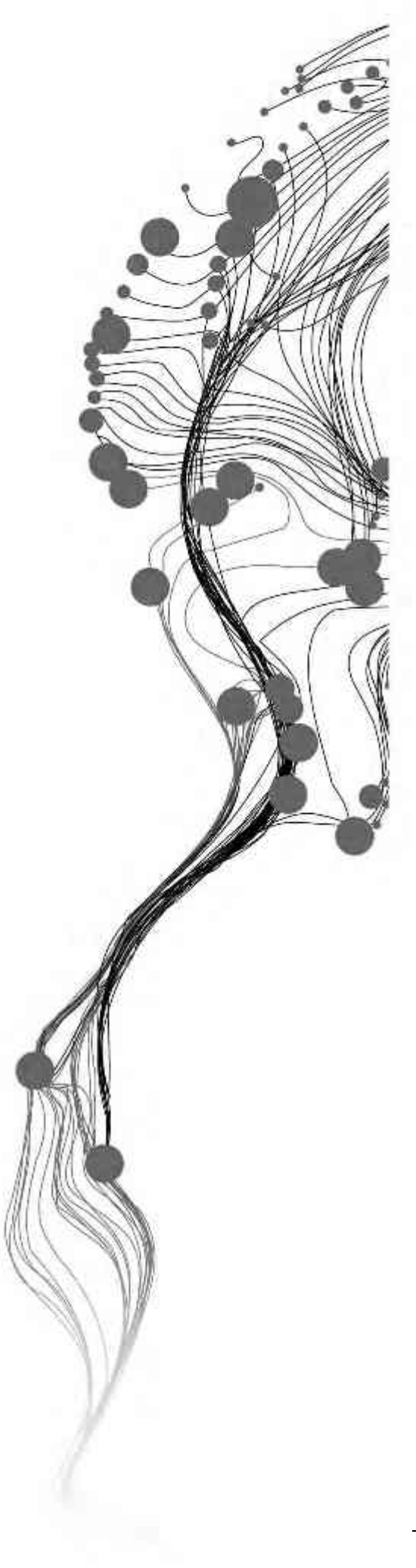
Comparison of deterministic and stochastic interpolation methods by assessing spatial variability in soil properties in a hilly terrain

VAIBHAV CHHIPA
March, 2018

SUPERVISORS:
Mr. Hari Shankar
prof.dr.ir. Alfred Stein
Mr. Justin George K
Advisor: Ms. Fakhereh Alidoost



Comparison of deterministic and stochastic interpolation methods by assessing spatial variability in soil properties in a hilly terrain

VAIBHAV CHHIPA
Enschede, The Netherlands, March, 2018

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-Information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Mr Hari Shankar
prof.dr.ir.Alfred Stein
Mr Justin George K
Advisor: Ms Fakhereh Alidoost

THESIS ASSESSMENT BOARD:
Chair : dr.ir. R.A. de By
ITC Professor : prof.dr.ir. Alfred Stein
External Examiner: Prof. S.K. Ghosh, Indian Institute of Technology (IIT), Roorkee

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

Faith is the bird that feels the light when the dawn is still dark

- Rabindranath Tagore

ABSTRACT

Analysis of environmental variables require accurate sampling of locations. To study natural properties that are continuous across space, interpolation is required. This is because not all locations can be sampled due to various physical and financial constraints. Point data are usually collected from the field and the values at the unknown locations are interpolated. Interpolation is defined as the prediction of values between a range of values. The choice of the interpolation method depends on the objective of the study. Broadly, two types of interpolation methods are present – one which cannot derive the error values and is based on parametric equations, known as deterministic and the other which considers the spatial dependency between random variables, are called geostatistical. Both the methods had been explored in this research work, RBF which is a deterministic method and Bayesian kriging, regression kriging and copula-based interpolators which are the geostatistical methods.

Three soil parameters – pH, Electrical Conductivity and Total Organic Carbon were considered to find the best among all the mentioned interpolators. In terms of the aforementioned parameters, the soil in the study area was found to be acidic, without any salts and sufficient TOC content was present. Cressie's robust variogram estimator and the optimal pixel size for interpolation were also taken into account. Optimal sampling scheme was designed for each of the study areas. It was based on minimizing the kriging variance. 96 sampling points were considered in Langha-Tauli, and 7 were considered in Barwa. As the Bayesian kriging process considers the uncertainty in parameter values, it was used to check if the spatial information could be utilized from the first study area to the second. The mean error, mean square error and the residual variance values of 0.1466, 0.0772 and 0.7306 respectively were quite satisfactory in Barwa using Bayesian kriging as compared to ordinary kriging error values.

With regards to the application of the interpolation methods, regression kriging outperformed all the other methods in terms of the uncertainty measurements at the surface and sub-surface levels for all the soil parameters. The obtained mean error, mean squared error, root mean square error values for pH at the surface and sub-surface levels were 5.78×10^{-6}, 3.15×10^{-6}, 0.0018 and 1.42×10^{-7}, 4.22×10^{-7}, 0.0006 respectively. Similarly, the obtained values in the same sequence for electrical conductivity were 0.0001, 0.3328, 0.5768 and 0.0368, 805.1854, 28.3758 respectively for the surface and sub-surface levels. For TOC, the error values were 0.0031, 0.1594 and 0.3992 for the surface level and 0.0004, 0.0138 and 0.1174 for the sub-surface level.

Although copulas-based interpolators were believed to perform better than the other methods, they performed the worst. This may have been attributed to less skewness or near-normal distribution of data. The proceedings from this research work may be recommended for future government schemes wherein soil health needs to be assessed.

ACKNOWLEDGEMENTS

This research work wouldn't have been possible without the efforts and guidance of some of the people who tried to help me in times of need.

I would like to thank my supervisors, Mr Hari Shankar, prof.dr.ir. Alfred Stein and Mr Justin George K who gave me direction on how to go about the research. I would especially like to thank Ms Fakhereh Alidoost, who took her time in explaining me the new field of copulas. She was quite prompt in replying to my queries. As the study was an interdisciplinary one, field visits and work in the laboratory were required. For this, I am grateful to Mr Yogesh Ghotekar, Mr Gyandeep and Mr Vicky who helped me out in conducting the lab experiments. Though the task was huge and time-consuming, I was helped in every possible way by them.

I would also like to thank Dr Sameer Saran, Course Director and Dr Valentyn Tolpekin, Course Coordinator at ITC who helped us in getting through the structure of the course. They gave us advice on matters other than academics as well. One of the reasons why our stay at ITC was quite pleasant was because of Dr Valentyn Tolpekin.

I would like to thank IIRS and ITC for giving me this opportunity to study a course which I was passionate about.

I greatly appreciate the company and experiences with my batchmates without whom this journey would have been dull. Special thanks to Krishnakali and Surbhi who advised and supported me at the 'not so good times' and to continue doing my work. Additionally, I would like to thank a Russian lady, without whom this research wouldn't have been possible. Also, music helped me in getting through the course.

Finally, I would like to thank my parents and family, who were always there for me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

According to Merriam - Webster (2017) interpolation is defined as "the process of calculating an approximate value based on the value that is already known". The prediction value at a location may be calculated as the weighted average of the observation value. The predicting parameter is considered as a random variable taking into account all possible realizations at that location When samples are collected from different locations, they are assumed to be drawn from one particular realization of the random experiment (Schabenberger & Pierce, 2001). It is like a photograph being taken of some object in the space-time continuum. The weights may have a fixed equation or they may define the dependence structure, depending on the objective and the interpolation method being used. Interpolation methods may be divided into three types – deterministic, geostatistical and their combination. Methods that depend upon certain parameters for prediction of values are defined as deterministic such as Inverse Distance Weighting (IDW) and Radial Basis Functions (RBF). Methods that additionally consider random functions, including the spatial dependence between points, are called geostatistical methods. Particular examples – are Simple Kriging and Cokriging (Sluiter, 2008). The dependence structure may depend on spatial coordinates or external variable values. The external variable values may help in improving the prediction process (Pebesma, 2006). Some of these methods used in the field of environmental sciences were compared by Li & Heap (2014).

Accurate information regarding soil properties is required to address issues related to land and soil quality. If a hillslope is to be used for farming purposes [- terrace farming], then a soil quality and soil health analysis need to be carried out. This governs the type of trees/crops that may grow in that area. Since the soil of the whole area cannot be analysed, representative soil samples need to be collected to study these effects. A soil survey can be a tedious and costly task. Therefore, an optimal sampling scheme needs to be designed for an efficient and economical collection of samples.

A soil study is particularly important in hilly terrains as it is difficult to collect soil information for a whole area due to accessibility issues. Interpolation methods can be used to study the spatial variability in soil properties since they can address the spatial variation in point data values. The need for monitoring changes and assessment of deterioration of soil quality has been presented through a selection of indicators in Arshad & Martin (2002).

Multiple methods had been employed to perform soil mapping such as by airborne gamma radiometric data (Cook et al., 1996). Although this was used for identifying the presence of material spatially, it did not consider its presence at the surface and sub-surface levels. Even satellites had been employed for studying soil properties such as soil moisture (Wagner et al., 2007) and organic matter content as well as the presence/absence of organic soil (Poggio et al., 2013). Not all properties could be assessed by means of these methods. Moreover, these measurements were not too accurate. Thus, the requirement for taking representative samples on the ground remains and the subsequent use of interpolation methods.

The following research aimed to predict soil information within a hilly terrain. Reducing the uncertainty measurements was required to get an accurate measure of the values of soil parameters. Topographically similar features had been identified as study areas to understand the transfer of spatial information from

one area to another. Deterministic and stochastic interpolation methods were explored and compared to best model the prediction surface.

1.1. Research identification

This research involved soil sampling and the application of interpolation methods to generate a continuous surface of the soil parameters. Values at unknown locations may be required for various objectives. For this, optimal sampling schemes were designed such that minimum variation in predicted values is there. The interpolation methods may be used for getting the values at different magnitudes. The parameters considered for this research were pH, Total Organic Carbon (TOC) and Electrical Conductivity (EC). These parameters had been identified by Jones (2016) and Arshad & Martin (2002) as primary indicators for the assessment of soil health. Also, the Department of Agriculture under the Government of India promoted them as the physical and basic parameters for assessing the soil health by providing Soil Health Cards (SHCs) to the farmers (National portal of India, 2017). Also, the parameters were tested for correlation among themselves and between each other. External variables such as elevation data and its derivatives were employed in geostatistical methods to make better sampling schemes.

1.2. Research objectives

The main objective of this study is to perform a comparison of deterministic and geostatistical interpolation methods on two hillslopes of Sitlarao watershed area in Dehradun region of India.
The specific objectives are:

i. Conduct a literature review of different sampling strategies and a comparison of interpolation methods.
ii. Develop a suitable sampling strategy for collecting soil data.
iii. Apply different interpolation methods for the sample point data.
iv. Critically analyse the soil properties variability and its effects in the study area.
v. Make a solid comparison of the interpolation methods.

1.3. Research questions

With reference to the objectives mentioned above, following are the questions that need to be answered:

Specific objective 1

i. What are the different sampling strategies and the differences between them?
ii. What are the different interpolation methods and the differences between them?

Specific objective 2

i. What are the criteria to select a suitable sampling strategy for collecting the data?
ii. How many samples need to be collected for statistically significant analysis?

Specific objective 3

i. Which interpolation methods need to be applied to the sample data and why?
ii. What external variables/covariates needs to be used for application with respect to geostatistical methods?
iii. How to obtain the values of covariates at unvisited locations?

Specific objective 4

i. What is the effect of the presence of soil parameter values on soil health in the study area?
ii. Can a correlation be made between the topographical features and the soil parameter value?

Specific objective 5

i. What measures of uncertainty are to be used to analyse the quality of interpolation methods and why?

1.4. Innovation aimed at

This research tried to identify the best interpolation method amongst various deterministic and geostatistical methods for a hilly terrain. The soil parameter values may have had highly varied values depending upon the terrain structure which increased the complexity in the application of these methods. Copulas which have been previously used in the field of financial mathematics were used as an interpolator. This has been a recent application in the field of spatial statistics. Also, no previous application of copula-based interpolator had been applied in a hilly terrain in Indian soils.

Additionally, the use of spatial information from one study area to another was examined. The information collected from an area could be utilized for another area with similar characteristics without losing many details. Then it could help in designing an optimal sampling scheme and thereafter prediction surfaces without any prior survey.

1.5. Structure of the thesis

The thesis is structured in the following manner:

1. **Chapter 2** establishes the literature review of the works done by researchers in the past and the theory behind different methods used in the research. Section 2.1 explains the spatial sampling strategies and the usage of the Spatial Simulated Annealing method for generating optimal sampling schemes. The theory behind the usage of spatial information from one area to another i.e. the Bayesian kriging method is mentioned in Section 2.2. The equations for identifying the right pixel size for interpolation is given in Section 2.3. Different interpolation methods – Radial Basis Function, Regression kriging and copula based interpolator have been explained in Section 2.4. The different uncertainty measurements used in the study are as mentioned in Section 2.5.
2. **Chapter 3** describes about the topographical and the geographical characteristics of the two study areas.
3. **Chapter 4** discusses the methodology part of the research. The data that had been used are mentioned in Section 4.1. The approach followed for generating the optimal sampling schemes are explained in Section 4.2. Whereas, Section 4.3 gives the steps for the laboratory tests that had been performed. Section 4.4 mentions the pixel size that had been used in the study for interpolation in the two study areas. The application of the robust variogram estimation and fitting is given in Section 4.5. The implementation of the Bayesian kriging method for studying the usage of spatial information from one area to another is given in Section 4.6. Section 4.7 explains the different steps followed in the application of different interpolation methods and the method used for the comparison of them.
4. **Chapter 5** discusses the results of the study wherein Section 5.1 shows the sampling scheme obtained for the two study areas. The descriptive statistics based on the data collected from the two study areas are mentioned in Section 5.2. The results of the Bayesian kriging implementation are as given in Section 5.3. The generated maps using different interpolation methods are presented in Section 5.4.The values for the uncertainty measurements for different interpolation methods are mentioned in Section 5.5.

5. **Chapter 6** discusses the obtained results and makes correlations with the previous studies. Section 6.1 and Section 6.2 discusses about the optimal sampling scheme and the descriptive statistics with soil health, respectively. The findings from Bayesian kriging are discussed in Section 6.3. The findings from the application of different interpolation methods are discussed in Section 6.4.
6. **Chapter 7** gives the conclusion of the research work and states some of the recommendations as future scope of the study.

2. LITERATURE REVIEW

Environmental variables can be best modelled by taking representations from the real world. This is due to various economic and geographical constraints. Various density distribution functions may be studied in turn to assess the variation in soil information. Soil variables have been generally found to follow a positively skewed distribution. Some of the most commonly used distributions for their modelling are – Poisson, Weibull, Gamma, Exponential, and Log-normal (Becker et al., 1992). For the density distribution functions to be generated, point samples are required. Representative samples require careful sampling design for generating accurate and precise data (United States Environmental Protection Agency, 2002). Since data forms the backbone of any mathematical analysis, designing a good sampling scheme becomes a necessity.

2.1. Spatial sampling

Various sampling methods and statistical techniques for soil-survey data have been defined by Webster & Oliver (1990). They mention the advantages and disadvantages of each of those methods. Also, they explain the greater efficiency of stratified and unaligned sampling methods over simple random sampling methods for soil survey. The choice of the sampling method usually depends on the desired objective of the study. Sampling considers various objectives – such as for independent and identically distributed (iid) population, considering correlation and heterogeneity, which have been mentioned by Wang et al. (2012). They explain various design (such as simple and stratified random, systematic and two-step random sampling) and model-based sampling methods. Design based sampling methods are those that have their population (single realization of the random experiment) unknown but fixed. Model-based sampling methods have been described as those that have their population unfixed but as a set of values (superpopulation) representing a single realization of the random experiment. They involve minimizing a single objective function. 3 criteria for objective functions – minimization of estimation error variance, equal spatial coverage for irregular polygons and equal coverage in feature space have been discussed.

2.1.1. Soil and geostatistics

The uncertainties associated with predicted values might be present due to instrumental or measurement errors. The deterministic interpolation methods cannot report these error values. The stochastic processes [geostatistical processes] may be considered to get the uncertainty measures at these locations. Lark (2012) concluded that geostatistics and soil science were closely interlinked. Although there are lots of factors at play for soil processes, he was hopeful for a further development of soil processes being linked to statistical distributions. Geostatistics has been used in the past for generating optimal sampling schemes. It was found that they generated more efficient schemes in terms of cost [measurement time] than the traditional sampling schemes (Xiao et al., 2005).

Yfantis et al. (1987) found out that among the square, equilateral and hexagonal grids, equilateral grid sampling scheme gave the most reliable estimate of the variogram. The variogram defines the spatial dependence of a random variable. It is visualised as the variance between spatial observations of a random variable at different lag/distance classes. The variogram estimator ($\hat{\gamma}_k$) was defined in the following manner (Müller, 1999; Matheron, 1963):

$$2\hat{\gamma}_k = \frac{1}{N_{H_k}} \sum_{s_i,\, s_i+h \in H_k} (z(s_i) - z(s_i + h))^2 ; \tag{2.1}$$

In equation 2.1, H_k denotes the distance bins containing all the point pairs; N_{H_k} denotes the number of point pairs falling in each bin and h denotes the lag distance. Also, $z(s_i)$ and $z(s_i + h)$ denotes the observation values at locations s_i and $s_i + h$ respectively.

2.1.2. The spatial simulated annealing method

Annealing [in metallurgy] is the process of heating a metal above a certain recrystallization temperature and then cooling either rapidly or slowly depending on the desired product. The lattice structure starts to come to its equilibrium state with cooling, hence increasing the workability of the metal. Simulated Annealing (SA) is a similar process to find the global minima/maxima wherein perturbations analogous to heating and cooling processes are given to the mathematical function. It was first proposed by Metropolis et al. (1953), which later came to be known as the Metropolis criterion. Spatial Simulated Annealing (SSA) is the extension of SA method in the geographical domain. Van Groenigen et al. (1999) explained this in the following way.

A collection of possible sampling schemes S^n consisting of *n* observations was considered. An objective/fitness function $\phi(S_i) \in S^n$ was defined which was to be minimized. Initially, a random sampling scheme S_0 was taken and then random perturbations added to it such that a new sampling scheme S_{i+1} was generated. It had a probability $P_c(S_i \rightarrow S_{i+1})$ of being accepted which was defined in the form of Metropolis criterion as:

$$P_c(S_i \rightarrow S_{i+1}) = 1, \qquad if\ \phi(S_{i+1}) \leq \phi(S_i)$$

$$P_c(S_i \rightarrow S_{i+1}) = \exp\left(\frac{\phi(S_i) - \phi(S_{i+1})}{c}\right), \qquad if\ \phi(S_{i+1}) > \phi(S_i); \tag{2.2}$$

Here, c denotes the control parameter which decreases as the optimization progresses. The random perturbations were added to the sample points such that the points moved to the new location in random direction and at a random distance $h \in (0, h_{max})$. The distance h_{max} was initially considered to be half the length of the study area in the two dimensions. It gradually decreased with each optimization step.
van Groenigen (1997) showed that SSA could be used for generating optimal sampling schemes. He found out that it gave better sampling scheme than the equilateral triangular grid.

2.1.3. Optimal sampling schemes

In the case of model-based sampling methods, SSA has been used in the past for generating optimal sampling scheme with minimal kriging variance as the criterion (Van Groenigen et al., 1999; van Groenigen & Stein, 1998). They used ordinary kriging (OK) variance as the objective function to be minimized. SSA with Minimization of the Mean of Shortest Distances (MMSD) was used as a criterion for determining the global minima as the sampling configuration was changed in each iteration. Even spreading of points over an area was achieved through MMSD. Also, regression kriging (RK) variance has been used as the objective function in cases where there were a lot of constraints involved (Szatmári et al., 2015).

The generation of the objective function with regards to kriging variance requires the variogram to denote the variation of soil properties. For a fair computation of the variogram, at least 100 samples need to be collected, 150 samples for satisfactory and 225 for a reliable computation of a normally distributed isotropic variable (Webster & Oliver, 1992). The variable whose properties does not depend on the direction is isotropic.

Guidelines for collecting soil samples and their description have been provided by the Food and Agricultural Organization (FAO) of the United Nations (Jahn et al., 2006). These help in the proper management and handling of the collected soil samples.

2.2. Usage of spatial information from one study area to another

For generating prediction surfaces, spatial information may be used from one area to another if the two of them are found to be similar in properties. This may be done to save on any additional costs for sampling as well as for conducting laboratory tests. A procedure was developed by Cui et al. (1995) for generating continuous surfaces of soil parameters by using the Bayesian form of kriging. They had compared whether Bayesian kriging performed any better than ordinary kriging. The results of their study led them to conclude that although ordinary kriging performed better for a large number of observations, Bayesian kriging predicted values of approximately the same precision as ordinary kriging for a smaller number of observations.

2.2.1. Bayesian Kriging (BK)

Bayesian kriging involves specification of prior distributions for the parameters instead of them being estimated. These distributions are updated regularly based on the data, using Markov chain Monte Carlo (MCMC) simulation. Thus, leading to posterior distributions for each of the parameters. The advantage of BK over other forms of interpolation methods is that it quantifies the uncertainty in the estimation of model parameter values (Verdin et al., 2015). Diggle & Ribeiro (1999) explained the Bayesian form of kriging in the following manner: -

Considering the spatial observations $Z(s_1) \dots Z(s_n)$ as being the single realisation of a random variable Z at the set of locations $s_i, i \in [1, n]$ and $s_i \in \mathbb{R}^d$ with positive d - dimensional volume. The model considers the variable Z being a "noisy" version of a latent spatial process, the signal $Q(s), s$ denoting the vector of locations $s_1 \dots s_n$. The "noises" are assumed to be Gaussian and conditionally independent given $Q(s)$. According to the given definitions and assumptions, the model is specified in a hierarchical scheme. The signal $Q(s)$ is considered to be decomposed into a sum of latent processes $T_k(s)$ scaled by σ_k^2. Thus, the model is written as follows:

$$\text{Level 1}: Z(s) = X(s)\beta + Q(s) + \varepsilon(s)$$
$$= X(s)\beta + \sum_{k=1}^{K} \sigma_k T_k(s) + \varepsilon(s); \quad (2.3)$$

$$\text{Level 2}: T_k(s) \sim \mathcal{N}\left(0, R_k(\phi_k)\right), T_1 \dots T_K \text{ are mutually independent and}$$
$$\varepsilon(s) \sim \mathcal{N}(0, \tau^2 \mathrm{I}); \quad (2.4)$$

$$\text{Level 3}: (\beta, \sigma^2, \phi, \tau^2) \sim pr(\cdot), \text{ a prior distribution} \quad (2.5)$$

where, the model components are described as:

1. $Z(s)$ is a random vector stating the sample location measurements;
2. $X(s)\beta = \mu(s)$ is the expectation of $Z(s)$. $X(s)$ is the matrix of fixed covariates at locations s. β is a vector parameter. If, there are no covariates, $X(s) = 1$ and the mean becomes a constant value at all the locations. In geostatistical terms, the term **trend** refers to the mean part of the model $X(s)\beta$;

3. $T_k(s)$ is the random vector at sample locations, of a standardised latent stationary spatial process T_k. It has zero mean, variance one and correlation matrix $R_k(\phi_k)$. The elements of $R_k(\phi_k)$ are given by a correlation function $\rho_k(h;\phi_k)$. If the process is isotropic this parameter is denoted by ϕ_k and h is reduced to a scalar h i.e. the Euclidean distance between two locations.T_k refers to a **structure in a variogram**;
4. σ_k is a scale parameter. The value σ_k^2 corresponds to the **partial sill** of a variogram;
5. $\varepsilon(s)$ denotes the error vector at the sample locations s. It has zero mean and variance τ^2 at the sample locations. The **nugget effect** in a variogram is denoted by τ^2;
6. In a Bayesian approach to inference, the specification of the prior for the model parameters is given in the third level. Conjugate priors are taken into account. These refer to the same family of distributions in the posterior as the ones specified in prior.

Now, considering the probability distribution of Z by the function $pr(z|\vartheta)$, indexed by the unknown vector parameter $\vartheta = (\beta, \sigma^2, \phi, \tau^2)$. z is the sample observed and $L(\vartheta|z) \equiv pr(z|\vartheta)$, $L(\cdot)$ is a function of ϑ and is called the likelihood function. In the Bayesian approach, variable Z and parameters ϑ are considered as random with joint distribution $pr(z,\vartheta) = pr(z|\vartheta)pr(\vartheta)$. Here, $pr(\vartheta)$ is the prior distribution and | denotes conditionality. Bayes' Theorem (Weisstein, n.d.) updates the prior knowledge about the parameters using the relation:

$$pr(\vartheta|Z) \propto pr(\vartheta)pr(Z|\vartheta); \tag{2.6}$$

The distribution $pr(\vartheta|Z)$ is called posterior distribution which forms the basis for Bayesian inference of model parameters.

Let z_0 and $pr(z_0|z)$ denote the vector of prediction locations and the predictive distribution respectively. The predictive distribution may be written as follows:

$$\begin{aligned} pr(z_0|z) &= \int pr(z_0,\vartheta|z)d\vartheta \\ &= \int pr(z_0|z,\vartheta)pr(\vartheta|z)d\vartheta; \end{aligned} \tag{2.7}$$

In equation (2.7), $pr(z_0|z,\vartheta)$ refers to the conditional distribution with weights given by the posterior distribution $pr(\vartheta|z)$. In terms of the Bayesian inference, the predictive distribution may also be written as:

$$pr(z_0|z) = \int \frac{pr(z,z_0|\vartheta)pr(\vartheta)}{\int pr(z|\vartheta)pr(\vartheta)d\vartheta} d\vartheta; \tag{2.8}$$

2.3. Right pixel size for interpolation

Before performing interpolation, the scientific justification for choosing the grid resolution (pixel size, in case of raster images) needs to be presented. One should not randomly choose the grid resolution without any sound proof. Hengl (2006) explained methods to choose grid resolution based on various aspects. According to him, no ideal pixel size existed, but it could be chosen in such a way that compliance with the input datasets may be maintained. The equations for the range of resolutions and a possible compromise were as given in Table 2-1.

Table 2-1: Summary equations to select grid resolution (Hengl, 2006).

Aspect	Coarsest legible resolution	Finest legible resolution	Recommended compromise
Working scale	$\leq SN \times 0.0025$	$\geq SN \times 0.0001$	$= SN \times 0.0005$
GPS positioning error	$\leq 1.8 \times r_{E(P=99\%)}$	$\geq \bar{r}_E \times \sqrt{\pi}$	$= 1.8 \times r_{E(P=95\%)}$
Size of reference objects	$\leq \frac{\sqrt{\bar{a}}}{4}$	$\geq \frac{\sqrt{w_{MLD}}}{2}$	$= \frac{\sqrt{a_{MLD}}}{4}$
Inspection density	$\leq 0.1 \times \sqrt{\frac{A}{N}}$	$\geq 0.05 \times \sqrt{\frac{A}{N}}$	$= 0.0791 \times \sqrt{\frac{A}{N}}$
Distance between points	$\leq \frac{\bar{h}_{ij}}{2}$	$\geq h_{ij(P=5\%)}$	$= 0.25(0.5) \times \sqrt{\frac{A}{N}}$
Spatial dependence structure	$\leq \frac{h_R}{2}$	$\geq h_{ij(P=5\%)}$	$= h_R \times m^{\frac{-1}{3}}$
Complexity of terrain	$\leq \frac{A}{\sum l}$	$\geq \frac{w_{MLD}}{2}$	$= \frac{A}{2 \times \sum l}$

SN is scale factor, r_E is positioning error, $\bar{r}_E$ is average positioning error, $\bar{a}$ is average size of delineations, a_{MLD} is area of the minimum legible delineation, w_{MLD} is width of narrowest legible delineation, A is surface of study area in m^2, N is number of sampled points in study area, h_{ij} is spacing between closest point pairs, $\bar{h}_{ij}$ is average spacing between closest points, h_R is range of spatial dependence, m is number of point pairs within range of spatial dependence, and l is the total length of contours

2.4. Interpolation methods

Values at all locations cannot be sampled. They can only be predicted otherwise it would become impracticable to collect samples at each and every location. Different interpolation methods are used to generate a continuous surface. Jasiński (2016) demonstrated its usage for modelling environmental data such as air temperature and SO_2 (Sulphur dioxide) concentration. He concluded that the interpolation methods can be used satisfactorily for modelling environmental data. Also, he suggested that interpolation is preceded by an assessment of the modelling accuracy, for different ways of filling in the unknown values, for getting the best results. 25 of these methods had been compared by Li & Heap (2014) and the similarities amongst each other discussed. They also stated the software packages that may be used for performing interpolation. Some of the deterministic (radial basis function), as well as geostatistical (regression kriging, copulas as interpolators) interpolation methods, have been discussed below.

2.4.1. Radial Basis Functions (RBF)

RBFs have been used in the past for multivariate interpolation (Lazzaro & Montefusco, 2002). It is a mathematical function whose values depends on the distance from an absolute centre. The basis is a set of elements in the vector space which are linearly independent. All the other vectors may be written as a linear combination of these vectors. Wright (2003) explained their usage in generating continuous surfaces by stating them as a generalised version of the multiquadric equations given by Hardy (1971). He gave the following definition for the basic RBF method:

Given n distinct data points $\{x_j\}_{j=1}^{n}$ and their corresponding data value $\{f_j\}_{j=1}^{n}$, the basic RBF interpolant was given as,

$$s(x) = \sum_{j=1}^{n} \lambda_j \phi(\|x - x_j\|); \quad (2.9)$$

where $\phi(r) = \phi\|x - x_j\|, r \geq 0$ is some radial function. Coefficients λ_j are determined from the conditions $s(x_j) = f_j, j = 1, \ldots, n$ leading to the following linear equation:

$$[A][\lambda] = [f]; \quad (2.10)$$

Here, the entries of A is described by $a_{j,k} = \phi(\|x_j - x_k\|)$. Also, $\|\cdot\|$ refers to the norm of the equation. Generally Euclidean norm is used for RBF. Some of the most commonly used radial basis functions are as given in Table 2-2. The parameter ε is a fixed non – zero value used for controlling the shape of functions.

Table 2-2: Commonly used RBFs (Wright, 2003)

Type of basis function	$\boldsymbol{\phi(r), r \geq 0}$
Infinitely smooth RBFs	
Gaussian	$e^{-(\varepsilon r)^2}$
Inverse quadratic	$\frac{1}{1 + (\varepsilon r)^2}$
Inverse multiquadric	$\frac{1}{\sqrt{1 + (\varepsilon r)^2}}$
Multiquadric	$\sqrt{1 + (\varepsilon r)^2}$
Piecewise smooth RBFs	
Linear	r
Cubic	r^3
Thin Plate Spline (TPS)	$r^2 \log r$

2.4.2. Regression Kriging (RK)

Spatial observations $Z(s_1), \ldots, Z(s_n)$ of a random variable Z are not the same as being observed n times over, but the variables at locations $s_i, i \in [1, n]$ observed once. Random variable value $Z(s_0)$ is usually considered by taking the distribution of all possible realizations at that location (Schabenberger & Pierce, 2001). When these observations have a constant spatial mean at all locations, these are termed to be stationary. It is not always reasonable to assume that the mean is constant. It may vary with respect to covariates or coordinates.

To address the non-stationarity of mean in ordinary kriging, regression kriging is used. Hengl et al. (2007) explained this in the following manner.

In a geostatistical approach, predictions at unknown locations are usually given as the weighted average of the observations:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i \cdot z(s_i); \tag{2.11}$$

Here, $\hat{z}(s_0)$ denotes the prediction value. The observation values at different locations is given by the data values $z(s_1), z(s_2), \dots, z(s_n)$. RK uses the values of the auxiliary variable at unknown locations to predict the predictor variable values. In RK, regression is used to fit the explanatory variation and simple kriging with expected value 0 is used to fit the residuals (unexplained variation) (Hengl et al., 2004):

$$\begin{aligned} \hat{z}(s_0) &= \hat{m}(s_0) + \hat{e}(s_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k . q_k(s_0) + \sum_{i=1}^{n} \lambda_i \cdot e(s_i); \end{aligned} \tag{2.12}$$

In equation 2.12, $\hat{m}(s_0)$ is the fitted drift, $\hat{e}(s_0)$ is the interpolated residual, $\hat{\beta}_k$ are estimated drift model coefficients or the regression coefficients, $q_k(s_0)$ is the predictor at location s_0, λ_i are kriging weights determined by the spatial dependence structure i.e. the variogram parameters (Matheron, 1969) of the residual where $e(s_i)$ is the residual at location s_i. The regression coefficients $\hat{\beta}_k$ are estimated from the samples by either the ordinary least squares (OLS) method or the generalized least squares (GLS), the latter being more optimal. This takes into account the spatial correlation between observations (Cressie, 2015):

$$\hat{\beta}_{GLS} = (q^T . C^{-1} . q)^{-1} . q^T . C^{-1} . z; \tag{2.13}$$

In equation 2.13, $\hat{\beta}_{GLS}$ is the vector of estimated regression coefficients, C is the covariance matrix of the residuals, q is a matrix of predictors and z is the vector of measured values of the predictor variable. In matrix notation, the equation for the predicted value at location s_0 is written as follows (Christensen, 2001):

$$\hat{z}(s_0) = q_0^T . \hat{\beta}_{GLS} + \lambda_0^T . \left(z - q . \hat{\beta}_{GLS}\right); \tag{2.14}$$

In equation 2.14, q_0 is the vector of $p + 1$ predictors, and λ_0 is the vector of n kriging weights used to interpolate the residuals. The RK prediction error variance is given as follows:

$$\sigma_{RK}^2(s_0) = (C_0 + C_1) - c_0^T . C^{-1} . c_0 + (q_0 - q^T . C^{-1} . c_0)^T . (q^T . C^{-1} . q)^{-1} . (q_0 - q^T . C^{-1} . c_0); \tag{2.15}$$

In the above equation, $C_0 + C_1$ is the total sill value of the variogram and c_0 is the vector of covariances of residuals at locations with unknown values.

The estimation of the residuals is an iterative process wherein the drift model is firstly estimated using OLS. Next, the covariance function of the residuals is used to obtain the GLS coefficients, which are further used to calculate the residuals and then the covariance function and so on (Hengl et al., 2007).

2.4.3. Copulas

Copulas were first proposed by Sklar (1959), who described them in the following form:

Let $H(x_1, \dots, x_n)$ be a joint n – variate distribution function with margins $F_1(x_1), \dots, F_n(x_n)$. Then there exists a n - dimensional copula C_n such that $\forall\ x_1, \dots, x_n$ in $\mathbb{R}$,

$$H(x_1, \dots, x_n) = C_n(F_1(x_1), \dots, F_n(x_n)); \tag{2.16}$$

Copulas may be defined as functions that join the joint distribution function to their one – dimensional margins (Nelsen, 2006). In other words, their one – dimensional marginals are uniform in the interval $[0,1]$.

Geostatistical methods like kriging require the data to be normally distributed for giving the best results. In the real world, data may not always be normally distributed. Although data transformations may be applied, still it may not be distributed normally. The process of interpolation using copulas does not necessarily require the data to be normally distributed, which forms the strength of the model. They help by separating the dependence structure from the marginal distribution (Bárdossy & Li, 2008). Thereby, the margins can be estimated separately and the dependence structure is explained by copulas. The interest in these generated out of their capability to model non – Gaussian distributions (Kazianka & Pilz, 2010). They have been used in the past for predicting groundwater quality parameters (Bárdossy & Li, 2008), precipitation at different time scales (Bárdossy & Pegram, 2013) and soil properties (Marchant et al., 2011). Copulas as interpolators performed better than most of the other geostatistical methods. Analogous to the semivariance values in a variogram, interpolation using copulas have a copula structure in a correlogram (plot of correlation with distance classes/lags). Some of the most commonly used copula families are as shown in Table 2-3.

Table 2-3: Some commonly used copula functions. (Nelsen, 2006; Li, 1999; Demarta & McNeil, 2005)

Copula family	Cumulative distribution function ($C(u, v)$)	Domain
Gaussian	$\phi_2(\phi^{-1}(u), \phi^{-1}(v), \rho)$	$-1 \leq \rho \leq 1$
Student's t	$\int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^2\lvert P\rvert}} \left(1 + \frac{x' P^{-1} x}{\nu}\right)^{-\frac{\nu+2}{2}} dx$	$\nu > 0$
Clayton	$\left[max\left(u^{-\theta} + v^{-\theta} - 1, 0\right)\right]^{-\frac{1}{\theta}}$	$\theta \in [-1, \infty) \backslash \{0\}$
Gumbel	$e^{-[(-\ln u)^\theta + (-\ln v)^\theta]^{\frac{1}{\theta}}}$	$\theta \in [1, \infty)$
Frank	$-\frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right)$	$\theta \in (-\infty, \infty) \backslash \{0\}$
Independence	$u \times v$	$\theta \in (-\infty, \infty)$

u and v are the random variables, θ is the probability mass/parameter value, ϕ_2 is the bivariate normal distribution function with correlation coefficient ρ, ϕ^{-1} is the inverse of a univariate normal distribution function, t_ν is the degree of freedom of ν, t_ν^{-1} denotes the inverse of Student t distribution, $\Gamma(\cdot)$ is the gamma function, P is the correlation matrix and x is the integral variable

The copula family is selected by using the maximum likelihood method. Prediction at unvisited location s_0 can be obtained by calculating the mean or median (Gräler, 2014a):

$$\hat{z}_{mean}(s_0) = \int_0^1 F^{-1}(u).\, c_{k+1}(u|F(x_1), \dots, F(x_k))du; \tag{2.17}$$

$$\hat{z}_{median}(s_0) = F^{-1}(C_{k+1}^{-1}(0.5|(x_1), \dots, F(x_k))); \tag{2.18}$$

In the above equations, $\hat{z}(s_0)$ represents the random variable value at location s_0, that follows the distribution $H(x_0|x_1, \dots, x_k)$ conditioned under the observed values of the k nearest neighbours $x_1, \dots, x_k$. F is the marginal cumulative distribution function, k denotes the nearest neighbours to the point at location s_0, c_{k+1} is the conditional density of the copula C_{k+1} given as:

$$c_{k+1}(u_0|u_1, \dots, u_k) = \frac{c_{k+1}(u_0, u_1, \dots, u_k)}{c_k(u_1, \dots, u_k)}; \tag{2.19}$$

Here, $u_i, i \in [0, k]$ represents the variables. The copula density reflects the strength of dependence of the variables.

2.5. Comparison of interpolation methods

Many comparative studies of different interpolation methods have been performed for different soil parameters such as pH (Liu et al., 2013) and OC (organic carbon) content (Piccini et al., 2014). Study for comparison of interpolation methods in complex terrain has also been performed (Yao et al., 2013). Studies have also been conducted for comparing methods in spatial-temporal context (Adhikary & Dash, 2017).

2.5.1. Measures of uncertainty

Cui et al., (1995) mentioned the following three statistics for measuring the uncertainty in predicted values – the mean error (ϵ), the mean squared error (MSE), and the variance of the reduced error (σ_{RE}^2). Additionally, the root mean squared error (RMSE) has also been mentioned. They are defined as follows:

$$\epsilon = \frac{1}{n}\sum_{i=1}^{n}(\hat{z}(s_i) - z(s_i)); \tag{2.20}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{z}(s_i) - z(s_i))^2; \tag{2.21}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{z}(s_i) - z(s_i))^2}; \tag{2.22}$$

$$\sigma_{RE}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{z}(s_i) - z(s_i))^2}{var(\hat{Z}(s_i) - Z(s_i))}; \tag{2.23}$$

In the above equations, n denotes the total number of observations, $\hat{z}(s_i)$ is the prediction and $z(s_i)$ is the observation at the i^{th} test point. For assessing the certainty of the value to the observed value, the mean error should be close to zero, the MSE and RMSE value should be small and the variance of the reduced error should be close to one.

In addition to that, the coefficient of determination R^2 value, that defines the amount of variance explained by the model is given as follows (Coster, n.d.):

$$R^2 = 1 - \frac{SSE}{SST}; \tag{2.24}$$

In the equation 2.24, SSE denotes the sum of squares of the residuals and SST denotes the total sum of square errors. Its value ranges from 0 to 1 with a value closer to 1 indicating that a large amount of variance is explained by the model.

3. STUDY AREA

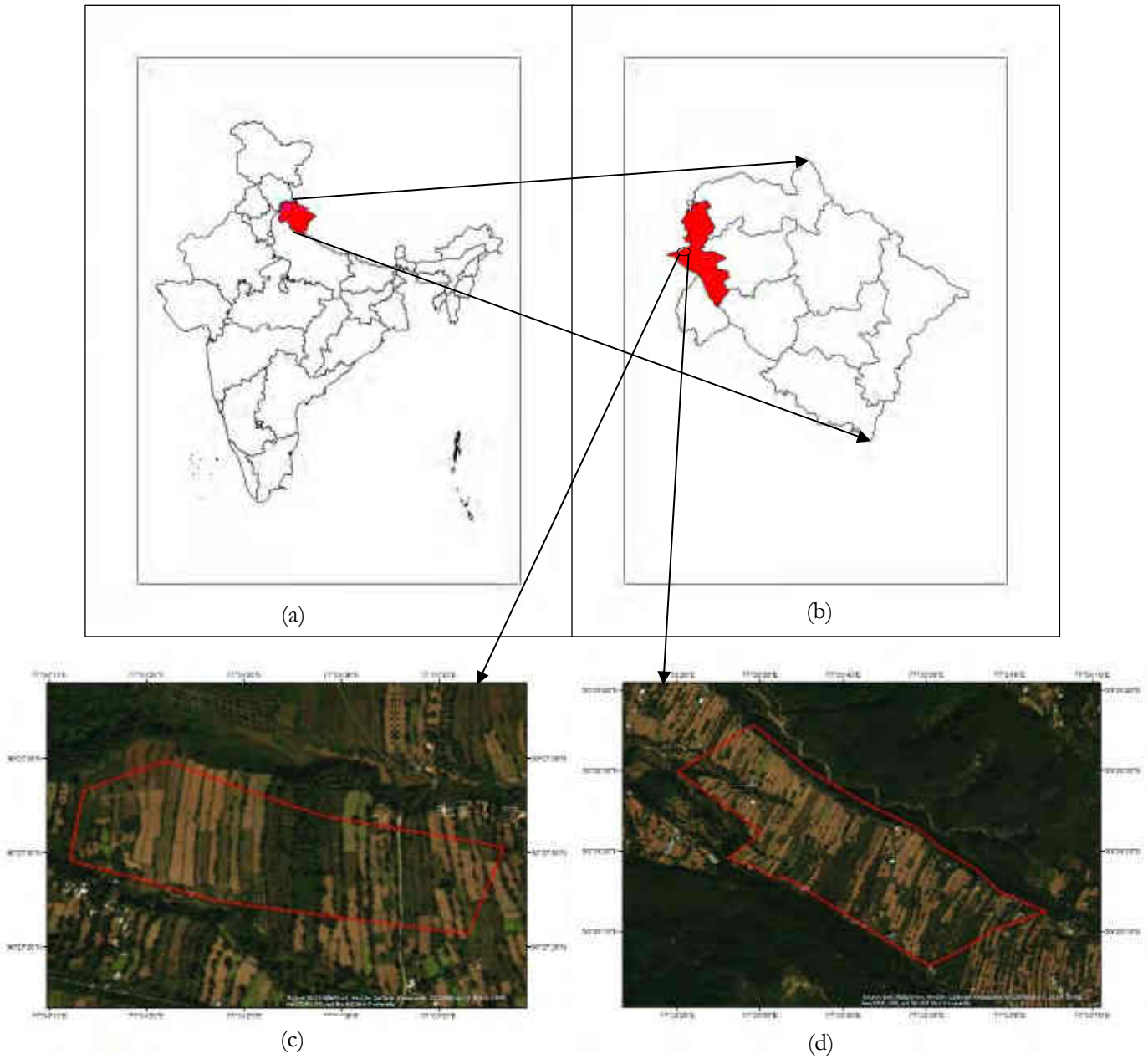


Figure 2-1: Study Area - (a) India; (b) Uttarakhand; (c) Langha – Tauli (in red boundary); (d) Barwa (in red boundary). Image Source – (a) and (b) Indian Institute of Remote Sensing, (c) and (d) Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN and the GIS User Community

The study area lies within the Sitlarao watershed area in the western part of Dehradun district of Uttarakhand. It belongs to the Asan river system which is a tributary of Yamuna river and covers an area of 8.05 km² . The climate of the area is humid sub-tropical with mean temperature ranging from 15 °C in winter to 35 °C in summer. The soil texture is predominantly sandy loam to loam (Kumar & Singh, 2016). Langha-Tauli lies in the north – western direction from Dehradun city at a distance of about 45 km, in the Dehradun district of Uttarakhand state of India. The first study area, as shown in Figure 3-1(c) covers an area of approximately 0.4 km². It lies between 30°28´5˝ N and 30°28´35˝ N latitude and, 77°53´20˝ E and 77°54´4˝ E longitude. Also, due to non – availability of any irrigation systems except at a few locations,

farmers were quoted as saying that their crops were dependant on rainfall. The terrain was observed to be consisting of many stones. This observation was made during the site visit for collecting soil samples. The elevation in the area varied from 668 m in the North-West direction, forming the downslope region, and 774 m in the South-East direction, the upslope region. Both the stated elevation values were above mean sea level.

Barwa lies in the South-East direction to Langha-Tauli. The chosen study area, as shown in Figure 3-1(d) had an area of approximately 0.1 km². It lies between 30°27′25.6″ N and 30°27′34.8″ N latitudes and, 77°53′20″ E and 77°54′38.3″ E longitudes. The elevation values varied from 798 m in the North-West direction to 868 m in the South-East direction.

The highest slope value observed in the Langha-Tauli was 40.52° whereas it was 37.74° in Barwa. A gradual decrease in elevation was observed from the South-East direction to North-West for both the study areas. Both were observed to be topographically flat wherever the land was used for agricultural purposes, with sudden rises and falls at regular intervals. This was a typical case of the farming style in hilly areas. Figure 3-2 (a) shows the slope of the first study area of Langha – Tauli, whereas Figure 3-2 (b) shows the gradient in Barwa.

(a) (b)

Figure 3-2: Slopes of (a) Langha-Tauli, and (b) Barwa

Figure 3-3: Collection of soil samples in the first study area - Langha-Tauli

4. METHODOLOGY

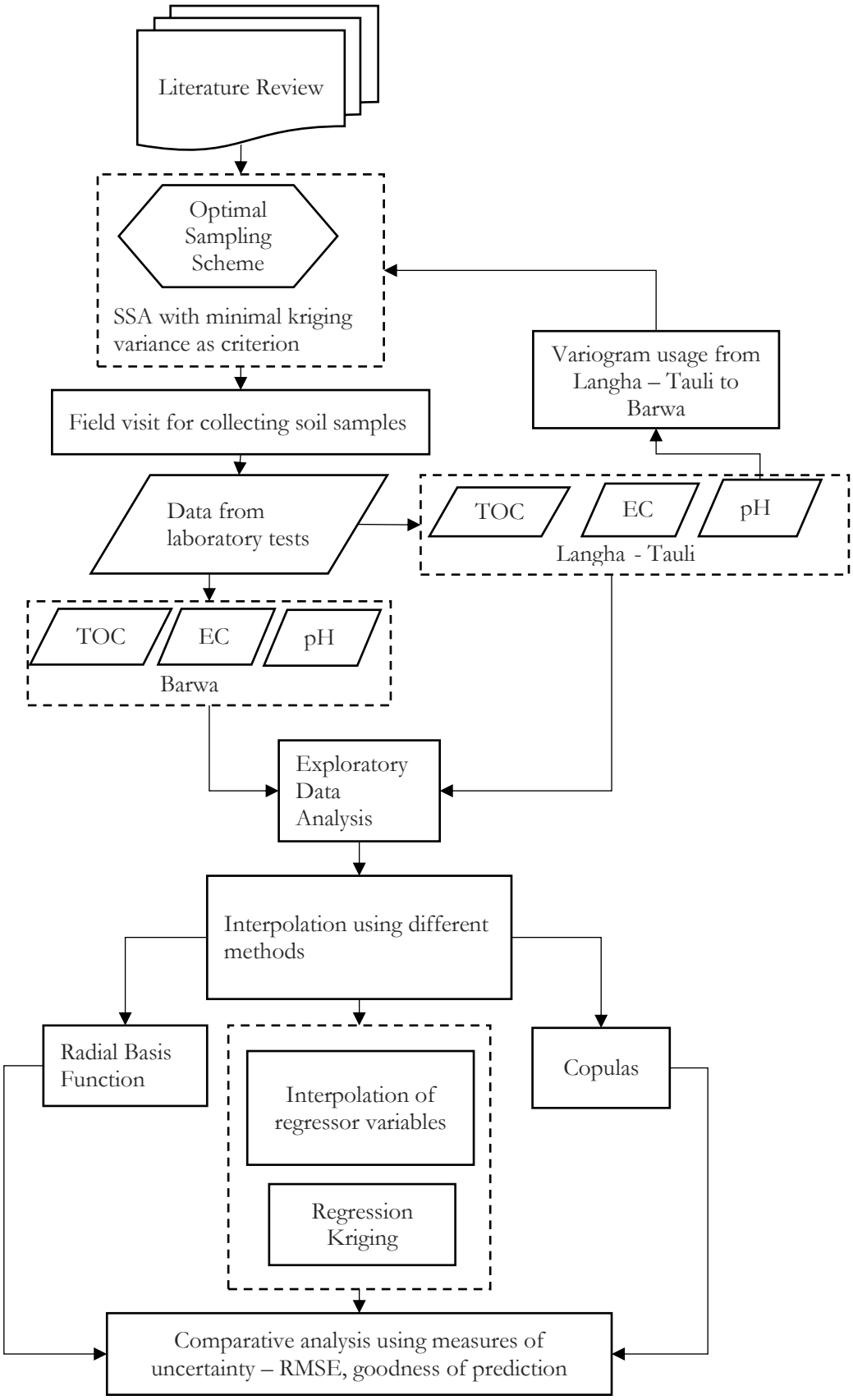


Figure 4-1: Methodological Flowchart

The overall research methodology is as given in Figure 4-1.

4.1. Data used

The data for the study was obtained through the analysis of collected soil samples in laboratory. The slope map was generated from DEM (Digital Elevation Model) from Cartosat – 1 satellite. The spatial resolution of the same was 10 m and the vertical accuracy is 8 m (Muralikrishnan et al., 2011). Also, the boundary of the study area was digitized by the researcher.

The selected parameters (pH, EC and TOC) were so chosen according to the suggestions by Jones (2016) and Arshad & Martin (2002). These were indicated as one of the primary indicators of soil health.

4.2. Sampling Strategy

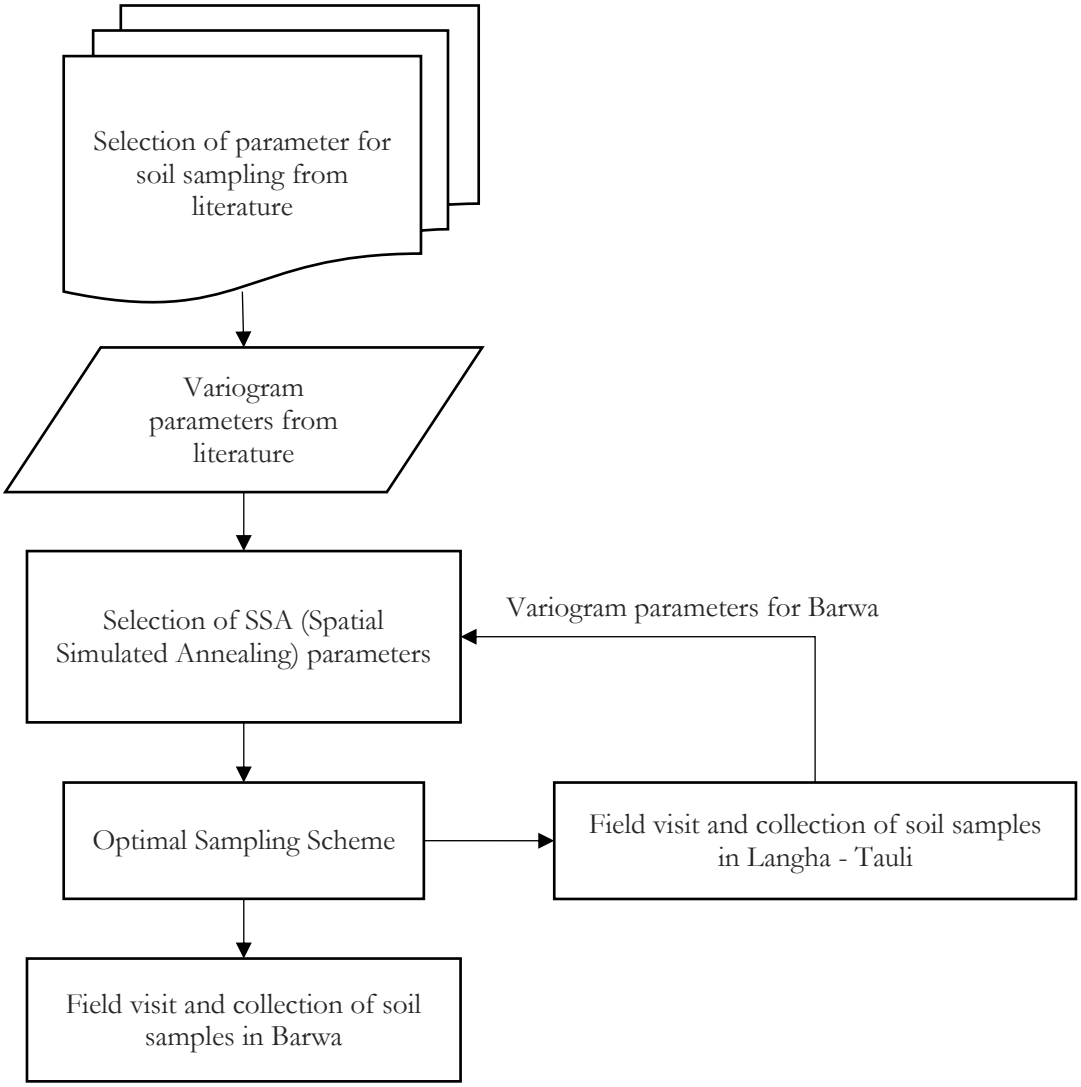


Figure 4-2: Sampling Strategy Flowchart

The sampling scheme was designed, and the soil samples were collected according to the steps in the flowchart as shown in Figure 4-2.

4.2.1. Selection of parameter for designing sampling scheme

pH was selected as the soil parameter whose variogram parameters (i.e. nugget, partial sill and range) were considered for designing the sampling scheme. This was because of its relationships with other parameters – EC and TOC.

EC involves a measure of the flow of ions in a solution, and pH is a measure of the hydrogen (H^+) or hydroxyl (OH^-) ions in a solution. Since H^+ ion is the most mobile cation (Moore, 1999) therefore, pH had a direct relationship with EC. According to Pietri & Brookes (2008) as the soil pH decreased, soil TOC generally decreased.

4.2.2. Selection of variogram parameters

Since no prior geostatistical information was present for the study area, it was necessary to pick the information from either a similar area or an averaged-out variogram parameter values. The latter was chosen for this study. McBratney & Pringle (1999) chose 19 different variograsm from journal articles for pH and plotted the average variogram of them. For calculating the mean of the variogram, fourth root transformation of the data was performed to bring them to normality. This was done because the variogram followed a chi-square distribution. The 4th roots minimize the Pearson's index of kurtosis for a chi-square variable (Goria, 1992). The average exponential variogram model was selected based on the AIC (Akaike Information Criterion) value. AIC describes the relative quality and depends on the number of estimated parameters and the maximum likelihood value of the models. These values were used for fitting the variogram in the first area, Langha – Tauli.

Initially, the variogram parameters were required for designing the sampling scheme in Barwa as explained in the following sections. Therefore, the variogram parameter values used for the second area, Barwa were based on the variogram fit of the pH data from Langha – Tauli. The selected variogram parameter values are as shown in Table 4-1.

Table 4-1: Model and model parameter values for Langha - Tauli (McBratney & Pringle, 1999) and Barwa

Area	Model selected	Nugget	Partial sill	Range (m)
Langha - Tauli	**Exponential**	0.0358	0.0841	62.073
Barwa	**Matérn**	0.1572	0.0951	47.808

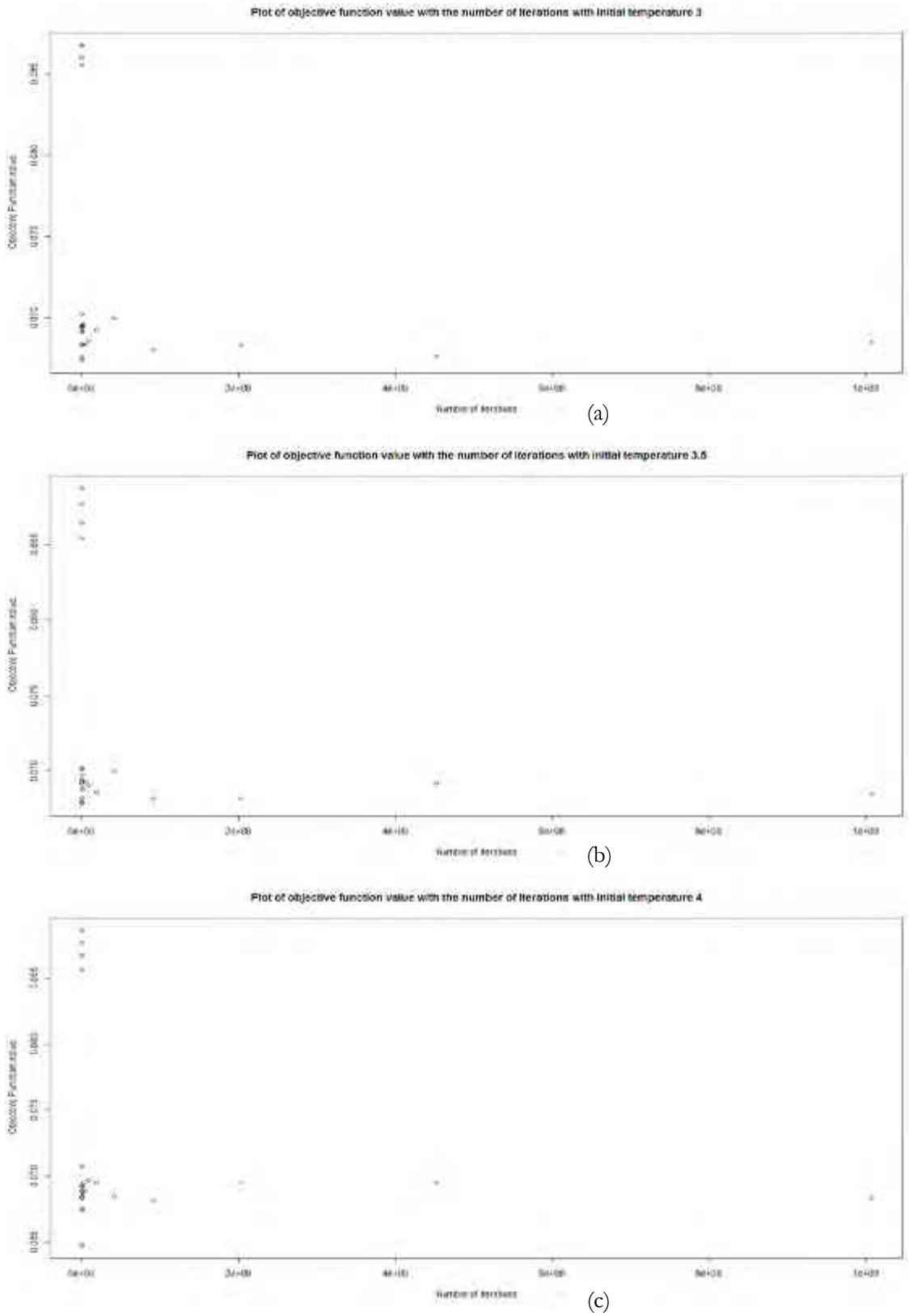
4.2.3. Selection of SSA parameters

The slope values of the study area were used as the covariate in RK for calculating the kriging variance (objective function value).

According to Tso & Mather (2001), the initial temperature for SSA was usually set to a value of 2 or 3. So, three different initial temperature values of 3, 3.5 and 4 and a different number of iterations were chosen to find out the sampling configuration with the least kriging variance/objective function value for Langha – Tauli and Barwa. Figure 4-3 (a-c) depicts the plots of objective function value with the number of iterations for Langha - Tauli. Also, the maximum distance that the sample point may move in the horizontal, as well as the vertical direction, was taken as half the area size (van Groenigen et al., 1999).

For the first area, Langha – Tauli, the initial temperature for SSA algorithm was chosen as 4 with the number of iterations being 550 as per the minimum objective function value of 0.06475. Similarly, the initial temperature for Barwa was selected as 3.5 with the number of iterations being 127 and a minimum objective function value of 0.23163. The objective function values against the number of iterations are as shown in Appendix A for both the study areas. Since the selected initial temperature was quite low, a

higher probability of acceptance as 0.95 was taken. This was because a low initial temperature value with a low initial acceptance value resulted in the algorithm to behave as greedy algorithm wherein local minima may be selected (Samuel-Rosa et al., 2017).



(a)

(b)

(c)

Figure 4-3: Graphical plots of objective function v. number of iterations for Langha-Tauli with initial temperature (a) 3, (b) 3.5 and (c) 4

4.2.4. Field visit and collection of soil samples

The field visit was done during the first week of November 2017 for Langha – Tauli, just after the summer crops (Kharif crops) had been harvested. Whereas, the same was done during the first week of January 2018 for Barwa. Saplings had already started growing in Barwa while the samples were collected. Firstly, the top thin layer of soil consisting of grass or stones was cleared. Then the soil was collected in plastic bags using an auger till 15 cm which formed the surface sample. The soil sample for the sub – surface layer (15 – 30 cm) was further collected.

Precautions were taken while collecting the soil samples. A sample point (number 66, in Langha – Tauli area) was falling inside the valley region. So, no soil sample collection was performed at that point due to inaccessibility. Also, if any sample point fell within the crop region, samples were collected from the vicinity so as to not disturb the crops growing in the field.

4.3. Performance of chemical tests

The soil samples were put to dry in open air for a week. They were further passed through a sieve of 2 mm aperture and the soil was stored in polypropylene (PP) containers. They were numbered according to the sites visited. The tests for pH and EC were conducted according to Singh et al. (n.d.) and Ghotekar (2016).

4.3.1. pH

The soil pH was determined through a soil – water suspension prepared in 1:2 ratio. The following procedure was followed: -

i. 20 g of soil sample was taken in a 100 mL beaker.
ii. 40 mL of distilled water was added to it, the solution was stirred well for about 3 minutes with a glass rod and kept still for half an hour.
iii. The solution was again stirred just before immersing the electrodes of the pH meter and the reading was noted.

4.3.2. Electrical Conductivity

i. After the pH reading was taken, the solution was kept aside for another half an hour until a clear supernatant liquid was obtained.
ii. The conductivity of the supernatant liquid was determined with the help of the conductivity meter. The unit of measurement was μS/cm.

4.3.3. Total Organic Carbon

i. The soil samples were made to pass through a sieve of 0.2 mm aperture.
ii. A small quantity (~ 30 – 50 mg) of soil sample was measured in 2 different ceramic boats.
iii. The first boat was kept in the combustion tube for TC (Total Carbon) measurement, whereas the second one was kept in the IC (Inorganic Carbon) combustion tube in the Solid Sample Module of the TOC analyzer.
iv. The TOC value was obtained as the difference of the values of TC and IC. It was measured as percentage content of the soil.

4.4. Right pixel size for interpolation

The optimal grid resolution according to the equations mentioned in Section 2.3 for Langha – Tauli and Barwa region were calculated as shown in Table 4-2. The recommended compromise in grid resolutions as

suggested by Hengl (2006) was considered for further processing. Considering the aspects of inspection density, the distance between sample points and the complexity of terrain for interpolation using RBF, RK and interpolation using copulas, a grid resolution of 16 m for Langha –Tauli was considered. It was 20 m for Barwa which was used in the case of BK.

For determining the minimum contour interval for the optimal pixel size, the legacy National Map Accuracy Standard (NMAS) of 1947 of the United States Geological Survey (USGS) was considered (ASPRS Map Accuracy Standards Working Group, 2015). It states that the minimum contour interval is twice the vertical accuracy of the DEM. Since, the vertical accuracy of CartoDEM (Cartosat – 1 DEM) was 8 m (Muralikrishnan et al., 2011), the minimum contour interval was taken as 16 m.

Table 4-2: Recommended pixel size for interpolation in Langha – Tauli.

Aspect	Recommended compromise
Langha-Tauli	
Inspection density	$0.0791 \cdot \sqrt{\frac{A_1}{N_1}} = 5.11\ m$
Distance between points	$0.25(0.5) \cdot \sqrt{\frac{A_1}{N_1}} = 8.06\ m$
Complexity of terrain	$\frac{A_1}{2 \cdot \sum l_1} = 34.08\ m$
Barwa	
Inspection density	$0.0791 \cdot \sqrt{\frac{A_2}{N_2}} = 9.45\ m$
Distance between points	$0.25(0.5) \cdot \sqrt{\frac{A_2}{N_2}} = 14.94\ m$
Complexity of terrain	$\frac{A_2}{2 \cdot \sum l_2} = 35.74\ m$

$A_1 = 0.4\ km^2 = 400000\ m^2, N_1 = 96,\ \sum l_1 = 0.58\ km = 5869\ m;\ A_2 = 0.1\ km^2 = 100000\ m^2, N_2 = 7,$
$\sum l_2 = 0.14\ km = 1399\ m$

4.5. Robust variogram estimation and fitting

A variogram in geostatistical methods describes the dependence structure of the random variable. For normal – like distributions which have heavier tails, a robust estimation of variogram had been discussed by Cressie & Hawkins (1980). Robustness against outliers and non – normal values had been considered. They concluded that the arithmetic mean of the fourth root of $(Z_{t+h} - Z_t)^2$ gave a robust estimate of the variogram. Here, Z_t is the value of the random variable Z at location t and h is the lag distance. As observed in Figure 4-4, the soil parameter values were found to be positively skewed and had slightly heavier than normal tails. Variogram parameters were estimated both by the conventional moment's method as well as the Cressie's robust variogram estimation. Also, different bin widths were experimented from 10 to 1000 with an increment by 5. The estimated variogram with the bin width having minimum sum of square error was selected for further analysis.

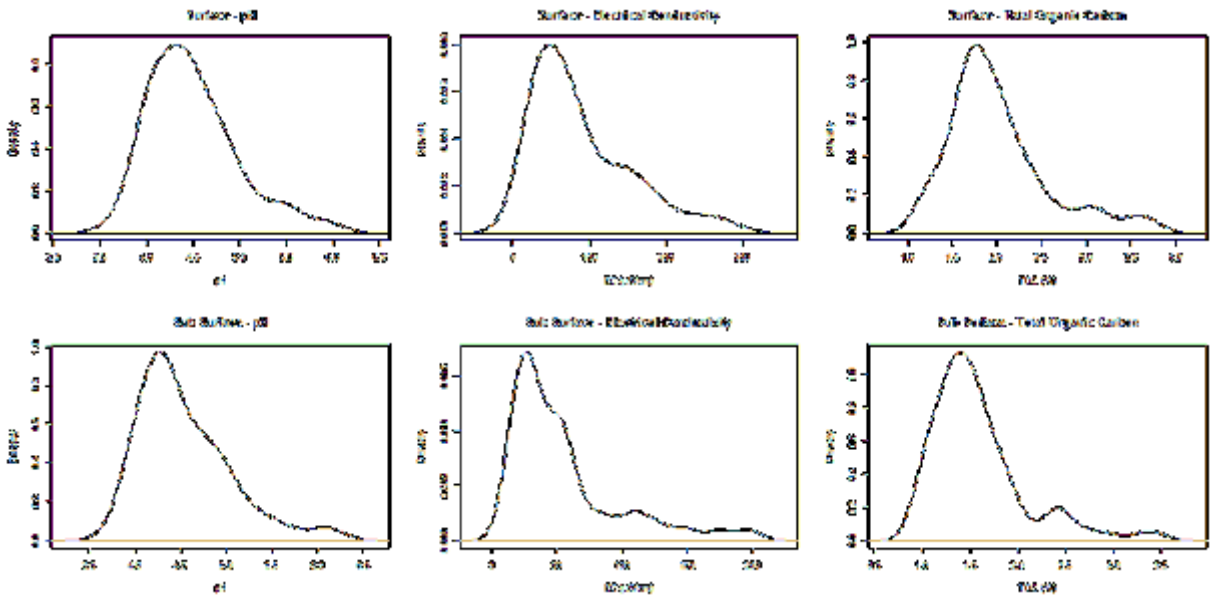


Figure 4-4: Density functions of soil parameters for Langha - Tauli area

For fitting the soil variogram models, the Matérn model was used (Minasny & McBratney, 2005). This was because the smoothness parameter in the Matérn function may be adjusted such that it represents different variogram models. It may be considered as a generalization of many theoretical variogram models. Minasny & McBratney (2005) observed in their study that the smoothness parameter within the range of 0.25 – 0.5 was considered to be rough (unsmooth) whereas, a value of 3 suggested a smooth process.

4.6. Using Bayesian kriging to extend spatial information from one area to another

The methodology developed by Cui et al. (1995) was followed for spatial information extension. The following steps were followed: -

1. Out of the 96 observations in Langha – Tauli area, 33 pseudo-random (since the samples were randomly picked using an algorithm, the term pseudo-random was used) observations were selected. Varying widths/lag distances were experimented and the one with the least sum of square error was selected for variogram estimation. After the variogram was fitted to the subset, the range and inverse of partial sill values were stored in a variable.
2. 59 iterations were run for the above step.
3. Different distributions functions were fitted to the range and inverse partial sill values. The best fit of the distribution function was assessed using the goodness of fit statistic – Anderson Darling statistic (Anderson & Darling, 1952) and the goodness of fit criteria – Akaike Information Criterion (AIC) (Akaike, 1974). The distribution with the least statistical value was considered to be the best fit. Since AIC measures the relative quality of models, Anderson Darling statistic, which gives the absolute quality was given preference.
4. As shown in Table 4-3, the best fit for the inverse sill was found to be the chi-square distribution and the log-normal distribution for range values.

Table 4-3: Goodness of fit statistic/criterion values for various distributions for range and inverse sill values

Goodness of fit statistic/criterion	Log-Normal	Gamma	Normal	Weibull	Exponential	Chi-Square
Inverse sill						
Anderson Darling statistic	5.332	2.789	4.424	2.812	2.754	1.803
AIC	336.939	308.850	356.713	309.959	308.490	303.366
Range						
Anderson Darling statistic	2.722	7.917	17.879	4.139	Infinite	-
AIC	789.664	846.139	1222.81	815.391	984.164	-

5. 'krige.bayes' function (Diggle & Ribeiro, 2002) from the 'geoR' package (Ribeiro & Diggle, 2016) of the R language (R Core Team, 2017) was used to perform Bayesian kriging. The surface level dataset of pH was considered for the interpolation process. Box-Cox transformation (Box & Cox, 1964) of data was performed with varying lambda values. The lambda value with the highest log likelihood value was selected for data transformation. For the specifications of the model control parameters, a Matérn covariance model and λ = -2.5 was selected. Scaled inverse chi-square distribution with degree of freedom value as 58 was set as the prior for sill (σ^2). Since, there was no provision for providing a log normal distribution as the prior for range (ϕ), an approximation to it was considered. The function allowed for describing a user defined discrete distribution. So, equally spaced numbers from 5 to 100 with an increment of 5 were defined as the support points and their corresponding probability values were stored. The log normal distribution function was defined with the meanlog (mean) value of 4.56 and sdlog (standard deviation) value of 1.97. The parameter values were estimated by maximum likelihood estimation. The probability values were scaled down by dividing each of them by the sum of all the probabilities such that the sum of probabilities was 1. This became the prior distribution for ϕ. Default value was accepted for the relative nugget value (τ^2/σ^2).
6. The obtained posterior distribution from Langha-Tauli was considered as the prior distribution for Barwa. A Matérn covariance model and λ = 4 was selected for transforming the data to normality in Barwa. BK was performed using the mentioned parameters.
7. The variogram was estimated, fitted and ordinary kriging (OK) performed for Barwa using the already collected sample data. ϵ, MSE and VRE were calculated as the measures of uncertainty.
8. The parameters for prior of Barwa were adjusted in such a way that the expectation of the distribution of partial sill values was the same as the partial sill value obtained while performing OK. Similarly, the parameters of prior for range values were adjusted such that their mean matched the range obtained from OK.
9. BK was again performed for Barwa with the modified prior distribution and the uncertainty measures were compared with the previous BK interpolation and OK.

4.7. Interpolation and Comparison

The following interpolation methods were chosen and the methodology adopted to obtain the final soil parameter surfaces.

4.7.1. RBF interpolation

RBF interpolation was performed using ArcMap 10.1 (ESRI, 2012). The 'Geostatistical Wizard' tool was used to find the optimal kernel function and the parameter values. It determines the parameter value by finding the value with the minimum root mean square prediction error (RMSPE). Table 7 shows the parameters selected for RBF interpolation for different soil variables for the surface as well as subsurface layers for Langha - Tauli.

Table 4-4: Kernel functions and parameter values for different soil parameters for Langha - Tauli

	Soil variables	**Kernel Function**	**Kernel Parameter**
Surface	pH	Spline with Tension	0.1865
	EC	Inverse Multiquadric	$1.1754 \cdot e^{-38}$
	TOC	Spline with Tension	0.2236
Sub - Surface	pH	Spline with Tension	0.2236
	EC	Inverse Multiquadric	$1.1754 \cdot e^{-38}$
	TOC	Inverse Multiquadric	12.3142

4.7.2. Regression kriging interpolation

Regression kriging was performed using the 'gstat' package (Pebesma, 2004; Gräler et al., 2016) in R language. The generation of the interpolated maps for the surface as well as subsurface layers was performed by the following steps. Here, the pixel size was considered as 16 m for Langha – Tauli as mentioned in Section 4.4. Variogram parameters were estimated according to the methodology mentioned in Section 4.5.

1. While performing RK, covariate values were required at prediction locations. OK, being a simple geostatistical model as compared to others was preferred for this operation. Similar pixel size as the one used for RK was considered.
2. Box-Cox transformation was performed for each of the soil variables to coerce them to normality. The λ (lambda) value was allowed to vary from -6 to 6 with a difference of 0.1. Log likelihood values were calculated in each of the iterations. The λ value with the maximum log likelihood value was selected for data transformation. Histograms and quantile – quantile (q-q) plots against the normal distribution were plotted to check the normality of the dataset. For linear geostatistical models such as kriging, normality becomes a requirement for the interpolation process to give the best results (Pebesma, 2006). Even the predictor variable was transformed to normality for checking the best possible linear model fits.
3. A linear model was then fitted to the target variable with its predictor variable. Different combinations of the target, as well as predictor variables, were considered as either the original or the transformed dataset. The corresponding R^2 (Coster, n.d.) and the adjusted R^2 (Frost, 2013) values were then compared. The combination of the target and the predictor variable with the highest statistical values were further selected for kriging.
4. Cross-validation of the target values was performed with the relevant combination between the target and predictor variables.
5. ϵ (Mean error), MSE, RMSE and R^2 values were calculated for each of the target variables for surface and sub – surface levels. These were used for comparing the better of the interpolation methods.

4.7.3. Interpolation using copulas

The R packages used for performing interpolation using copulas were – rgdal (Bivand et al., 2017), gstat, geoR, sp (Pebesma & Bivand, 2005; Bivand et al., 2008), RColorBrewer (Neuwirth, 2014), raster (Hijmans, 2017), VineCopula (Schepsmeier et al., 2018), spcopula (Gräler, 2014b), vines (Gonzalez-Fernandez & Soto, 2016) and fitdistrplus (Delignette-Muller & Dutang, 2015). In addition to the aforementioned R packages, functions for performing interpolation with covariates were also considered (Alidoost & Stein, 2016). The same pixel size was considered as for the cases of RK and BK. The prediction grid was then clipped to the study area polygon to constraint the interpolation process to the study area. A similar methodology was followed for the surface as well as subsurface level data. The following methodology was followed to implement the interpolation method without considering the covariates:

1. Different theoretical distribution functions: log-normal, gamma, normal, weibull and exponential were fitted to the target variable (interpolated variable). For identifying the distribution that the target variable followed, Cullen and Frey graphs (Cullen & Frey, 1999) were plotted. The distributions lying close to the observation were checked for fitness. The goodness of fit statistics (Anderson Darling statistic) and criteria (AIC) were obtained for each of the aforementioned distributions. As mentioned in Section 4.6, Anderson Darling statistic (AD) was given preference. The cumulative inverse distribution functions were used to transform the data into a uniform distribution. The quantiles of the data were obtained using the inverse cumulative distribution function. The goodness of fit statistics for the variables at surface and subsurface levels are as shown in Table 4-5.

Table 4-5: Goodness of fit statistics/criteria for the target variable

Goodness of fit statistic/criteria		**Log-normal**	**Gamma**	**Normal**	**Weibull**	**Exponential**
Surface						
pH	AD	0.618	0.761	1.096	2.516	-
	AIC	100.734	102.372	106.379	122.991	-
EC	AD	0.752	0.828	2.542	0.864	-
	AIC	787.197	788.696	817.968	790.680	-
TOC	AD	0.609	0.984	2.062	2.328	-
	AIC	117.121	120.737	132.524	135.666	-
Sub – Surface						
pH	AD	1.430	1.614	2.024	3.642	-
	AIC	99.713	102.279	108.124	129.356	-
EC	AD	0.707	1.595	4.791	1.930	4.286
	AIC	687.256	697.844	744.882	704.382	716.582
TOC	AD	1.0468	1.582	2.991	3.539	18.887
	AIC	70.405	76.626	93.911	98.716	208.621

2. As shown in Table 4-5, log-normal distribution was selected as the marginal distribution for all the soil parameters. This was in concordance with the least Anderson Darling statistic and AIC value among different distributions.
3. Spatial bins were calculated for different lag distance classes with their respective correlation measures. Kendall's Tau (Kendall, 1938) was used as the correlation method for this case. This was because of the need for correlation coefficient being independent of the marginal

distributions (Li, 1999). The number of bins was fixed such that the number of point pairs was approximately 100 or greater than that. Also, a decreasing trend in the correlation values with distance was preferred. Low correlation value with each lag distance was the reason behind less number of bins. Thus, a correlogram was generated for the soil parameters.

4. The next step involved fitting a correlation function to the correlogram. In this case, a cut-off bin was set such that the correlation value showed an increase in value with an increasing distance class/bin. This was done because correlation value generally decreases with distance and a sudden increase would indicate deviation from this rule. Also, the degree of function was set as either 1 or 2 depending on the cut-off bin. As less number of curve parameters needed to be estimated, a lower degree function was selected.
5. Spatial copulas were then fitted to the correlogram at corresponding bins. For this, the log likelihood values for given copula families was calculated for each lag/distance class. The families of copulas considered for fitting were – normal copula, t – copula, frank copula, clayton copula, gumbel copula and an independence copula (Table 2-3, Section 2.4.3). The copula family with the highest log likelihood value was assigned to the corresponding bin until the cut-off bin. The independence copula was assigned to the remaining bins. Gumbel and Clayton copulas are used only for bins with positive correlation values. Therefore, their values were coerced to zero if found to be negative. The spatial copulas were thus constructed.
6. The non – spatial dependence structure of the random variables was considered subsequently. C – vine structures that considered the conditional dependence structure was created (Bedford & Cooke, 2002). For fitting the vine structure, a local neighbourhood of size 8 around the sample point was defined. The conditional copula density was calculated. Copulas in the vine structure were fitted to the data and the pair – copula families were selected based on AIC value. The parameters for vines were estimated by maximum likelihood method.
7. Finally, the spatial copula and the vine copula structures were joined together into one superclass. The neighbourhood for prediction locations was defined and the interpolation process for the target variable was performed.

For executing the interpolation process considering the covariates i.e. the other two soil parameters than the one being predicted, the following methodology was adhered to:

1. For obtaining the covariate values at prediction locations, similar methodology as in Section 4.7.2 was followed i.e. using OK.
2. The rank transformation was performed for the combined dataset of the sampled and prediction locations for each of the variables. The ranks of sampled and prediction locations were then further separated to their respective variables.
3. An appropriate copula family was selected for the combinations of the ranks of the target variable with each of the covariates. The selection criteria were AIC and the parameters were estimated using the maximum likelihood method.
4. For fitting the non – spatial vine structures, the covariate ranks that were stored at the prediction locations were added to the existing neighbourhood. The existing neighbourhood was created while performing interpolation without covariates. This was defined for the case concerning the sample locations. The conditional density for the covariates' copula families was calculated. These were then appended to the conditional density calculated for the target variable. The copulas were then fitted to the vine structure and the parameter values were estimated using the maximum likelihood method.
5. The covariate copulas, spatial copulas and the vine copulas were then joined together into one superclass. A prediction neighbourhood for covariates was created and appended to the existing

prediction neighbourhood. Lastly, the interpolation process using covariates was executed for the target variable.

4.7.4. Comparison of interpolation methods

The measures of uncertainty as mentioned in Section 2.5.1 were calculated by performing leave one out cross-validation (LOOCV) of the full dataset. The obtained values were then compared and assessed accordingly.

5. RESULTS

5.1. Optimal sampling scheme

The 'spsann' package (Samuel-Rosa et al., 2017) of R language was used to perform the optimization process. Figure 5-1 displays the obtained optimized sampling schemes for Langha – Tauli and Barwa. In total, 96 sampling points were present in Langha – Tauli and 7 points in Barwa.

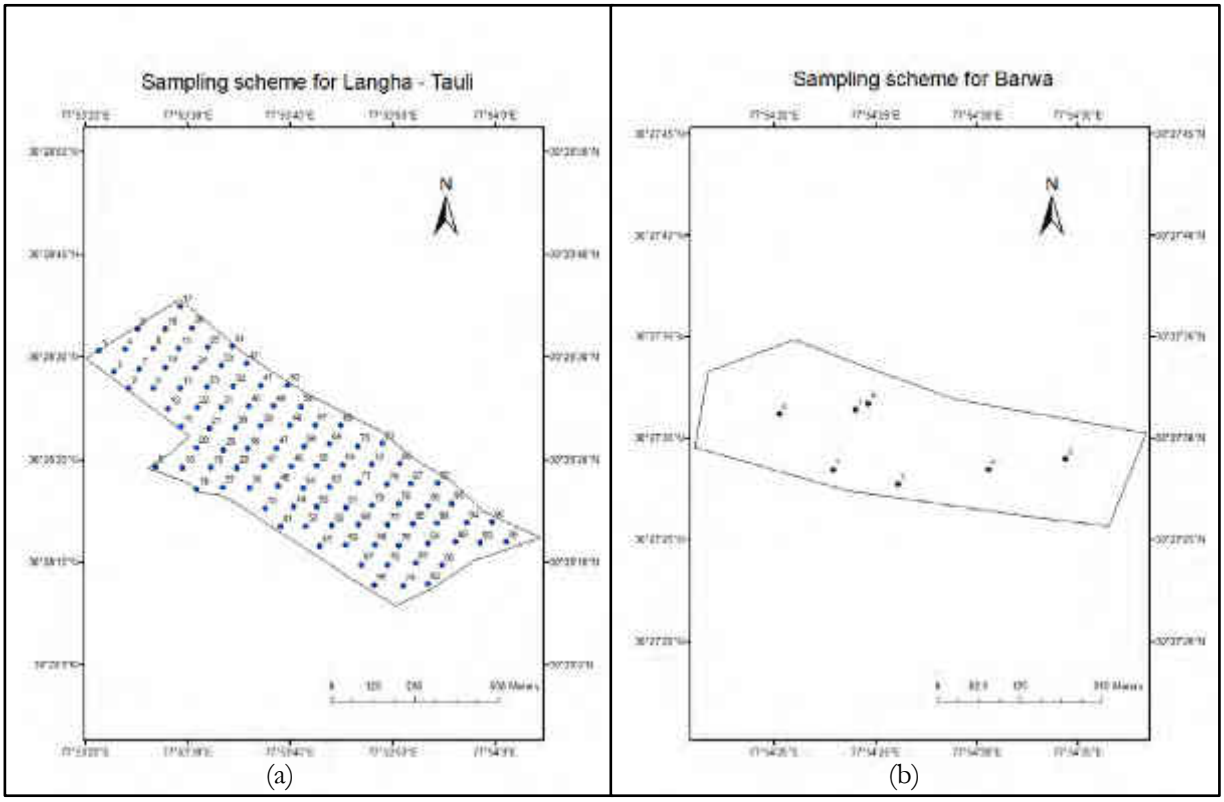


Figure 5-1: Optimal sampling schemes for (a) Langha - Tauli and (b) Barwa

5.2. Descriptive statistics

It can be observed from Table 5-1 and Figure 4-4 that the density curves of all the variables were slightly positively skewed with the skewness coefficients ranging from 0.9194 for surface level pH data to 1.6251 for sub – surface level EC data. It also indicated a deviation from normality since a skewness coefficient close to zero denotes a normal distribution. The data was found to be rather homogeneous with the highest variation in sub surface level TOC data having a mean value of 1.58 % and ranging from 0.96 % to 3.44 % (the range of values varying approximately from twice the standard deviation away from the mean at the lower limit to four times the standard deviation away from mean at the upper limit). For the other soil variables, the range of variation was approximately four times the standard deviation. Also, the sub surface level data had more variation as compared to the surface level data.

The descriptive statistics for Barwa are as shown in Table 5-2. The density curves in Barwa were positively skewed for the soil variable EC and negatively skewed for pH and TOC at the surface and sub surface levels. The variable pH at sub surface level was found to be the most skewed with a skewness coefficient

of -0.9674. The least skewed or the variable closest to normality was TOC at sub surface level with a skewness coefficient of -0.0659. The variables were observed to be similarly varying across their respective mean values with respect to their standard deviations.

Table 5-1: Descriptive statistics of the dataset (96 sample points) in Langha - Tauli

	Range of values	**Mean**	**Standard Deviation**	**Skewness Coefficient**
Surface				
pH	3.627 – 5.966	4.4944	0.4754	0.9194
EC (µS/cm)	17.02 – 287.70	88.9166	64.1478	1.1498
TOC (%)	1.151 – 3.752	2.0075	0.5601	1.1574
Sub Surface				
pH	3.726 – 6.153	4.5406	0.4972	1.1176
EC (µS/cm)	13.04 – 198	54.6914	42.5808	1.6251
TOC (%)	0.956 – 3.446	1.5801	0.4902	1.5534

Table 5-2: Descriptive statistics of the dataset (7 sample points) in Barwa

	Range of values	**Mean**	**Standard Deviation**	**Skewness Coefficient**
Surface				
pH	3.661 – 4.803	4.3126	0.3534	-0.4951
EC (µS/cm)	25.40 – 67.79	39.2228	16.2949	0.5825
TOC (%)	1.8235 – 2.6213	2.3119	0.3089	-0.3852
Sub Surface				
pH	3.666 - 4.573	4.257	0.2932	-0.9674
EC (µS/cm)	20.94 - 104.20	47.9728	34.0284	0.6377
TOC (%)	1.7476 - 2.4694	2.1112	0.2747	-0.0659

5.3. Using spatial information from one area to another - Bayesian kriging implementation

The interpolated and variance maps for pH in Langha – Tauli are as shown in Figure 5-2 (a-b). The predicted pH values ranged from 3.74 to 5.73 whereas the variance spanned from 0 to 0.26. Higher prediction values have been displayed in a lighter shade of red as the pH value close to 7 is considered neutral and lower values indicated an acidic nature. Areas of higher acidic nature were found in pockets of areas. These were mainly concentrated in the North-West direction. Although the north-central part of the study area was observed to have higher pH values, the relatively higher variance was present at that location. In a broader perspective, the low variance was observed after performing BK. The sampling locations had a lower variance value.

While considering the posterior distribution of Langha – Tauli as the prior distribution in Barwa, similar maps were generated for the second study area. These maps are as shown in Figure 5-3 (a-b). The predicted pH values ranged from 3.21 to 4.71. The variance was not very high in the second study area with values ranging from 0 to 0.12. Lower pH values and relatively higher variance was observed in the western part of Barwa. The maps, after updating the prior are presented in Figure 5-4 (a-b). The range of predicted values reduced after updating the prior. The pH value varied from 4.22 to 4.32 with a relatively higher upper variance value of 0.16 than the previous case. The lower variance value was 0.14.

The uncertainty measurements as shown in Section 2.5.1 for Langha – Tauli and Barwa using OK and BK are as provided in Table 5-3. The ϵ and MSE values were lesser in case of interpolation by OK as compared to BK. The σ^2_{RE} value was also much closer to 1 for OK . OK initially performed better than BK, with the mean error of 0.0096, the mean squared error of 0.1519 and the residual variance of 1.4076 against the respective values of 0.4078, 0.3064 and 10.4199 using BK for Barwa. The degree of freedom (df), after updating the chi-square distribution was 1 for the partial sill values. Whereas, the parameter values were 0.007 as the mean and 0.016 as the standard deviation of the log-normal distribution for range. The mean error ϵ, MSE and σ^2_{RE} with the updated priors for BK were better than OK.

Table 5-3: Values for mean error ϵ, the mean squared error MSE and the residual variance σ^2_{RE} with ordinary and Bayesian kriging for different parameter values. Here L(a,b) is the log-normal distribution with parameter a as meanlog value and b as the standard deviation of log value. Area 1 = Langha-Tauli, Area 2 = Barwa

Method of interpolation	σ^{-2}	ϕ	ϵ	MSE	σ^2_{RE}
Ordinary (Area 1)	6.58	220	0.0019	0.2035	1.1641
Bayesian (Area 1)	χ^2_{58}	L(4.56,1.97)	0.0336	0.2045	1.9860
Ordinary (Area 2)	0.103	11.3	0.0096	0.1519	1.4076
Bayesian (Area 2)	χ^2_{153}	L(-6.46,4.24)	0.4078	0.3064	10.4199
Bayesian (Area 2, after updating prior)	χ^2_1	L(0.007,0.016)	0.1466	0.0772	0.7306

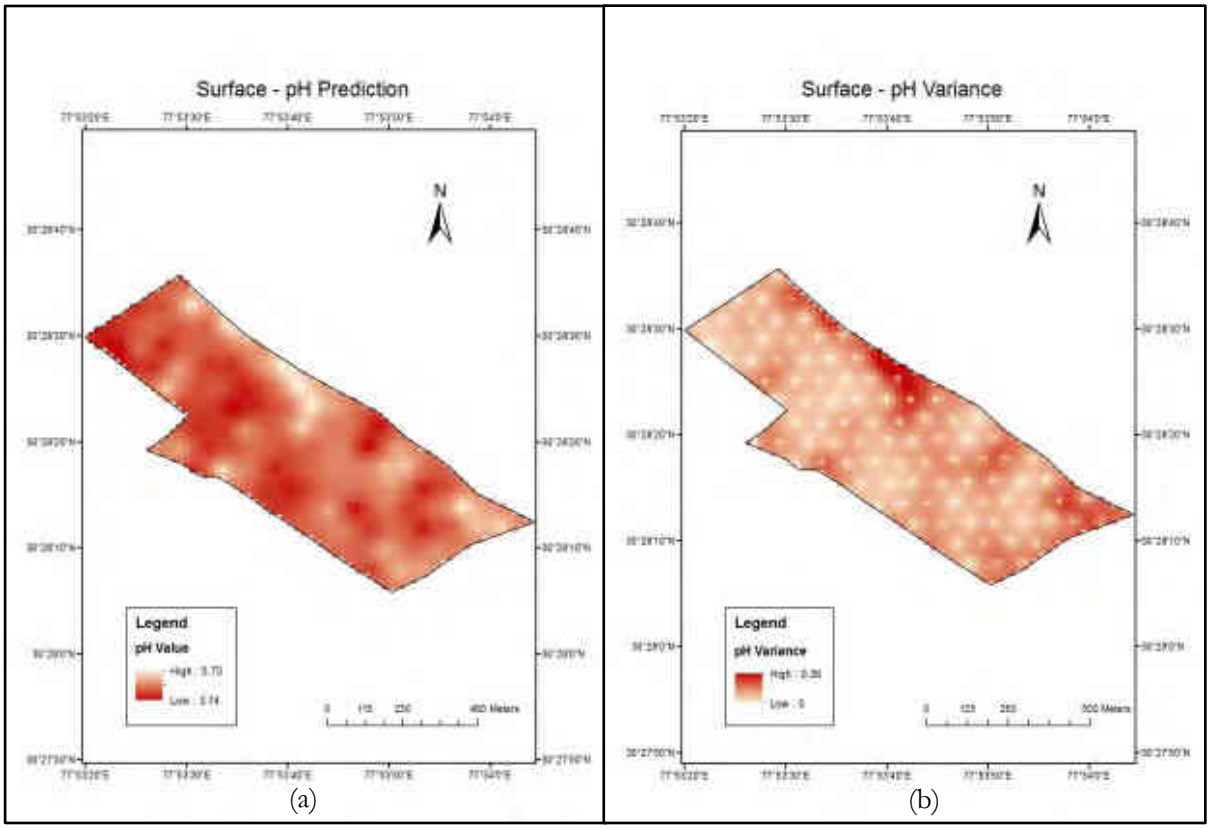
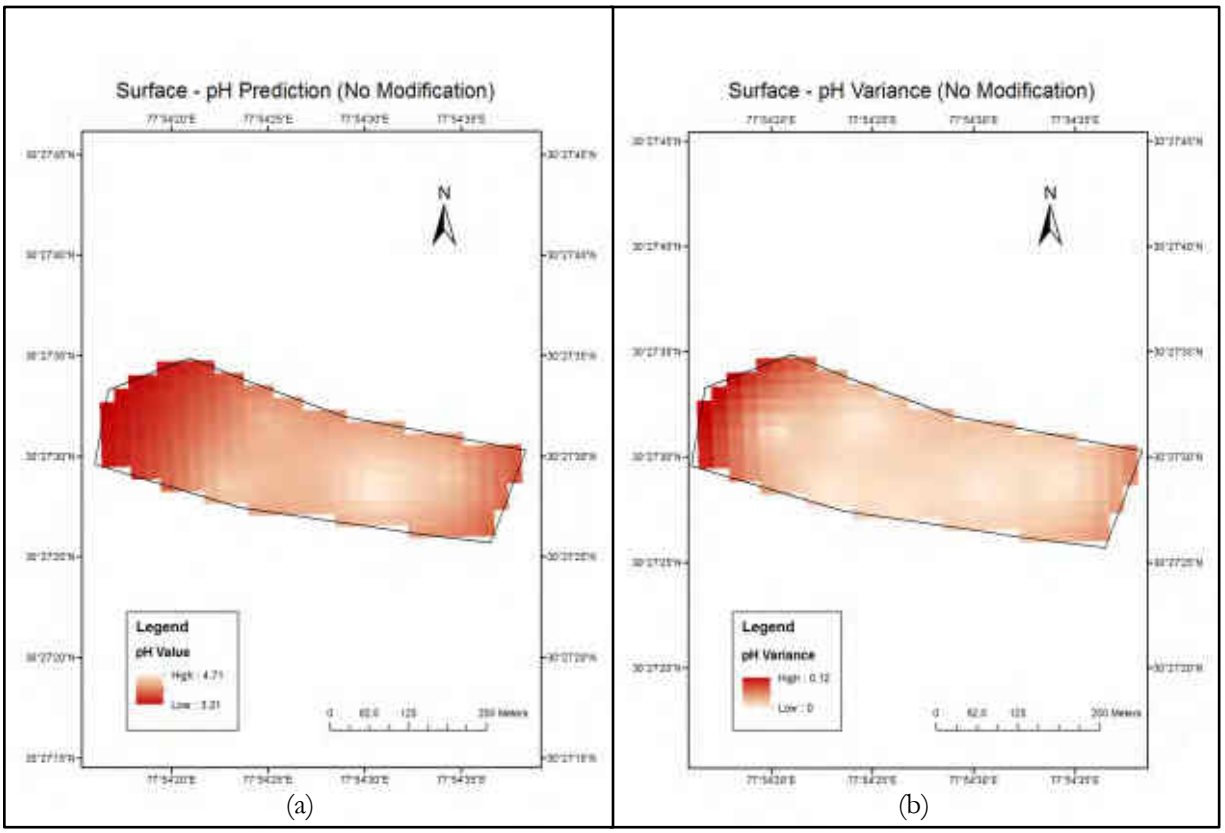
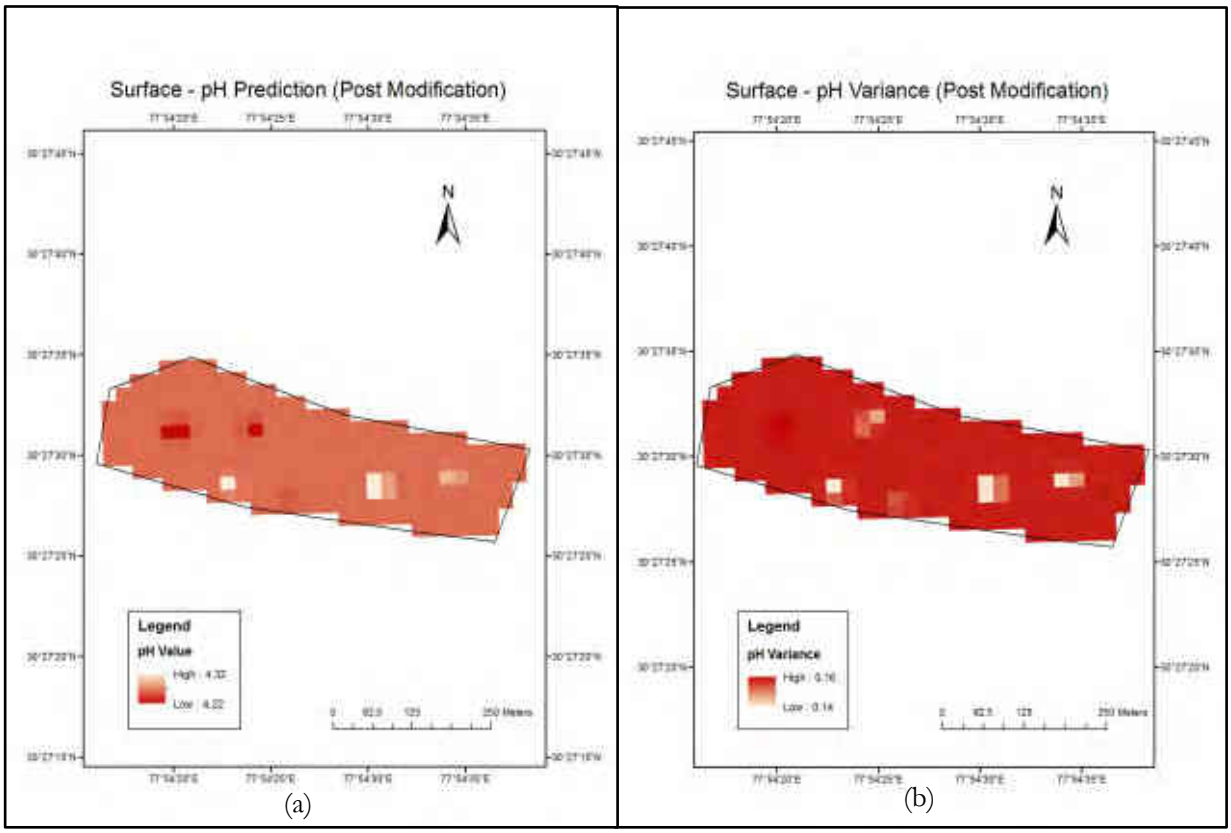


Figure 5-2: Surface level (a) interpolation and (b) variance map, for pH in Langha-Tauli using Bayesian kriging. The pixel size is 16 m and the projected coordinate system is Universal Transverse Mercator (UTM) 44 N.



(a) (b)

Figure 5-3: Surface level (a) interpolation and (b) variance map, for pH in Barwa using Bayesian kriging. The pixel size is 20 m and the projected coordinate system is Universal Transverse Mercator (UTM) 44 N.



(a) (b)

Figure 5-4: Surface level (a) interpolation and (b) variance map, for pH in Barwa using Bayesian kriging post updating the prior. The pixel size is 20 m and the projected coordinate system is Universal Transverse Mercator (UTM) 44 N.

5.4. Interpolation maps

All the interpolated maps had their values displayed in shades of red with the lower value being shown by the lighter shade and the higher value by the darker shade. This was excluding the maps of pH where the lighter shade meant a higher pH value and the darker shade, lower pH value.

5.4.1. RBF

The interpolation maps for surface and sub-surface level for different soil parameters in Langha – Tauli are displayed in Figure 5-5 (a-f).

The predicted values for pH ranged from 3.71 to 5.80 at surface level and between 3.81 and 5.87 at sub surface level. Lower values of pH were observed in the majority portion of the study area, particularly in the North – West direction of Langha – Tauli. This observation was similar for the surface as well as the sub surface level pH prediction map. Areas of high values were mainly observed in the north-central part except for a few pockets in the whole region.

The spatial variation for EC was almost similar for the surface and the sub surface level map. The range was between 37.69 μS/cm and 140.73 μS/cm for surface level EC data, whereas it was between 22.28 μS/cm and 83.58 μS/cm for sub surface level data. The change in values didn't appear to be gradual and a sudden change in values was observed. Lower values in their respective ranges for surface and sub surface level maps were noted in the south-central part of the area. The higher values were concentrated at the edges, particularly in the northern and northern – central part of the map.

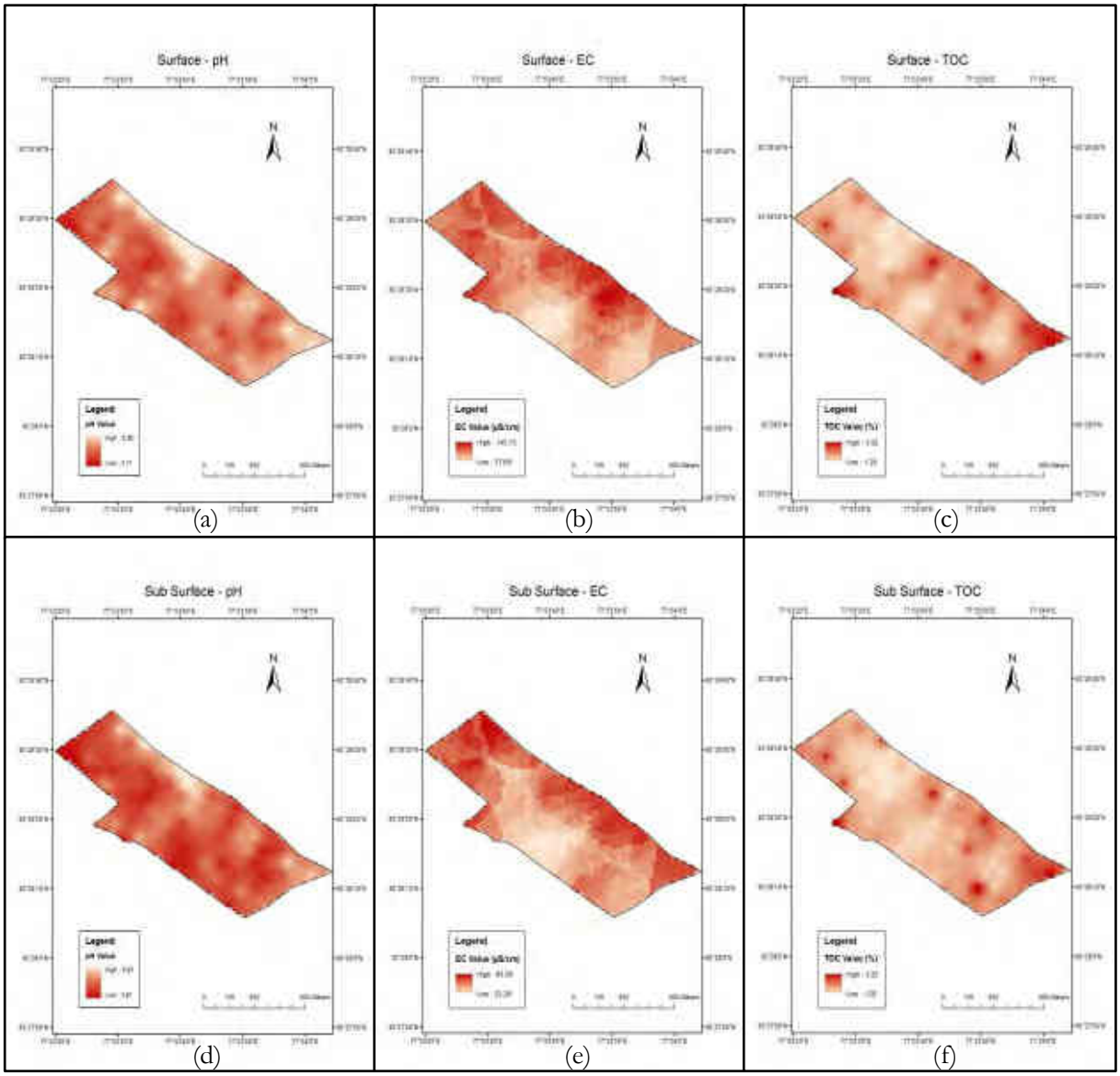


Figure 5-5: Surface (a-c) and sub-surface (d-f) level interpolation maps for pH, EC and TOC respectively in Langha – Tauli using RBF as interpolator. The pixel size is 16 m and the projected coordinate system is UTM 44 N.

The TOC value varied from 1.24 % to 3.52 % at surface level and from 1.02 % to 3.25 % at sub surface level. They were higher in the south – eastern part i.e. the upslope of the study area. Small portions of high concentration were observed for the surface and the sub surface level.

5.4.2. Regression kriging

The interpolation and variance maps for surface and sub-surface level for different soil parameters in Langha – Tauli are as shown in Figure 5-6 (a-f) and Figure 5-7 (a-f) respectively. The spatial variation for all the soil variables was gradual. No sudden changes were visible in the interpolated maps. Also, the variance did not change much across space and when it changed, the difference was quite small. The higher variance value was present at the edges of the study area.

The pH value varied from 4.06 to 4.73 for the surface and from 4.09 to 4.91 at the sub surface level. Lower values were present in the north – western corner of the study area for the surface as well as sub surface level. The variance was same across the area except at the edges of the surface level map. Although

the value in the legend of the map states the low and high value as 1, very small variations in the value may be there. The variance for the map at the sub surface level was same everywhere.
The predicted EC values varied from 48.89 µS/cm to 95.38 µS/cm and 32.36 µS/cm to 90.70 µS/cm at the surface and sub surface levels respectively. High EC value was present in the north – eastern part of the area. The variance value at the surface level ranged from 1.351 µS/cm to 1.404 µS/cm. It ranged from 628.319 µS/cm to 770.762 µS/cm at the sub surface level. The pattern of spatial variation was quite dissimilar for the two strata for EC.

The surface level TOC value had a range from 1.60 % to 2.76 %. Whereas it was from 1.15 % to 2.32 % at the sub surface level. Higher TOC value was observed at the upslope portion i.e. the eastern edge of the study area. Little difference in variance values was there. It varied from 0.146 % to 0.182 % at the surface and from 1.014 % to 1.017 % at the sub surface level. Higher variance value was observed at the edges.

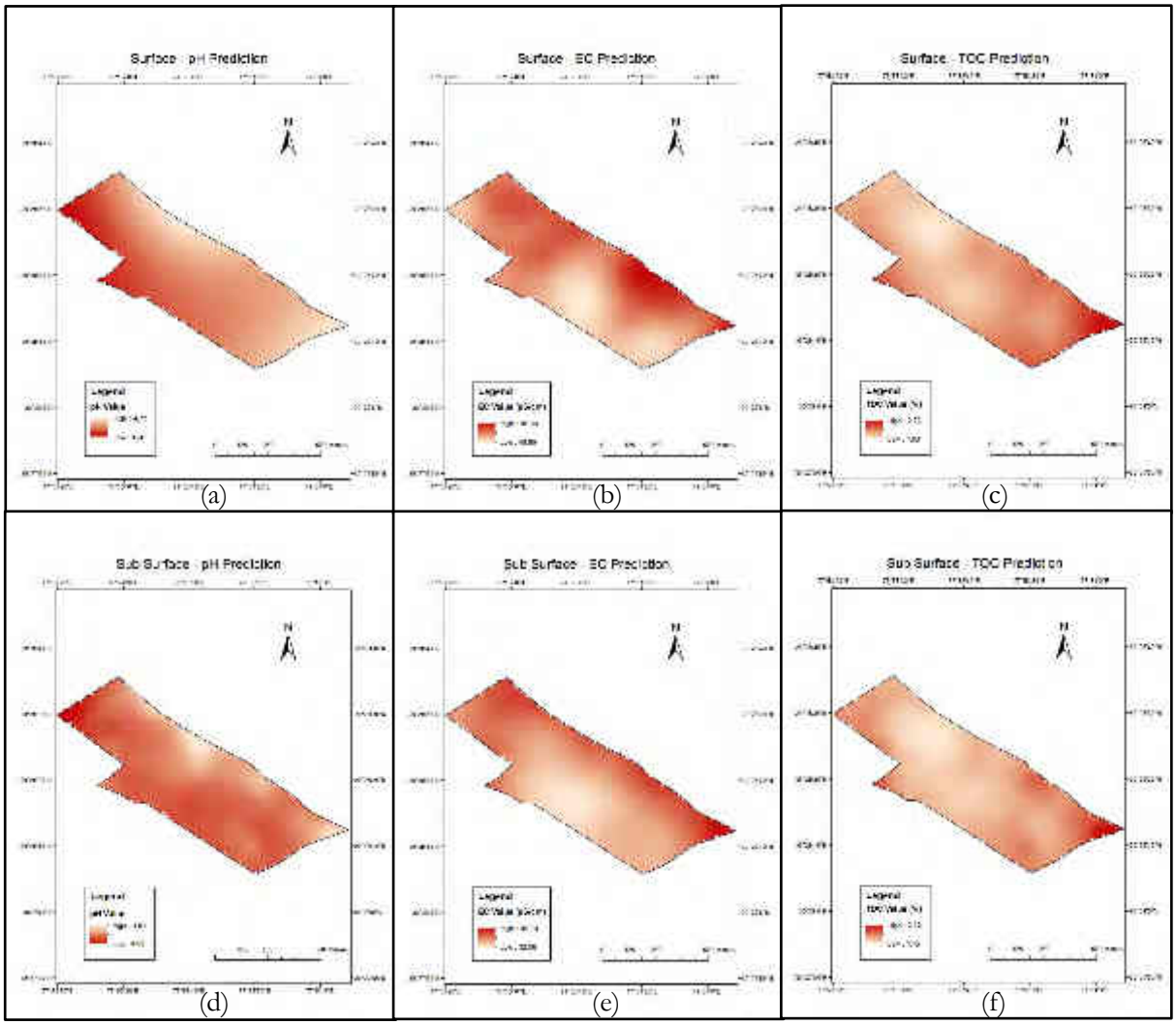


Figure 5-6: Surface (a-c) and sub-surface (d-f) level interpolation maps for pH, EC and TOC respectively in Langha – Tauli using RK. The pixel size is 16 m and the projected coordinate system is UTM 44 N.

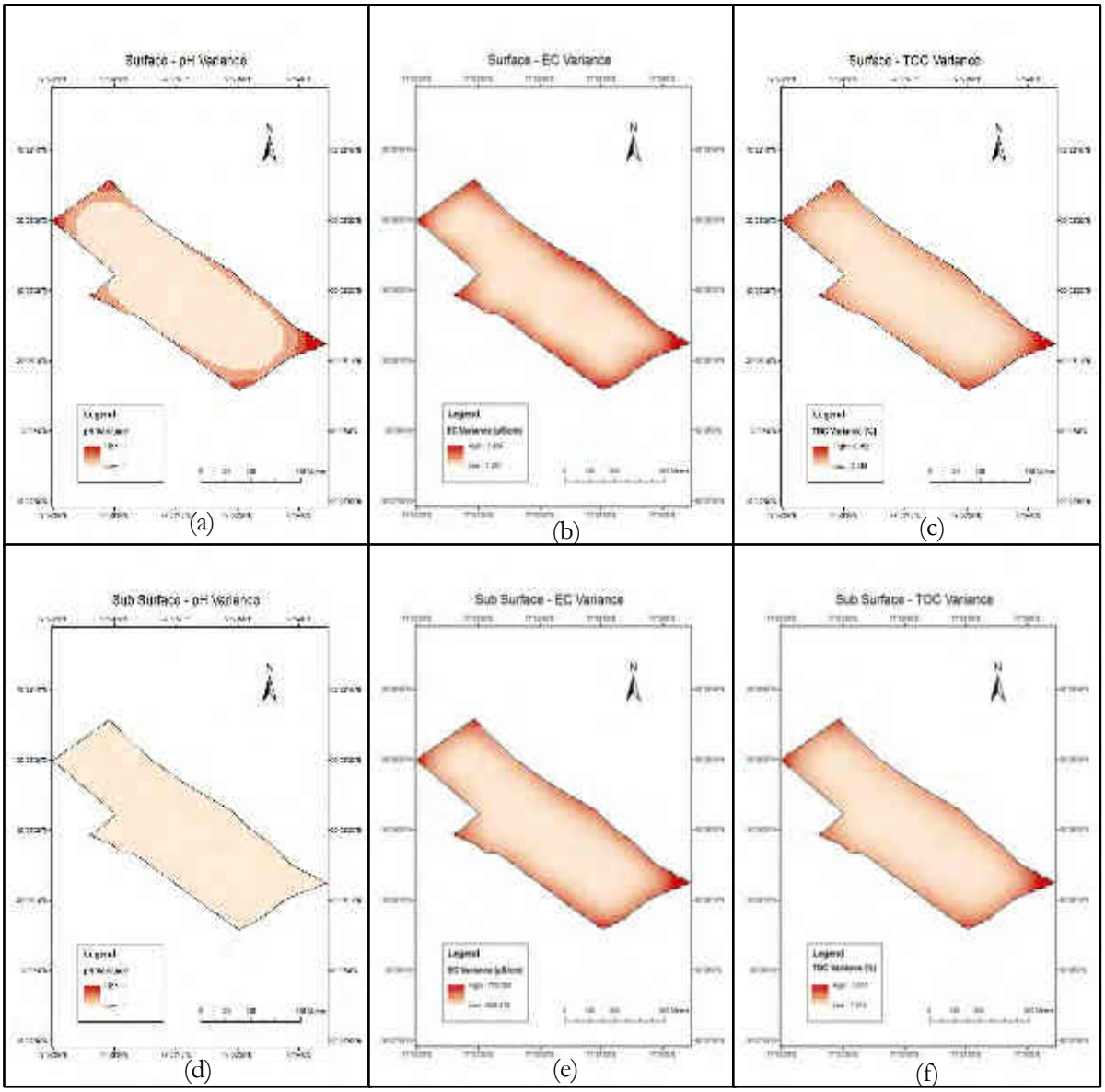


Figure 5-7: Surface (a-c) and sub-surface (d-f) level variance maps for pH, EC and TOC respectively in Langha – Tauli using RK. The pixel size is 16 m and the projected coordinate system is UTM 44 N.

5.4.3. Interpolation using copulas

The interpolation maps for surface and sub-surface level for different soil parameters in Langha – Tauli are as shown in Figure 5-8 (a-f) and Figure 5-9 (a-f) respectively. Maps, generated with copulas as interpolators, without using covariates are shown in Figure 5-8 whereas Figure 5-9 shows the maps when covariates had been used. The covariates comprised of soil parameters apart from the target variable being interpolated.

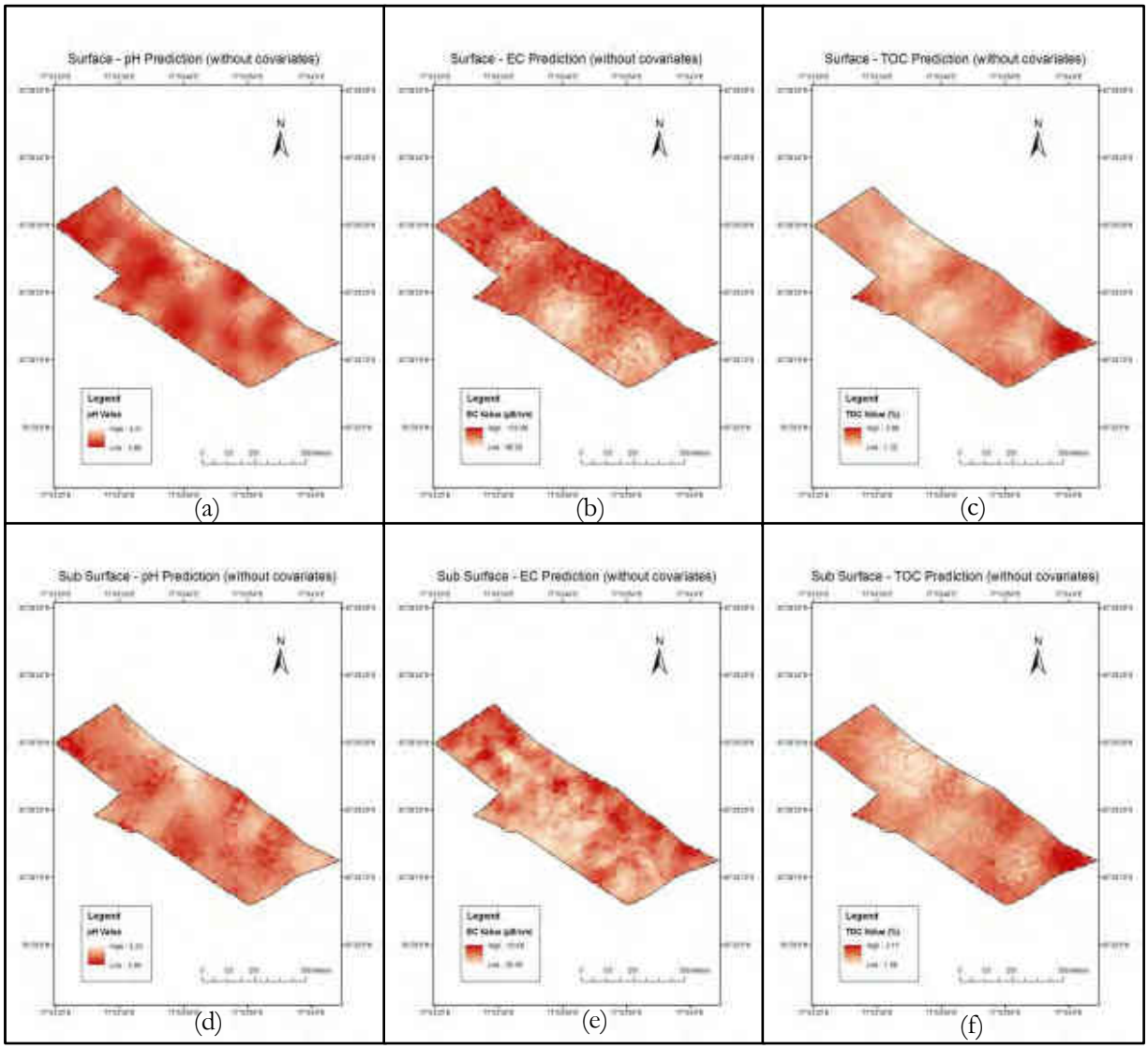


Figure 5-8: Surface (a-c) and sub-surface (d-f) level interpolation maps for pH, EC and TOC respectively in Langha – Tauli using copulas as interpolators. In this case, only the variable to be interpolated had been used. The pixel size is 16 m and the projected coordinate system is UTM 44 N.

For the case of interpolation using copulas without covariates as shown in Figure 5-8, the range of values for pH was 3.89 to 5.51 and 3.84 to 5.23 at the surface and sub surface level respectively. The spatial pattern at the surface and sub surface levels was slightly similar. It indicated soil area with high acidic nature spread across the study area in large pockets mostly concentrated on the downslope region i.e. the north – western direction. Most of the area at the sub surface level was found to be less acidic as compared to the surface level. The EC value varied from 66.26 µS/cm to 105.08 µS/cm at the surface and 36.49 µS/cm to 70.60 µS/cm at the sub surface level. A higher EC value was observed along the northern border of the area at the surface level. Higher EC values for the sub surface level were found in small

portions. For the predicted TOC values, the range was from 1.32 % to 2.88 % for surface and 1.06 % to 2.17 % at the sub surface level. The higher values were observed along the eastern edge of the map or the upslope part of the study area for both strata – surface and sub surface of the soil. Values were lower along the downslope region of the area.

The interpolated results using copulas as interpolators with covariates are as shown in Figure 5-9. The value ranged between 3.88 and 5.52 for the surface, and between 3.84 and 5.34 for sub surface level. Lower values were observed in the north – western direction or the downslope part of the area. Also, higher values were observed along a small strip on the northern border. The observation was the same for both surface and sub surface level values. The higher values were present in small patches for the surface level values whereas it was a similar case for low values at the sub surface level. For the predicted EC values, the lower and upper limit was 52.91 µS/cm and 191.05 µS/cm for the surface, and 26.43 µS/cm and 107.44 µS/cm for sub surface level respectively. Higher values were observed in the north – western direction for both the strata. Apart from them, the predicted values were higher on the northern border (downward slope) of the region for the surface level. A patch of the area had high value in case of sub surface level. The TOC values varied between 1.16 % and 2.75 %, whereas they were between 0.96 % and 2.46 % for the surface and sub surface level. Higher values were particularly observed in the north – western edge and the eastern edge of the study area at the surface and sub surface level. Additionally, values were found to be higher on the eastern edge of the study area for the surface level data.

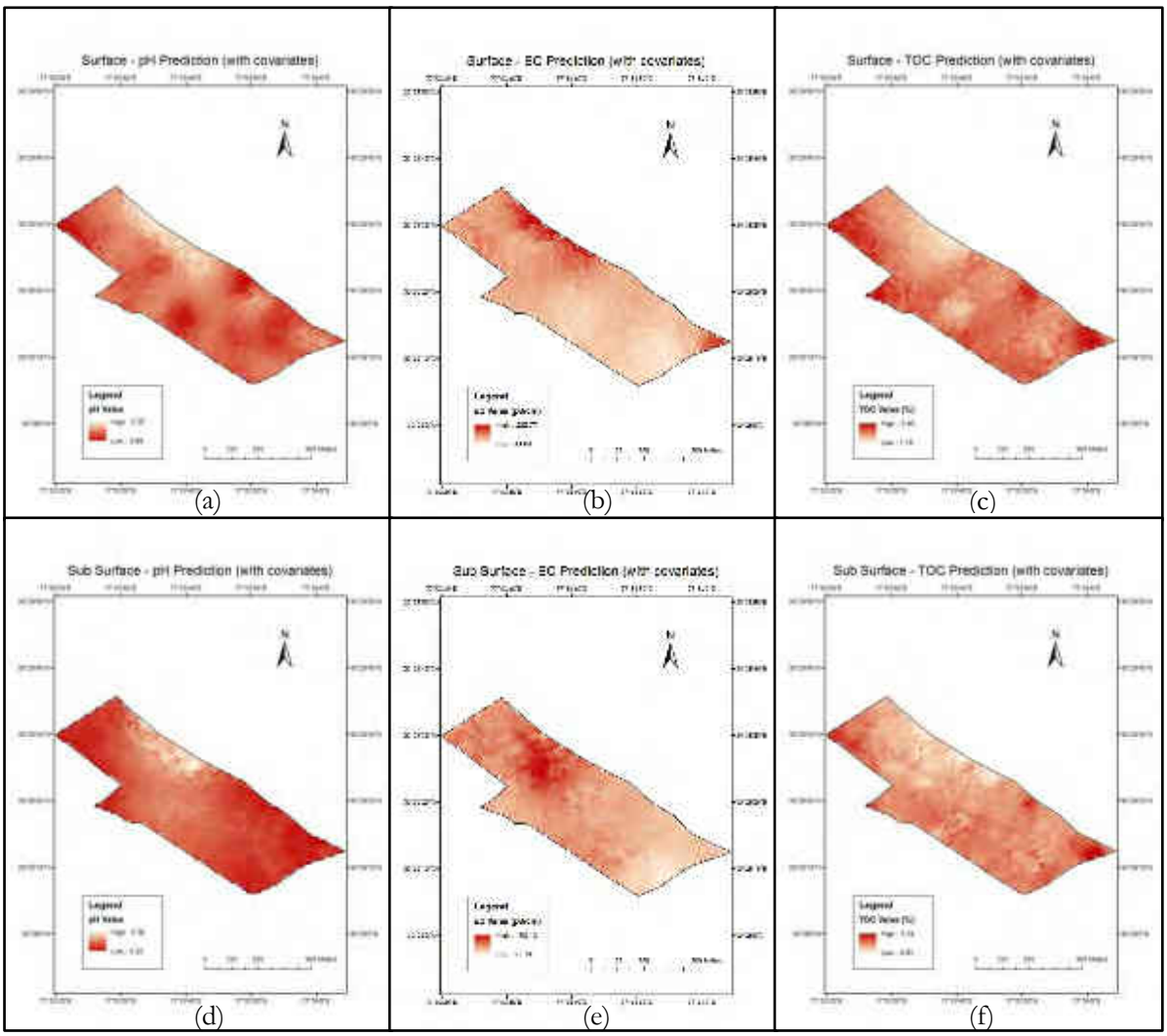


Figure 5-9: Surface (a-c) and sub-surface (d-f) level interpolation maps for pH, EC and TOC respectively in Langha – Tauli using copulas as interpolators. In this case, the covariates had also been used for interpolation. The pixel size is 16 m and the projected coordinate system is UTM 44 N.

5.5. Measures of uncertainty

The measures of uncertainties for different interpolation methods that had been used are mentioned in Table 5-4 and Table 5-5.

For the surface level statistics for pH, the absolute value of ϵ ranged from 5.78×10^{-6} in case of RK to 0.0672 for interpolation using copulas with covariates. The MSE and RMSE values for the pH values were 3.15×10^{-6} and 0.0018 respectively for interpolation by RK at the lower range to 0.2344 and 0.4841 respectively at the upper range limit for interpolation using copulas with covariates. Also, the R^2 value was found to be the least for the case of interpolation with covariates at 0.0025 and the highest for RK having a value of 0.4074. This implied that RK was able to explain 40.74 % of the variance when interpolation of pH data was performed. Similarly, for EC, the ϵ value varied from a minimum of 0.0001 using RK to a maximum value of 8.9865 using copulas as interpolators with covariates. The lower MSE and RMSE values were 0.3328 and 0.5768 respectively. These values were achieved when RK was used. Respectively, the highest MSE and RMSE values were 4568.373 and 67.5897. These were obtained while performing interpolation using copulas without covariates. The highest R^2 value of 0.3533 was achieved through RK whereas the lowest value of 0.0008 was obtained by using copulas with covariates as an interpolator. The least ϵ value for TOC was 0.0031 obtained using RK, whereas it was the highest in case of interpolation through copulas with covariates having a value of 0.1111. The MSE and RMSE values were least when RK method was used having a value of 0.1594 and 0.3992 respectively. The highest values for MSE and RMSE were 0.3250 and 0.5701 respectively, when copulas with covariates as an interpolator was used. RK was able to explain the maximum variance with the R^2 value of 0.4864 whereas the worst performance was for the case of copulas with covariates as interpolators with the R^2 value of 0.0927.

For the predicted values at sub surface level, RK performed the best among all the interpolators for pH with the least absolute ϵ, MSE and RMSE values of 1.42×10^{-7}, 4.22×10^{-7} and 0.0006 respectively. The R^2 value was the highest with a value of 0.3068. The worst performer as an interpolator was when it was done using copulas with covariates in terms of MSE, RMSE and R^2. The corresponding MSE and RMSE values were 0.3070 and 0.5541. The R^2 value was 0.0002. The highest ϵ value was observed in the case of copulas without covariates as interpolator with an absolute value of 0.0051. In case of EC predictions, interpolation was the best when RK was used with the corresponding ϵ, MSE, RMSE and R^2 values of 0.0368, 805.1854, 28.3758 and 0.5512. Interpolation using copulas with covariates performed the worst with the ϵ value of 6.1647, MSE value of 2130.901 and RMSE value of 46.1616. The least R^2 value of 0.0058 was observed for the case of RBF as an interpolator. For the predicted TOC values, the best among all the interpolation methods was RK with ϵ, MSE, RMSE and R^2 values of 0.0004, 0.0138, 0.1174 and 0.5329 respectively. The worst performer in terms of interpolation was when copulas with covariates were used. The ϵ value was 0.0933, MSE and RMSE values were 0.3064 and 0.5535 respectively and the R^2 value was 0.0437.

Table 5-4: The mean error ϵ, the mean squared error MSE, the root mean squared error RMSE and R^2 value of soil parameters for various interpolation methods for surface level

Interpolation Methods		**Soil parameter**	**ϵ**	**MSE**	**RMSE**	**R^2**
RBF		pH	-0.0048	0.1879	0.4334	0.1616
		EC	-3.7950	4022.9152	63.4264	0.0252
		TOC	-0.0142	0.2692	0.5188	0.1389
RK		pH	-5.7822e-06	3.1506e-06	0.0018	0.4074
		EC	0.0001	0.3328	0.5768	0.3533
		TOC	0.0031	0.1594	0.3992	0.4864
Copulas	without covariates	pH	0.0187	0.2144	0.4630	0.0503
		EC	-0.1856	4568.373	67.5897	0.0103
		TOC	0.0488	0.2759	0.5252	0.1197
	with covariates	pH	0.0672	0.2344	0.4841	0.0025
		EC	8.9865	4360.239	66.0321	0.0008
		TOC	0.1111	0.3250	0.5701	0.0927

Table 5-5: The mean error ϵ, the mean squared error MSE, the root mean squared error RMSE and R^2 value of soil parameters for various interpolation methods for sub-surface level

Interpolation Methods		**Soil parameter**	**ϵ**	**MSE**	**RMSE**	**R^2**
RBF		pH	-0.0042	0.2115	0.4598	0.1398
		EC	-3.8389	1849.4841	43.0056	0.0058
		TOC	-0.0366	0.2216	0.4707	0.0792
RK		pH	-1.4234e-07	4.2187e-07	0.0006	0.3068
		EC	0.0368	805.1854	28.3758	0.5512
		TOC	0.0004	0.0138	0.1174	0.5329
Copulas	without covariates	pH	-0.0051	0.2428	0.4927	0.0196
		EC	1.2174	1877.853	43.3342	0.0118
		TOC	0.0436	0.2204	0.4694	0.0822
	with covariates	pH	0.0006	0.3070	0.5541	0.0002
		EC	6.1647	2130.901	46.1616	0.0471
		TOC	0.0933	0.3064	0.5535	0.0437

6. DISCUSSION

6.1. Optimal sampling scheme

As observed in Figure 5-1 (a), the sampling scheme closely matched with that of an equilateral grid for Langha-Tauli with an observed mean distance between points of 65.32 m. The sampling scheme had been obtained such that the minimum variance value was present in the study area. Also, the slope of the study area was gentle and almost flat at some locations with sharp falls. Although, the slope data had been considered while generating the optimized scheme, equally distributed points in the form of the equilateral grid were derived. This was in conformance with the sampling grid suggested by Yfantis et al. (1987). Due to unavailability of prior information in Barwa, the values of the variogram parameters were taken post analysis of soil samples in Langha-Tauli. Since sufficient observation locations were not present for Barwa, the obtained sampling scheme was sub-optimal and was just an instance of the many possible sampling scheme sets.

6.2. Descriptive statistics and soil health

Soil variables in the actual world do not ideally follow normality and are usually positively skewed. It had been stated by Becker et al. (1992), which was confirmed by the findings presented in Table 5-1 for Langha-Tauli. Variables such as pH and TOC were noted to be negatively skewed for Barwa as stated in Table 5-2. This may have been due to the smaller number of samples that had been collected in the area.
Soils in hills were usually observed to be acidic (Reddy, 2011). Nearly 7 % of the total geographical area in the state of Uttarakhand was acidic and no area was observed with salinity as shown in Table 3 of Chapter 2 in Katyal et al. (2016). The soil samples were found to be acidic and without any salinity in nature for both the study areas of Langha-Tauli and Barwa. The mean TOC values for surface and sub surface level for Langha-Tauli and Barwa were well above the critical values of 1.5 – 2 % as prescribed by Lal (2016). The TOC content needed to be above the critical limit for the proper functioning of the soil.

Approximately 26 % of the land in Uttarakhand was observed to be degraded (Katyal et al., 2016). This was attributed to acidic and saline soils. In addition to that, low TOC values meant low water retention and use efficiency, resistance to climate change and heat wave, and low nutrient retention (Lal, 2016). Therefore, in terms of soil health, the soil was found to be non-degraded in terms of salinity and TOC values. It was considered degraded in terms of pH, as the soil was found to be highly acidic. The prevailing soil conditions and properties fall under the category of soil health. Continuous monitoring is required for assessing soil health. A Soil Health Card (SHC) scheme had been initiated by the Government of India, under which the analysed parameters considered in the research work as the physical/basic attributes were used for establishing soil health, apart from the biological and chemical ones (National portal of India, 2017).

6.3. Using spatial information from one area to another – a Bayesian kriging implementation

After comparing the results of OK and BK as stated in Table 5-3, OK outperformed BK. This observation was in agreement with the findings by Cui et al. (1995) for sufficient number of samples. Webster & Oliver (1992) had suggested that at least 100 data points needed to be present for a soil survey. Since the number of observations was adequate to properly estimate the variogram, the performance of OK over BK was better in Langha-Tauli.

One particular point to note in the cross-validation technique of BK results was that the predictions were performed for all the locations in Barwa. Only, the observation for which the uncertainty measurement was to be performed was excluded from BK process in the Bayesian kriging function of R language. This was done because of the presence of less number of observations. Cui et al. (1995) had assumed a chi-square distribution for the inverse of partial sill values and exponential distribution for the range parameter as prior for BK. Although a similar distribution function, i.e. chi-square was obtained for the inverse of partial sill values, log-normal distribution was fitted to the various obtained range values. This was based on the goodness of fit statistics and criteria. When the posterior distribution of Langha-Tauli was considered as the prior distribution for Barwa, the obtained results from OK were found to be better than BK. After the priors for inverse partial sill and range values were updated, BK was found to perform much better than the case without updating priors. Also, it performed better in terms of MSE, but slightly worse σ_{RE}^2 values were observed than OK. Thus, an association between two similar topographic features was tried to be established. Due to the constraints in research work, more observations in Barwa could not be collected.

6.4. Interpolation

Interpolation by all the stated methods – RBF, RK and copula-based were performed only for the first study area i.e. Langha-Tauli. It wasn't implemented in the second study area (Barwa) because a low number of observations were present. An exception to this was by BK. Although interpolation by RBF does not have any constraint for the minimum number of observations, RK requires a variogram to be estimated. Depending on the spatial properties of the variable, a minimum of 30 observations are generally required for satisfactorily determining a variogram (Warrick & Myers, 1987). It was a similar case with copulas wherein correlogram instead of variogram needs to be estimated. A sufficient number of point pairs of observations needs to be present in bins or lag distance classes for a correlogram to be reasonably estimated. Only 7 observations were present for Barwa. Therefore, even if the interpolation process would have been performed by RK and copula-based interpolators, the resultant variogram and correlogram would not have been accurate. But, BK was observed to generate better results as compared to OK with less number of samples in the second study area.

Interpolation using RBF produced decent results as compared to other geostatistical techniques since the data was found to be largely homogeneous as stated in Section 5.2. Also, RK outperformed RBF and copula-based interpolators as seen from the interpolated maps in Figure 5-6 (a-f) and the uncertainty measurements in Table 5-4 and Table 5-5 for surface and sub surface levels respectively. Kriging the residuals of the Box-Cox transformed variables helped in producing interpolated surfaces with low uncertainty measurements. The value of covariates at unvisited locations was obtained through OK. The method was used because it considered the spatial variation in data. Additionally, less number of parameters needed to be estimated for performing the interpolation process.

Although it had been established that copula-based interpolators performed better than other geostatistical techniques (Marchant et al., 2011; Bárdossy & Li, 2008), they did not perform better than RK and RBF for the research. The reason for this may be attributed to the data not being highly skewed or deviating much from normality.

Amongst all the interpolation methods, RK was able to efficiently model the soil variables when sufficient number of soil samples were present. This may be credited to the variables being accurately transformed to normality. BK was able to efficiently predict the values when less number of samples (7 in this case) were present.

7. CONCLUSIONS AND RECOMMENDATIONS

The main objective of the research work was to compare various interpolation methods by studying the variation in soil properties in a hilly terrain. For this, the study spanned various stages of research – from designing optimal sampling schemes and doing field visits to collecting samples and conducting experiments to generating continuous mathematical surfaces using interpolation methods.

The sampling scheme was based on the model-based method wherein minimization of error variance as an objective function had been used. Regression kriging was used as a method to generate the variance surface, which was supposed to be minimized. The elevation data from CartoDEM was used to generate the slope surface, which in turn was used as a covariate in the kriging process. Spatial simulated annealing was used as an optimizer for getting the combination of locations with the minimum variance across the whole space. This process was repeated for different starting initial temperature values and the number of iterations. Almost 100 locations (97 locations in total, 1 excluded due to non - accessibility) were considered for this procedure in Langha-Tauli. The soil samples were then collected post harvesting of the summer crops. While collecting soil samples, a location had to be excluded after the sampling scheme had been generated because of inaccessibility issues. This would have affected the optimal behaviour of the sampling scheme. The soil samples were then tested in the Central Analytical Laboratory of the Indian Institute of Remote Sensing. The variogram parameters of the obtained surface level pH data were considered for getting the optimal sampling scheme in case of Barwa. Soil samples in Barwa were collected when saplings had started growing, so the obtained analysis results may have some bias. This was because the soil samples were collected in such a way that the crop was not disturbed.

For assessing whether spatial information from one area can be utilized to another without actually conducting any previous soil sampling, Bayesian kriging was used as an interpolation method. The only previous knowledge of the second study area was its topography. This formed a limitation for proper analysis. Accurate delineation of the area may have been performed if the soil data regarding any possible previous major anthropogenic activity was available. Distribution functions were defined to model uncertainties in variogram parameters. These were derived from fixed width random subsets of the observations. The inverse of the partial sill and the range values were then checked for fitness distribution functions. Inverse chi-square and log-normal distribution gave the best fit for the inverse of partial sill and the range values respectively. The posterior distribution information from Langha-Tauli after performing Bayesian kriging was used as the prior distribution in Barwa. This prior was further updated such that its mean value matched with that of the variogram parameters and the Bayesian kriging was performed again using the new priors. The uncertainty measurements denoting the quality of interpolation showed an improvement for the case when updating the prior had been done. This was in comparison with the ordinary kriging results and when the priors for Barwa were used as the posteriors from Langha-Tauli. More observations from Barwa could be collected to test the better of the interpolation methods. Since more samples could not be collected, the research was limited to a single iteration of improvement of priors.

A deterministic method – interpolation using radial basis functions and 2 geostatistical methods – regression kriging and interpolation using copulas were used to assess the soil variability in the study areas. For interpolation by radial basis functions, the root mean squared prediction error value was taken as the criterion for choosing the kernel function. The soil variables were Box-Cox transformed to normality.

Various combinations of the soil variable and its transformation, to be predicted, and the covariates experimented. The combination which gave the least uncertainty measurements was chosen for the kriging process. Copula-based interpolation was performed with and without covariates and the uncertainty measures noted. The variables for which the prediction was not being performed were chosen as covariates. Gaussian, as well as non-Gaussian copulas, were utilized for modelling the dependency structure. Contrary to the established proofs in literature, copulas did not perform well for the concerned study area. Regression kriging performed the best among all the interpolators at the surface level and sub-surface levels. Copulas are not a universal solution to the problem of deriving the unknown values. In the past, the formula that led to the economic crisis of 2008 has been attributed to the Gaussian copula function (Salmon, 2009).

Similar methodology may be followed for any other study area because of the generic attribute of the research. Only the results may vary, as the interpolation process is data dependent.

Recommendations

In addition to using the DEM data, remote sensing data may be utilized to get a better understanding of the soil properties. In addition to that, proper pre-survey of the study area is recommended so that inaccessibility issues such as those observed in the research may be avoided.

Since the study involved multiple dependent methods, an error propagation study may also be conducted to quantify their impact. One such case is in the case of regression kriging, where ordinary kriging had been used to interpolate the covariate values at the prediction locations. Uncertainty may propagate with each successive method used.

A Soil Health Card scheme had been launched by the Government of India for helping the farmers know their soils better and work on improving soil health. The research work may be utilized to efficiently and economically design a sampling scheme. Based on the optimal sampling scheme, the soil samples may be collected, tested and a continuous surface is generated. Based on this, the fertilizer recommendations and soil amendments required for the land may be suggested.

Publications

As of the submission date of the thesis, communication with the journal 'Geoderma', which is a global journal of soil science was commenced for publication.

LIST OF REFERENCES

Aciego Pietri, J. C., & Brookes, P. C. (2008). Relationships between soil pH and microbial properties in a UK arable soil. *Soil Biology and Biochemistry*, *40*(7), 1856–1861. https://doi.org/10.1016/j.soilbio.2008.03.020

Adhikary, P. P., & Dash, C. J. (2017). Comparison of deterministic and stochastic methods to predict spatial variation of groundwater depth. *Applied Water Science*, *7*(1), 339–348. https://doi.org/10.1007/s13201-014-0249-8

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alidoost, F., & Stein, A. (2016). Interpolation of daily mean air temperature data via spatial and non-spatial copulas. In *Spatial Statistics 2017* (pp. s1–s21). Lancaster, United Kingdom. Retrieved from https://research.utwente.nl/en/publications/interpolation-of-daily-mean-air-temperature-data-via-spatial-and-

Anderson, T. W., & Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, *23*(2), 193–212. https://doi.org/10.1214/aoms/1177729437

Arshad, M. A., & Martin, S. (2002). Identifying critical limits for soil quality indicators in agro-ecosystems. *Agriculture, Ecosystems & Environment*, *88*(2), 153–160. https://doi.org/10.1016/S0167-8809(01)00252-3

ASPRS Map Accuracy Standards Working Group. (2015). ASPRS Positional Accuracy Standards for Digital Geospatial Data. *Photogrammetric Engineering & Remote Sensing*, *81*(3), 1–26. https://doi.org/10.14358/PERS.81.3.A1-A26

Bárdossy, A., & Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research*, *44*(7), 1–15. https://doi.org/10.1029/2007WR006115

Bárdossy, A., & Pegram, G. (2013). Interpolation of precipitation under topographic influence at different time scales. *Water Resources Research*, *49*(8), 4545–4565. https://doi.org/10.1002/wrcr.20307

Becker, M., Christensen, B. T., Horne, D. J., Khind, C. S., Ladha, J. K., Pareek, R. P., … Wallis, M. G. (1992). *Advances in Soil Science*. (B. A. Stewart, Ed.) (Vol. 20). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-2930-8

Bedford, T., & Cooke, R. M. (2002). Vines : A New Graphical Model for Dependent Random Variables. *The Annals of Statistics*, *30*(4), 1031–1068. Retrieved from http://www.jstor.org/stable/1558694

Bivand, R., Keitt, T., & Rowlingson, B. (2017). rgdal: Bindings for the "Geospatial" Data Abstraction Library. Retrieved from https://cran.r-project.org/package=rgdal

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R. Use R* (Vol. 1). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-78171-6

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. Retrieved from http://www.jstor.org/stable/2984418

Christensen, R. (2001). *Linear Models for Multivariate Time Series, and Spatial Data* (second edi). New York: Springer. Retrieved from http://www.springer.com/us/book/9781475741032

Cook, S., Corner, R., Groves, P., & Grealish, G. (1996). Use of airborne gamma radiometric data for soil mapping. *Australian Journal of Soil Research*, *34*(1), 183–194. https://doi.org/10.1071/SR9960183

Coster, A. (n.d.). Goodness-of-fit statistics. Retrieved March 2, 2018, from http://web.maths.unsw.edu.au/~adelle/Garvan/Assays/GoodnessOfFit.html

Cressie, N. A. C. (2015). *Statistics for Spatial Data* (Revised ed). New York: Wiley Interscience. Retrieved from https://www.wiley.com/en-us/Statistics+for+Spatial+Data%2C+Revised+Edition-p-9781119114611

Cressie, N., & Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, *12*(2), 115–125. https://doi.org/10.1007/BF01035243

Cui, H., Stein, A., & Myers, D. E. (1995). Extension of spatial information, bayesian kriging and updating of prior variogram parameters. *Environmetrics*, *6*(4), 373–384. https://doi.org/10.1002/env.3170060406

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic Techniques in Exposure Assessment*. Springer US. Retrieved from http://www.springer.com/in/book/9780306459566

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus : An R Package for Fitting Distributions. *Journal of Statistical Software*, *64*(4). https://doi.org/10.18637/jss.v064.i04

Demarta, S., & McNeil, A. J. (2005). The t Copula and Related Copulas. *International Statistical Review*, *73*(1), 111–129. https://doi.org/10.1111/j.1751-5823.2005.tb00254.x

Diggle, P. J., & Ribeiro, P. J. (1999). *Bayesian Inference in Gaussian Model-based Geostatistics*. Lancaster. Retrieved from http://www.leg.ufpr.br/geoR/geoRdoc/bayeskrige.pdf

Diggle, P. J., & Ribeiro, P. J. (2002). Bayesian Inference in Gaussian Model-based Geostatistics. *Geographical and Environmental Modelling*, *6*(2), 129–146. https://doi.org/10.1080/1361593022000029467

ESRI, Environmental Systems Research Institute (2012). ArcGIS Release 10.1. Redlands, California.

Frost, J. (2013). Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables. Retrieved February 22, 2018, from http://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables

Ghotekar, Y. (2016). *Central Analytical Laboratory Manual* (Version 2).

Gonzalez-Fernandez, Y., & Soto, M. (2016). vines: Multivariate Dependence Modeling with Vines. Retrieved from http://cran.r-project.org/package=vines

Goria, M. N. (1992). On the fourth root transformation of chi-square. *Australian Journal of Statistics*, *34*(1), 55–64. https://doi.org/10.1111/j.1467-842X.1992.tb01043.x

Gräler, B. (2014a). *Developing spatio - temporal copulas*. Doctoral thesis, Universität Münster. Retrieved from http://www.graeler.org/publications/Diss_Benedikt_Graeler.pdf

Gräler, B. (2014b). Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, *10*, 87–102. https://doi.org/10.1016/j.spasta.2014.01.001

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. *The R Journal*, *8*(1), 204–218. Retrieved from https://journal.r-project.org/archive/2016/RJ-2016-014/RJ-2016-014.pdf

Hardy, R. L. (1971). Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysical Research*, *76*(8), 1905–1915. https://doi.org/10.1029/JB076i008p01905

Hengl, T. (2006). Finding the right pixel size. *Computers & Geosciences*, *32*(9), 1283–1298. https://doi.org/10.1016/j.cageo.2005.11.008

Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, *33*(10), 1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, *120*(1–2), 75–93. https://doi.org/10.1016/j.geoderma.2003.08.018

Hijmans, R. J. (2017). raster: Geographic Data Analysis and Modeling. Retrieved from https://cran.r-project.org/package=raster

Jahn, R., Blume, H. P., Asio, V. B., Spaargaren, O., & Schad, P. (2006). *Guidelines for Soil Description*. *Food and Agriculture Organization of the United Nations* (Vol. 1). Rome. Retrieved from http://www.fao.org/docrep/019/a0541e/a0541e.pdf

Jasiński, R. (2016). The use of interpolation methods for the modelling of environmental data. *Desalination and Water Treatment*, *57*(3), 964–970. https://doi.org/10.1080/19443994.2014.1002282

Jones, S. (2016). National Soil Health Measurements to Accelerate Agricultural Transformation. Retrieved August 17, 2017, from http://soilhealthinstitute.org/national-soil-health-measurements-accelerate-agricultural-transformation/

Katyal, J., Chaudhari, S., Dwivedi, B., Biswas, D., Rattan, R., & Majumdar, K. (Eds.). (2016). *Soil Health: Concept, Status and Monitoring* (30th ed.). Retrieved from https://www.isss-india.org/images/doc_file/Bull30.pdf

Kazianka, H., & Pilz, J. (2010). Geostatistical modeling using non-Gaussian copulas. In *Accuracy 2010 - Proceedings of the 9th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Leicester. Retrieved from http://www.spatial-accuracy.org/system/files/img-X03141001_0.pdf

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*(1/2), 81–93. https://doi.org/10.2307/2332226

Kumar, S., & Singh, R. P. (2016). Spatial distribution of soil nutrients in a watershed of Himalayan landscape using terrain attributes and geostatistical methods. *Environmental Earth Sciences*, *75*(6), 473. https://doi.org/10.1007/s12665-015-5098-8

Lal, R. (2016). Soil health and carbon management. *Food and Energy Security*, *5*(4), 212–222. https://doi.org/10.1002/fes3.96

Lark, R. M. (2012). Towards soil geostatistics. *Spatial Statistics*, *1*, 92–99. https://doi.org/10.1016/j.spasta.2012.02.001

Lazzaro, D., & Montefusco, L. B. (2002). Radial basis functions for the multivariate interpolation of large scattered data sets. *Journal of Computational and Applied Mathematics*, *140*(1–2), 521–536. https://doi.org/10.1016/S0377-0427(01)00485-X

Li, D. X. (1999). On Default Correlation: A Copula Function Approach. *SSRN Electronic Journal*, (99), 31. https://doi.org/10.2139/ssrn.187289

Li, J., & Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, *53*, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008

Liu, Z. P., Shao, M. A., & Wang, Y. Q. (2013). Large-scale spatial interpolation of soil pH across the Loess Plateau, China. *Environmental Earth Sciences*, *69*(8), 2731–2741. https://doi.org/10.1007/s12665-012-2095-z

Marchant, B. P., Saby, N. P. A., Jolivet, C. C., Arrouays, D., & Lark, R. M. (2011). Spatial prediction of soil properties with copulas. *Geoderma*, *162*(3–4), 327–334. https://doi.org/10.1016/j.geoderma.2011.03.005

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246

Matheron, G. (1969). *Le krigeage universel*. (l'Ecole Nationale Supérieure des Mines de Paris, Ed.), *Les cahiers du Centre de morphologie mathématique de Fontainebleau, Fascicule 1*. Retrieved from http://www.cg.ensmp.fr/bibliotheque/public/MATHERON_Ouvrage_00131.pdf

McBratney, A. B., & Pringle, M. J. (1999). Estimating Average and Proportional Variograms of soil properties and their potential use in Precision Agriculture. *Precision Agriculture*, *1*(2), 125–152. https://doi.org/10.1023/A:1009995404447

Merriam - Webster. (2017). Interpolation | Definition of interpolation by Merriam - Webster. Retrieved November 29, 2017, from https://www.merriam-webster.com/dictionary/interpolation

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Minasny, B., & McBratney, A. B. (2005). The Matérn function as a general model for soil variograms. *Geoderma*, *128*(3–4), 192–207. https://doi.org/10.1016/j.geoderma.2005.04.003

Moore, W. J. (1999). *Physical Chemistry* (5th ed.). Orient Blackswan. Retrieved from https://books.google.co.in/books?id=nIggbG9i8qEC&lpg=PA435&ots=OPVPgZ2kGP&dq=ionic mobilities of hydrogen and hydroxyl in water&pg=PA435#v=onepage&q=ionic mobilities of hydrogen and hydroxyl in water&f=false

Müller, W. G. (1999). Least-squares fitting from the variogram cloud. *Statistics & Probability Letters*, *43*(1), 93–98. https://doi.org/10.1016/S0167-7152(98)00250-8

Muralikrishnan, S., Narender, B., Reddy, S., & Pillai, A. (2011). *Evaluation of Indian National DEM from Cartosat-1 Data. Indian Space Research Organization-NRSC* (Vol. 1). Hyderabad. Retrieved from http://bhuvan-noeda.nrsc.gov.in/download/download/tools/document/CartoDEMReadme_v1_u1_23082011.pdf

National portal of India. (2017). Soil Health Card. Retrieved March 3, 2018, from https://www.india.gov.in/spotlight/soil-health-card#tab=tab-1

Nelsen, R. B. (2006). *An Introduction to Copulas* (2nd ed.). New York, NY: Springer New York. https://doi.org/10.1007/0-387-28678-0

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. Retrieved from https://cran.r-project.org/package=RColorBrewer

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, *30*(7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Pebesma, E. J. (2006). The Role of External Variables and GIS Databases in Geostatistical Analysis.

Transactions in GIS, *10*(4), 615–632. https://doi.org/10.1111/j.1467-9671.2006.01015.x
Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13. Retrieved from https://cran.r-project.org/doc/Rnews/Rnews_2005-2.pdf
Piccini, C., Marchetti, A., & Francaviglia, R. (2014). Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment. *Ecological Indicators*, *36*, 301–314. https://doi.org/10.1016/j.ecolind.2013.08.009
Poggio, L., Gimona, A., & Brewer, M. J. (2013). Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. *Geoderma*, *209–210*, 1–14. https://doi.org/10.1016/j.geoderma.2013.05.029
Reddy, K. S. (2011). *Vision 2030*. Bhopal. Retrieved from http://www.iiss.nic.in/vision/vision2030.pdf
Ribeiro, P. J., & Diggle, P. J. (2016). geoR: Analysis of Geostatistical Data. Retrieved from https://cran.r-project.org/package=geoR
Salmon, F. (2009). Recipe for Disaster: The formula that killed Wall Street. Retrieved March 1, 2018, from https://www.wired.com/2009/02/wp-quant/
Samuel-Rosa, A., Heuvelink, G., Vasques, G., & Anjos, L. (2017). *spsann - Optimization of Sample Patterns Using Spatial Simulated Annealing*. Rio de Janeiro. Retrieved from https://cran.r-project.org/package=spsann
Schabenberger, O., & Pierce, F. (2001). *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press. https://doi.org/10.1201/9781420040197
Schepsmeier, U., Stoeber, J., Brechmann, E. C., Gräler, B., Nagler, T., & Erhardt, T. (2018). VineCopula: Statistical Inference of Vine Copulas. Retrieved from https://cran.r-project.org/package=VineCopula
Singh, D., Chhonkar, P. K., & Dwivedi, B. S. (2010). *Manual on Soil, Water and Plant Analysis*. New Delhi: Westville Publishing House.
Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, *8*, 229–231.
Sluiter, R. (2008). *Interpolation methods for climate data literature review*. *KNMI, R&D Information and Observation Technology*. De Bilt. Retrieved from https://www.snap.uaf.edu/sites/default/files/files/Interpolation_methods_for_climate_data.pdf
Szatmári, G., Barta, K., & Pásztor, L. (2015). An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. *Hungarian Geographical Bulletin*, *64*(1), 35–48. https://doi.org/10.15201/hungeobull.64.1.4
R Core Team (2017). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/
Tso, B., & Mather, P. M. (2001). *Classification methods for remotely sensed data* (2nd ed.). CRC Press. Retrieved from https://www.crcpress.com/Classification-Methods-for-Remotely-Sensed-Data-Second-Edition/Mather-Tso/p/book/9781420090727
United States Environmental Protection Agency. (2002). *Guidance on Choosing a Sampling Design for Environmental Data Collection*. *Office of Environmental Information*. Washington. Retrieved from https://www.epa.gov/sites/production/files/2015-06/documents/g5s-final.pdf
van Groenigen, J. W. (1997). Spatial Simulated Annealing for optimizing sampling - Different optimization criteria compared. In A. Soares, J. Gómez-Hernandez, & R. Froidevaux (Eds.), *Geostatistics for Environmental Applications* (pp. 351–361). Lisbon: Kluwer Academic Publications. Retrieved from https://link.springer.com/chapter/10.1007/978-94-017-1675-8_29
van Groenigen, J. W., Siderius, W., & Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, *87*(3–4), 239–259. https://doi.org/10.1016/S0016-7061(98)00056-1
van Groenigen, J. W., & Stein, A. (1998). Constrained Optimization of Spatial Sampling using Continuous Simulated Annealing. *Journal of Environment Quality*, *27*(5), 1078. https://doi.org/10.2134/jeq1998.00472425002700050013x
Verdin, A., Rajagopalan, B., Kleiber, W., & Funk, C. (2015). A Bayesian kriging approach for blending satellite and ground precipitation observations. *Water Resources Research*, *51*(2), 908–921. https://doi.org/10.1002/2014WR015963
Wagner, W., Naeimi, V., Scipal, K., Jeu, R., & Martínez-Fernández, J. (2007). Soil moisture from operational meteorological satellites. *Hydrogeology Journal*, *15*(1), 121–131.

https://doi.org/10.1007/s10040-006-0104-6
Wang, J.-F., Stein, A., Gao, B.-B., & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, *2*(1), 1–14. https://doi.org/10.1016/j.spasta.2012.08.001
Warrick, A. W., & Myers, D. E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, *23*(3), 496–500. https://doi.org/10.1029/WR023i003p00496
Webster, R., & Oliver, M. A. (1990). *Statistical Methods in Soil and Land Resource Survey*. Oxford University Press.
Webster, R., & Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, *43*(1), 177–192. https://doi.org/10.1111/j.1365-2389.1992.tb00128.x
Weisstein, E. W. (n.d.). Bayes' Theorem. Retrieved February 6, 2018, from http://mathworld.wolfram.com/BayesTheorem.html
Wright, G. B. (2003). *Radial Basis Function Interpolation: Numerical and Analytical Developments*. Doctoral Thesis, University of Colorado. Retrieved from https://amath.colorado.edu/faculty/fornberg/Docs/GradyWrightThesis.pdf
Xiao, X., Gertner, G., Wang, G., & Anderson, A. B. (2005). Optimal Sampling Scheme for Estimation Landscape Mapping of Vegetation Cover. *Landscape Ecology*, *20*(4), 375–387. https://doi.org/10.1007/s10980-004-3161-z
Yao, X., Fu, B., Lü, Y., Sun, F., Wang, S., & Liu, M. (2013). Comparison of Four Spatial Interpolation Methods for Estimating Soil Moisture in a Complex Terrain Catchment. *PLoS ONE*, *8*(1), e54660. https://doi.org/10.1371/journal.pone.0054660
Yfantis, E. A., Flatman, G. T., & Behar, J. V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology*, *19*(3), 183–205. https://doi.org/10.1007/BF00897746

APPENDIX A

Table A- 1: Table showing objective function value with the number of iterations for different initial temperature values for Langha – Tauli

Number of iterations	Objective function value for initial temperature of 3	Objective function value for initial temperature of 3.5	Objective function value for initial temperature of 4
10	0.08603	0.08771	0.08771
23	0.08669	0.08645	0.08864
50	0.08555	0.08537	0.08568
111	0.08676	0.08871	0.08670
247	0.06915	0.07016	0.07069
550	0.06756	0.06791	0.06475
1224	0.06918	0.06922	0.06844
2728	0.06741	0.06939	0.06831
6078	0.07023	0.07022	0.06879
13542	0.06952	0.06823	0.06883
30173	0.06831	0.06941	0.06926
67233	0.06942	0.06920	0.06755
149810	0.06839	0.06798	0.06917
333811	0.06954	0.06974	0.06741
743810	0.06941	0.06888	0.06942
1657384	0.06960	0.06877	0.06867
3693046	0.06836	0.06934	0.06891
8228983	0.06857	0.06911	0.06966
18336131	0.06926	0.06864	0.06952
40857260	0.06997	0.07000	0.06843
91039690	0.06802	0.06823	0.06816
202858079	0.06831	0.06824	0.06955
452016040	0.06762	0.06924	0.06949
1007199229	0.06851	0.06851	0.06832

Table A- 2: Table showing objective function value with the number of iterations for different initial temperature values for Barwa

Number of iterations	Objective function value for initial temperature of 3	Objective function value for initial temperature of 3.5	Objective function value for initial temperature of 4
10	0.28928	0.24651	0.28824
16	0.24252	0.29101	0.27742
24	0.56566	0.27491	0.27972
36	0.31717	0.28608	0.29569
55	0.27776	0.29782	0.29975
83	0.29007	0.30331	0.28241
127	0.27530	0.23163	0.28689
194	0.25460	0.25959	0.25948
295	0.25606	0.25139	0.26045
450	0.25486	0.25765	0.25466

687	0.26321	0.25368	0.25646
1049	0.25918	0.25754	0.25162
1600	0.30240	0.25570	0.25451
2443	0.25743	0.25883	0.25672
3728	0.25730	0.25470	0.25434
5690	0.26044	0.26284	0.46900
8686	0.26091	0.25562	0.25683
13258	0.26325	0.25831	0.26073
20236	0.23821	0.26008	0.25984
30889	0.24497	0.25905	0.25627
47149	0.25753	0.23718	0.26076
71969	0.25643	0.25636	0.25355
109855	0.25996	0.26160	0.24975
167684	0.23784	0.67197	0.23253
255955	0.26192	0.25805	0.61102
390694	0.25595	0.25679	0.26349
596363	0.25708	0.25467	0.25619
910299	0.25793	0.25406	0.25768
1389496	0.26119	0.25232	0.25403
2120951	0.25812	0.25850	0.25773
3237458	0.68943	0.26170	0.23949
4941714	0.26020	0.23714	0.25509
7543121	0.25320	0.25019	0.25478
11513954	0.25317	0.25997	0.67089
17575107	0.25619	0.25880	0.25749
26826958	0.25494	0.25630	0.51409
40949151	0.25941	0.25642	0.25276
62505520	0.25768	0.61003	0.26011
95409548	0.25314	0.26491	0.26288
145634848	0.53655	0.25729	0.25261
222299649	0.26012	0.25925	0.25585
339322178	0.51571	0.26056	0.26164
517947468	0.25761	0.25930	0.25609
790604322	0.25528	0.26041	0.26248
1206792641	0.26346	0.26495	0.25080
1842069970	0.25486	0.25373	0.25848