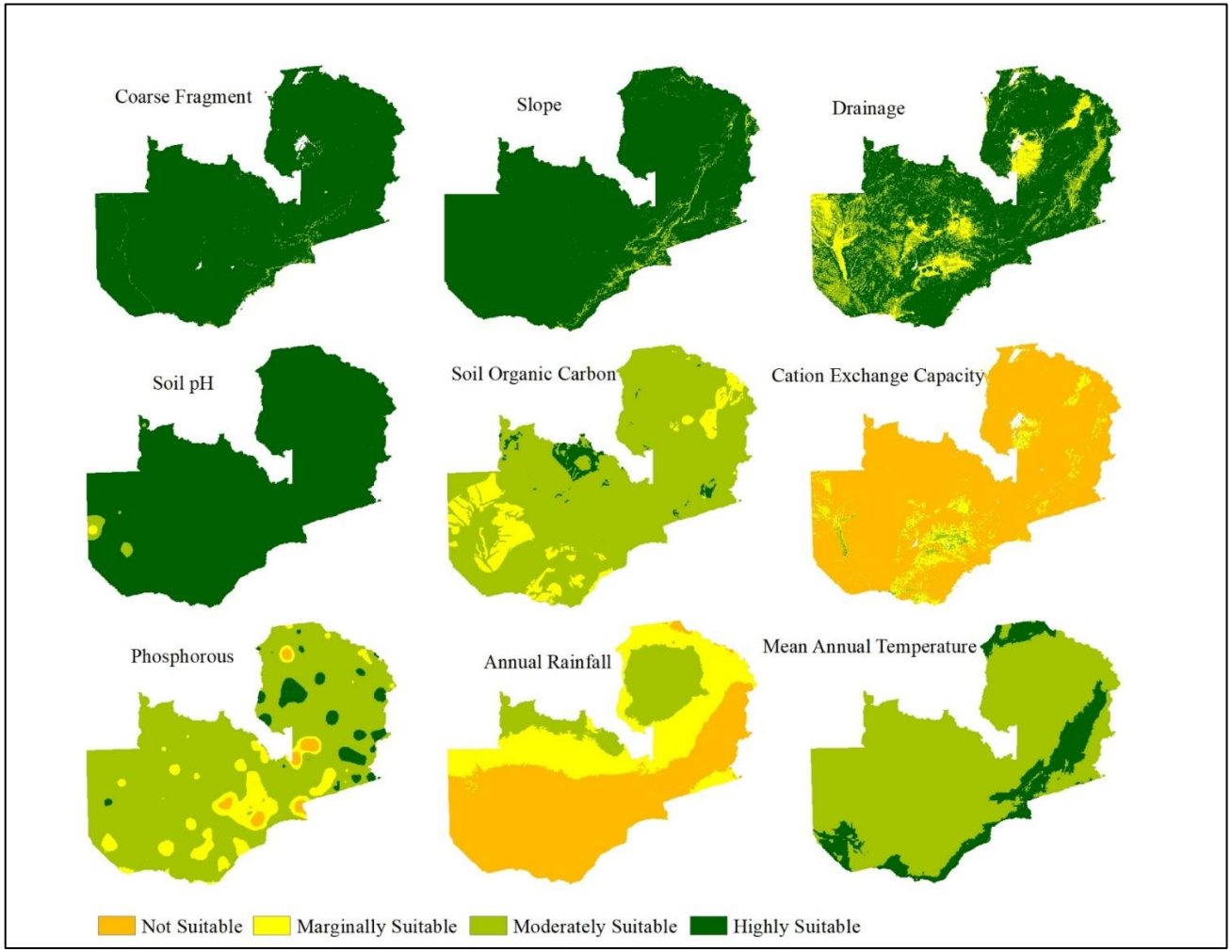



*Figure 18: Rainfed Upland Rice Suitability ratings for each criterion*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*

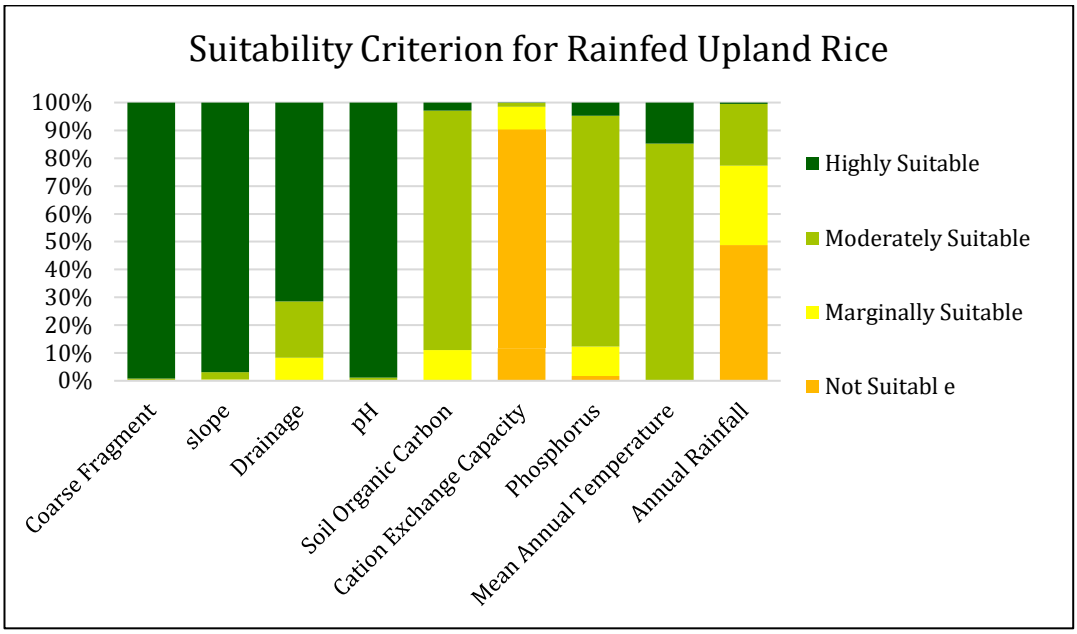



*Figure 19: Proportions of suitability rating in each suitability criterion for rainfed upland rice*

Figures 20, 23, 26 and 29 are the respective suitability maps of Zambia for rainfed paddy, rainfed upland rice, irrigated paddy rice and irrigated upland rice produced using weighted overlay (described in section 3.2.8) of the suitability criterion maps. The resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988). Figures 20, 23, 26 and 29 are the suitability maps overlayed with the protected area, this shows that some of the suitable areas are not available for production as they fall under urban, national parks and forest reserves. Figures 21, 24, 27 and 30 show the area proportions of suitability classes for total area, area under urban, area under national parks, area under forest reserves and potential area for rainfed and irrigated paddy as well as upland rice.

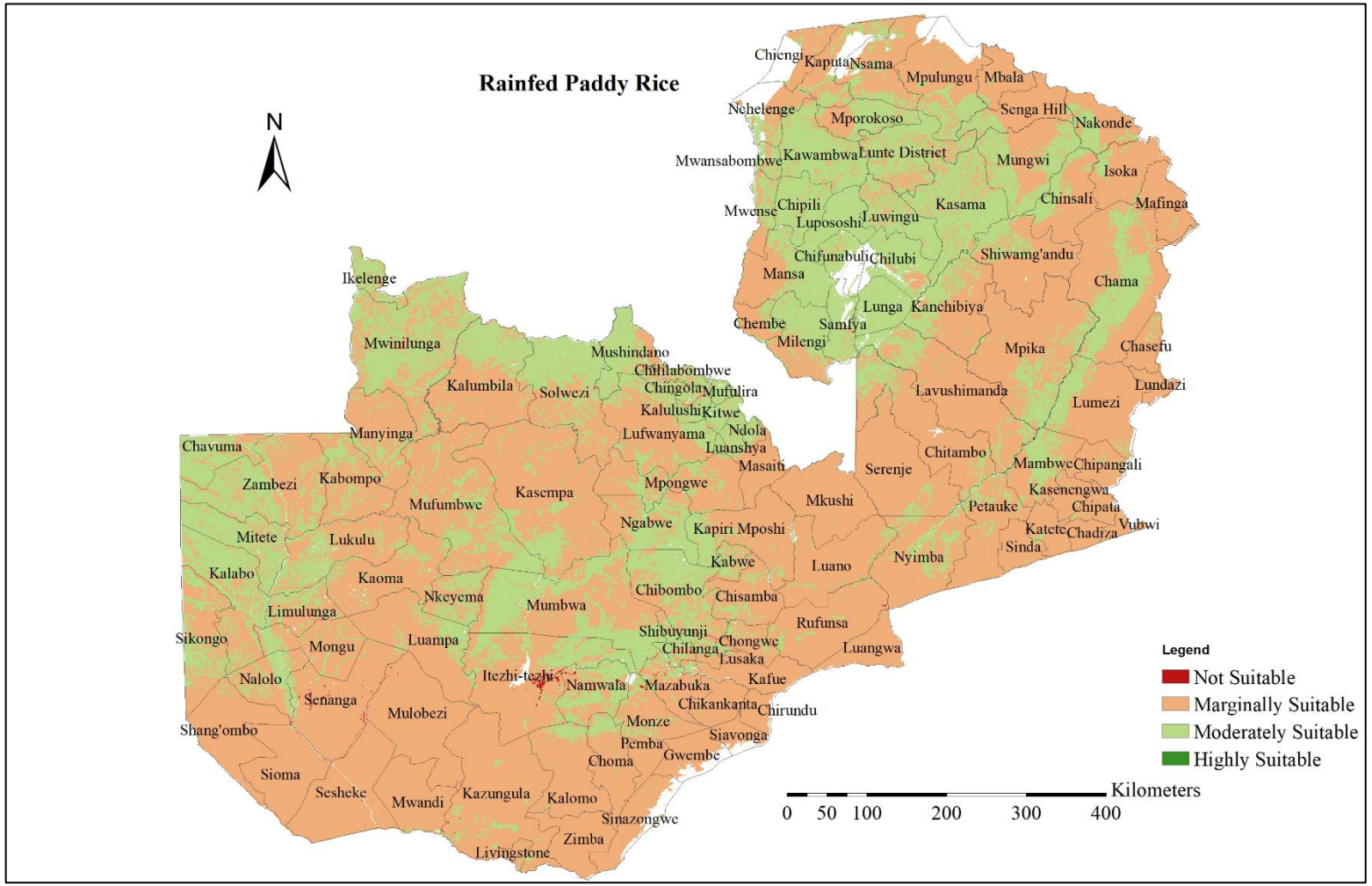



*Figure 20: Suitability map for rainfed paddy rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988)*

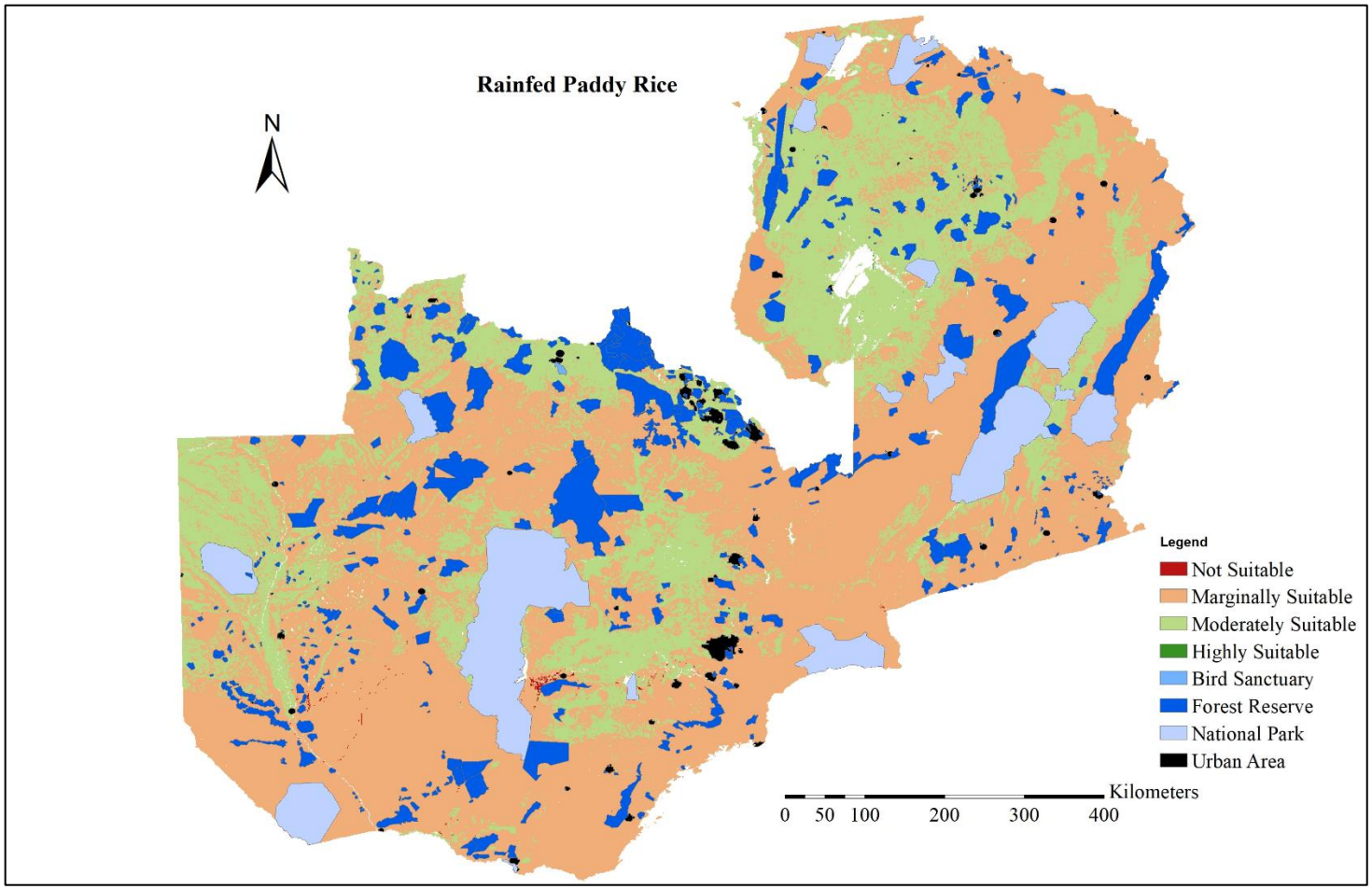
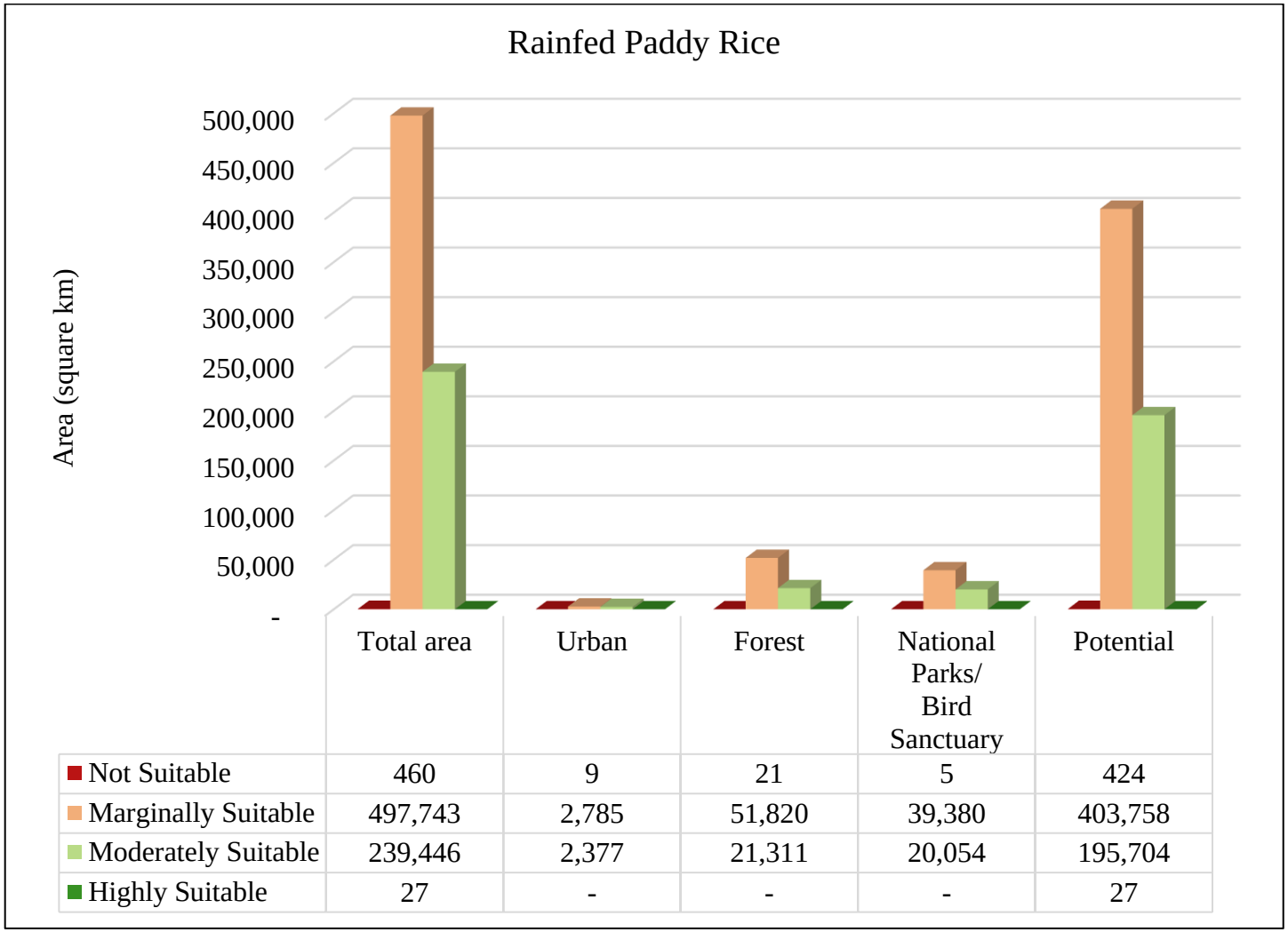



*Figure 21: Protected area over the suitability map of rainfed paddy rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988)*




| | Total area | Urban | Forest | National Parks/ Bird Sanctuary | Potential |
|---|---|---|---|---|---|
| Not Suitable | 460 | 9 | 21 | 5 | 424 |
| Marginally Suitable | 497,743 | 2,785 | 51,820 | 39,380 | 403,758 |
| Moderately Suitable | 239,446 | 2,377 | 21,311 | 20,054 | 195,704 |
| Highly Suitable | 27 | - | - | - | 27 |

*Figure 22: Rainfed paddy rice area Proportion of Suitability classes*

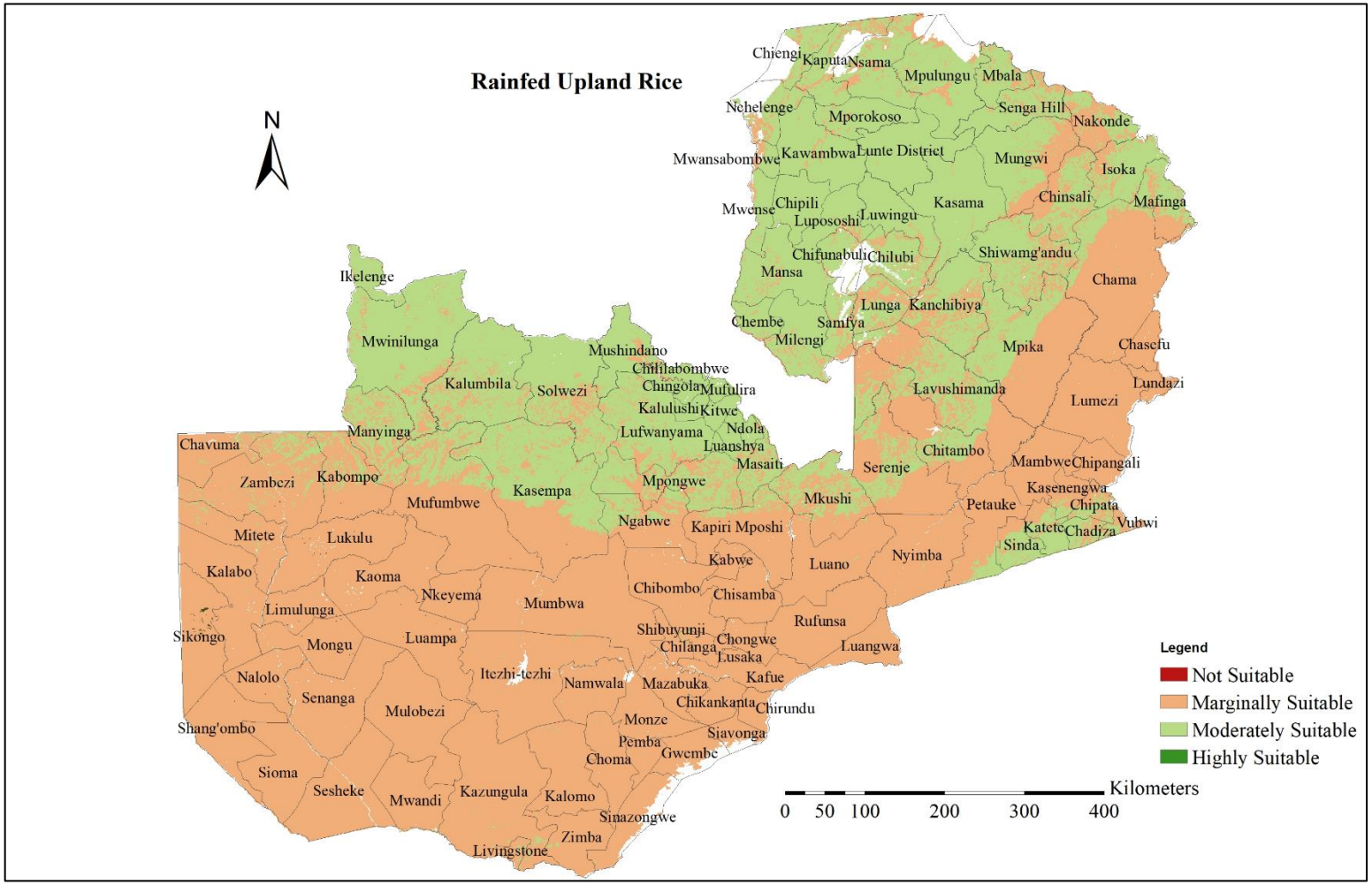



*Figure 23: Suitability map for rainfed upland rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*

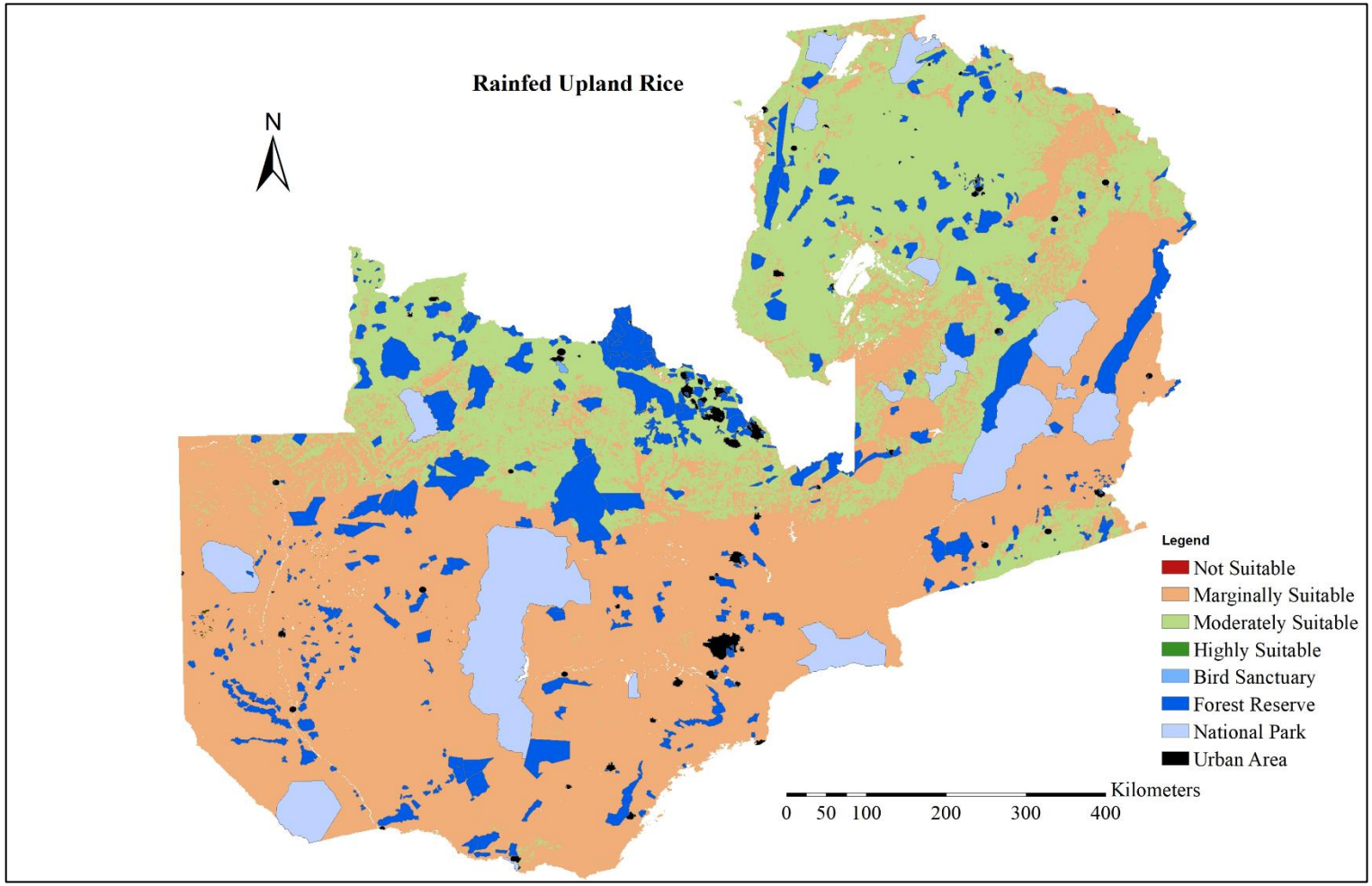
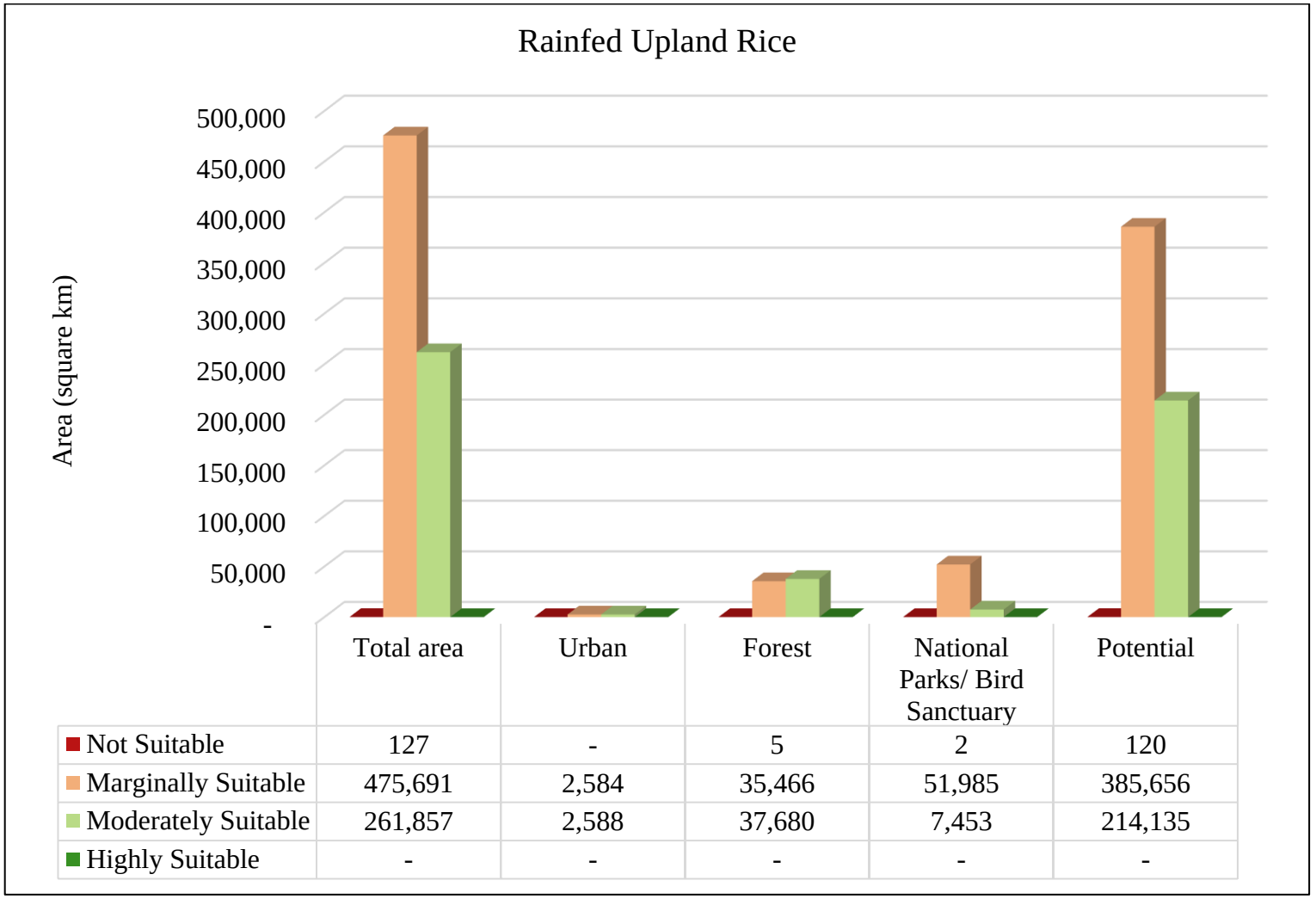



*Figure 24: Protected area over the suitability map of rainfed upland rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*




| | Total area | Urban | Forest | National Parks/ Bird Sanctuary | Potential |
|---|---|---|---|---|---|
| Not Suitable | 127 | - | 5 | 2 | 120 |
| Marginally Suitable | 475,691 | 2,584 | 35,466 | 51,985 | 385,656 |
| Moderately Suitable | 261,857 | 2,588 | 37,680 | 7,453 | 214,135 |
| Highly Suitable | - | - | - | - | - |

*Figure 25: Rainfed upland rice Area Proportion of Suitability classes*

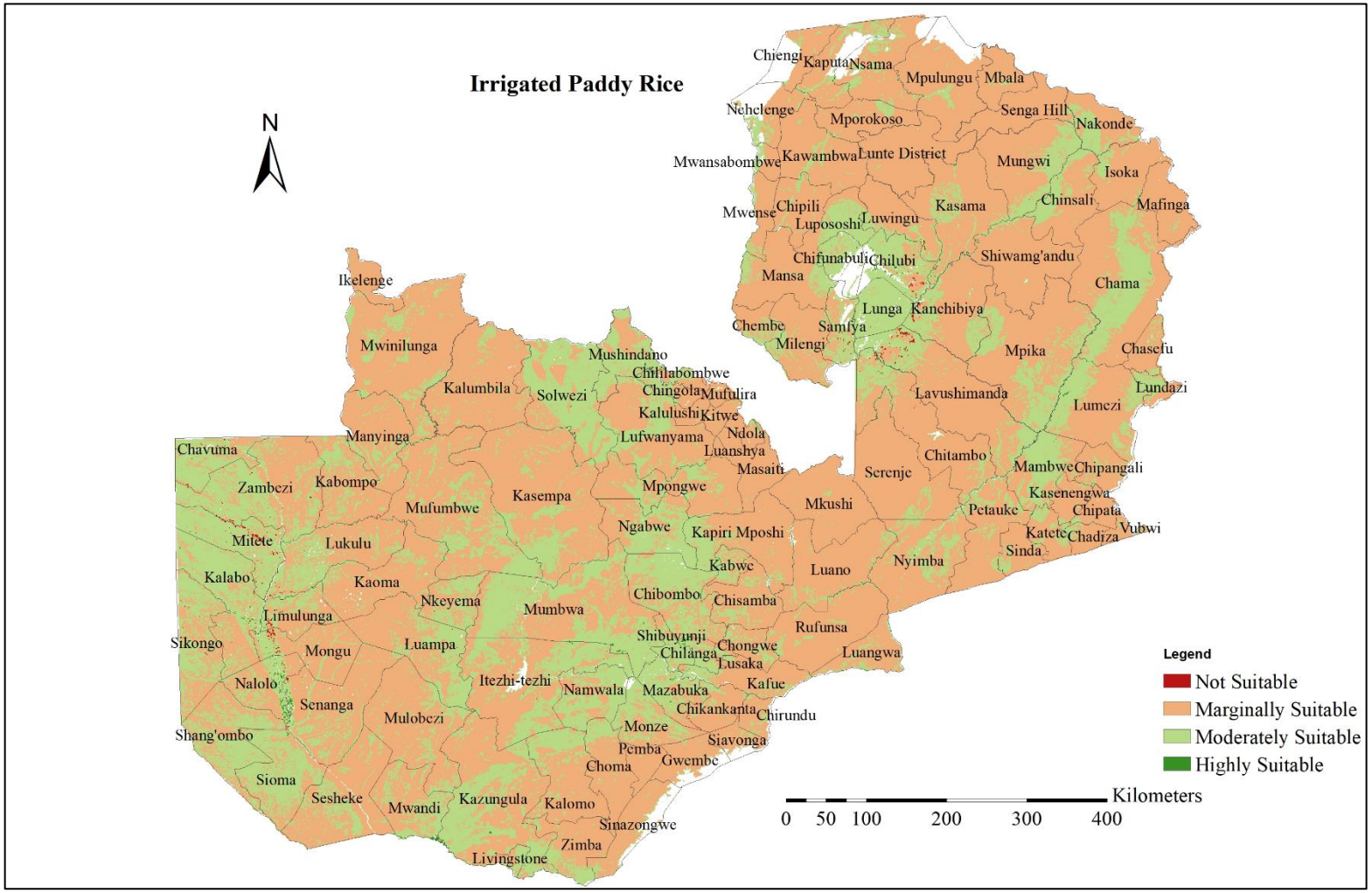



*Figure 26: Suitability map for irrigated paddy rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*

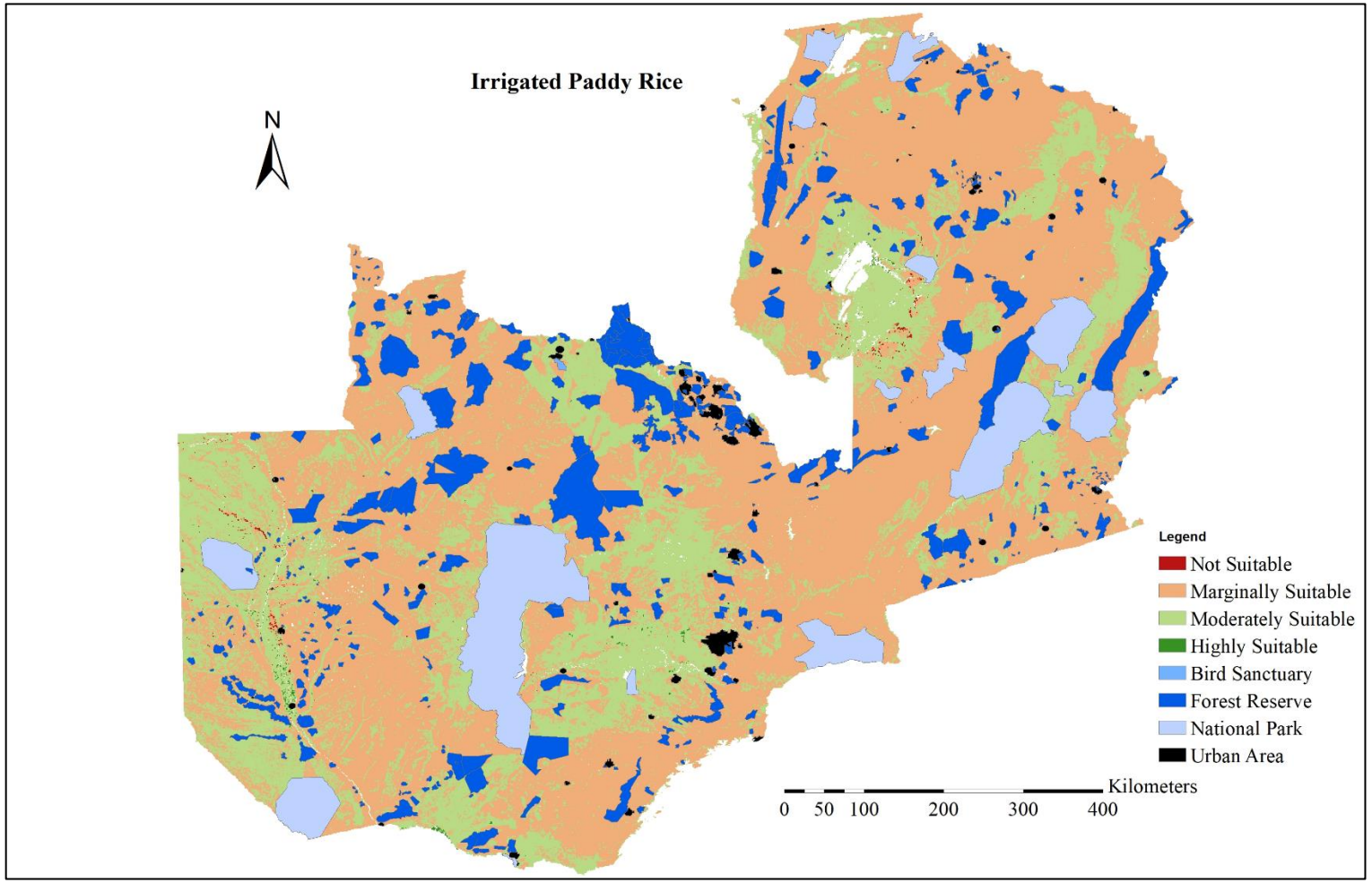
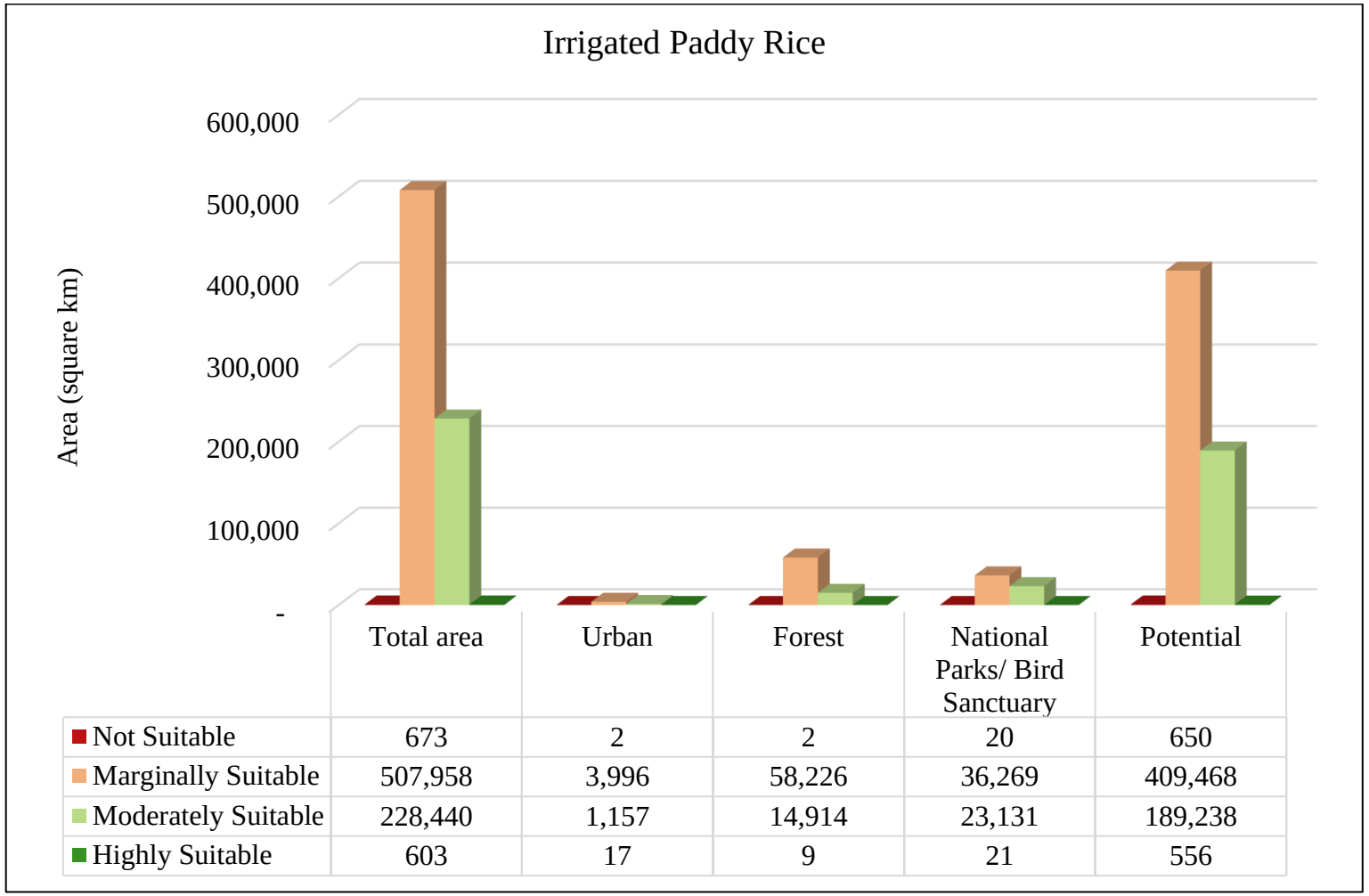



*Figure 27: Protected area over the suitability map of irrigated paddy rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*




| | Total area | Urban | Forest | National Parks/ Bird Sanctuary | Potential |
|---|---|---|---|---|---|
| Not Suitable | 673 | 2 | 2 | 20 | 650 |
| Marginally Suitable | 507,958 | 3,996 | 58,226 | 36,269 | 409,468 |
| Moderately Suitable | 228,440 | 1,157 | 14,914 | 23,131 | 189,238 |
| Highly Suitable | 603 | 17 | 9 | 21 | 556 |

*Figure 28: Irrigated paddy rice Area Proportion of Suitability classes*

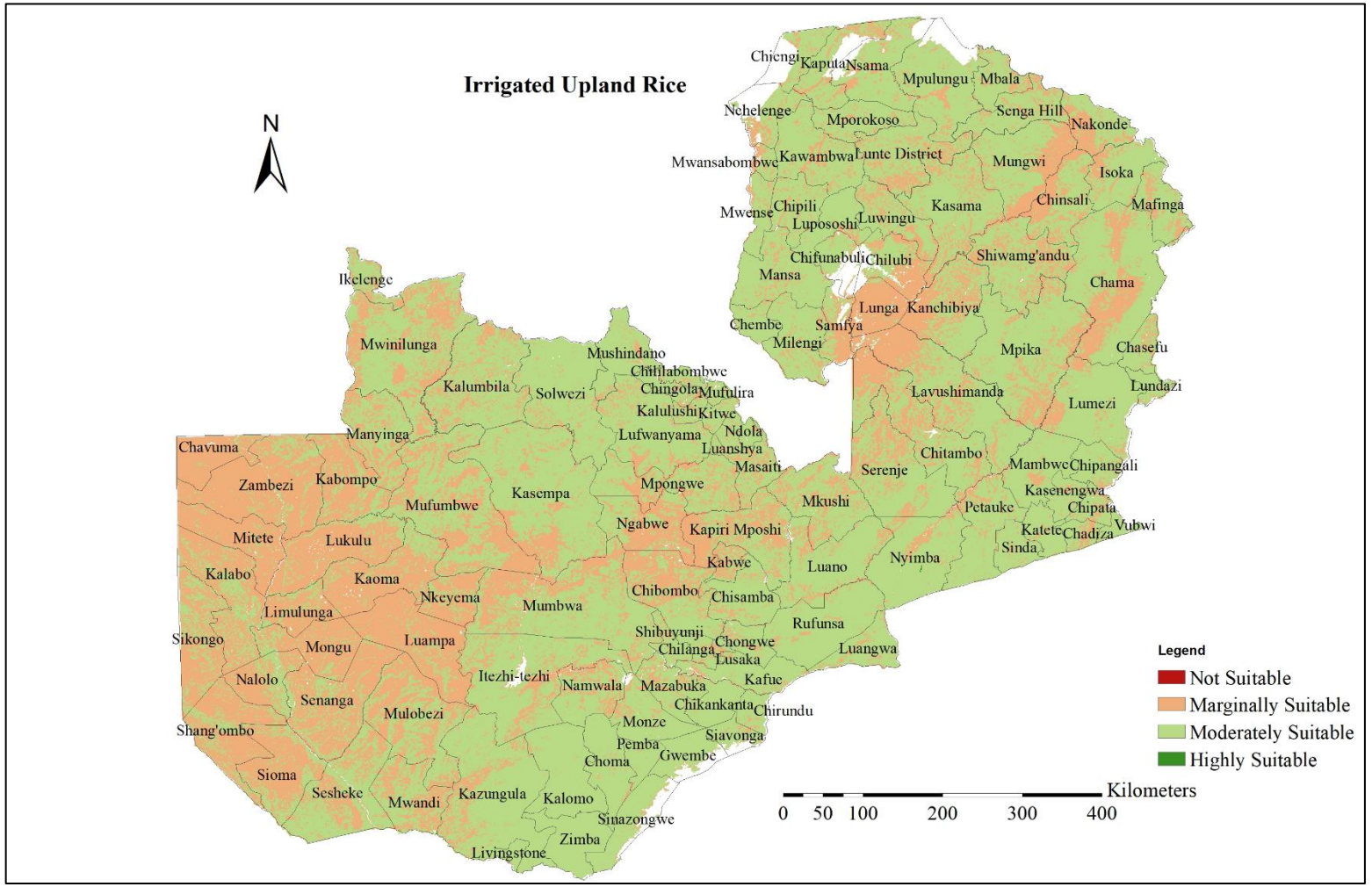



*Figure 29: Suitability map for irrigated upland rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*

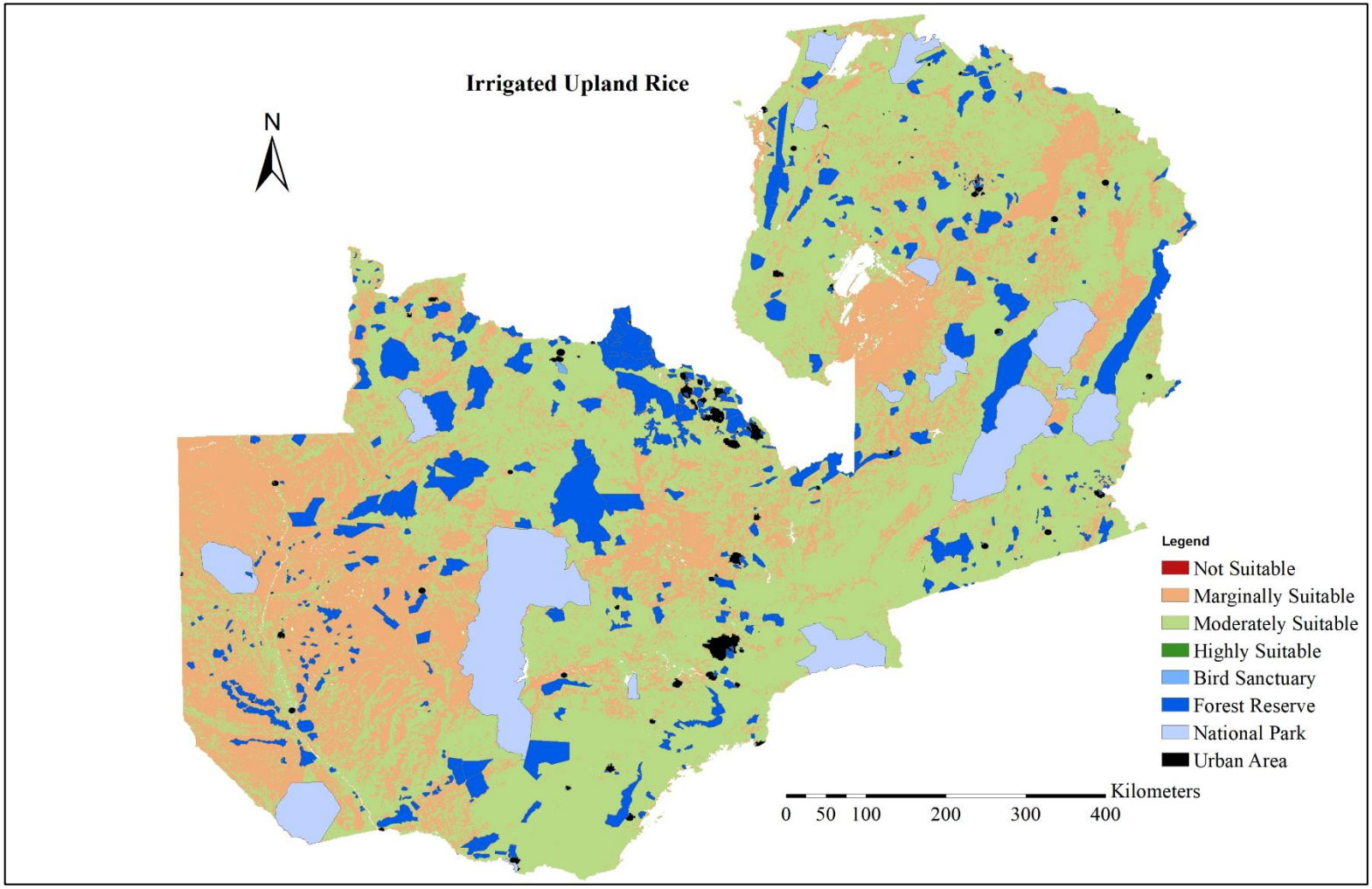
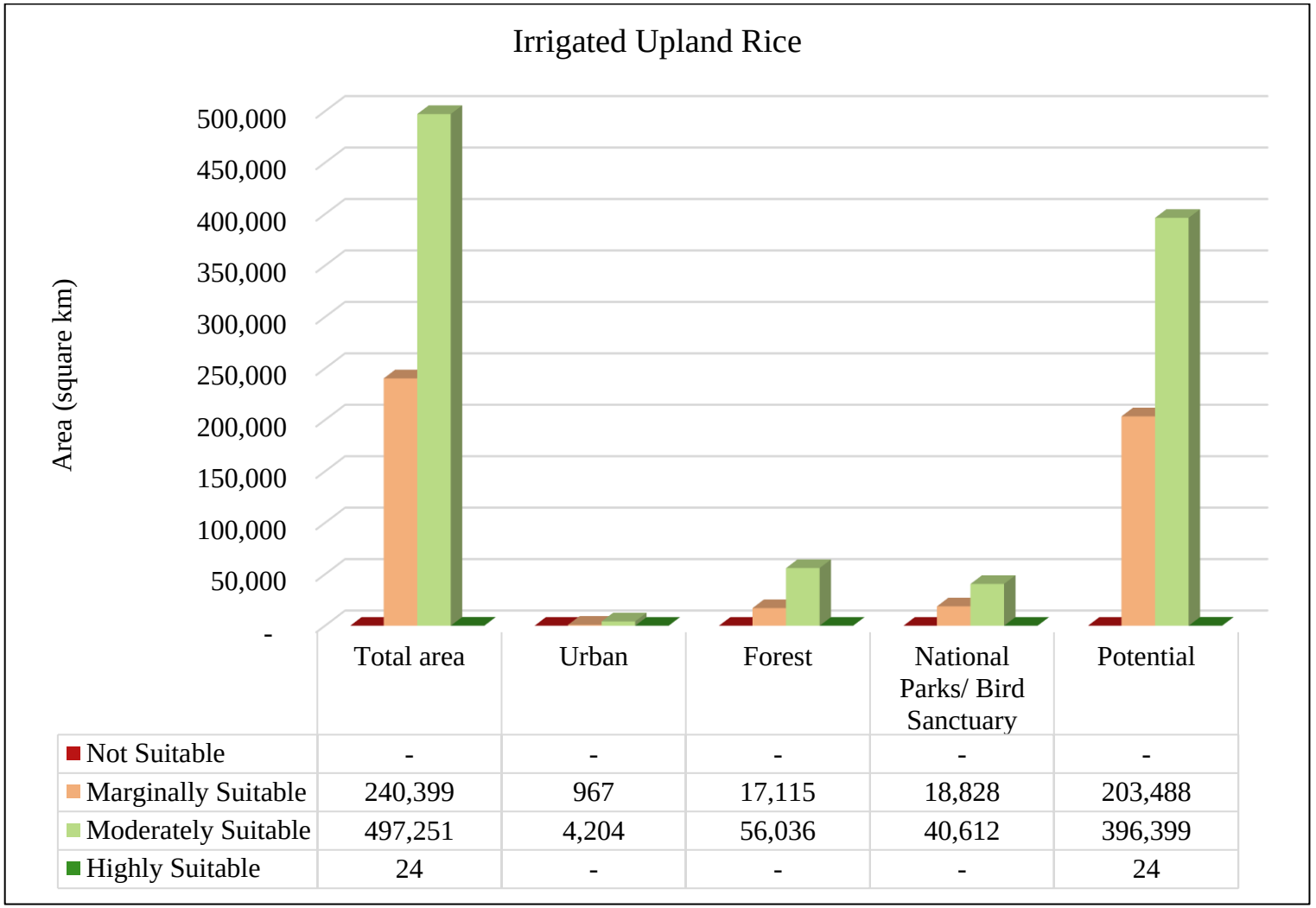



*Figure 30: Protected area over the suitability map of irrigated upland rice in Zambia*
*Note: resolution of the raster maps is 1km which translates to a scale of 1:2,000,000 according to Tobler (1988).*




| | Total area | Urban | Forest | National Parks/ Bird Sanctuary | Potential |
|---|---|---|---|---|---|
| Not Suitable | - | - | - | - | - |
| Marginally Suitable | 240,399 | 967 | 17,115 | 18,828 | 203,488 |
| Moderately Suitable | 497,251 | 4,204 | 56,036 | 40,612 | 396,399 |
| Highly Suitable | 24 | - | - | - | 24 |

*Figure 31: Irrigated upland rice Area Proportion of Suitability classes*

### 4.4. Statistical Evaluation of the Rainfed Paddy Rice Suitability Map

Tables 32, 33 and 34 show the observed and expected counts for each cell of the crop presence and suitability class contingency tables for households in category A, B and C, respectively. Also shown are the $\chi^2$ statistic and associated p-value under the null hypothesis of random association. For households in category A and B the value of the statistic, is large, and the probability of a value this large or larger under the null hypothesis is small (p=0.0002 for category A and p=0.0046 for category B). This is evidence to reject the null hypothesis. This is not the case for Category C (p=0.025), so the null hypothesis is retained in this case. For both category A and B, it is observed that there are fewer households with rice present observed than expected under the null hypothesis in the marginally suitable class, and more in the moderately suitable class. This is consistent with the suitability classes' being informative about land suitability for rice production.

*Table 30: Table of observed, expected, deviation, chi-square and p-values from independence for rice farmers in category A and suitability classes.*

| | | | Suitability class | | | |
|---|---|---|---|---|---|---|
| | | | Moderately suitable | Marginally Suitable | Not Suitable | |
| crop presence | Rice | Observed | 89.00 | 103.00 | 3.00 | **195** |
| | | Expected | 51.45 | 143.36 | 0.19 | |
| | | O - E | 37.55 | -40.36 | 2.81 | |
| | No rice | Observed | 734.00 | 2190.00 | 0.00 | **2924** |
| | | Expected | 771.55 | 2149.64 | 2.81 | |
| | | O - E | -37.55 | 40.36 | -2.81 | |
| | | | **823** | **2293** | **3** | **3119** |
| | | | **$\chi^2$ = 86.329** | | **p-value = 0.0002** | |

*Table 31: Table of observed, expected, deviation, chi-square and p values from independence for rice farmers in category B and suitability classes.*

| | | | Suitability class | | | | |
|---|---|---|---|---|---|---|---|
| | | | Highly Suitable | Moderately suitable | Marginally suitable | Not Suitable | |
| crop presence | Rice | Observed | 0.00 | 47.00 | 163.00 | 0.00 | **210** |
| | | Expected | 0.08 | 45.15 | 166.61 | 0.16 | |
| | | O - E | -0.08 | 1.85 | -3.61 | -0.16 | |
| | No rice | Observed | 1.00 | 502.00 | 1863.00 | 0.00 | **2366** |
| | | Expected | 0.92 | 503.85 | 1859.39 | 1.84 | |
| | | O - E | 0.08 | -1.85 | 3.61 | -1.84 | |
| | | | **1** | **549** | **2026** | **0** | **2576** |
| | | | | $\chi^2$ **= 22.578** | | **p-value = 0.004599** | |

*Table 32: Table of observed, expected, deviation, chi-square and p values from independence for rice farmers in category C and suitability classes.*

| | | | Suitability class | | | |
|---|---|---|---|---|---|---|
| | | | Moderately suitable | Marginally Suitable | Not Suitable | |
| crop presence | Rice | Observed | 31 | 99 | 1 | **131** |
| | | Expected | 26.9769 | 103.951 | 0.07194 | |
| | | O - E | 4.02306 | -4.9511 | 0.92806 | |
| | No rice | Observed | 344 | 1346 | 0 | **1690** |
| | | Expected | 348.023 | 1341.05 | 0.92806 | |
| | | O - E | -4.0231 | 4.95113 | -0.9281 | |
| | | | **375** | **1445** | **1** | **1821** |
| | | $\chi^2$ **= 13.801** | | **p-value = 0.025** | | |

# CHAPTER FIVE

# 5.0. DISCUSSION

## 5.1. Spatial Interpolation of Exchangeable Acidity (pH), Soil Phosphorus (P) and Soil Organic Carbon (SOC)

The mapped soil pH by all approaches are shown in Figure 12. The optimum pH ($CaCl_2$) for plant growth is between 5.2 and 7.5. Below the pH of 5.2, the levels of Aluminum, Manganese and Copper are toxic for plant growth as Phosphorous and Magnesium are not available for plant uptake. Above pH of 7.5, the interactions between Calcium, Magnesium and Potassium have a negative impact on root absorption. Copper, Iron, Manganese, Zinc, Boron and Phosphorous are deficient (Lake, 2000). The maps in Figure 12 show pH values less than 5.2 in the western and northern parts of the country meaning we expect these areas to have challenges of Aluminum, Manganese and Copper toxicity as well as Phosphorous and Magnesium deficiencies. In the Southern parts of the country the pH values, according to all the maps in Figure 12, range from 5.2 to 7.5 which are optimal for plant growth. There are few areas in the southern part of Zambia with pH above 7.5. Similar spatial variations were observed by Chapoto et al., (2016). The southern parts where the pH is high is a valley area, the northern parts receive high rainfall therefore prone to leaching of basic ions and the western parts despite receiving the same amount of rainfall as the eastern parts, the area is characterized by Kalahari sand. The results show a similar spatial pattern for soil pH as that presented in the SoilGrids map (www.soilgrids.org) of Hengl et al., (2017). The main difference is that the maps in this study show low pH values in the west of the country, whereas the SoilGrids map shows larger values there. Results are more plausible pedologically given the parent material, and it has been long established that the soils formed over the Kalahari sands of western Zambia are weakly to extremely acidic (Brammer, 1976) because they are sandy. A more thorough assessment of the SoilGrids predictions using the RALS data would be of interest.

For soil pH, predictions by the EBLUP from the LMM with the only fixed effect a constant mean (equivalent to ordinary kriging) were better than other predictions in the sense that the mean and median errors were closest to zero and the mean square error and expected square error were the smallest. This is unexpected given the

evidence provided in the model-fitting stage for a significant relationship between soil pH and the selected covariates. The result for soil phosphorus and soil organic carbon were expected to result in better predictions from the LMM which included covariates as fixed effects. However, one may note for soil pH, soil phosphorus and soil organic carbon that the difference between the correlated random variance in the LMM with selected predictors as fixed effects and the corresponding variance in the LMM with a constant mean the only fixed effect is very small. The fact that a covariate is significantly related to a soil property does not necessarily mean that it will allow improved prediction of that property relative to a model without that covariate. That is because the corresponding fixed effect coefficient must be estimated, and this estimation is a source of error in the prediction. Furthermore, Zimmerman et al., (1999) found that ordinary kriging performed better than universal kriging (UK, equivalent to the EBLUP with some covariates) with a spatially clustered data set, while UK performed better when the data were not clustered. This may be because, in a strongly clustered data set, the effective degrees of freedom with which the fixed effects coefficients are estimated may be relatively small.

The use of random forests to include the environmental covariates in spatial prediction was less successful than the LMM and EBLUP, with larger values of ESE. This could be due to over-fitting. It is notable that the residuals from the fitted RF at the calibration data points showed no spatial dependence for soil pH and soil Phosphorus, while the RF prediction errors at the validation points (Tables 26 and 27) do show spatial dependence. This could arise because the RF algorithm, given its flexibility and ability to fit non-linear relationships, generates a model which closely fits the variations within the training data set, but which is not representative of the relationship between the predictor variables and target variable in the underlying population. This would lead both to marked bias in models of the random variation based on the residuals, as can also occur with ordinary least squares (Lark et al., 2006) and also in poor performance of the RF on a separate validation data set. These data may also provide a problem for the RF methodology because of the strong spatial clustering. If some data from a cluster are used in the development of trees while others are in the OOB subset then the assessment of the model and the value of the predictors may be over-optimistic. A predictor variable over fitted to a clustered data set might well fail to predict effectively at independent validation points. This

emphasizes the importance of a genuinely independent validation of spatial predictions (Brus et al., 2011).

Spatial clustering of the observations may also be a contributing factor to the small p-values attributed to the entirely random, although spatially autocorrelated, null predictors which were evaluated with soil pH. This gives reason for caution when interpreting RF output. It is consistent with the findings of Wadoux et al., (2020) that the RF algorithm may select as predictor variables covariates which are not related to the target properties of interest by any direct or indirect causal relations. A spatially dependent predictor variable of this nature may indeed support spatial prediction of a variable to which it has no underlying relationship, but if this is the case then one might prefer to use a properly-designed set of orthogonal polynomial basis functions for the model rather than arbitrary variables. Furthermore, with strongly spatially clustered data, it is even more likely that an uncausally-related predictor will result in poor predictions at independent validation sites.

Due to presence of legacy data, there is an opportunity for use of this data in digital soil mapping. As legacy data sets may originate in local surveys, or from networks of experimental stations, they may show marked spatial clustering, as do the RALS data because of their two-stage cluster sampling design. It is noted that such clustering may cause difficulties for the RF algorithm but that it is also important to account for it when dividing data into prediction and validation subsets. There is a risk of bias in the validation of a map if validation and training data are drawn from common clusters.

### 5.2. Land Suitability Evaluation for Rice Production

The method used to obtain weights to combine the different factors was based on a pair-wise rating method, tested for consistency. However, it reveals an underlying hierarchical structure of these factors in terms of their implied importance as suitability determinants for rice production. The factors, shown in Tables 5 and 6, can be divided in three categories. First are the climate factors (Rainfall and Temperature); second are the topographic and soil physical factors (coarse fragment, drainage and slope) and finally, soil chemical factors (pH, OC, Phosphorus and CEC). And as shown by the weights in Table 12 and 13 three things can be observed. First, coarse fragment and temperature have the lowest weights of 0.019 and are dominated by soil physical and chemical factors with higher weights. Second, CEC and rainfall have a higher weight

of 0.239 dominating all soil physical factors and chemical factors except for drainage whose weight is 0.235 which is close to that of these two factors. Third, CEC dominates all other soil chemical properties, and all soil chemical properties dominate all other factors except for drainage and rainfall which are equally important as CEC. In rice production, water availability is extremely important and it is determined by rainfall and soil water-holding capacity (Moormann and Van Breemen 1978). It is therefore unsurprising to see factors such as rainfall and those that reflect soil water holding capacity such such as Drainage and organic matter dominate other factors.

The nine maps for separate factors for rainfed paddy rice, showed that much land is not suitable or marginally suitable for paddy rice production as judged from CEC, rainfall and slope. The CEC of soil is restrictive mainly observed in the western part of the country, where soils formed in Kalahari sand cover have limited clay content. Large slopes were mainly observed in the eastern part of the country along the margins of valley areas.

Based on the weighted factors, the overall suitability map for rainfed paddy rice showed about 27 percent of the study area was highly and moderately suitable. For rainfed upland rice, only 29 percent of the study area was highly and moderately suitable. For irrigated paddy rice, only 26 percent of the study area was highly and moderately suitable. And for irrigated upland rice, 54 percent of the study area was highly and moderately suitable.

With only about 27 percent, 29 percent and 26 percent, of land area potentially suitable for rainfed paddy, rainfed upland and irrigated paddy rice production, respectively. Taking into consideration competition with other staple crops, there is limited potential for rainfed paddy, rainfed upland and irrigated paddy rice production in Zambia. Most of the land has limitations which are severe for production of these production methods and will reduce productivity as well as increasing requirements for inputs. However, with 54 percent of land area potentially suitable for irrigated upland rice, this would help expand the production of rice beyond the limited potential suitable area for rainfed paddy, rainfed and upland rice. This result agrees with Mutale et al., (2010) whose study recommended that the New Rice for Africa (NERICA) upland varieties be explored for possible cultivation in upland Zambia. With such limited potential, there is also a need to improve the productivity of rainfed paddy rice among the few farmers growing the crop by investing resources in training them in agricultural practices that

will help reduce the limitations such as increasing organic matter content of their fields by adding manure and practicing conservation agriculture as this will help increase the CEC as well as soil moisture retention of their soils.

The validation of the suitability map using RALS 2012 data showed that, among farmers in category A, B and C, there were fewer rice presence than expected under random association in the marginally suitable class and more in the moderately suitable class, and that the difference between the observed numbers and expected numbers under random association was statistically significant (P= 0.0002,0.0046 and 0.025 respectively) meaning there is a greater chance of rice being grown on moderately suitable land than would be expected by chance alone

This land suitability assessment was carried out using soil pH, soil Phosphorus and soil organic carbon maps produced from legacy soil data collected during the RALS 2012 survey as well as a secondary data analysis using free access secondary data available on soilgrid, worldclim and USGS websites. Because surveys are costly and time consuming, such data can be used to address important problems without incurring the high cost and time consuming of a new survey.

As noted in the methods section, the division of the ranges of values for the SSF were based on a range of studies from across Tropical and Subtropical conditions. They may therefore be used in comparable evaluations in other settings in the Tropics or Subtropics, although they should be updated wherever possible from the results of new studies or systematic reviews. However, the pairwise comparison matrix, A, was based primarily on local expert opinion. The validation results give some evidence that this opinion was soundly-based. However, the elicited pairwise matrix produced here should not be applied outside Zambia without care, and local experts should first be asked to review the comparisons made in Table 7, and to amend these in the light of local experience. The methods section provides sufficient information on how the consistency of an amended matrix can be tested.

As observed above, this land suitability assessment is based on expert judgement, and also on the assumption that overall land suitability can be treated as a weighted linear combination of contributions from multiple factors. The validation described above suggests that this assessment is of value, at least as a provisional guide, but further work is needed to develop such assessments and to refine them. These might use

process models, or surveys of actual paddy rice yields at locations across Zambia or an analysis of proxy variables for crop yield, e.g., from remote sensor data, both to compare these between the suitability classes obtained here, but also to explore other non-linear effects of multiple factors, possibly using a modelling method such as boundary line analysis (Lark et al., 2020). Furthermore, this study has considered biophysical factors which might control land suitability, but farmers' decisions are not based only on biophysical limitations (Rossiter, 1995). A proposal is made, that given the need to make assumptions about the joint effects of multiple factors, there is an argument for not combining biophysical and socio-economic factors into a single multi-criterion index of suitability. Therefore, further research, that economic surveys are undertaken to record local commodity prices (rice and alternatives), input and labor costs, historical experience of rice production, contemporary attitudes to rice as a crop, knowledge of rice production among local extension officers. These could be focused on areas where the analysis presented here suggests that biophysical factors are conducive to the production of paddy rice, but where the RALS data show marked differences in the extent to which local farmers choose to produce the crop. This would enable identification of key socio-economic factors that may limit rice production where the physical environment is suitable for it.

# CHAPTER SIX

## 6.0. CONCLUSION

This study has shown when carrying out digital soil mapping for a large area across the country, one can successfully map the spatial variability of soil properties such as soil pH, soil phosphorus and soil organic carbon using legacy data from already existing surveys without incurring the high cost of soil sampling and analysis. However,

The spatial variability of Soil pH, Soil Phosphorus and Soil organic carbon was successfully mapped and can be utilized by different stakeholders such as farmers, policy makers, researchers as well as farmers for agricultural activities such as land suitability assessment not only for rice production but other crops as well.

It also successfully demonstrates how to use a model-based approach to validate the performance of spatial models from highly clustered spatial distribution legacy data by splitting the data set at cluster level. Then MSE and ESE can be used to validate the performance of the models for spatial prediction of soil properties. The ESE, computed from the model, in all the cases (soil pH, soil phosphorus and soil organic carbon) was not very different from the MSE computed directly although in each case it exceeded the mean square error was expected. This could be because when data was split at cluster level, the clusters are reasonably balanced (similar numbers of observations in each), and are themselves independently and randomly. The model-based method to quantify prediction uncertainty is, nonetheless, a more general approach for use with validation data from locations not selected by probability sampling.

The empirical best linear unbiased predictor (EBLUP) with the only fixed effect a constant mean performed better than the other methods used for predicting soil pH and the EBLUP with fixed effect performed better than other methods used for predicting soil organic carbon and soil Phosphorus. Random forests had the largest model-based estimates of the expected squared errors in all predictions. The soil pH, soil phosphorus and soil organic matter maps produced using the best performing models, were used as factors in land suitability assessment of paddy and upland rice under rainfed and irrigation conditions.

In addition to the soil pH, soil phosphorus and soil organic matter maps. This study integrated free access secondary data available on soilgrid, worldclim and USGS websites in the land suitability assessment. Because surveys are costly and time consuming, this study has shown that such data can be used to address important problems without incurring the high cost and time consuming of a new survey.

The suitability for rainfed paddy, rainfed upland, irrigated paddy and irrigated upland rice production was found and areas that are highly and moderately suitable identified. Potential land classified as highly and moderately suitable was 27 percent for rainfed paddy rice, 29 percent for rainfed upland rice, 25 percent for irrigated paddy rice and 54 percent irrigated upland rice. This result shows limited potential to develop production of rainfed paddy, rainfed upland a well as irrigated paddy rice in Zambia but great potential for irrigated upland rice production. These findings have implications for the Ministry of Agriculture managing the National Rice Development Strategy (NRDS) of Zambia, which should also explore the irrigation of upland rice production in Zambia as this would help expand the potential production area of rice.

# CHAPTER SEVEN

## 7.0. RECOMMENDATIONS

Ministry of Agriculture managing the National Rice Development Strategy (NRDS), should explore the irrigation of upland rice production in Zambia in areas identified to be suitable as this would help expand the potential production area of rice.

When mapping spatial variability of soils from legacy data with highly clustered spatial distribution, it is recommended to split the data into prediction and validation subsets at cluster.

Random forest algorithm appeared susceptible to wholly random "null predictors" which were simulated. This is taken as the first example where such null predictors have been selected alongside pedologically plausible ones. This approach can be used by pedometricians to generate insight from their analyses, as well as predictions.

## REFERENCES


Agbeshie, A.A. and R. Adjei, 2019. Land suitability of the Nkrankwanta Lowland for rice cultivation in the Dormaa West District, Ghana. Advances in Research, pp.1-15. Doi:10.9734/air/2019/v20i430162

Agidew, A.A., 2015. Land Suitability Evaluation for Sorghum and Barley Crops in South Wollo Zone of Ethiopia. J. Econ. Sustain. Dev. 6, 14–26.

Altmann, A., Tolosi, L., Sander, O. & Lengauer, T. 2010. Permutation importance: a corrected feature importance measure, Bioinformatics 26:1340-1347. doi:10.1093/bioinformatics/btq134

Ambarwulan, W., Santoso, P. B., Sabiham, S., Hikmat, M., 2016. Remote sensing and land suitability analysis to establish local specific inputs for paddy fields in Subang, West Java. Procedia Environmental Sciences 33, 94-107. https://doi.org/10.1016/j.proenv.2016.03.061

Arai, Y. and Sparks, D.L., 2007. Phosphate reaction dynamics in soils and soil components: A multiscale approach. Advances in agronomy, 94, pp.135-179.

Archer, K.J. and Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Computational statistics & data analysis, 52(4), pp.2249-2260. DOI:10.1016/j.csda.2007.08.015

Aune, J.B., Sekhar, N.U., Esser, K., Tesfai, M., 2014. Opportunities for Support to System of Rice Intensification in Tanzania, Zambia and Malawi Opportunities for Support to System of Rice.

Ayoade, M.A. Theor Appl Climatol., 2017. Suitability assessment and mapping of Oyo State, Nigeria, for rice cultivation using GIS, 129(3-4), 1341–1354. https://doi.org/10.1007/s00704-016-1852-4

Behrens, T., Scholten, T., 2007. A Comparison of Data-Mining Techniques in Predictive Soil Mapping, in: Lagacherie, P., McBratney, A, B., Voltz, M. (Eds.), Developments in Soil Science. Elsevier B.V., p. 353.

Brady, N.C. and Weil, R.R., 2014. The nature and properties of soil. (14th New Int. Ed.).

Brammer, H. 1976. Soils of Zambia. Department of Agriculture, Land Use Branch. Lusaka.

Breiman, L., 1996. Bagging Predictors. Machine learning, 26(2), pp.123-140.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62(3), pp.394-407.

Burke, W.J., Frossard, E., Kabwe, S. and Jayne, T.S., 2019. Understanding fertilizer adoption and effectiveness on maize in Zambia. Food policy, 86, p.101721. DOI:10.1016/j.foodpol.2019.05.004

Central Statistical Office, Ministry of Agriculture and Livestock, and Indaba Agricultural Policy Research Institute (CSO/MAL/IAPRI)., 2015. Rural Agricultural Livelihoods Survey. Lusaka, Zambia: CSO/MAL/IAPRI.

Central Statistical Office (CSO)., 2012. Rural Agricultural Livelihoods Survey; Instruction Manual for Listing, Sample Selection, and Largest Maize Field Data Collection**.**

Central Statistical Office (CSO)., 2018. Crop Forecast Survey 2017/18 https://www.zamstats.gov.zm

Chai, X., Shen, C., Yuan, X., Huang, Y., 2008. Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. Geoderma, 148(2), 159-166, https://doi.org/10.1016/j.geoderma.2008.09.018

Chapoto, Antony; Chabala, Lydia M.; and Lungu, Olipa N. 2016. "A Long History of Low Productivity in Zambia: Is it Time to Do Away with Blanket Recommendations?" Zambia Social Science Journal: Vol. 6: No. 2, Article 6. Available at: https://scholarship.law.cornell.edu/zssj/vol6/iss2/6

Chinene, V.R.N., 1991. The Zambian land evaluation system (ZLES). Soil use and management, 7(1), pp.21-29.

Chirwa, M., Mrema, J.P., Mtakwa, P.W., Kaaya, A.K. and Lungu, O.I., 2016. Evaluation of Soil Fertility Status and Land Suitability for Smallholder Farmers' Groundnut and Maize Production in Chisamba District, Zambia.

International Journal of Plant and Soil Science, 10(4), 1-18. https://dx.doi.org/10.9734/IJPSS/2016/25161

Chisci, G., 2009. Rice (Oryza sativa L.). In Manual of Methods for Soil and Land Evaluation 197-202. CRC Press.

Chizhuka, F. 2009. A Study of the Rice Value-chain in Zambia. Lusaka, Zambia: CUTS Africa Resource Centre.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Bï¿½hner, J. 2015: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model Dev., 8, 1991-2007, doi:10.5194/gmd-8-1991-2015.

Costantini, E.A.C., 2009. Manual of Methods for Soil and Land Evaluation. Taylor & Francis Group, LLC.

Cummings, M.P., Myers, D.S. and Mangelson, M., 2004. Applying permutation tests to tree-based statistical models: extending the R package rpart. In Tech Rep CS-TR-4581, UMIACS-TR-2004-24, Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland.

De Data, S.K., 1981. Principles and practices of rice production (No. 633.18 D2622p Ej. 2 019781). John Wiley & Sons.

Dengiz, O., Özyazici, M.A., Sağlam, M., 2015. Multi-criteria assessment and geostatistical approach for determination of rice growing suitability sites in Gokirmak catchment. Paddy Water Environ, 13: 1. https://doi-org.ezproxy.nottingham.ac.uk/10.1007/s10333-013-0400-4

Diggle, P.J. and Ribeiro Jr., P.J., 2007. Model-based Geostatistics. Springer, New York.

ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017). http://maps.elie.ucl.ac.be/CCI/viewer/download.php. Accessed on 20th July 2019.

ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

FA0., 1976. A Framework for Land Evaluation. FA0 Soil Bulletin No. 32, FAO, Rome.

FAO., 2006. Guidelines for soil description. http://www.fao.org/3/a-a0541e.pdf

Fick, S.E., R.J. Hijmans., 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areaIntl. J. Climatology 37 (12): 4302-4315. https://www.worldclim.org

Fischer, G., van Velthuizen, H., Shah, M., Nachtergaele, F., 2002. Global Agro-ecological Assessment for Agriculture in the 21st Century: Methodology and Results. IIASA, Laxenburg and FAO, Rome, Itlay.

Fissore, C., Dalzell, B.J., Berhe, A.A., Voegtle, M., Evans, M. and Wu, A., 2017. Influence of topography on soil organic carbon dynamics in a Southern California grassland. *Catena*, *149*, pp.140-149.

Foster, D.P. and Stine, R.A., 2008. α-investing: a procedure for sequential control of expected false discoveries. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(2), pp.429-444.

Gashu, D., Lark, R.M., Milne, A.E., Amede, T., Bailey, E.H., Chagumaira, C., Dunham, S.J., Gameda, S., Kumssa, D.B., Mossa, A.W., Walsh, M.G., 2020. Spatial prediction of the concentration of selenium (Se) in grain across part of Amhara Region, Ethiopia. Sci. Total Environ. p.139231.

Golden, B. L.,Wang, Q., 1990. An Alternative Measure of Consistency. In: B. L. Golden, A. Wasil & P.T. Harker (eds.) Analytic Hierarchy Process: Applications and Studies, 68-81, New-York: Springer Verlag.

Government of the Republic of Zambia (GRZ), 1991. Exploratory Soil Map of Zambia (1: 1.000, 000).

Government of the Republic of Zambia (GRZ) and UNDP., 2009. Adaptation to the effects of drought and climate change in Agro-ecological Regions I and II in Zambia. Project document. Ministry of Agriculture and Cooperatives. Accepted by: Ministry of finance and National Planning, and UNDP.

Hengl, T., 2003. Pedometric mapping: bridging the gaps between the conventional and pedometric approaches. Wageningen University.

Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B. and Guevara, M.A., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS one, 12(2). https://doi.org/10.1371/journal.pone.0169748

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Tamene, L., Tondoh, J.E., 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS One 10, 1–26. DOI:10.1371/journal.pone.0125814

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B. and Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, p.e5518. DOI:10.7717/peerj.5518

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

ISRIC –World Soil Information., 2017. SoilGrids250m: Global gridded soil information. Available https://soilgrids.org

Jahn, R., Blume, H.P., Asio, V.B., Spaargaren, O., Schad, P., 2006. Guidelines for soil description. FAO.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, pp. 3-7). New York: springer. DOI:10.1007/978-1-4614-7138-7.

Jones, P.G. and Garrity, D.P., 1986. Toward a classification system for upland rice growing environments. In Progress in upland rice research. Proceedings of the 1985 Jakarta Conference. Los Baños (Philippines): International Rice Research Institute. p (pp. 107-116).

Jones Jr, J.B., 2012. Plant nutrition and soil fertility manual. CRC press.

Jones, D.L., Nguyen, C. and Finlay, R.D., 2009. Carbon flow in the rhizosphere: carbon trading at the soil–root interface. Plant and soil, 321(1-2), 5-33. https://doi-org.ezproxy.nottingham.ac.uk/10.1007/s11104-009-9925-0

Kampen, B., Heuvelink, B, G., Brus, J, D., Stoorvogel, J, J., 2010. Pedometric mapping of soil organic matter using a soil. Eur. J. Soil Sci. 61, 333–347. DOI:10.1111/j.1365-2389.2010.01232.x

Kitanidis, P.K. 1987. Parametric estimation of covariances of regionalized variables. Water Resources Bulletin, 23, 557–567.

Kienast-Brown, Suzann Libohova, Z., Janis, B., 2010. Digital Soil Mapping, in: Soil Survey Manual. USDA-NRCS, pp. 429–436. DOI:10.1007/978-90-481-8863-5

Kleinman, P.J., Srinivasan, M.S., Dell, C.J., Schmidt, J.P., Sharpley, A.N. and Bryant, R.B., 2006. Role of rainfall intensity and hydrology in nutrient transport via surface runoff. Journal of environmental quality, 35(4), pp.1248-1259.

Lake, B., 2000. Understanding soil pH. Acid Soil Action. Leaflet, (2).

Lark, R.M., 2017. Controlling the marginal false discovery rate in inferences from a soil dataset with α-investment. European Journal of Soil Science, 68(2), pp.221-234.

Lark, R. M., Ander, E. L., & Broadley, M. R. (2019). Combining two national-scale datasets to map soil properties, the case of available magnesium in England and Wales. European journal of soil science, 70(2), 361-377.

Lark, R.M. and Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. European Journal of Soil Science, 55(4), pp.799-813. DOI: 10.1111/j.1365-2389.2004.00637.x

Lark, R.M., Cullis, B.R. and Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science, 57(6): 787-799. DOI:10.1111/j.1365-2389.2005.00768.x

Lark, R.M., Gillingham, V., Langton, D., Marchant, B.P. 2020. Boundary line models for soil nutrient concentrations and wheat yield in national-scale datasets. European Journal of Soil Science (in press). https://doi.org/10.1111/ejss.12891

Lark, R.M. and Webster, R., 2006. Geostatistical mapping of geomorphic variables in the presence of trend. Earth Surface Processes and Landforms: The Journal of

the British Geomorphological Research Group, 31(7), pp.862-874. DOI: 10.1002/esp.1296

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

Li, J., Heap, A.D., Potter, A., Huang, Z. and Daniell, J.J., 2011. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. Continental Shelf Research, 31(13), pp.1365-1376.

Lindsay, W.L. and Moreno, E.C., 1960. Phosphate phase equilibria in soils. Soil Science Society of America Journal, 24(3), pp.177-182.

Makungwe, M., Chabala, L.M., Chishala, B.H. and Lark, R.M., 2021. Performance of linear mixed models and random forests for spatial prediction of soil pH. Geoderma, 397, p.115079.

Massawe, B. H., Kaaya, A. K., Slater, B., 2019. Involving small holder farmers in the agricultural land use planning process using Analytic Hierarchy Process in rice farming systems of Kilombero Valley, Tanzania.

Masoud, J., Agyare, W. A., Forkuor, G., Namara, R., Ofori, E., 2013. Modeling inland valley suitability for rice cultivation. ARPN Journal of Engineering and Applied Sciences 8(1), 9-19.

Malczewski, J., 1999. GIS and multicriteria decision analysis. John Wiley & Sons. Ministry of Agriculture and Cooperatives. 2011. National Rice Development Strategy (2011 - 2015). Ministry of Agriculture and Cooperatives.

McCauley, A., Jones, C. and Jacobsen, J., 2009. Soil pH and organic matter. Nutrient management module, 8(2), pp.1-12.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey q. Geoderma 97, 293–327.

Minasny, B. and McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma, 140(4), pp.324-336. DOI:10.1016/j.geoderma.2007.04.028

Ministry of Agriculture., 2016. Second National Rice Development Strategy (2016 - 2020). Ministry of Agriculture.

Mohammed, H. G., 2011. Land Suitability Evaluation for Some Selected Land Use Types in the Institute for Agricultural Research Farm, Zaria, Nigeria. (PhD thesis). Ahmmadu Bello University.

Mokarram, M., Aminzadeh, F., 1996. GIS-Based Multicriteria Land Suitability Evaluation Using Ordered Weighted Averaging with Fuzzy Quantifier: A Case Study in Shavur Plain, Iran. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,38, 508–512.

Moonjun, R., Farshad, A., Shrestha, D, P., Vaiphasa, C., 2010. Artificial Neural Network and Decision Tree in Predictive Soil Mapping of Hoi Num Rin Sub-Watershed, Thailand, in: Boettinger, J, L., Kienast-Brown, S., Howell, D, W., Moore, A, C., Hartemink, A, E. (Eds.), Digital Soil Mapping (Bridging Research, Environmental Application, and Operation). Springer Science+Business Media B.V.

Moormann, F. R., Van Breemen, N., 1978. Rice: soil, water, land. Int. Rice Res. Inst.

Moreno, J., Sánchez, F., (2007). Applicability of Knowledge Based and Fuzzy Theory Oriented Aproaches to Land Suitability for Upland Rice and Rubber, as Compared to the Farmers' Perception: A Case Study of Lao PDR. ITC.

Munene, P., Chabala, L. and Mweetwa, M., 2017. Land Suitability Assessment for Soybean (Glycine max (L.) Merr.) Production in Kabwe District, Central Zambia. Journal of Agricultural Science, *9*(3), 1-16. http://dx.doi.org/10.5539/jas.v9n3p74

Mutale, C., Lungu, D.M., Muuka, F.P., Books, R.U.F.O.R.U.M., OER, R., SCARDA, R. and Tenders, R.U.F.O.R.U.M., 2010. Adaptability of rice cultivars to different ecologies in western province of Zambia. In Second RUFORUM Biennial Regional Conference on" Building capacity for food security in Africa", Entebbe, Uganda, 20-24 September 2010, 421-424. RUFORUM.

Mwila, G, P., Ng'uni, D., Phiri, A., 2008. Country Report on the State of Plant Genetic Resources for Food and Agriculture.

Ojara, M.A., Olivier A., Aribo L., Bob A. O., Wasswa P., 2017. Predicting Suitability of Upland Rice for Adoption as Food Security and Poverty Alleviation Crop in Uganda. Jornal of Geography and Earth Sciences, 5(1), 26-40. Doi:10.15640/jges.v5n1a2

Olaleye, A.O., Ogunkunle, A.O., Sahrawat, K.L., Osiname, O.A. and Ayanlaja, S.A., 2002. Suitability evaluation of selected wetland soils in Nigeria for rainfed rice cultivation. Tropicultura, 20(3), pp.97-103.

Omutu, T. C. 2020. soilassessment: Assessment Models for Agriculture Soil Conditions and Crop Suitability. https://CRAN.R-project.org/package=soilassessment

Patterson, H. D. and Thompson, R. 1971. Recovery of inter block information when block sizes are unequal. Biometrika 58, 545-554.

Rawlins, B.G., Lark, R.M., O'donnell, K.E., Tye, A.M. and Lister, T.R., 2005. The assessment of point and diffuse metal pollution of soils from an urban geochemical survey of Sheffield, England. Soil use and management, 21(4), pp.353-362. DOI: 10.1079/SUM2005335

R Core Team., 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, v3.6.2. Vienna, Austria. URL http://www.R-project.org/

Ribeiro, P.J., Diggle, P.J., 2001. geoR: A package for geostatistical analysis. R-News 1, 15–18

Ribeiro Jr, P.J., Diggle, P.J., Ribeiro Jr, M.P.J. and Suggests, M.A.S.S., 2007. The geoR package. *R news*, *1*(2), pp.14-18.Saaty, T. L., 1980. The analytic hierarchy process McGraw-Hill. New York, 324.

Rossiter, D.G., 1995. Economic land evaluation: why and how. Soil Use Manag. 11 (3), 132–140. https://doi.org/10.1111/j.1475-2743.1995.tb00511.x

Saaty T.L., 1988. What is the Analytic Hierarchy Process. In: Mitra G., Greenberg H.J., Lootsma F.A., Rijkaert M.J., Zimmermann H.J. (eds) Mathematical Models for Decision Support. NATO ASI Series (Series F: Computer and Systems Sciences), vol 48. Springer, Berlin, Heidelberg. https://doi-org.ezproxy.nottingham.ac.uk/10.1007/978-3-642-83555-1_5

Sekulić, A., Kilibarda, M., Heuvelink, G., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote Sens. 12(10), 1687, https://doi.org/10.3390/rs12101687

Schlather, M., Malinowski, A., Menck, P.J., Oesting, M., Strokorb, K., 2015. Analysis, simulation and prediction of multivariate random fields with package random fields. J. Stat. Softw. 63(8), 1-25, https://doi.org/10.18637/jss.v063.i08

Shen, J., Yuan, L., Zhang, J., Li, H., Bai, Z., Chen, X., Zhang, W. and Zhang, F., 2011. Phosphorus dynamics: from soil to plant. Plant physiology, 156(3), 997-1005. https://doi-org.ezproxy.nottingham.ac.uk/10.1104/pp.111.175232

Singha, C., Swain, K. C., 2016. Land suitability evaluation criteria for agricultural crop selection: A review. Agric. Reviews, 37(2), 125–132.

Smeck, N.E., 1985. Phosphorus dynamics in soils and landscapes. *Geoderma*, *36*(3-4), pp.185-199.

Stein, M.L. 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), p.25. DOI:10.1186/1471-2105-8-25

Styger, E., 2014. Rice production diagnostic for Chinsali (Chinsali District, Northern Province) and Mfuwe (Mwambe District, Eastern Province), Zambia (No. July). COMACO / Center for Sustainable Development.

Styger, E., Uphoff, N., 2014. The System of Rice Intensification (SRI): Revisiting Agronomy for a Changing Climate. Climate-smart, Practice Brief.

Suheri, N. A., Mujiyo, M., Widijanto, H., 2018. Land Suitability Evaluation for Upland Rice in Tirtomoyo District, Wonogiri Regency, Indonesia. SAINS TANAH – Journal of Soil Science and Agroclimatology, 15(1), 46.

Swallow, W. H. and Monahan, J. F. 1984. Monte-Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components. Technometrics 26, 47-57

Sys, C., Van Ranst, E., Debaveye, J., Beernaert. F., 1993. Land Evaluation. Part III: crop requirements. Agricultural Publications, 7, 1-191. GADC, Brussels, Belgium. https://www.researchgate.net/publication/324330469_Land_Evaluation_Part_3_Crop_Requirements

Tanasă, I.C., Niculită, M., Roşca, B. and Pîrnău, R., 2010. Pedometric techniques in spatialisation of soil properties for agricultural land evaluation. *Bulletin of* University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Agriculture, 67(1).

Tobler, W., 1988. Resolution, Resampling, and All That, [in:] H. Mounsey, R. Tomlinson (eds.), Building Data Bases for Global Science.

United States Geological Survey (USGS)., 2019 NASA Shuttle Radar Topography Mission (SRTM3) data available on the World Wide Web https://earthexplorer.usgs.gov (accessed 10 July, 2019)

Veldkamp, W.J., Muchinda, M. and Delmotte, A.P., 1984. Agro-climatic zones in Zambia. Soil Survey Bulletin (Zambia).

Verbeke, G., Molenberghs, G. 2000. Linear mixed models for longitudinal data. Springer-Verlag, New York.

Viscarra Rossel, R.A., Webster, R. and Kidd, D., 2014. Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. Earth Surface Processes and Landforms, 39(6), pp.735-748.

Wadoux, A.M.C., Samuel-Rosa, A., Poggio, L. and Mulder, V.L., 2020. A note on knowledge discovery and machine learning in digital soil mapping. European Journal of Soil Science, 71(2), pp.133-136.

Webster, R. and Oliver, M.A., 2007. Geostatistics for environmental scientists. John Wiley & Sons.

Wright, M. N. and Ziegler, A. 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1-17. DOI:10.18637/jss.v077.i01

Yohannes, H. and Soromessa, T., 2018. Land suitability assessment for major crops by using GIS-based multi-criteria approach in Andit Tid watershed, Ethiopia. Cogent Food & Agriculture, 4(1), p.1470481.

Zimmerman, D.L. and Zimmerman, M.B., 1991. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. Technometrics, 33(1), pp.77-91. DOI:10.1080/00401706.1991.10484771

Zimmerman, D., Pavlik, C., Ruggles, A. and Armstrong, P, M. 1999. An Experimental Comparison of Ordinary and Universal Kriging and Inverse Distance Weighting. Mathematical Geology 31: 375–390. DOI:10.1023/A:100758650743

# APPENDICES

## APPENDIX 1: Detailed RALS 2012 Sampling Procedure

The first stage involved identifying the Primary Sampling Unit (PSU) which is one or more Standard Enumeration Areas (SEAs) each comprising a minimum of 30 agricultural households. The SEA is the smallest area with well-defined boundaries identified on census sketch maps. The second stage involved listing and identification of agricultural households in selected SEAs. The listed agricultural households were then stratified into three categories - A, B and C (CSO/MAL/IAPRI, 2015). Category C comprised households with 5 to 19.99 ha of land under crops, grown one or more special crops, raising ≥ 50 cattle, ≥20 pigs, ≥ 30 goats and or ≥ 50 chickens. Category B comprised agricultural households with 2 to 4.99 hectares area under crop. Category A comprised households with 0 to 1.99 hectares of land under crop or owning livestock numbers less than those specified in category C.

Systematic sampling from the household list comprised by the enumerators in the SEA was to select 20 households distributed across the three strata. Where all the three categories had adequate numbers of households listed, the sample household distribution was C=10, B=5 and A=5. Where there were shortfalls in category C, all households in this category where selected and the difference from 20 was equally allocated to categories B and A. If the difference from 20 could not be equally allocated to the two categories, category B was allocated one more sample household than category A. Where there was no household in category C, 10 sample households were allocated to category B, and 10 to category A. Where there was no household in category C and less than 10 in category B, all were included in the sample and the allocation for category A was increased to make up for the shortfall from the required number of 20 sample households. Where all households fall in category A, all the required 20 sample households were selected from that category. For each stratum, systematic sampling was done to select the required households. First the sampling interval was calculated by dividing the total number of households in the category by the sample number. Then the random start number was selected by randomly selecting a column from the table of random numbers. Starting from the top of that column, the first random number between 1 and the number of households in the category was selected, inclusive as the first corresponding selected household in the sample. To add the next household number, the sampling interval was added to the chosen random

number and this procedure was repeated to add remaining households of the sample (CSO, 2012).

## APPENDIX 2: Journal Publications

Geoderma 397 (2021) 115079

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

# Performance of linear mixed models and random forests for spatial prediction of soil pH

Mirriam Makungwe [a,*], Lydia Mumbi Chabala [a], Benson H. Chishala [a], R. Murray Lark [b]

[a] Department of Soil Science, University of Zambia, School of Agricultural Sciences, 32379 Lusaka, Zambia
[b] School of Biosciences, University of Nottingham, Sutton Bonington, Nottinghamshire LE12 5RD, UK

ARTICLE INFO

Handling Editor: Budiman Minasny

*Keywords:*
Linear mixed models
REML-EBLUP
Random forests
Spatial prediction of soil pH

ABSTRACT

Digital soil maps describe the spatial variation of soil and provide important information on spatial variation of soil properties which provides policy makers with a synoptic view of the state of the soil. This paper presents a study to tackle the task of how to map the spatial variation of soil pH across Zambia. This was part of a project to assess suitability for rice production across the country. Legacy data on the target variable were available along with additional exhaustive environmental covariates as potential predictor variables. We had the option of undertaking spatial prediction by geostatistical or machine learning methods. We set out to compare the approaches from the selection of predictor variables through to model validation, and to test the predictors on a set of validation observations. We also addressed the problem of how to robustly validate models from legacy data when these have, as is often the case, a strongly clustered spatial distribution. The validation statistics results showed that the empirical best linear unbiased predictor (EBLUP) with the only fixed effect a constant mean (ordinary kriging) performed better than the other methods. Random forests had the largest model-based estimates of the expected squared errors. We also noticed that the random forest algorithm was prone to select as "important" spatially correlated random variables which we had simulated.

## 1. Introduction

Soil maps describe the spatial variation of soil types and provide important information on spatial variation of soil properties (Kempen et al., 2010). Mapping of soil properties is important as it provides policy makers with a synoptic view of the state of the soil, and agricultural stakeholders with information about where soil problems might occur (Lark et al., 2019). Soil maps are generated using various soil mapping methods which can be divided into conventional and pedometric approaches (Kienast-Brown et al., 2010; Hengl, 2003).

Conventional soil survey represents soil variation in terms of profile classes and corresponding map legend units. It can provide a basis for spatial prediction of soil properties and may also serve as a structure for recording substantial information on soil management and for systematizing knowledge of the distribution of soils in the landscape. Conventional approaches were based largely on manual processes which are costly and time consuming (Kienast-Brown et al., 2010) mainly because of long fieldwork periods (Moonjun et al., 2010). Pedometric approaches are based on the application of mathematical and statistical methods for the primary purpose of predicting the values of soil properties where these have not been observed directly (McBratney et al., 2000). A well-established statistical approach to doing this is the application of model-based geostatistics (Stein, 1999; Diggle and Ribeiro, 2007). In this approach the variation of the soil is represented in a linear mixed model (LMM) as a combination of fixed effects (which may be a constant unknown mean, or a function of predictive covariates such as remote sensor data), and random effects, including Gaussian random fields which exhibit spatial correlation. The parameters of the LMM model can be estimated by Residual Maximum Likelihood (REML) method developed by Patterson and Thompson (1971), which allows parameters of the random effects to be estimated with small bias arising from uncertainty in the fixed effects (Kitanidis, 1987; Swallow and Monahan, 1984; Zimmerman and Zimmerman, 1991; Lark and Cullis, 2004). When the model is fixed, values of the soil property at unsampled sites can be obtained by the empirical best linear unbiased predictor (EBLUP) (Stein, 1999; Lark et al., 2006; Lark and Webster, 2006; Minasny and McBratney, 2007).

There has been a growing interest in the potential of machine learning methods (e.g. Breiman, 2001) as an alternative to statistical

* Corresponding author: Department of Soil Science, School of Agricultural Sciences, University of Zambia, 32379 Lusaka, Zambia.
*E-mail addresses:* mirriammakungwe.tolopu@gmail.com (M. Makungwe), lchabala@unza.zm (L.M. Chabala), bchishala@unza.zm (B.H. Chishala), Murray.Lark@nottingham.ac.uk (R.M. Lark).

https://doi.org/10.1016/j.geoderma.2021.115079
Received 31 October 2020; Received in revised form 3 March 2021; Accepted 5 March 2021
Available online 2 April 2021

Geoderma Regional 27 (2021) e00438

Contents lists available at ScienceDirect

Geoderma Regional

journal homepage: www.elsevier.com/locate/geodrs

# Assessing land suitability for rainfed paddy rice production in Zambia

Mirriam Makungwe [a,*], Lydia Mumbi Chabala [a], Michiel Van Dijk [b,d], Benson H. Chishala [a], R. Murray Lark [c]

[a] Department of Soil Science, School of Agricultural Sciences, University of Zambia, P.O. Box 32379, Lusaka, Zambia
[b] Ecosystems Services and Management (ESM), International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria
[c] School of Biosciences, University of Nottingham, Sutton Bonington, Nottinghamshire LE12 5RD, UK
[d] Wageningen Economic Research, 528, Pr. Beatrixlaan 582, 2595 BM Den Haag, the Netherlands

ARTICLE INFO

*Keywords:*
Land suitability
Multi-criteria evaluation
Paddy rice
Multiple soil classes

ABSTRACT

Rice is one of the staple food crops and is a profitable smallholder cash crop in Zambia. It has the potential to contribute significantly to increased incomes and employment among rural producers. However, rice is the only staple crop in the country for which domestic production does not meet or exceed domestic demand. Low productivity is one of the factors that contribute to this. One necessary step towards addressing this problem is the identification of land with greatest potential for rice production, as well as the identification of land-based limitations which might be overcome by improved management. The aim of this study was to develop a land suitability index for rainfed paddy rice production reflecting expert opinion and published studies based on climatic, topographic and soil properties. Land suitability was evaluated using a method which accounts for important multiple factors, and which considers their joint effect in terms of a hierarchical model of constraints. The suitability classes were ranked according to the FAO land suitability classification as: Highly Suitable (S1), Moderately Suitable (S2), Marginally Suitable (S3), Currently Not Suitable (N2), and Permanently Not Suitable (N1). Results showed that there is limited potential for rainfed paddy rice production in Zambia with < 20% of the land classified as either highly or moderately suitable. Therefore, the potential of irrigated and upland rice production in Zambia needs to be assessed as this would help expand the potential production area of rice.

## 1. Introduction

Rice, in addition to maize, cassava, sorghum, millet, wheat, sweet and Irish potato, is one of the staple food crops (Styger, 2014) in Zambia. It is a profitable smallholder cash crop with the potential to contribute significantly to increased incomes and employment among rural producers (Chizhuka, 2009). The current status of rice is evidence of its growing importance. The annual demand for rice rose steadily from below 20,000 tones to almost 70,000 tones for the period 2003–2017 as illustrated in Fig. 1 (CSO, 2018).

However, the demand for rice exceeds production, making rice the only crop in Zambia with a deficit. To meet this deficit, the country has imported between 5000 and 20,000 tons of milled rice annually (Ministry of Agriculture, 2016). In response, the government through the Ministry of Agriculture, developed the National Rice Development Strategy (NRDS) in 2016, whose overall objective was to increase local rice production by at least 50% and to enhance its competitiveness on the market by the year 2020. However, to date the national average yield of rice has not increased, neither has the area planted, although the staple requirement continues to increase (Table 1).

Poor yield is one of the factors that has contributed to Zambia's inability to meet the increasing demand for rice through local production. Average rice yields are 1.3 t/ha (CSO/MAL/RALS, 2015) which is quite low when compared to other Eastern and Southern African countries such as South Africa, Kenya, Uganda and Zimbabwe where national average yields were 2.61, 5.24, 2.30 and 2.26 t/ha respectively for the year 2013 (Ministry of Agriculture, 2016).

Apart from soil constraints (Aune et al., 2014), poor water management is also one of the factors that limits rice yields (Styger and Uphoff, 2016). Most of the rice grown in Zambia is rainfed paddy rice and this limits its cultivation to flooded or semi-flooded lowland environments (Mutale et al., 2010). With frequent occurrence of droughts, floods and other extreme weather conditions due to climate change, farmers generally find it difficult to improve production and

* Corresponding author.
*E-mail addresses:* mirriammakungwe.tolopu@gmail.com (M. Makungwe), lchabala@unza.zm (L.M. Chabala), Michiel.vandijk@wur.nl (M. Van Dijk), bchishala@unza.zm (B.H. Chishala), Murray.Lark@nottingham.ac.uk (R.M. Lark).

https://doi.org/10.1016/j.geodrs.2021.e00438
Received 24 May 2021; Received in revised form 29 August 2021; Accepted 30 August 2021
Available online 2 September 2021