Studies in Computational Intelligence 839

Shrutilipi Bhattacharjee
Soumya Kanti Ghosh
Jia Chen

Semantic Kriging for Spatio-temporal Prediction

Springer

Studies in Computational Intelligence

Volume 839

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are submitted to indexing to Web of Science, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink.

More information about this series at http://www.springer.com/series/7092

Shrutilipi Bhattacharjee ·
Soumya Kanti Ghosh · Jia Chen

Semantic Kriging for Spatio-temporal Prediction

Springer

Shrutilipi Bhattacharjee
Department of Electrical and Computer Engineering
Technical University of Munich (TUM)
Munich, Bayern, Germany

Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur, West Bengal, India

Jia Chen
Department of Electrical and Computer Engineering
Technical University of Munich (TUM)
Munich, Bayern, Germany

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-981-13-8663-3 ISBN 978-981-13-8664-0 (eBook)
https://doi.org/10.1007/978-981-13-8664-0

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Dedicated to all young geospatial scientists

Preface

Modeling climatological dynamics and weather patterns have been studied extensively in *remote sensing* (*RS*) and *geographic information system* (*GIS*). These analyses are crucial for various geospatial applications, such as climatological trend analysis, prediction and forecasting, urban growth modeling. The meteorological parameters, closely related to earth surface, play important roles for any climatological study. Prediction of these parameters is one of the crucial preprocessing steps involved in most of the analyses. The geostatistical spatial interpolation methods are reported to be the most efficient choice for predicting those parameters which are derived from the satellite (raster) imagery. These methods facilitate improved modeling of spatial autocorrelation/proximity, hence producing minimal error. It is also observed that the interdependencies between the meteorological and terrestrial dynamics play a critical role in the proximity estimation. The semantic modeling of these land–atmosphere interactions and analyzing the associations between different factors are obvious for the betterment in the prediction process.

This book focuses on the semantic land–atmosphere interaction modeling for the meteorological parameters that are correlated and influenced by the terrestrial dynamics. A new spatial interpolation method is presented, namely *semantic kriging* (*SemK*), which is capable not only to model the terrestrial land-use/land-cover (*LULC*) distribution, but also to incorporate this property into the existing interpolation method to make the prediction process more pragmatic and accurate. It is a novel approach to extend any spatial interpolation method (for meteorological parameters) with contextual/semantic *LULC* knowledge of the terrain. A hierarchical ontology-based approach has been adopted to quantify the same. To blend this semantic knowledge into the interpolation process, the most popular interpolation method reported in the literature, i.e., *ordinary kriging* (*OK*), has been extended further in the semantic domain. The *SemK* is categorized as a geostatistical univariate spatial interpolation method, which aims to minimize the variance of estimation error.

Munich, Germany Shrutilipi Bhattacharjee
Kharagpur, India Soumya Kanti Ghosh
Munich, Germany Jia Chen
December 2018

Acknowledgements

This work is mainly carried out under the INSPIRE Fellowship Scheme by the Department of Science and Technology (DST), New Delhi, India. The doctoral curricular of the first author is supported by this fellowship. It is also partially funded by the German Excellence Initiative and the European Union Seventh Framework Program under grant agreement n° 291763. Currently, both the first and third authors are supported by this fellowship.

We are indebted to Prof. Pabitra Mitra from the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India, and Prof. Shashi Shekhar from the Department of Computer Science, University of Minnesota, USA, for their valuable comments and suggestions to improve this work.

We would like to thank the staff at Springer, in particular Mr. Aninda Bose for technical help and support.

Munich, Germany Shrutilipi Bhattacharjee
Kharagpur, India Soumya Kanti Ghosh
Munich, Germany Jia Chen

Contents

About the Authors

Dr. Shrutilipi Bhattacharjee is a Postdoctoral Fellow (PDF) in the Department of Electrical and Computer Engineering, Technical University of Munich, Germany. She completed her B.Tech. from West Bengal University of Technology, India; M.Tech. from National Institute of Technology, Durgapur, India, and Ph.D. from Indian Institute of Technology Kharagpur, India. She has worked as a Fulbright Fellow at the University of Minnesota, USA. Her research interests include geoscience, remote sensing, environmental modelling, semantic analysis, spatial data mining, and spatial statistics. She has published numerous papers in international journals and at conferences. She is also a reviewer for a number of journals and conferences. She is a young professional member of IEEE (including GRSS and WIE) and ACM.

Prof. Soumya Kanti Ghosh is a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Kharagpur, India. He has been awarded the National Geospatial Chair Professorship by the Department of Science and Technology, Government of India in 2017. He completed his M.Tech. and Ph.D. in Computer Science and Engineering at IIT Kharagpur. Prior to IIT Kharagpur, he worked for Indian Space Research Organization, Government of India, as Scientist in the area of Satellite Remote Sensing and GIS. He has more than 15 years of teaching experience and supervised more than 10 Ph.D. theses. His research interests include spatial informatics, spatial data science, geographic information systems, and cloud computing. He has published numerous papers in international journals and conference proceedings. He is a member of IEEE and ACM.

Prof. Dr.-Ing Jia Chen is a Professor at the Technical University of Munich and an Associate in the Department of Earth and Planetary Sciences at Harvard University. She completed her Master's degree in Engineering at University Karlsruhe and Ph.D. at the Technical University of Munich, after which she also worked as Postdoctoral Fellow in Environmental Science & Engineering at Harvard University.

She has published over 100 papers in international journals and conferences and has also filed 12 patents. She is also an active reviewer for several international journals and a member of IEEE Photonics Society, EGU, VDE and VDI.

Acronyms

1D	One dimension
2D	Two dimension
3D	Three dimension
AK	Akima’s interpolator
ARIMA	Autoregressive integrated moving average
BB	Bounding box
BK	Block kriging
BN	Bayesian network
Cl	Classification
CTF	Causality testing framework
DAG	Directed acyclic graph
DM	Data mining
FBN	Fuzzy Bayesian network
FB-SemK	Fuzzy Bayesian semantic kriging
GC	Granger causality
GIS	Geographic information system
IDS	Inverse distance squared
IDW	Inverse distance weighting
IK	Indicator kriging
KED	Kriging with external drift
LM	Linear regression models
LST	Land surface temperature
LULC	Land-use/land-cover
MAE	Mean absolute error
MCK	Markov cube kriging
ML	Machine learning
MSI	Moisture stress index
NDBI	Normalized difference built-up index
NDVI	Normalized difference vegetation index
NDWI	Normalized difference water index

NN	Nearest neighbors
OK	Ordinary kriging
OPT	Optimal
PSNR	Peak signal-to-noise ratio
RMSE	Root-mean-square error
RoI	Region of interest
RS	Remote sensing
SemK	Semantic kriging
SemK_mod	Modified semantic kriging
SI	Spatial importance
SK	Simple kriging
SS	Semantic similarity
ST-IDW	Spatio-temporal inverse distance weighting
STIE	Space–time interpolation environment
ST-NN	Spatio-temporal nearest neighbors
ST-OK	Spatio-temporal ordinary kriging
ST-RevSemK	Spatio-temporal reverse semantic kriging
ST-SemK$_{NSep}$	Non-separable spatio-temporal semantic kriging
ST-SemK$_{Sep}$	Separable spatio-temporal semantic kriging
ST-SemK	Spatio-temporal semantic kriging
ST-SK	Spatio-temporal simple kriging
ST	Spatio-temporal
ST-TPS	Spatio-temporal thin plate splines
ST-UK	Spatio-temporal universal kriging
TPS	Thin plate splines
TSA	Trend surface analysis
UK	Universal kriging

Symbols

$\frac{1}{a}$	Inverse of a
$\hat{a}$	Predicted/estimated value of a
a^i	ith power of a
$\mathbf{a}$	Vector a
a_i	ith value of $\mathbf{a}$
$\lvert a \rvert$	Absolute value of a
$\sqrt{a}$	Square root of a
$\int f(a)$	Definite integral of function f of a variable a
$P(a)$	Marginal probability of a
$E\{a\}$	Expected value (or mean) of discrete random variable a
$\gamma(a)$	*Semivariance* for lag interval a
$\sigma(a)$	Error in estimation of random field value at point a
$\lvert A \rvert$	Cardinality of set A
A^T	Transpose of matrix A
A^{-1}	Inverse of matrix A
A^{-H}	*Hadamard* inverse of matrix A
A_{ij}	Value at ith row and jth column of matrix A
$*$	Multiplication
$\times$	Matrix multiplication
$\circledast$	Dot product between matrices
$\circ$	*Hadamard* product
$- \cdot - \cdot -$	*Hadamard* division
$=$	Equals to (equality)
$\neq$	Not equals to (inequality)
$<$	Less than
$>$	Greater than
$\leq$	Less than or equal to
$\geq$	Greater than or equal to
$\approx$	Approximately
$a \rightarrow b$	a is parent of b in a *DAG*

$a \Rightarrow b$	*a* implies *b*
$a\|b$	*a* given *b*
$(a, b]$	Half-closed interval bounded by the limit points *a* and *b*
$\theta_{\mathbf{a},\mathbf{b}}$	A plane angle (in geometry) between **a** and **b**
$\in$	Belongs to
$\ni$	Such that
$\because$	Because
∞	Infinity
$\sum$	Summation operator
$\prod$	Product operator
$\bigcup$	Set union operator
$\mathbb{Z}^+$	Set of positive integers
$\mathbb{R}^+$	Set of positive real numbers
cos	Cosine
cos^{-1}	Inverse cosine

List of Figures

List of Tables

Algorithms

Chapter 1
Introduction

Abstract Advancement of data capturing technology in the field of remote sensing (RS) and geographic information system (GIS) has introduced a significant amount of research challenges. It facilitates enormous availability of spatial data (both in the form of raster and vector) from different sources. This monograph primarily focuses to deal with the incomplete raster satellite imagery of the meteorological parameters and the geostatistical spatial interpolation methods to be applied for the prediction of these parameters.

Advancement of data capturing technology in the field of *remote sensing* (*RS*) and *geographic information system* (*GIS*) has introduced a significant amount of research challenges. It facilitates enormous availability of spatial data (both in the form of raster and vector) from different sources. However, proper staging of this data is necessary as most of the geospatial repositories contain missing and erroneous information. This monograph primarily focuses to deal with the raster satellite imagery, more specifically on the meteorological parameters related to the terrain that can be derived from these satellite (raster) imageries. Examples of these parameters include *land surface temperature* (*LST*), other meteorological indices such as *normalized difference vegetation index* (*NDVI*), *moisture stress index* (*MSI*), *normalized difference water index* (*NDWI*), *normalized difference build up index* (*NDBI*), etc. In case of climatological applications and weather analysis, the meteorological parameters are considered to be the important factors for land–atmospheric interaction analysis, as well as for modeling climate dynamics.

Though data capturing technology of these satellite data have been improved a lot in the past few decades, still these imageries may consist of some missing pixels, line gaps due to faulty sensors, fallacious/negligent post-processing, etc. [7]. This work is mainly using the Landsat-7 ETM+ satellite imagery and Fig. 1.1 shows one raw (band) image of the same (path:138, row:44).[1] It is showing a raw image with some line gaps (line of missing pixels). A zoomed view of a portion is shown to the left

[1] URL: http://landsat.usgs.gov/; Accessed on: July 22, 2016.

S. Bhattacharjee et al., *Semantic Kriging for Spatio-temporal Prediction*, Studies in Computational Intelligence 839,
https://doi.org/10.1007/978-981-13-8664-0_1

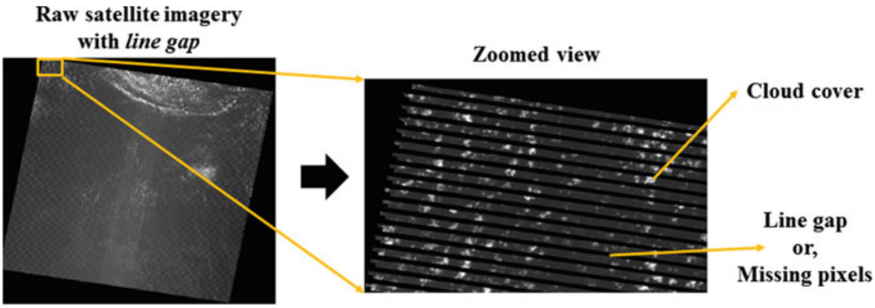
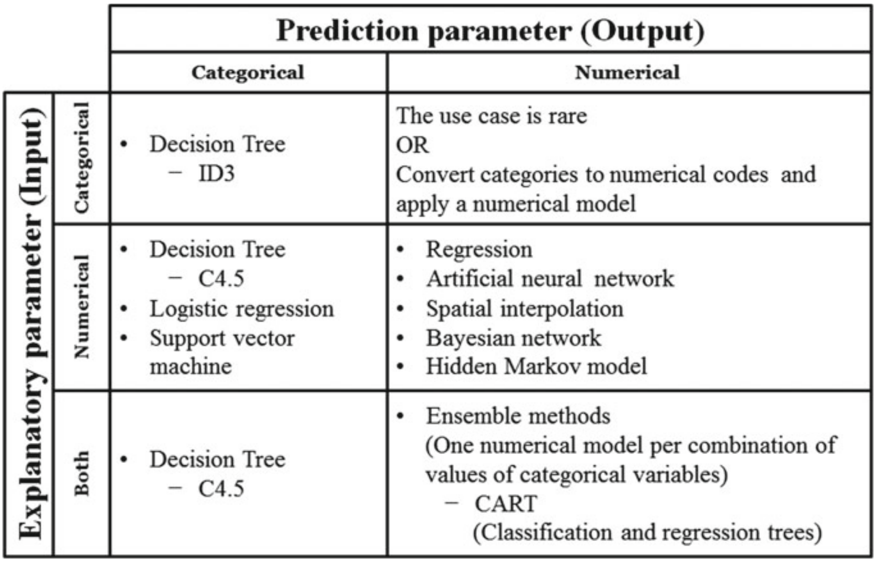

Raw satellite imagery with *line gap*
Zoomed view
Cloud cover
Line gap or, Missing pixels

Fig. 1.1 Satellite imagery with line gap and cloud cover

where these line gaps are more clearly visible, along with some cloud covers (small white patches). The cloud covers are another concern associated with the satellite imagery. It creates an obstacle to calculate the parameter value at those pixels that are beneath the cloud cover.

In this scenario, prediction of spatial parameters with highest degree of accuracy has attracted a significant amount of research interest in this field of study. Further, huge availability of time-series data has invoked some obvious research aspects such as spatio-temporal prediction, forecasting of these parameters and other correlated events such as forecasting of urban landscape, urban planning for future, etc. In this context, there have been reported many approaches for the prediction of missing and erroneous spatial parameters and their forecasting. These are broadly classified into two types: (a) *data mining* (*DM*) and *machine learning* (*ML*) based approaches and (b) spatial interpolation based approaches [15]. Some ensemble techniques have also been reported for the same, combining the advantages of individual categories. Some of the important and well-known prediction techniques that are reported in the literature of spatial analyses include *artificial neural network*, *Bayesian network*, *spatial interpolation* based approaches (*kriging*, *IDW*, etc.), *support vector machines*, *decision trees*, *hidden Markov model*, *regression*, etc. In the context of this study on spatial analysis, mainly prediction, these prediction techniques are broadly divided with respect to the type of explanatory (input) and the predicted (output) parameters involved for that technique. Two types of parameters have been considered here: categorical and numerical. Based on parameter type, Fig. 1.2 introduces the types of different techniques for spatial prediction.

According to the state of the art, the spatial interpolation is reported to be the most efficient choice for the prediction of meteorological parameters [13] that are derived from the satellite raster imagery and highly correlated with the terrestrial dynamics. The geostatistical spatial interpolators are often considered to be the most appropriate methods, which yield minimal error in estimation [14]. One of the major reasons behind this can be stated as the meteorological parameters are distinctive in nature and most of them can be treated as random field parameters showing high spatial autocorrelation [16]. In this context, the geostatistical interpolators are capable to deal with this spatial property efficiently [7], than other *machine learning* based approaches. This monograph mainly concentrates on the geostatistical spatial interpolation methods for the prediction of meteorological parameters.



		Prediction parameter (Output)	
		Categorical	Numerical
Explanatory parameter (Input)	Categorical	• Decision Tree – ID3	The use case is rare OR Convert categories to numerical codes and apply a numerical model
	Numerical	• Decision Tree – C4.5 • Logistic regression • Support vector machine	• Regression • Artificial neural network • Spatial interpolation • Bayesian network • Hidden Markov model
	Both	• Decision Tree – C4.5	• Ensemble methods (One numerical model per combination of values of categorical variables) – CART (Classification and regression trees)

Fig. 1.2 Types of prediction methods for spatial analysis

1.1 Overview of Spatial Interpolation

The basic objective behind the interpolation process of meteorological parameters is to accurately predict a missing pixel or a group of pixels of a surface in the region of interest (*RoI*). In spatial interpolation, the values of the prediction parameter at some pixels in a gridded surface are considered to be missing. Consider the scenario depicted in Fig. 1.3, where the big outer square box represents the *RoI* boundary and each small square represents a pixel. Each hollow pixel (□) is a missing pixel, and small box with a ● inside represents a sampled location (⊡), i.e., the location where the parameter value is present. These points (⊡) can be considered as the interpolating points or the known sampled locations that are utilized to measure each of the missing pixels or prediction point/location or unsampled location (□). In general, each of the sampled locations is assigned an influence value (weight) with respect to the prediction location, in terms of its *Euclidean* distance. This weight evaluation process is such that higher distant interpolating point will have less impact on the prediction point and vice-versa. Hence, assigned weight to the sampled location is inversely proportional to its distance from the prediction point. The spatial interpolation process evaluates the weight vector for N number of sampled location. For most of the interpolation processes, mainly for the *kriging*, these N dimensional vectors are considered as $N \times 1$ dimensional weight matrices.

Differences exist among the spatial interpolation methods based on how the dependencies among the sampled and unsampled locations are measured. This spatial

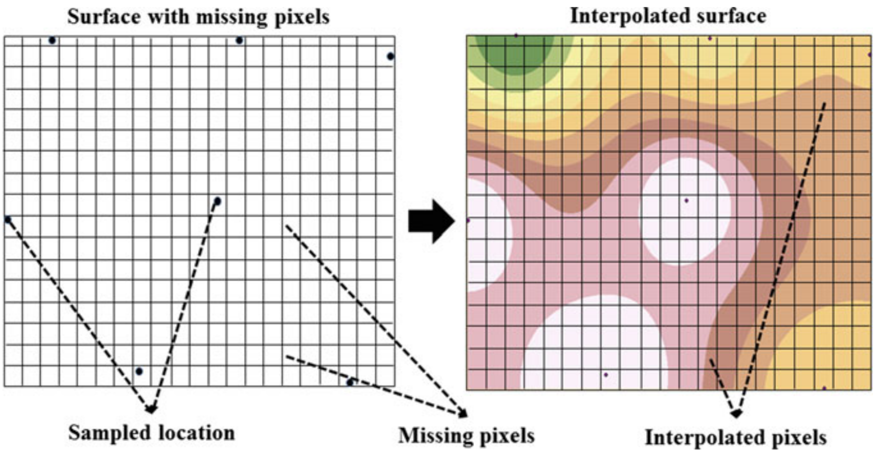


Fig. 1.3 Example interpolated surface

dependency is generally to as spatial autocorrelation. The geostatistical interpolation methods, based on regression analysis, exhibit better performance than other coexisting methods as these methods can model spatial autocorrelation/proximity within the *RoI* and can incorporate that into the regression process. The concept of spatial proximity can be stated by the Tobler's law: "*everything is related to everything else, but near things are more related than distant things*" [16]. It is also referred to as "*first law of geography*". The geostatistical interpolators capture this relatedness/proximity between the locations under consideration most efficiently [7]. However, the non-geostatistical methods do not model complete autocorrelation model for the *RoI*. Hence, for these type of methods, the spatial proximity among the sampled locations are not evaluated, making the interpolation process unpragmatic for real-life applications.

The popular approaches for regression-based spatial interpolation include *kriging* [9], which represents the full family of geostatistical interpolation methods. Some popular members of this *kriging* family for the univariate and multivariate analysis are: *ordinary kriging* (*OK*), *simple kriging* (*SK*), *universal kriging* (*UK*), *kriging with external drift* (*KED*), etc. The popular examples of non-geostatistical interpolation methods, which are most frequently used and compared in the literature, include *inverse distance weighting* (*IDW*), *nearest neighbors* (*NN*), *thin plate spline* (*TPS*), etc. [18]. Though these non-geostatistical interpolation methods are not very pragmatic process to handle spatial autocorrelation, still they are widely used mainly because of modeling simplicity. Among all these spatial interpolation methods, the *ordinary kriging*, followed by *inverse distance weighting* are the two most frequently used, compared, and mostly recommended interpolation techniques [14].

1.2 Research Issues and Challenges in Remote Sensing based Prediction

Although regression-based interpolation methods show better performance for the prediction of meteorological parameters, some important dynamics of the terrain are found to be neglected in the regression process. For the spatial interpolation of the parameters that are nearby to the earth surface and highly influenced by the terrestrial dynamics, the land–atmospheric interaction modeling is crucial for their analysis. One of the obvious examples of these parameters is *land surface temperature* (*LST*), which is highly influenced by the land-use/land-cover (*LULC*) distribution of the terrain. For example, a *building* may absorb and emit more heat than a *waterbody*. Hence, the former will have more impact on the *land surface temperature* of its nearby locations, than a *waterbody*. Many literature, technical reports from several organizations have acknowledged this fact. For the existing interpolation methods, the spatial autocorrelation is modeled depending on the location-based distribution of the sample points, which is the function of *Euclidean* distance. However, the local knowledge of the *RoI* such as representative *LULC* of the sampled and unsampled locations, play an important role for modeling land–atmospheric interaction and to achieve better precision in prediction. Consider the same gridded surface (as depicted in Fig. 1.3), with the *LULC* information of the terrain that is depicted in Fig. 1.4.

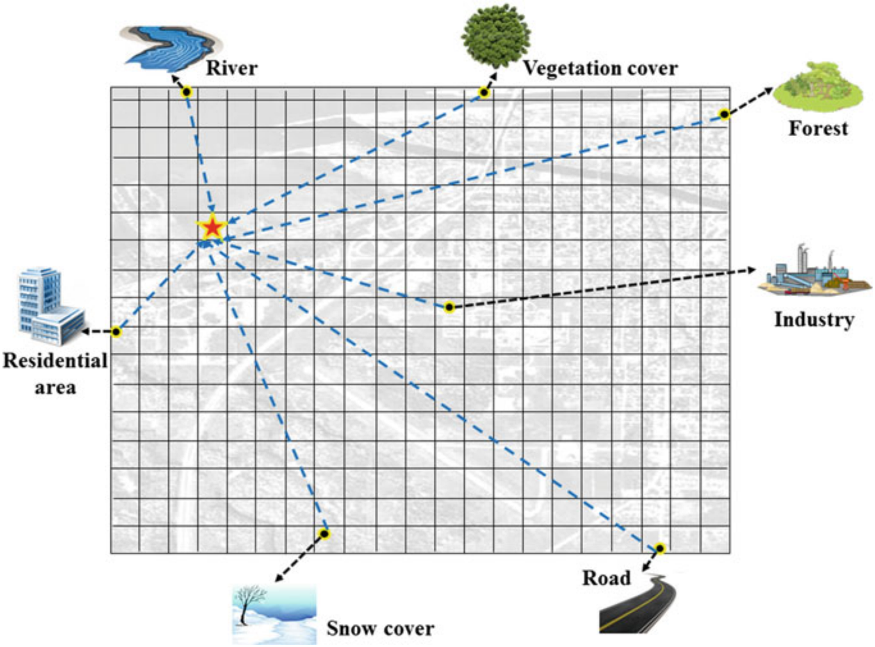


Fig. 1.4 Spatial *RoI* with *LULC* information

Here, each of the sampled locations is not just a point x_i that is represented as (X_i, Y_i) in 2D space. It is also represented by the *LULC* type it corresponds to (f_i) such as *river*, *vegetation cover*, *road*, *industry*, etc. as (X_i, Y_i, f_i). Hence, this surface is more informative than the normal gridded surface. Thus, if the pixel/location that is represented by a red star is the point to be predicted by other sampled locations, the influence of the sampled locations on the prediction point is not only the function of *Euclidean* distance but also the semantic distance between their representative *LULC* classes [7]. Hence, quantification of this semantic contextual knowledge of the terrain, and finding contextual correlation between every pair of the sampled locations is the primary research motivation of this monograph. Hence, the broad scope of research for this monograph can be stated as follows: assuming the fact that the *LULC* distribution of the terrain influences the land–atmospheric interaction for the meteorological parameters (mainly *LST*), the prediction of these parameters should include this contextual knowledge for the estimation process to produce more pragmatic and accurate prediction model.

1.3 Contributions

The existing spatial interpolation methods suffer from the shortcoming of the lack of terrain knowledge, which influences the meteorological parameters significantly. With this scope of further improvement, the broad outline of this monograph can be defined further. This work focuses on the semantic land-atmospheric interaction modeling for the meteorological parameters, where the terrestrial dynamics or the *LULC* distribution is considered as the *semantic* property of a terrain. It is attempted to incorporate this knowledge into an existing interpolation method, to propose a more pragmatic and accurate interpolation model for the meteorological parameters. In this regard, the most popular univariate geostatistical interpolation method, *ordinary kriging* (*OK*) [14] has been extended further with *LULC* modeling. The newly proposed spatial interpolation method is named as: *semantic kriging* (*SemK*) [2]. This monograph is an excerpt from a PhD thesis in the same approach [2]. This basic *SemK* is further extended to an a-posterior probabilistic approach, a spatio-temporal approach, and for forecasting.

The main purpose of this monograph can be stated as: the proposed *SemK*-based framework presents a novel approach to extend any spatial interpolation method (for meteorological parameters) with contextual/semantic *LULC* knowledge of the terrain. According to Fig. 1.2, the existing spatial interpolation methods belong to the category [Numerical, Numerical] only, i.e., both the input and output parameters are numerical. However, the newly proposed *SemK* belong to the category of [Both, Numerical], i.e., it can process contextual knowledge for input as well. Though *OK* has been considered as the base method for this extension, any geostatistical method could have been chosen for this purpose. Therefore, the major contributions of this monograph are as follows:

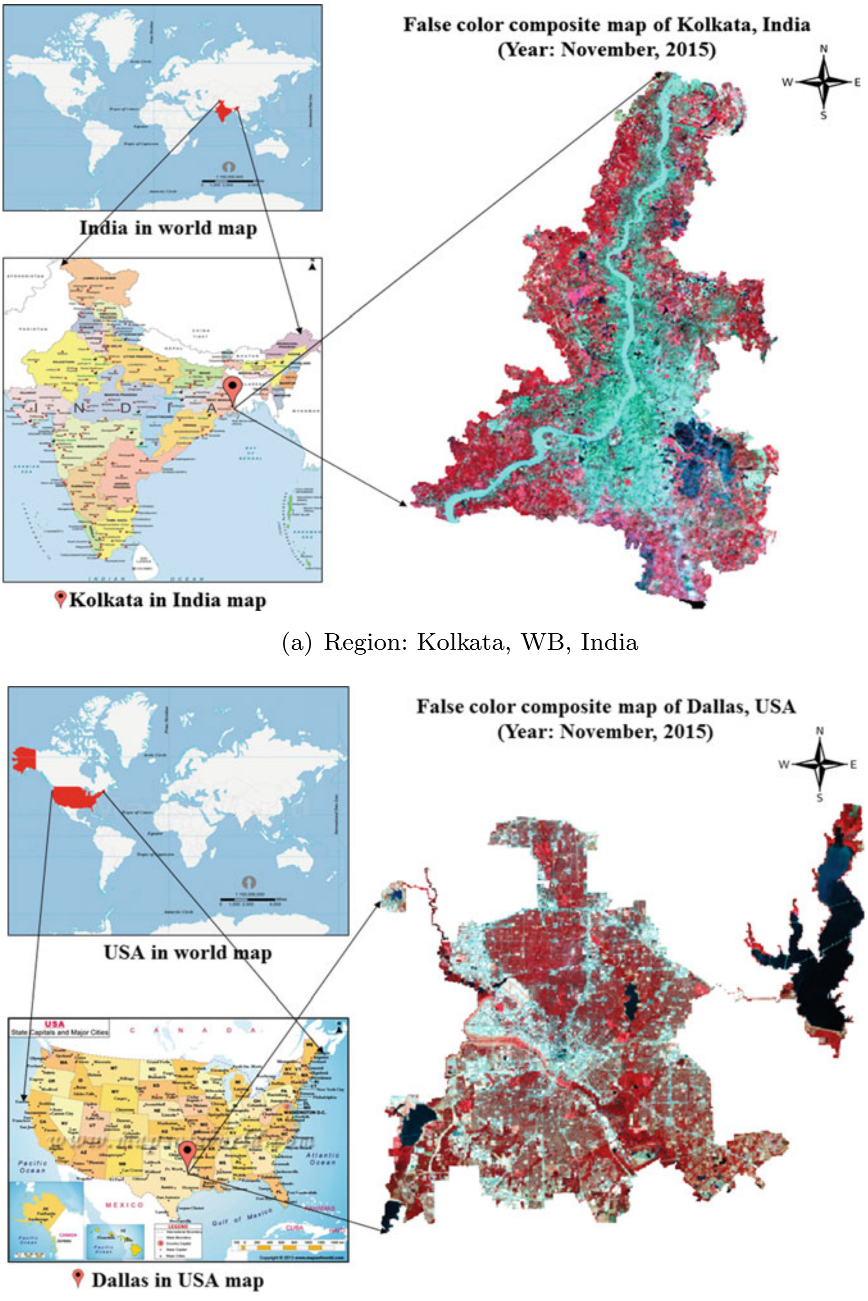
- proposing a spatial interpolation method (*SemK*) for the prediction of meteorological parameters by incorporating the *LULC* knowledge of the terrain.
- improving the basic *SemK* approach by extending its a-priori correlation analysis with a-posterior probabilistic correlation analysis between *LULC* classes (*FB-SemK*).
- extending the notion of spatio-temporal *SemK* process for a forecasting application, which involves the analysis of *LULC* distribution. A new multivariate variant of basic spatio-temporal *SemK* approach is proposed for forecasting urban landscape or future *LULC* distribution pattern (*ST-RevSemK*).

1.4 Specifications of Empirical Study

An empirical study has been conducted with *land surface temperature* [17] data that is derived from satellite imagery. This experimentation is carried out for all the variants of *SemK* process, with some modifications of the basic experimental specifications. The basic common specifications of the experimental study are maintained throughout this monograph. The following subsections present different aspects of the experimental setup.

1.4.1 Regions of Interest

Two spatial regions have been considered as the *RoI*s for this study. One is Kolkata, West Bengal (WB), India, with central coordinate: $22°34'N$ $88°22'E$. This metropolitan area is depicted through Fig. 1.5a. Kolkata, WB, India is subject to a tropical wet-and-dry climate. The annual mean temperature is 26.8 °C and monthly mean temperatures are 19–30 °C. Another *RoI* is Dallas, Texas (TX), USA (central coordinate: $32°46'33''N$ $96°47'48''W$), which is a major city in the state of Texas and is the largest urban center of the fourth most popular metropolitan area in the USA. This metropolitan area is depicted through Fig. 1.5b. Dallas, TX, USA has a humid and hot subtropical climate with mean of temperatures about 39 °C at summer and heat-humidity indexes soaring to as high as 47 °C. Both of the *RoI*s have diverse types of *LULC* distribution for the whole region. Some example common *LULC* classes for both the *RoI*s are given as follows: *residential*, *commercial*, *agriculture*, *lakes*, etc. Five zones from each of these two *RoI*s have been considered for each of the chapters to carry out the estimation of *LST* with different interpolation methods. These zones are considered to reduce the effect of force generalization of spatial autocorrelation for the whole region. It will facilitate accurate modeling of spatial proximity for each of the local zones individually. All these *RoI*s are typically presented in 1:50,000 scale throughout the monograph with Landsat ETM+ datasets.



(a) Region: Kolkata, WB, India

(b) Region: Dallas, TX, USA

Fig. 1.5 *RoI* for empirical study

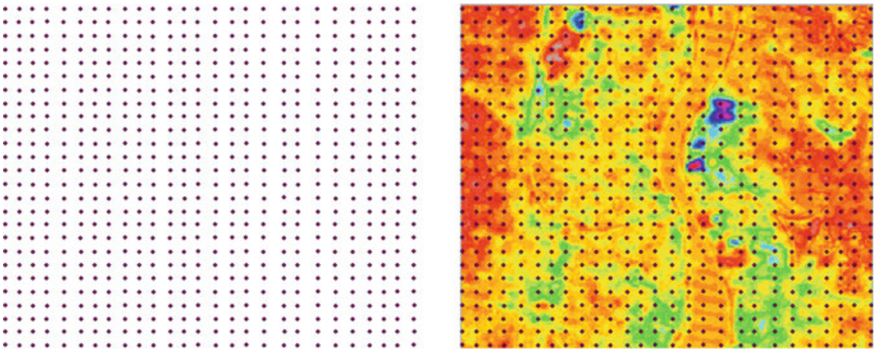
1.4.2 Source of Experimental Dataset

The spatial and spatio-temporal prediction of *land surface temperature* (*LST*) has been considered as the main focus of this experimental study. For this analysis, the *LST* and *LULC* data have been considered as the input to the prediction algorithms. Several other meteorological indexes are also needed to be derived in the intermediate steps of processing *LST* data. The *LULC* forecasting application that is considered for the *ST-RevSemK* framework, require *NDVI* and *MSI* data as well, as the secondary parameters in the multivariate model. All these data can be derived by processing raw satellite imagery. For this purpose, the Landsat-7 ETM+ satellite data is considered that is provided by US Geological Survey[2] (USGS). For the spatial and time-series interpolation, 11 years' dataset (2005–2015) have been considered for both the *RoI*s. To desists the seasonal effect, each of the dataset is considered within the same period (mid-October to mid-November) for each year.

1.4.3 Specifications of Dataset

The Landsat-7 ETM+ satellite imagery for both the *RoI*s are considered from USGS. The metropolitan area of both Kolkata, WB, India and Dallas, TX, USA are depicted in Fig. 1.5a and b, respectively. The pictures are depicted through the standard FCC (false color composite) representation of satellite imagery. Each of the imagery consists of seven spectral bands with spatial resolutions 30 m (for band 1–5, 7) and 60 m (for band 6). Each band is considered to be the raw image information by processing of which the derived parameter value of each of the pixels can be measured. The derivation processes of different meteorological parameters are presented in the following subsections. Once the *LST* value of each of the pixels is derived from the satellite imagery, the values for some pixels are assumed to be missing, where the interpolation is supposed to be carried out with different methods. The actual data for each pixel is needed for comparing the performance of the estimation process with the actual pixel value. In this regard, for each of the zones that are considered from both the *RoI*s, square grids of sampled locations are considered. Here, the number of sampled locations (26×27) are approximately $\frac{1}{70}$ fraction of the total number of pixels (205×241) of each zone. Figure 1.6b depicts one example zone from Kolkata, WB, India on which the square gridded sampled locations (refer Fig. 1.6a) are overlaid, and rest of the locations are assumed to be missing. For this empirical evaluation, all the actual and predicted surfaces for *LST* and *LULC* are represented through two randomly chosen standard symbologies of ArcGIS 10.1 [11], [high ▬ low] and [▬], respectively.

[2] URL: https://www.usgs.gov; Accessed on: July 22, 2016.



(a) Square gridded sampled locations

(b) Sampled locations overlaid on selected zone

Fig. 1.6 Square gridded sampled locations for empirical study

1.4.4 Data Processing Tools

Two standard raster image analysis tools have been considered for this study: ERDAS Imagine [12] and ArcGIS 10.1 [11]. For the extraction of different meteorological parameters such as *LST*, *NDVI*, *MSI*, and *LULC* data from the raw band information of the satellite imagery, the mathematical models for the respective parameters are implemented in ERDAS Imagine. The interpolation processes of the existing methods (such as *NN*, *IDW*, *UK*, *OK*, *TPS*) that are considered for comparison study in this monograph, are carried out with predefined tools in ArcGIS 10.1. The proposed *semantic kriging* and its variants are implemented in MATLAB R2013a. The same experimental specifications (search radius, maximum number of interpolating points, etc.) are considered for each of the interpolation methods for individual experimentation.

For the extraction of meteorological indexes, first, all the seven bands of the satellite imagery of a region are stacked one above other using *Layer Stack* operation in ERDAS Imagine. Consider Fig. 1.7, where image ① represents seven bands overlaid on each other and ② represents the layer stacked imagery, represented though FCC scheme (red: band 4, green: band 3, blue: band 2). This stacked image is useful for extracting different parameters. Images ③, ⑤, and ⑦ are the instances of the mathematical model for the extraction of *NDVI*, *LST* and *MSI*, respectively. The raster imagery for these parameters are shown through images ④, ⑥, and ⑧, respectively. The details of these mathematical models are presented in the following section.

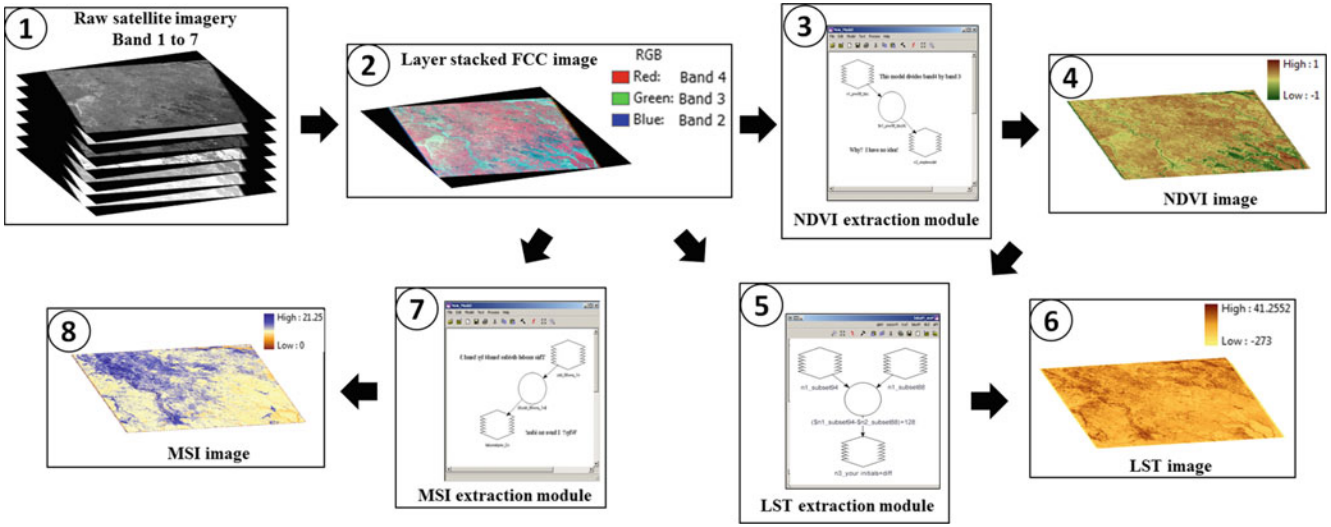


Fig. 1.7 Meteorological parameter extraction from satellite imagery

1.4.5 Extraction of Spatial Parameters

The details of the mathematical models, for the derivation of meteorological parameters and *LULC* distribution of the terrain from the layer stacked FCC image, are described in this section.

1.4.5.1 Extraction of NDVI

The *normalized difference vegetation index* (*NDVI*) identifies the photosynthetic affinity or "*greenness*" of the vegetation through the reflective proprieties of the chlorophyll and mesophyll layers within plants. The *NDVI* value of a given pixel always ranges from minus one (−1) to plus one (+1). A zero means no vegetation, hence, soil, barren surfaces (rock and soil) is observed to have *NDVI* values close to zero. The values close to +1 (0.8–0.9) indicates the highest possible density of green leaves. The water, snow, ice, and clouds are normally associated with negative values. The mathematical formulation of this parameter with respect to Landsat-7 ETM+ satellite imagery is given as follows [10]:

$$NDVI = \frac{Band4 - Band3}{Band4 + Band3} \tag{1.1}$$

1.4.5.2 Extraction of MSI

The *moisture stress index* (*MSI*) is a reflectance measurement that is sensitive for increasing leaf water content. It has direct impact on several applications such as canopy stress analysis, productivity prediction and modeling, fire hazard condition analysis, and studies of ecosystem physiology. The *MSI* is an inverted measure, relative to other water *VI*s. The higher values indicate more water stress and less water content. The value of this index ranges from 0 to 3. The common range for green vegetation is 0.4–2. Considering Landsat-7 ETM+ satellite imagery, the mathematical formulation of this parameter is given as follows [10]:

$$MSI = \frac{Band5}{Band4} \tag{1.2}$$

1.4.5.3 Extraction of LST

All substances in earth surface emit electromagnetic radiation at a temperature above absolute zero (0°K). The temperature of the earth materials and high- temperature phenomena can be estimated based on the thermal emission from these materials. Landsat-7 ETM+ data have the best spatial resolution in thermal band among other commercial satellites. The thermal imaging in this data is determined by two bands, one band is 6.1, where the acquisition the low gain and the second band is 6.2, for which the acquisition will be high gain. The Planck's radiation equation can be applied to convert measured spectral radiance to kinetic temperature (T_K) or the *land surface temperature* (*LST*). For Landsat-7 ETM+ satellite data, it is calculated using *NDVI* derived emissivity models [17]. For this method, the *NDVI* is categorized in three classes of emissivity: *NDVI* < 0.2 is bare soil, *NDVI* > 0.5 is vegetation and *NDVI* between 0.2 and 0.5 represents the category of mixed pixel, which contains combination of vegetation, soil, rock, etc. The kinetic temperature of a given pixel can be given as follows:

$$E = a + b * \ln(NDVI) \tag{1.3}$$

$$T_K = \frac{T_R}{E^{\frac{1}{4}}} \tag{1.4}$$

where E and *NDVI* are the average thermal emissivity and average *normalized difference vegetation index* for individual surface covers, respectively. The a and b are the two constants ($a = 1.0094$ and $b = 0.047$ for a correlation coefficient of 0.941 at 0.01 level of significance), T_R is the radiant temperature [17].

1.4.5.4 Extraction of LULC

The image processing tool ERDAS Imagine [12] is used in order to carry out the classification of the Landsat-7 ETM+ satellite imagery. The supervised classification scheme has been chosen for both the *RoI*s that is carried out using domain experts' knowledge. For the empirical study in this monograph, the considered *LULC* classes are *waterbodies*, *wetlands*, *built-up*, *grassland*, *wastelands*, *agriculture*, *forest*, etc. (refer first level of ontology in Fig. 3.2).

1.4.6 Error Metrics and PSNR

Performance of each of the interpolation methods is specified by two standard error measurement metrics: *mean absolute error* (*MAE*) and *root mean square error* (*RMSE*). The evaluation criteria and physical significance of each metric is discussed in [13]. If *MAE* is closer to zero and *RMSE* is smaller than others, the prediction model can be considered as better than others. If RMSE >1, the method underestimates the primary parameter, else overestimates the primary parameter for prediction.

The mathematical formulations of both *MAE* and *RMSE* are presented in Table 1.1. For these two metrics, N is the number of interpolating points, $\hat{Z}(x_i)$ is the predicted value, and $Z(x_i)$ is the actual or observed value at ith prediction point [13]. For the pixel-by-pixel comparison of the predicted imagery with respect to the actual

Table 1.1 Error metrics and their specifications

Error metric	Definition
Mean absolute error (MAE)	$\frac{\sum_{i=1}^{N} \lvert Z(x_i) - \hat{Z}(x_i) \rvert}{N}$
Root mean square error (RMSE)	$\sqrt{\frac{\sum_{i=1}^{N} [Z(x_i) - \hat{Z}(x_i)]^2}{N}}$

surface, the *peak signal-to-noise ratio* (*PSNR*)[3] is measured against each prediction method. The *PSNR* is a standard metric for predicted image analysis, which is a ratio between the maximum possible power of a signal and the power of corrupting noise. Generally, higher *PSNR* indicates that the prediction is of higher quality. The *PSNR* (in decibel (dB)) is defined as follows:

$$PSNR = 20log_{10}\left(\frac{MAX_I}{RMSE}\right) \tag{1.5}$$

where MAX_I is the maximum possible pixel value of the image.

1.5 Organization of the Monograph

This monograph is organized in seven chapters as follows:

- Chapter 1 discusses the motivation, the scope of the work and summarizes the contributions of this monograph. The need of prediction for the meteorological parameters, overview of the existing spatial interpolation methods, the need for semantic modeling of the *LULC* distribution of the terrain for spatial interpolation are discussed in details. One of the major aspects of this monograph is the empirical analysis with derived *land surface temperature* data for evaluating the performance of the proposed methods and all its variants. The detailed common specifications of the empirical study are presented in this chapter, which are preserved in each of the chapters of this monograph.
- Chapter 2 reviews the existing spatial and spatio-temporal interpolation approaches. This chapter presents an extensive literature survey of different categories of the spatial interpolation methods. Based on their popularity or frequency of being compared in different environmental applications, a popularity graph is proposed. Based on this graph, several groups of these interpolation methods are suggested. The most frequently compared group has been chosen for the comparison with the proposed methodology, whereas, the most frequent method (*ordinary kriging*) in the literature is considered to be extended further to address the scope of this monograph (i.e., incorporating *LULC* knowledge in spatial interpolation process).

[3] URL: https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio; Accessed on: July 22, 2016.

It also reviews the importance of *LULC* analysis and *Bayesian network* learning for spatial prediction.

- Chapter 3 presents the proposed spatial interpolation method *semantic kriging* (*SemK*). Two proposed metrics are formulated, *semantic similarity* and *spatial importance*, which quantify the contextual domain knowledge (*LULC* distribution of the terrain). The theoretical error analysis, empirical and information content based performance evaluations are presented further, in comparison with its base method *ordinary kriging* (*OK*) and others, to prove its efficacy in prediction. The theoretical performance evaluation of *SemK* provides some basic characteristics, evaluates the impact of the granularity of semantic knowledge hierarchy (ontology), its relationship with *OK*, etc.
- Chapter 4 identifies further scope of improvement of spatial *SemK*. From Chap. 3, it is observed that *SemK* presumes the correlation analysis between every pair of *LULC* classes to be a-priori. That is, the effect of other nearby *LULC* classes is ignored while measuring the *spatial importance* between any pair of *LULC* classes. This analysis can be improved further by introducing the a- posterior correlation analysis, by evaluating conditional probability based correlation scores. This new variant of *SemK*, i.e., *FB-SemK* is further compared empirically with other methods and *SemK* as well, to check whether this variant is at all providing further improvement over *SemK*.
- Chapter 5 presents a multivariate variant of spatio-temporal *SemK* for urban landscape modeling. For this application, the spatio-temporal forecasting of *LULC* distribution pattern has been chosen as the case study. The notion of separable spatio-temporal *SemK* is considered for multivariate extension, which learns and models the past behaviors of multiple meteorological parameters and their correlation with *LULC*, for its forecasting. One *causality testing framework* is also proposed as an approach for the preprocessing of meteorological data. It checks and selects those meteorological parameters that are actually causal to the *LULC* pattern, by pruning the rest. This framework is empirically tested with different combinations of parameters' drift and further identifying the best drift.
- Chapter 6 concludes the monograph by summarizing the major contributions and the significance of the proposed work. It presents the interrelationships among *SemK* and its other variants. This chapter also identifies some future research directions that can deploy *SemK* framework (or its variants) for forecasting environmental events.

1.6 Further Discussions

This chapter first identifies the requirement of predicting meteorological parameters in the field of *remote sensing* and *GIS*. It then indicates the limitations of the existing prediction methods. Though land–atmospheric interaction modeling is important to analyze the parameters that are influenced by the earth surface, the population prediction or spatial interpolation methods do not model this interaction yet. For example, *land surface temperature* is highly influenced by the *LULC* distribution of the terrain. Therefore, prediction model for this parameter must incorporate this contextual knowledge. With these limitations of the existing methods, this chapter defines the scope of this monograph and highlights the contributions of the monograph. This chapter also describes the specifications of the empirical experimentation, which have been followed for each of the chapters in this monograph. For example, the region of interest to carry out the empirical experimentation, the metrics considered to evaluate the performance of the interpolation methods, data processing tools, the meteorological parameters' extraction process, etc., are presented in this chapter. Finally, a brief outline of the organization of the monograph has been listed. The next chapter presents a detailed literature survey on different categories of spatial interpolation methods, *LULC* modeling for meteorological data, probabilistic analysis of prediction methods, etc.

References

1. Bhattacharjee S (2015) Prediction of meteorological parameters: a semantic kriging approach. In: 23rd ACM SIGSPATIAL international conference on advances in geographic information systems (ACM SIGSPATIAL 2015) PhD Symposium. ACM, p 1
2. Bhattacharjee S (2016) Semantic kriging: a semantically enhanced approach for spatial interpolation. PhD thesis, Indian Institute of Technology (IIT) Kharagpur, India
3. Bhattacharjee S, Ghosh SK (2015a) Performance evaluation of semantic kriging: a Euclidean vector analysis approach. IEEE Geosci Remote Sens Lett 12(6):1185–1189
4. Bhattacharjee S, Ghosh SK (2015b) Spatio-temporal change modeling of LULC: a semantic kriging approach. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci 1:177–184
5. Bhattacharjee S, Ghosh SK (2016) Measuring semantic similarity between land-cover classes for spatial analysis: an ontology hierarchy exploration approach. Innov Syst Softw Eng 12(3):193–200
6. Bhattacharjee S, Ghosh SK (2017) Semantic kriging. In: Encyclopedia of GIS, vol 2, Springer International Publishing, Cham, pp 1868–1879
7. Bhattacharjee S, Mitra P, Ghosh SK (2014) Spatial interpolation to predict missing attributes in GIS using semantic kriging. IEEE Trans Geosci Remote Sens 52(8):4771–4780
8. Bhattacharjee S, Das M, Ghosh SK, Shekhar S (2016) Prediction of meteorological parameters: an a-posteriori probabilistic semantic kriging approach. In: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, p 38
9. Cressie N (2015) Statistics for spatial data. Wiley
10. Eklundh L, Hall K, Eriksson H, Ardö J, Pilesjö P (2003) Investigating the use of landsat thematic mapper data for estimation of forest leaf area index in southern sweden. Can J Remote Sens 29(3):349–362
11. ESRI (2012) Arcgis, 10.1
12. Imagine E (2006) Erdas imagine tour guides. Leica Geosyst 730
13. Li J (2008) A review of spatial interpolation methods for environmental scientists. Record (Australia. Geoscience Australia), Geoscience Australia
14. Li J, Heap AD (2011) A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol Inf 6(3):228–241
15. Li J, Heap AD, Potter A, Daniell JJ (2011) Application of machine learning methods to spatial interpolation of environmental variables. Environ Model Softw 26(12):1647–1659

16. Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Econ Geogr 46:234–240
17. Van de Griend A, Owe M (1993) On the relationship between thermal emissivity and the normalized difference vegetation index for natural surfaces. Int J Remote Sens 14(6):1119–1131
18. Zimmerman D, Pavlik C, Ruggles A, Armstrong MP (1999) An experimental comparison of ordinary and universal kriging and inverse distance weighting. Math Geol 31(4):375–390

Chapter 2
Spatial Interpolation

Abstract This chapter presents a background study on spatial interpolation methods, its extended variants for spatio-temporal interpolation and some probabilistic approaches for different spatial analyses. It also focuses on the issues of modeling terrestrial dynamics for meteorological applications. Modeling *LULC* knowledge of the terrain, evaluating semantic associations between them and enabling interoperability among the spatial data sources, have been studied extensively for spatial applications. For spatial interpolation methods and its variants, one frequency of comparison graph or a popularity graph is proposed, depicting their frequency of being chosen for comparative analysis in 85 selected articles. This facilitates us to identify the most popular, moderately popular, least popular groups of spatial interpolation methods. The most popular group members can further be chosen for the empirical comparison with the proposed approach. A brief description of each of those methods (of the most popular group) is also presented here.

2.1 Introduction

Continuous remote sensing data or raster surfaces play a significant role in several geospatial applications such as prediction and forecasting, urban planning, risk assessment, environmental decision- making, etc. However, in case of satellite raster imagery, it is not always readily available. In most of the cases, these imageries contain a large amount of missing pixels, line gaps due to faulty sensors, erroneous post processing, security reasons, etc. Beside satellite imagery, the other sources capture the environmental, meteorological, terrestrial data in some point locations. However, for different real-life applications, it is indispensable to obtain continuous data over the surface of the study region to make effective, accurate decisions to obtain justified end results [37, 64]. Spatial interpolation is one of the most effective choices to generate continuous field data from some observed locations. The state of the art identifies the interpolation as the most popular and widely used technique for

S. Bhattacharjee et al., *Semantic Kriging for Spatio-temporal Prediction*, Studies in Computational Intelligence 839,
https://doi.org/10.1007/978-981-13-8664-0_2

predicting missing pixels. Spatial interpolation is also necessary for the following situations [15, 59]:

- the discretized surfaces have different types of resolution, cell size/shape, and orientation.
- the continuous surfaces are represented by different data models other than that is suitable for the given application.
- the discretized surface do not extend to the complete region of interest.

These are the situations when spatial interpolation methods provide mathematical models to predict the missing environmental parameters at unsampled locations. Literature report many interpolation methods that are applied in various disciplines [107], e.g., *agricultural science*, *meteorology water resources*, *ecology*, *marine science*, and many other fields. A detailed review in spatial interpolation in reported by Li et al. [59], which is further extended in [60]. According to [59], the spatial interpolation methods are broadly classified into three categories: non-geostatistical, geostatistical, and ensemble/combined methods. The geostatistical methods formalize the complete spatial autocorrelation model and utilize it for the interpolation process. The non-geostatistical ones do not model the autocorrelation property of the terrain, or may sometime model a partial autocorrelation model. The geostatistical methods are considered to be the most pragmatic and popular over others. These methods are further divided into two categories: univariate and multivariate. In geostatistics, the methods that can utilize multiple secondary parameters' information in the prediction process are “multivariate" methods, whereas the methods that do not consider auxiliary parameter are referred to as “univariate” methods. Some combined methods are also reported in the literature, in which different *data mining* (*DM*) and *machine learning* (*ML*) techniques are combined together with spatial interpolation methods in order to achieve better estimation accuracy. However, there exist other classification criteria as well, with respect to which the interpolation approaches are categorized as follows:

- *Point interpolation versus areal interpolation*: Spatial data is generally available the form of points (or pixels) or a region (a combination of pixels). The point interpolation predicts the parameter value at a certain point, which can be determined from the nearby point estimations. For areal interpolation, the estimation is carried out for a whole region on an average.
- *Global interpolation versus local interpolation*: The global methods considers all the available data in the region of interest and prediction is carried out with a general trend model for the whole region. On the other hand, local methods operate a small range of the search window from the whole region of interest and estimation is carried out considering local trend models around the prediction point.
- *Exact interpolation versus inexact interpolation*: Interpolation techniques that yield the predicted parameter value of a point exactly the same as the observed value is called an exact method. Otherwise, it is called the inexact interpolation. All the interpolation methods try to converge from inexact to exact solution.

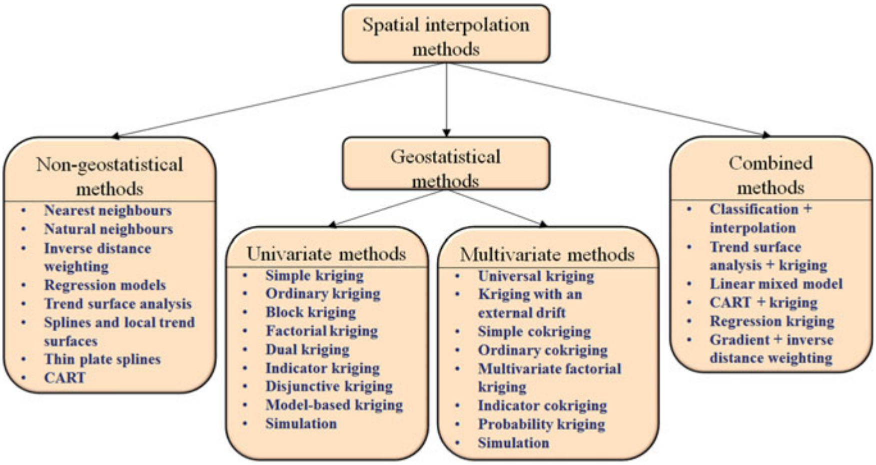


Fig. 2.1 Categorical hierarchy of spatial interpolation methods

- *Deterministic interpolation versus stochastic interpolation*: Stochastic methods deals with the randomness of the interpolating points and associated errors. For these methods, uncertainties are represented as the estimated variances. Deterministic methods do not assess errors of the predicted values.
- *Gradual interpolation versus abrupt interpolation*: Depending on some criteria such as simple distance relations, minimization of variance, curvature and imposition of smoothness, the spatial interpolation methods are categorized into the following two categories. If the interpolation approach produces a discrete and abrupt surface, it is called abrupt interpolation. The methods which produce a gradual and smooth surface are referred to as gradual interpolation.
- *Linear interpolation versus nonlinear interpolation*: Linear interpolation methods assume that the samples are normally distributed. On the other hand, the nonlinear methods are carried over the transformed values of the observed data.
- *Irregular interpolation versus regular interpolation*: An unsampled location can be interpolated based on either the regular gridded samples or irregular samples. The regular grid system has several advantages over irregular one.

The spatial interpolation methods can be extended further with time-series data to model spatio-temporal interpolation methods [23]. For spatio-temporal interpolation, “time” is another dimension, which is considered to model autocorrelation among the sampled locations. For spatio-temporal analysis, each of the sampled locations is characterized by their coordinate locations and measurement time instance, whereas, in case of purely spatial methods, all the sample points are measured in the same time instance. Figure 2.1 presents a hierarchy of popular spatial interpolation methods as per their category.

Software packages: The software packages that are currently available for environmental data interpolation are given as follows: ArcGIS, R-Analysis of Spatial

Data, uDIG, gvSIG, VolPack, GRASS GIS, SAGA GIS, IDRISI Taiga, TNTMips, Quantum GIS, HidroSIG, MapWindow GIS, Surfpack, GrADS, GSLIB, JUMP GIS, etc. [3].

2.2 Objectives to Review Spatial Interpolation Methods

This chapter aims to describe some of the popular spatial interpolation methods and their state of the art. This work attempts to choose some interpolation methods that are widely used in the literature and use them for comparative study. For this purpose, a popularity graph of spatial interpolation methods is presented, which is an extension of the study provided by Li et al. [59, 60]. This may also help us choose the best method, which can be extended with terrestrial *LULC* information modeling [7]. Therefore, the granular objectives are stated as follows:

- to categorize the spatial interpolation methods and understanding each of these categories.
- to present a popularity graph of the interpolation methods to choose the most popular one to be extended further for the proposed method.
- to choose the most popular group, which can be considered for comparison study.
- to describe the selected methods in details and present their state of the art.
- summarization of their drawback and define the scope of the proposed work.

2.3 Spatial Interpolation Methods and Their Popularity

In 2008, Li et al. [59] have reported a graph representing the frequencies with which few popular spatial interpolation methods are considered in 51 reviewed literature. Further, this frequency analysis is extended in [60] by the same authors considering more recent studies. Similarly, in this chapter, we have extended this study with a few more recent works and compared 85 articles that are applied for different environmental applications. The new popularity graph is shown in Fig. 2.2.

Therefore, the spatial interpolation methods are broadly classified into four groups with respect to their frequency in various literature. The first group consists of the most frequently reviewed and compared methods, *ordinary kriging* (*OK*), followed by *inverse distance weighting* (*IDW*), *nearest neighbors* (*NN*), *universal kriging* (*UK*), and *thin plate splines* (*TPS*). The next group consist of *linear regression model* (*LM*), *kriging with external drift* (*KED*), *simple kriging* (*SK*), *trend surface analysis* (*TSA*), *splines*, etc., which are the methods with medium frequency. The third group includes *classification* (*Cl*), *akima's interpolator* (*AK*), *block kriging* (*BK*), *indicator kriging* (*IK*), etc., which were less frequently compared methods. The last group consists of the remaining methods that are very less frequently considered (not included in this work).

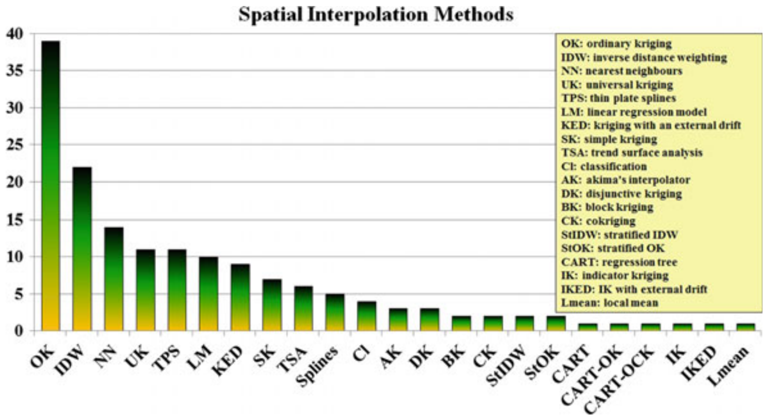


Fig. 2.2 Spatial interpolation methods and their popularity [6]

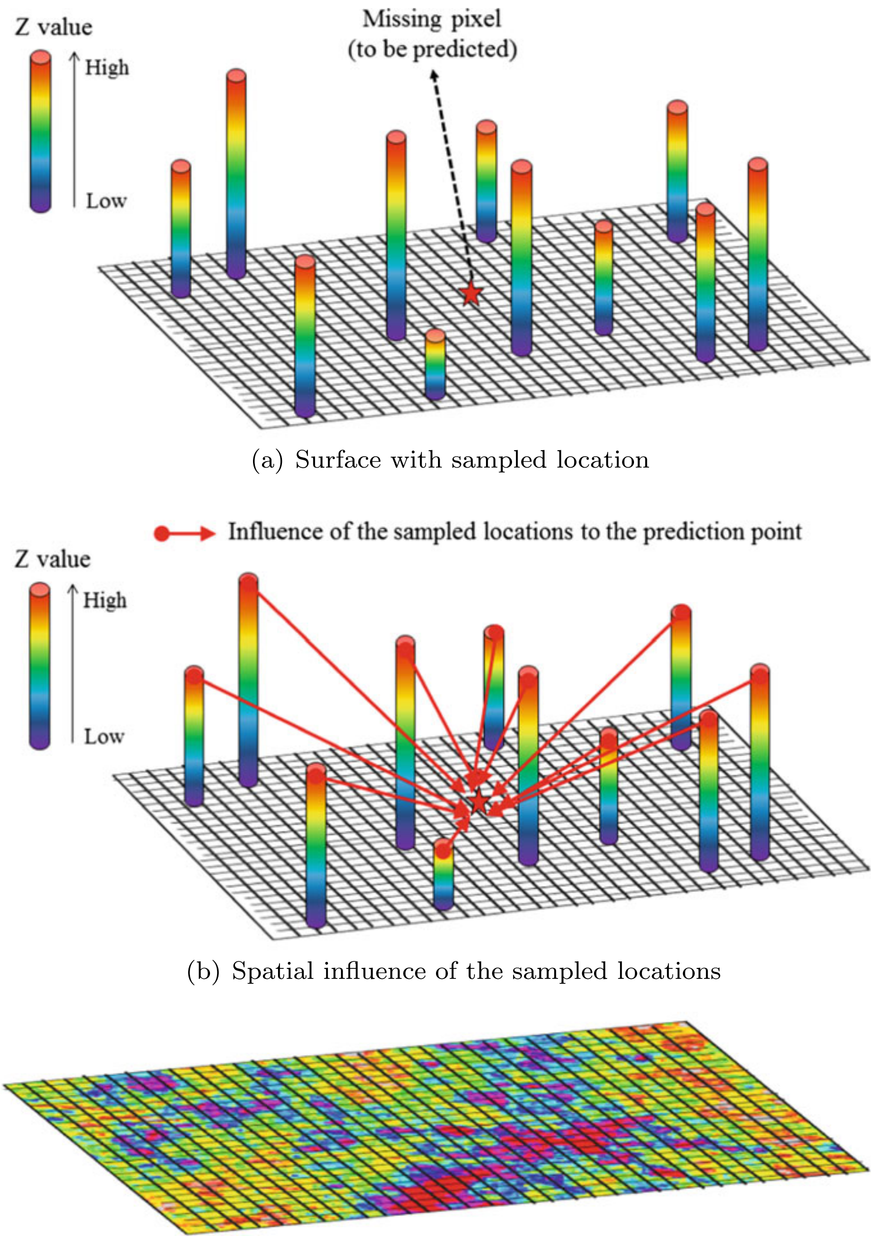
As shown in the frequency graph of Fig. 2.2, the *ordinary kriging* method is the most frequently considered method in the literature of spatial interpolation. For the proposed work, the *ordinary kriging* has been extended further (in proposed *SemK*) with semantic knowledge for the prediction of meteorological parameters. The methods belonging to the first group, i.e., *OK*, *IDW*, *NN*, *UK*, and *TPS* are considered in this monograph for comparison study.

2.4 Popular Techniques of Spatial Interpolation

The basic idea of spatial interpolation (discussed in Chap. 1) is depicted in Fig. 2.3. A gridded surface is depicted in the figure, where some points are sampled (represented as 3D pipes). The missing pixel is represented by a red star (★). The influence of the sampled locations to the prediction point is shown in Fig. 2.3b using red arrow (⟶), where higher the length of the arrow, lesser the influence. For spatial interpolation, this influence is based on *Euclidean* distance. By evaluating the optimal influence (weight), the interpolation method predicts each of the missing pixel in the grid and produces a complete surface (as shown in Fig. 2.3c). Therefore, pragmatic evaluation of this optimal weight for each interpolating point is the fundamental objective of spatial interpolation.

Several interpolation methods have been developed to be applied in various disciplines [55]. The estimation approaches of the spatial interpolation methods can be stated as the weighted average of the sampled values. The general estimation equation of spatial interpolation approach is given as follows:

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_i Z(x_i) \tag{2.1}$$



(a) Surface with sampled location

(b) Spatial influence of the sampled locations

(c) Interpolated surface

Fig. 2.3 Spatial interpolation process

Here, $\hat{Z}(x_0)$ represents the estimated value of the prediction parameter Z at the unsampled location x_0, $Z(x_i)$ is the actual value at the interpolating point x_i, w_i is the weight of the interpolating point x_i and N is the number of sample points [7]. Hence, for N interpolating points, a N dimensional vector is referred to as weight vector $\mathbf{W}$. In this work, each of the N dimensional vectors is considered as the matrix of dimension $[N \times 1]$. Some of the well-known interpolation methods (the first group in Fig. 2.2) are described in the following subsections, which have been used for the comparison study with the proposed methods in subsequent chapters. All of them use the variants of the above equation, modified as per their model.

2.4.1 Nearest Neighbors (NN)

The *nearest neighbors* (*NN*) method [58] is a type of non-geostatistical interpolation method [71]. This method predicts the parameter value of a missing pixel based on the nearest sampled location, by drawing perpendicular bisectors between the interpolating points. Therefore, it generates one Voronoi polygon with respect to each interpolating point. The point is assumed to be in the center of the polygon such that within each polygon, all the points are nearer than any other points of the rest of the polygons [59, 100]. The parameter value at prediction points within the polygon V_i is the actual value at the nearest interpolating point x_i, i.e., $\hat{Z}(x_0) = Z(x_i)$. Hence, the weights can be formulated as follows:

$$w_i = \begin{cases} 1 & : x_i \in V_i \\ 0 & : otherwise \end{cases}$$

For *NN*, all interpolating points within a particular polygon are assigned same weight [15, 100].

2.4.2 Inverse Distance Weighting (IDW)

Same as *NN*, this method (*IDW*) is also a non-geostatistical interpolation method. It is the second most applied interpolation method in environmental science [45, 69]. In *IDW*, the estimates are made based on sampled locations that are weighted with respect to *Euclidean* distance based proximity to the interpolation point. The weight assigned to each of the interpolating points is the inverse of its *Euclidean* distance from the prediction point. Therefore, the nearby points are assigned more weights compared to the distant points and vice versa. The sample points are assumed to be not related to each other. The basic method of *IDW* approach is known as *Shepard method* [84]. The estimated value $\hat{Z}(x_0)$ at the prediction point (x_0) is expressed as follows:

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \frac{w_i x_i}{\sum_{j=1}^{N} w_j} \tag{2.2}$$

Here, N is the number of sampled locations, x_i is the parameter value at the ith location, and w_i is the weight of the ith interpolating point. The assigned weight by *IDW* is expressed as follows:

$$w_i = \frac{1}{d(x_0, x_i)^p} \tag{2.3}$$

where $d(x_0, x_i)$ is the distance between the prediction point x_0 and the ith interpolating point x_i, N is the number of interpolating points, p is the power factor, defined as the rate of weight reduction with respect to increasing distance [74]. The p value depends on the dimension of the interpolation space. For example, for two-dimensional space, $p \leq 2$.

The *inverse distance squared* (*IDS*) [99] is a specialized form of the *IDW* interpolation method. For *IDS*, the specialized weight function is given as follows, where r is the search radius. For the experimental case studies, a specialized version of *IDW* is considered, for which $p \leq 2$.

$$w_i = \left(\frac{r - d(x_0, x_i)}{d(x_0, x_i)}\right)^2 \tag{2.4}$$

2.4.3 Ordinary Kriging (OK)

This method is type of univariate geostatistical interpolation method, named after Krige [83]. Among deterministic interpolation methods, *kriging* [79, 88] is the most popular approach based on linear regression. This method has been studied extensively for the past few decades [33, 88]. It represents the family of generalized least-square regression based interpolation methods. It aims to minimize mean squared error in prediction. It is considered to be better than the existing interpolation techniques due to its modeling of underlying spatial relationships among sampled locations. Unlike *IDW*, the sampled locations are not considered to be independent, but the underlying spatial autocorrelation impacts their behavior and relation. The variants of *kriging* [92] conforms to the following Eq. 2.5, with customized modifications of it:

$$\hat{Z}(x_0) - \mu = \sum_{i=1}^{N} w_i [Z(x_i) - \mu(x_0)] \tag{2.5}$$

Here, $\hat{Z}(x_0)$ is the predicted parameter value at point x_0, μ is the constant mean value over the region of interest (*RoI*) [96].

The *OK* assumes the stationarity of first moment of the prediction parameter, that is $E\{Z(x_i)\} = E\{Z(x_0)\} = \mu = \mu(x_0)$ and μ is unknown. The w_i is the assigned weight to the i^{th} sample point, N is the number of interpolating points that depends on the search window size. The w_i is measured from the experimental *semivariogram*. The *semivariance* $(\gamma(h))$ approximates the underlying relationships and the *Euclidean* distance based spatial autocorrelation. It is half of the *variance* of the difference between the parameter values of sample points that are h lag distance apart. The *semivariance* $(\gamma(h))$ of Z between two sample points that are h distance apart is given as follows:

$$\gamma(h) = \frac{\sum\limits_{i=1}^{N}[Z(x_i) - Z(x_i + h)]^2}{2M} \tag{2.6}$$

Here, $\gamma(h)$ represents the *semivariance* at lag interval h, $Z(x_i)$ is measured parameter value at a point x_i, $Z(x_i + h)$ is the measured parameter value at a sampled location which is h lag distance apart from x_i, M is the total number of pairs of the interpolating points that are h distance lag apart. In two-dimensional space, the *covariance* is a function of *Euclidean* distance between any pair of sampled locations, which is modeled through a *semivariogram* model. The experimental *semivariogram* represents a trend analysis plot of *semivariance* $(\gamma(h))$ with respect to lag distance (h) between known sampled locations. It reveals several important characteristics of the terrain [15]. One example experimental *semivariogram* of spatial region Kolkata, WB, India (year: 2015) is shown in Fig. 2.4. The fitted curve is exponential distribution. The important specifications of the *semivariogram* model are: nugget, range, partial sill, sill, etc. From this fitted model, the *semivariance* between every pairs of interpolating and interpolation points are estimated with respect to their respective spatial lag distance [7].

Let us consider $\epsilon(x_0)$ to be the error in estimation process of the prediction parameter value Z at the interpolation point x_0. Further, if $\hat{Z}(x_0)$ and $Z(x_0)$ are the predicted and the actual parameter values at x_0, then $\epsilon(x_0)$ is expressed as follows:

$$\epsilon(x_0) = \hat{Z}(x_0) - Z(x_0) \tag{2.7}$$

$$= \sum_{i=1}^{N} w_i Z(x_i) - Z(x_0) \tag{2.8}$$

where w_i is the assigned weight to the ith sample point and $Z(x_i)$ is the parameter value at x_i. For *OK* method, the stationarity of the random function implies that the expected value of error is zero. Thus the following holds:

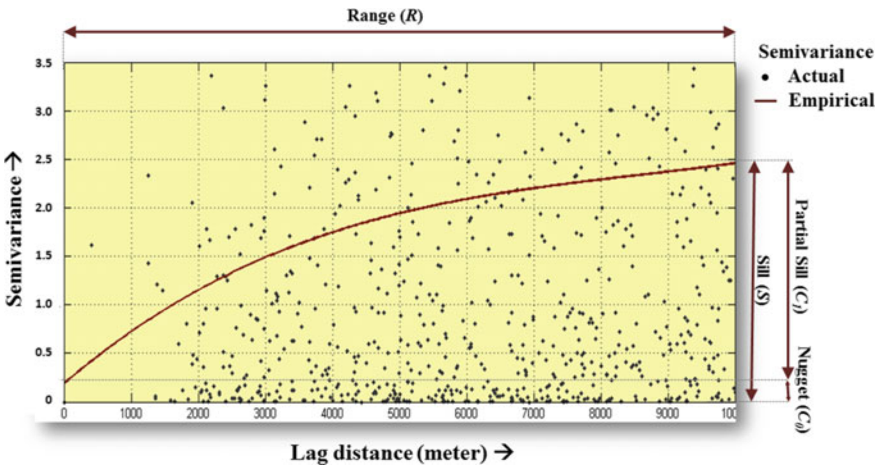


Fig. 2.4 An example experimental *semivariogram* [8]

$$E\left(\epsilon(x_0)\right) = 0 \tag{2.9}$$

$$\sum_{i=1}^{N} w_i E(Z(x_i)) - E(Z(x_0)) = 0 \tag{2.10}$$

$$\mu \sum_{i=1}^{N} w_i - \mu = 0 \tag{2.11}$$

$$\sum_{i=1}^{N} w_i = 1 \tag{2.12}$$

$$\mathbf{1}^T \mathbf{W} = 1 \tag{2.13}$$

Thus, the following equation can be considered as the general estimation approach by *ordinary kriging*, constrained by $\mathbf{1}^T\mathbf{W} = 1$, where the weight vector of size N, $\mathbf{W}$ is expressed as $[w_1 w_2 \cdots w_N]^T$.

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_i Z(x_i) \tag{2.14}$$

2.4.4 Universal Kriging (UK)

The *universal kriging* (*UK*) [88] belong to the group of multivariate geostatistical interpolation methods. In contradiction with *ordinary kriging* in terms of stationarity of the first-order moment, for *UK*, the *mean* is the function (β) of the coordinate location (X, Y) of the interpolating points, i.e., of each x_i [4]. This trend function can be in linear, quadratic, or higher form. For example, the Z value at the interpolating point x_i can be given (in linear and quadratic forms respectively) as follows:

$$\hat{Z}(x_i) = \beta_0 + \beta_1 x_i(X) + \beta_2 x_i(Y) + \delta(x_i) \tag{2.15}$$
$$\hat{Z}(x_i) = \beta_0 + \beta_1 x_i(X) + \beta_2 x_i(Y) + \beta_3 x_i(X)^2 + \beta_4 x_i(X, Y) + \beta_5 x_i(Y)^2 + \delta(x_i) \tag{2.16}$$

where β_is are the unknown trend coefficients, $\delta(x_i)$ is the stochastic component at the location x_i. Hence, assuming the trend of linear form and considering the number of interpolating points to be N, the estimated value at x_0, i.e., $\hat{Z}(x_0)$, with constraint $\mathbf{1}^T\mathbf{W} = 1$, is given as follows:

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_i Z(x_i) \tag{2.17}$$
$$= \beta_0 + \beta_1 \sum_{i=1}^{N} w_i Z(x_i)(X) + \beta_2 \sum_{i=1}^{N} w_i Z(x_i)(Y) + \sum_{i=1}^{N} w_i \delta(x_i) \tag{2.18}$$

2.4.5 Thin Plate Splines (TPS)

The *thin plate splines* (*TPS*) [11, 35] is a non-geostatistical interpolation method, which is formally known as *Laplacian smoothing splines*. This method was initially proposed in [98] for climatic data. A *thin plate spline* is mainly applied for the tabulated data that are arbitrarily spaced. The *TPS* can be regarded as a generalized natural cubic splines in one-dimensional space [26]. However, it can be applied for any dimensions with customized modifications.

In n-dimensional space, the aim of *TPS* is to select a function $f(x)$ that executes exact interpolation of the given sample points (x_i, y_i) and minimizes bending energy as given in the following equation:

$$E\{f(x)\} = \int_{\mathbb{R}^n} |D^2 f|^2 dX \tag{2.19}$$

Here, $|D^2 f|^2$ represents the sum of squares of the elements of matrix $D^2 f$. It is a matrix of second-order partial derivatives of f. The minor element of the hyper-volume is dX, given as $[dx_1 dx_2 \cdots dx_n]$ and x_is represents the components of x. A smoothing factor is introduced here for regularization of the interpolated surface [97]. A function f can be chosen such that it may approximate the interpolation process and does minimize the following Eq. 2.20 as follows:

$$E\{f(x)\} = \sum_{i=1}^{N} |f(x_i) + y_i|^2 + \rho \int_{\mathbb{R}^n} |D^2 f|^2 dX \tag{2.20}$$

Here, N is the total number of data points and ρ is the smoothing factor, always greater than 0 [26]. The spatio-temporal modeling of *TPS* that has been considered for the empirical comparison in this study has been specified in [35].

2.5 Background

This section presents the existing literature of relevant spatial and corresponding spatio-temporal interpolation methods. Mainly, the background study on five mostly recommended spatial interpolation methods are presented. The *kriging*, mainly the *ordinary kriging* and other methods, considered for the empirical comparison study with the proposed approach, are studied here. Further, the literature which identify the importance of *LULC* for different meteorological parameters are reported. Few studies focusing on the probabilistic analysis in several domains by Bayesian and fuzzy *Bayesian network* learning, are presented further.

2.5.1 Spatial Interpolation

Spatial interpolation methods are applied in various disciplines. The performance of the interpolation methods is data specific and also dependent on the application type. Many factors that affect the efficiency of the interpolation methods can be stated as follows:

- sample size, density, and spatial distribution
- distance from the prediction point
- type of surface/terrain
- data variance and normality
- stratification
- quality of auxiliary information
- grid size/resolution
- spatial proximity (autocorrelation) modeling approach, etc.

Li et al. [59] have reported a detailed review on spatial interpolation methods demonstrating each of their performance in environmental science. This study reports *ordinary kriging* to be the mostly applied and compared interpolation method, followed by *IDW*, *IDS*, and others. This analysis is taken forward in [60] resulting *ordinary kriging*, followed by *inverse distance squared* being the two mostly compared interpolation methods in the same field. They have also reported in their comparative study that the *nearest neighbors* is the most accurate method with error metric *relative* *root mean square error RRMSE*. Nalder et al. [69] have compared four interpolation methods, *nearest neighbors*, *inverse distance weighting*, *universal kriging*, and *ordinary kriging*, along with other methods. They have applied this method for monthly precipitation and temperature prediction in Western Canada. The performance is evaluated by *mean absolute error* metric and they have found that the *nearest neighbors* has performed better than *inverse distance squared*, followed by *universal kriging* and then *ordinary kriging*. Boer et al. [10] have compared four types of kriging (*OK*, *cokriging*, *regression kriging*, and *trivariate regression-kriging*) and three types of *TPS* (*bivariate TPS*, *trivariate TPS*, *partial TPS*) to estimate monthly mean precipitation and maximum temperature in Jalisco, Mexico. They have reported that the *trivariate TPS* and *partial TPS* produce comparable results and perform better than other approaches with respect to *mean square error*. Franzen et al. [31] and Weisz et al. [101] have reported that the *kriging* along with *inverse distance squared* to be the most widely considered methods in *GIS*. Several other works have also compared the two most frequently used methods (*OK* and *IDW*) for their own applications [73, 78]. In some cases, *kriging* outperforms *inverse distance weighting* [50, 78]. Kravchenko [50] has studied spatial autocorrelation to analyze the grid soil sampling with various sampling density considering these two processes. The *kriging* with known *semivariogram* has performed significantly better than *IDW* mostly studied applications [51]. Further, *IDW* shows better performance than *kriging* in some other studies as well [69]. Mueller et al. [68] have reported *IDW* based interpolation to perform generally equally or sometimes better than *kriging* considering optimal factors [52, 68]. However, some mixed results are also reported in the literature by [54, 67, 81]. Schloeder et al. [81] have reported the *OK* and *IDW* to perform similarly. Kuilenburg et al. [93] have analyzed and compared three methods, *OK*, *IDW* and *NN* for agricultural and soil science application. They have found that *OK* is the most preferred one over others. Laslett et al. [55] and Brus et al. [14] have also compared their performances to predict different soil properties and *OK* has reported minimal error. Odeh et al. [72] have considered *OK* and *UK* for predicting soil properties along with some other interpolation methods such as *isotopic cokriging*, *heterotopic cokriging*, *multi-linear regression*, and other *regression kriging* methods. Though other methods with covariates information performed better than *OK* and *UK*, however, *UK*'s performance was better than *OK* in terms of *RMSE*. Zimmerman et al. [108] have compared *OK*, *UK*, and two variants of *IDW* for generating interpolated surface with different data and sampling properties such as sampling pattern, surface type, noise level, and small-scale spatial correlation strength. They have found both the *kriging* methods are superior to the *IDW* variants. Teegavarapu et al. [91] have proposed some modifications to the weighting factors and surrogate measures of distances for *IDW* and compared these modified methods with *kriging*, revised *NN*, and other methods such as *artificial neural network* for missing rainfall events. They have observed that in terms of *RMSE*, the modified *IDW* performs better than other methods. Different variants of *nearest neighbor* method have been extensively studied by LeMay et al. in [58] for predicting a number of ground and Ariel variables. Hutchinson [43] has applied partial *TPS* and its variants for measuring annual mean rainfall and compared their performances. They have declared that the *trivariate spline* is the most efficient choice for their application. Harder et al. [36] have compared *spline* interpolation process with the interpolations by 21-term *least-squares polynomials* and indicated us with the superiority of the *spline* method.

Many recent works have also applied spatial interpolation techniques for prediction. Among those contemporary works, Yasrebi et al. [105] have compared *OK* and *IDW* to determine degree of spatial variability of soil chemical properties. The *OK* performed much better compared to the *IDW* method in this study. According to Karydas et al. [45], *IDW* variants are the most often considered techniques for several applications. Foster et al. [30] have reported that the *kriging* produces accurate results in many applications, however, other interpolation approaches such as *NN* also produce better results for reconstructing total electron content (*TEC*) of the ionosphere images. Rayitsfeld et al. [77] have examined a pair of methods for measuring rainfall from microwave links. One of them is *inverse distance weighting* and another is proposed in [66]. They found a comparable result for both the methodologies. Chen et al. [18] have examined the relation between interpolation accuracy of *IDW* and two significant factors of it, power and the search radius of influence for the prediction of the rainfall. Bhowmik et al. [9] have compared three interpolation methods, *spline*, *IDW*, and *kriging*, to generate continuous surfaces of temperature trends in Bangladesh (years from 1948 to 2007). They have found *OK* to be useful for maximum temperature trend analysis. Keblouti et al. [46] have tried to determine the most adequate rainfall interpolation technique and reported *inverse distance weighting* to be the best technique to characterize rainfall distribution. Similarly, four spatial interpolation methods, namely, *ordinary least squares*, *IDW*, *OK*, and *cokriging* are compared in article [19], along with their proposed *geographically weighted regression* method for forest canopy height prediction. Though *geographically weighted regression* dominates others, *inverse distance weighting* and *ordinary least squares* found to be better among the four techniques, with different sampling density. Xie et al. [103] and Phachomphon et al. [76] have studied the impact of some important interpolation techniques including *OK* and *IDW*, in the application of agriculture and soil science. The *OK* achieves the best ability to estimate the soil pollution trend in [103], however, *inverse distance weighting* outperforms *ordinary kriging* in the latter. Ruddick [80] has compared *IDW*, *NN*, and *OK* for Australian seascape prediction and found comparable results. Lu et al. [63] have developed an adaptive *IDW* method to find the best optimal and adaptive distance-decay factor and compared the method with *OK*. The adaptive *IDW* is found to perform better for predicting precipitation surface. Chaplot et al. [17] have conducted one interpolation experimentation of point height data with several methods such as *IDW*, *UK*, *OK*, *spline*, and others. They have found *IDW* to perform better than others. Brus et al. [13] have presented an optimization method of the sample pattern when the environmental parameter is interpolated using linearly related covariates. They have extended *universal kriging*, which has performed better than *OK* and others. Tait et al. [89] have presented a daily rainfall estimation method for the study region New Zealand (period 1960–2004) using *TPS* and verified several environmental factors. Gundogdu et al. [34] have examined *kriging* method for choosing an optimal experimental *semivariogram* model for the analyses of groundwater levels and reported that for their studied irrigation area, the rational quadratic empirical *semivariogram* model performed best. Hancock et al. [35] have proposed an automatic procedure for optimizing smoothness of *TPS* to minimize generalized cross-validation. They have tested their proposed method on temperature estimation in African and Australian continents and found optimal results. The *nearest neighbor* is compared against some ensemble methods including *kriging*, *IDW* and others, that are proposed by Stahl et al. in [87], for daily air temperature estimation over British Columbia, Canada. The variants of *IDW* performed better than other two. Nikolopoulos et al. [71] have compared three interpolation methods: *NN*, *OK*, and *IDW* and estimated the debris flow that triggers rainfall. They have reported *NN* to estimate with bias lesser than *OK* and *IDW* but with large estimation variances.

2.5.2 Spatio-Temporal Interpolation

Several studies have also been reported in the field of spatio-temporal interpolation of meteorological parameters from time-series data. Yang et al. [104] have reported a time-forward *kriging* for forecasting hourly spatio-temporal solar irradiance data for the study region Singapore. They have clearly presented the notion of three statistical properties, stationarity, full symmetry, and separability in the context of spatio-temporal analysis with Venn diagram. Liang et al. [62] presents a new interpolation method, *Markov cube kriging* (*MCK*), to address the scalability issue for handling huge spatio-temporal data. Agarwal [1] has presented a new spatio-temporal *kriging* model in his Master's thesis to predict the daily atmospheric temperature data in the USA. The model is compared with *ARIMA* (*Autoregressive integrated moving average*) model [106] and has been reported to perform better. Arslan [5] has considered spatio-temporal *ordinary kriging* and *indicator kriging* for predicting groundwater salinity at unobserved locations in Bafra Plain, Turkey. They have considered *OK* to analyze spatial variability of the salinity factor and the *indicator kriging* to analyze the salinity in terms of pollution threshold. Wentz et al. [102] have proposed a space–time interpolation approach, termed as *space–time interpolation environment* (*STIE*), based on two interpolation methods, one for the temporal and another for the spatial dimension to maximize the quality of the result. It reports 85.2% accuracy for estimating urban *LULC* growth in the region Phoenix, Arizona, which is better than using a single interpolation technique. Heuvelink et al. [40] have proposed a spatio-temporal prediction model in which the space–time variable is treated as a sum of trend model, considering independent stationarity in spatial, temporal, and space–time component (anisotropic). They have found stable spatial patterns during the studied time period. Spadavecchia et al. [85] have compared three *kriging*-based geostatistical models with a baseline *IDW* method. The methods are *SK*, *OK*, and *KED*. They have reported *KED* to perform better for the estimation of maximum and minimum *temperature* and *precipitation*, in terms of *MAE* and *RMSE*. Hengl et al. [38] have presented an interesting idea of predicting *land surface temperature* in Croatia from MODIS *LST* images. They have identified that the *land surface temperature* is a function of *location* (*latitude/longitude*), *LULC*, *orography*, *precipitation*, *industrial activities*, etc. Though they have addressed the fact that the *LST* to be the function of the representative *LULC*, this information is not incorporated into the prediction method. Dozier et al. [25] have used the daily time-series data to predict the daily albedo and snow cover in Sierra Nevada of California. They have applied *TPS* method for the same. Similarly, Hijmans et al. [41] have applied *TPS* for estimating monthly mean, maximum and minimum temperature, and precipitation in USA for datasets with different resolution and compared their performances. Srinivasan et al. [86] have established a new *kriging* approach for speeding up the interpolation process, and applied this method on ocean color data from the Chesapeake Bay region. They have reported significant performance improvement for the proposed approach over other standard kriging approaches. Kilibarda et al. [48] have applied spatio-temporal *regression kriging* approach to estimate the mean, minimum, and maximum temperatures in Europe. They have found the average accuracy in terms of *RMSE* to be very high in lower altitude. Carrera-Hernández et al. [16] have studied the spatio-temporal variation in rainfall and temperature in the Mexico Basin and also studied their relation with elevation. They have considered different interpolation methods such as *OK*, *KED*, *OK in a local neighborhood*, *block kriging with external drift*, and *KED in a local neighborhood*. They have observed that in each case, the prediction accuracy has improved while considering elevation as a secondary parameter.

2.5.3 Spatial Interpolation with Probabilistic Analysis

For a-posterior probabilistic analysis of meteorological and terrestrial data, several research works [24, 70] applying *Bayesian network* (*BN*) on spatial and spatio-temporal data have been reported till date. Le et al. [56] have developed an alternative spatial interpolation approach to *kriging* and presented its theoretical fundamentals in details. The research work developed by Coffino et al. [20] is a combination of *BN* and numeric atmospheric model to estimate weather patterns. The *BN* is used here to model the spatial and temporal dependencies among different weather stations for dealing with multivariate spatially distributed time series. Nandar [70] has developed a *Bayesian network* based probabilistic model for estimating rainfall in Myanmar. The *BN* is considered to investigate the spatial relation among meteorological variables. A *Bayesian belief network* (*BBN*) has been established by Dlamini in [24], considering abiotic, biotic, and human variables to determine the influencing factors of wild-fire in Swaziland. Hussain et al. [42] have proposed an extension of multivariate hierarchical Bayesian spatio-temporal interpolation for the accurate prediction of spatio-temporal precipitation for water resource management during monsoon. They have compared the proposed approach with the base or non-transformed one and observed that the proposed one provides more accuracy. Fuentes et al. [32] have proposed an improved model of spatial prediction by incorporating the posterior distribution of the ground truth data of SO_2 concentrations, by following a Bayesian analysis. They have obtained high-resolution SO_2 distribution. Kibria et al. [47] have presented a multivariate spatio-temporal prediction using Bayesian analysis for mapping $PM_{2.5}$ in Philadelphia and the proposed method found to perform well. Brown et al. [12] have adopted a multivariate spatial interpolation approach and model the uncertainty of the underlying technique using Bayesian analysis for monitoring different environmental factors. Le et al. [57] have developed an empirical Bayesian approach for spatial interpolation to estimate airborne pollutant concentrations over time.

In many cases, the *BN* learning method has been extended with fuzzy analysis to model the causal relationship among environmental variables better. It also deals with the uncertainties associated with the datasets in a better manner. Peng-Cheng et al. [61] have developed a fuzzy *BN* approach, which measures the causal relationships among human reliability and organizational factors qualitatively, as well as quantitatively. Ferreira et al. [29] have developed a unique method by integrating fuzzy logic and *BN* approach together which is able to evaluate and rank the suppliers more accurately compared to the existing alternatives. A hybrid inference approach by combining *BN* and fuzzy sets has been established by Tang and Liu in [90] and is named as *fuzzy Bayesian network* (*FBN*). The effectiveness of *FBN* has been proved already in the field of machine fault diagnosis.

2.5.4 Importance of LULC Analysis and Ontology

The importance of modeling *LULC* for meteorological parameters have been reported by a few studies in the literature. Hengl et al. [38] have described the *LST* to be highly influenced by the *LULC* distribution in the terrain. Many other works [53, 65, 82] have also stated that the terrestrial *LULC* distribution is impactful weather dynamics and climatic patterns analysis. A detrended *kriging* method is proposed by Janssen et al. [44] for air pollution measurement and is named as *RIO*. The authors have considered CORINE *LULC* data for the analysis. Petrişor et al. [75] have investigated whether a particular study region is influenced at macro-scale by change in *LULC* pattern. They have considered *OK* method to model the environmental changes and their impact on *LULC* pattern. As already discussed earlier, the *STIE* method [102], proposed by Wentz et al. is developed to predict *LULC* for estimating urban growth in the region Phoenix, Arizona. Hence, prediction of *LULC* and its analysis is one of the most challenging aspect in meteorological analysis. However, this knowledge is yet to be incorporated into the spatial interpolation process pragmatically. As far as our knowledge, the multivariate spatial interpolation of meteorological parameters such as*NDVI*, *LST*, etc., for the forecasting of the future *LULC* pattern is also a very novel approach in literature.

Many studies have reported ontology-based analysis of *LULC* in the field of *GIS*. Feng et al. [28] have applied a feature-based method for quantifying the semantic similarities among different classes of *LULC*. Based on this analysis, users can decide whether the given *LULC* information is acceptable for a particular application.

Comber et al. [22] have addressed the issue of properly understanding the semantic or meaning of the *LULC* classes coming from different datasets. Herold et al. [39] have focused on semantic interoperability among different *LULC* datasets in order to resolve terminological and conceptual incompatibilities. Another study by Ahlqvist [2] has addressed the semantic interoperability issues in *LULC* classifications model. Further, it has introduced an approach to conceptual spaces and rough fuzzy sets for evaluating semantic similarity between *LULC* classes. Comber et al. [21] have presented an approach to integrate time-series *LULC* information from diverse sources using domain experts' knowledge, when *LULC* classes may get fundamentally changed over time. They have relied on an ontology-based approach to integrate data from different sources. Varanka [95] has proposed semantics for complex *LULC*) ontology design patterns (*ODP*) for the topographic features as a data models. They have assessed the performance of the model with the USGS (United States Geological Survey) national map by assembling it into the proposed *ODP*. Similarly, to integrate the capabilities of the USGS national map and the semantic web, another research work on the semantic analysis of the topographic data for ontology and triples is presented in [94]. Kovacs et al. [49] have also investigated the need to consider topographic objects (complex *LULC*) from spatial, as well as from semantic perspective. This research work aims to model and represent the nonspatial and spatial entities that are semantically related in ontology.

2.6 Future of Spatial Interpolation

From the state of the art, it may be observed that many spatial interpolation methods have estimated different meteorological parameters for diverse applications. Many approaches are developed for the *temperature* estimation for different regions, all over the world. Further, though analysis of *LULC* classes have been studied for spatial applications, none of the studies have actually implemented this analysis to interpolate *LST* and other meteorological parameters more efficiently and accurately. It may be observed from the state of the art that *LULC* analysis is important for meteorological parameters as different classes influence the parameters in a varying manner [38, 44]. However, the combination of both the approaches (spatial interpolation and *LULC* analysis) is still lacking. According to to the article by Environmental Protection Agency (EPA), USA [27], the *LULC* classes such as *building*, *forest*, *industrial area*, *agricultural area*, etc., significantly influence the meteorological parameters. For example, a *water body* absorbs and emits less heat than a *building*. Thus a *building* in a certain location will increase the *LST* compared to a *water body*. However, in case of *moisture stress index* parameter, contrasting behavior is reported by these two *LULC* classes. Further, for the spatio-temporal prediction and forecasting, the *LULC* distribution of the same terrain may evolve over time. Hence, each of the time instance under consideration is having different semantics in terms of *LULC* distribution and influences the meteorological parameter distinctly. Therefore, spatial interpolation methods, for the applications such as meteorological parameter prediction, forecasting, etc., should incorporate the behavioral change in semantics or the knowledge of *LULC* distribution of the terrain. It may result in more pragmatic and accurate forecasting model.

Each of the existing spatial interpolation methods that has been discussed in this chapter, including the most popular *ordinary kriging*, models the experimental *semivariogram* model by analyzing the underlying spatial relationships among the sampled locations in terms of their *Euclidean* distances in 2D space. Therefore, it is independent of the influencing spatial *LULC* classes. With these specified limitations of the existing interpolation methods, this work attempts to propose a new univariate geostatistical interpolation model, which can blend the underlying spatial relationships of the terrain for the prediction of meteorological parameters. The most popular interpolation method, i.e., *ordinary kriging* (refer Fig. 2.2) has been chosen as the base method to be extended further to in spatio-semantic domain. It can be further extended for univariate and multivariate spatio-temporal prediction and forecasting as well. The proposed model can be directly applied for different spatial applications, which involves analysis of *LULC* distribution, for example, forecasting of urban landscape for change pattern analysis.

References

1. Agarwal A (2011) A new approach to spatio-temporal kriging and its applications. Master's thesis, Ohio State University
2. Ahlqvist O (2005) Using uncertain conceptual spaces to translate between land cover categories. Int J Geogr Inf Sci 19(7):831–857
3. Akkala A, Devabhaktuni V, Kumar A (2010) Interpolation techniques and associated software for environmental data. Environ Prog Sustain Energy 29(2):134–141
4. Armstrong M (1984) Problems with universal kriging. Math Geol 16(1):101–108
5. Arslan H (2012) Spatial and temporal mapping of groundwater salinity using ordinary kriging and indicator kriging: the case of Bafra plain, Turkey. Agric Water Manag 113:57–63
6. Bhattacharjee S, Ghosh SK (2017) Semantic kriging. In: Encyclopedia of GIS, vol 2. Springer International Publishing, Cham, pp 1868–1879
7. Bhattacharjee S, Mitra P, Ghosh SK (2014) Spatial interpolation to predict missing attributes in GIS using semantic kriging. IEEE Trans Geosci Remote Sens 52(8):4771–4780
8. Bhattacharjee S, Das M, Ghosh SK, Shekhar S (2016) Prediction of meteorological parameters: an a-posteriori probabilistic semantic kriging approach. In: Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, p 38
9. Bhowmik AK, Cabral P (2011) Statistical evaluation of spatial interpolation methods for small-sampled region: a case study of temperature change phenomenon in Bangladesh. In: Computational science and its applications-ICCSA 2011. Springer, Heidelberg, pp 44–59
10. Boer EP, de Beurs KM, Hartkamp AD (2001) Kriging and thin plate splines for mapping climate variables. Int J Appl Earth Observ Geoinf 3(2):146–154
11. Bookstein FL et al (1989) Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans Pattern Anal Mach Intell 11(6):567–585
12. Brown PJ, Le ND, Zidek JV (1994) Multivariate spatial interpolation and exposure to air pollutants. Can J Stat 22(4):489–509
13. Brus DJ, Heuvelink GB (2007) Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138(1):86–95

14. Brus DJ, de Gruijter JJ, Marsman BA, Visschers BA, Bregt AK, Breeuwsma A (1996) The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study. Environmetrics 7:1–16
15. Burrough PA, McDonnell RA, Lloyd CD (2015) Principles of geographical information systems. Oxford University Press, New York
16. Carrera-Hernández J, Gaskin S (2007) Spatio temporal analysis of daily precipitation and temperature in the basin of Mexico. J Hydrol 336(3):231–249
17. Chaplot V, Darboux F, Bourennane H, Leguédois S, Silvera N, Phachomphon K (2006) Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density. Geomorphology 77(1):126–141
18. Chen FW, Liu CW (2012) Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. Paddy Water Environ 10(3):209–222
19. Chen G, Zhao K, McDermid GJ, Hay GJ (2012) The influence of sampling density on geographically weighted regression: a case study using forest canopy height and optical data. Int J Remote Sens 33(9):2909–2924
20. Cofıno AS, Cano R, Sordo C, Gutierrez JM (2002) Bayesian networks for probabilistic weather prediction. In: 15th European conference on artificial intelligence, ECAI, Citeseer
21. Comber A, Fisher P, Wadsworth R (2004) Integrating land-cover data with different ontologies: identifying change from inconsistency. Int J Geogr Inf Sci 18(7):691–708
22. Comber A, Fisher P, Wadsworth R (2005) You know what land cover is but does anyone else? An investigation into semantic and ontological confusion. Int J Remote Sens 26(1):223–228
23. Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley
24. Dlamini WM (2010) A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. Environ Model Softw 25(2):199–208
25. Dozier J, Painter TH, Rittger K, Frew JE (2008) Time-space continuity of daily maps of fractional snow cover and albedo from MODIS. Adv Water Resour 31(11):1515–1526
26. Eberly D (2002) Thin plate splines. Geometric Tools Inc
27. EPA USA (2008) Cool roofs, reducing urban heat islands: compendium of strategies. Technical report, United States Environmental Protection Agency [EPA 2009a]. http://www.epa.gov/hiri/resources/compendium.htm. Accessed 18 Aug 2015
28. Feng CC, Flewelling DM (2004) Assessment of semantic similarity between land use/land cover classification systems. Comput Environ Urban Syst 28(3):229–246
29. Ferreira L, Borenstein D (2012) A fuzzy-Bayesian model for supplier selection. Expert Syst Appl 39(9):7834–7844
30. Foster MP, Evans AN (2008) An evaluation of interpolation techniques for reconstructing ionospheric TEC maps. IEEE Trans Geosci Remote Sens 46(7):2153–2164
31. Franzen DW, Peck TR (1995) Field soil sampling density for variable rate fertilization. J Product Agric 8(4):568–574
32. Fuentes M, Raftery AE (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61(1):36–45
33. Gelfand AE, Diggle P, Guttorp P, Fuentes M (2010) Handbook of spatial statistics. CRC Press
34. Gundogdu KS, Guney I (2007) Spatial analyses of groundwater levels using universal kriging. J Earth Syst Sci 116(1):49–55
35. Hancock P, Hutchinson M (2006) Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. Environ Model Softw 21(12):1684–1694
36. Harder RL, Desmarais RN (1972) Interpolation using surface splines. J Aircr 9(2):189–191
37. Hartkamp AD, De Beurs K, Stein A, White JW (1999) Interpolation techniques for climate variables. CIMMYT NRG-GIS Series
38. Hengl T, Heuvelink GB, Tadić MP, Pebesma EJ (2012) Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. Theor Appl Climatol 107(1–2):265–277
39. Herold M, Woodcock C, Di Gregorio A, Mayaux P, Belward AS, Latham J, Schmullius CC (2006) A joint initiative for harmonization and validation of land cover datasets. IEEE Trans Geosci Remote Sens 44(7):1719

40. Heuvelink G, Griffith DA (2010) Space-time geostatistics for geography: a case study of radiation monitoring across parts of Germany. Geogr Anal 42(2):161–179
41. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25(15):1965–1978
42. Hussain I, Spöck G, Pilz J, Yu HL (2010) Spatio-temporal interpolation of precipitation during monsoon periods in Pakistan. Adv Water Resour 33(8):880–886
43. Hutchinson MF (1995) Interpolating mean rainfall using thin plate smoothing splines. Int J Geogr Inf Syst 9(4):385–403
44. Janssen S, Dumont G, Fierens F, Mensink C (2008) Spatial interpolation of air pollution measurements using CORINE land cover data. Atmos Environ 42(20):4884–4903
45. Karydas CG, Gitas IZ, Koutsogiannaki E, Lydakis-Simantiris N, Silleos G et al (2009) Evaluation of spatial interpolation techniques for mapping agricultural topsoil properties in Crete. EARSeL eProceedings 8(1):26–39
46. Keblouti M, Ouerdachi L, Boutaghane H (2012) Spatial interpolation of annual precipitation in Annaba-Algeria-comparison and evaluation of methods. Energy Procedia 18:468–475
47. Kibria BG, Sun L, Zidek JV, Le ND (2002) Bayesian spatial prediction of random space-time fields with application to mapping PM2.5 exposure. J Am Stat Assoc 97(457), 112–124
48. Kilibarda M, Hengl T, Heuvelink G, Gräler B, Pebesma E, Perčec Tadić M, Bajat B (2014) Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. J Geophys Res Atmos 119(5):2294–2313
49. Kovacs K, Dolbear C, Goodwin J (2007) Spatial concepts and OWL issues in a topographic ontology framework. In: Proceeding of the GIS
50. Kravchenko A (2003) Influence of spatial structure on accuracy of interpolation methods. Soil Sci Soc Am J 67(5):1564–1571
51. Kravchenko A, Bullock DG (1999) A comparative study of interpolation methods for mapping soil properties. Agron J 91(3):393–400
52. Krivoruchkoa K, Gotwayb C (2003) Using spatial statistics in GIS. In: Proceedings of International Congress on Mod Sim, pp 713–736
53. Lambin EF, Geist HJ (2008) Land-use and land-cover change: local processes and global impacts. Springer Science & Business Media, Heidelberg
54. Lapen DR, Hayhoe HN (2003) Spatial analysis of seasonal and annual temperature and precipitation normals in Southern Ontario. Can J Great Lakes Res 29(4):529–544
55. Laslett GM, McBratney AB, Pahl PJ, Hutchinson MF (1987) Comparison of several spatial prediction methods for soil pH. J Soil Sci 38:325–341
56. Le ND, Zidek JV (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. J Multivar Anal 43(2):351–374
57. Le ND, Sun W, Zidek JV (1997) Bayesian multivariate spatial interpolation with data missing by design. J R Stat Soc Ser B (Stat Methodol) 59(2):501–510
58. LeMay V, Temesgen H (2005) Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Sci 51(2):109–119
59. Li J (2008) A review of spatial interpolation methods for environmental scientists. Record (Geoscience Australia)
60. Li J, Heap AD (2011) A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol Inf 6(3):228–241
61. Li PC, Chen GH, Dai LC, Zhang L (2012) A fuzzy Bayesian network approach to improve the quantification of organizational influences in HRA frameworks. Saf Sci 50(7):1569–1583
62. Liang D, Kumar N (2013) Time-space kriging to address the spatiotemporal misalignment in the large datasets. Atmos Environ 72:60–69
63. Lu GY, Wong DW (2008) An adaptive inverse-distance weighting spatial interpolation technique. Comput Geosci 34(9):1044–1055
64. Luo W, Taylor M, Parker S (2008) A comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales. Int J Climatol 28(7):947–959

65. Mahmood R, Quintanar AI, Conner G, Leeper R, Dobler S, Pielke RA Sr, Beltran-Przekurat A, Hubbard KG, Niyogi D, Bonan G et al (2010) Impacts of land use/land cover change on climate and future research priorities. Bull Am Meteorol Soc 91(1):37–46
66. Messer H, Zinevich A, Alpert P (2006) Environmental monitoring by wireless communication networks. Science 312(5774):713–713
67. Mueller T, Pierce F, Schabenberger O, Warncke D (2001) Map quality for site-specific fertility management. Soil Sci Soc Am J 65(5):1547–1558
68. Mueller T, Pusuluri N, Mathias K, Cornelius P, Barnhisel R, Shearer S (2004) Map quality for ordinary kriging and inverse distance weighted interpolation. Soil Sci Soc Am J 68(6):2042–2047
69. Nalder IA, Wein RW (1998) Spatial interpolation of climatic normals: test of a new method in the Canadian boreal forest. Agric Forest Meteorol 92(4):211–225
70. Nandar A (2009) Bayesian network probability model for weather prediction. In: International conference on the current trends in information technology (CTIT). IEEE, pp 1–5
71. Nikolopoulos E, Borga M, Creutin J, Marra F (2015) Estimation of debris flow triggering rainfall: influence of rain gauge density and interpolation methods. Geomorphology 243:40–50
72. Odeh IO, McBratney A, Chittleborough D (1995) Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67(3):215–226
73. Ozelkan E, Bagis S, Ozelkan EC, Ustundag BB, Yucel M, Ormeci C (2015) Spatial interpolation of climatic variables using land surface temperature and modified inverse distance weighting. Int J Remote Sens 36(4):1000–1025
74. Peralvo M, Maidment D (2003) Influence of DEM interpolation methods in drainage analysis. GIS Hydro 4
75. Petrişor AI, Ianoş I, Tălângă C (2010) Land cover and use changes focused on the urbanization processes in Romania. Environ Eng Manag J 9(6):765–771
76. Phachomphon K, Dlamini P, Chaplot V (2010) Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables. Geoderma 155(3):372–380
77. Rayitsfeld A, Samuels R, Zinevich A, Hadar U, Alpert P (2012) Comparison of two methodologies for long term rainfall monitoring using a commercial microwave communication system. Atmos Res 104:119–127
78. Reinstorf F, Binder M, Schirmer M, Grimm-Strele J, Walther W (2005) Comparative assessment of regionalisation methods of monitored atmospheric deposition loads. Atmos Environ 39(20):3661–3674
79. Ribeiro Sales MH, Souza CM, Kyriakidis PC (2013) Fusion of MODIS images using kriging with external drift. IEEE Trans Geosci Remote Sens 51(4):2250–2259
80. Ruddick R (2007) Data interpolation methods in the geoscience Australia seascape maps. Geoscience Australia, Canberra
81. Schloeder C, Zimmerman N, Jacobs M (2001) Comparison of methods for interpolating soil properties using limited data. Soil Sci Soc Am J 65(2):470–479
82. Sertel E, Ormeci C, Robock A (2011) Modelling land cover change impact on the summer climate of the Marmara region Turkey. Int J Global Warm 3(1):194–202
83. Sheikhhasan H (2006) A comparison of interpolation techniques for spatial data prediction. Master's thesis, Department Computer Science, Universiteit van Amsterdam, Amsterdam, The Netherlands
84. Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM national conference. ACM, pp 517–524
85. Spadavecchia L, Williams M (2009) Can spatio-temporal geostatistical methods improve high resolution regionalisation of meteorological variables? Agric Forest Meteorol 149(6):1105–1117
86. Srinivasan BV, Duraiswami R, Murtugudde R (2010) Efficient kriging for real-time spatio-temporal interpolation. In: Proceedings of the 20th conference on probability and statistics in the atmospheric sciences, pp 228–235

87. Stahl K, Moore R, Floyer J, Asplin M, McKendry I (2006) Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. Agric Forest Meteorol 139(3):224–236
88. Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer Verlag, New York
89. Tait A, Henderson R, Turner R, Zheng X (2006) Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. Int J Climatol 26(14):2097–2115
90. Tang H, Liu S (2007) Basic theory of fuzzy Bayesian networks and its application in machinery fault diagnosis. In: Fourth international conference on fuzzy systems and knowledge discovery, vol 4. IEEE, pp 132–137
91. Teegavarapu RS, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312(1):191–206
92. Ustrnul Z, Czekierda D (2005) Application of GIS for the development of climatological air temperature maps: an example from Poland. Meteorol Appl 12(01):43–50
93. Van Kuilenburg J, De Gruijter JJ, Marsman BA, Bouma J (1982) Accuracy of spatial interpolation between point data on soil moisture supply capacity, compared with estimates from mapping units. Geoderma 27:311–325
94. Varanka D, Carter J, Usery EL, Shoberg T (2011) Topographic mapping data semantics through data conversion and enhancement. In: Geospatial semantics and the semantic web. Springer, Boston, pp 145–162
95. Varanka DE (2011) Ontology patterns for complex topographic feature types. Cartogr Geogr Inf Sci 38(2):126–136
96. Wackernagel H, Oliveira VD, Kedem B (1997) Multivariate geostatistics. SIAM Rev 39(2):340–340
97. Wahba G (1990) Spline models for observational data, vol 59. SIAM
98. Wahba G, Wendelberger J (1980) Some new mathematical methods for variational objective analysis using splines and cross validation. Mon Weather Rev 108(8):1122–1143
99. Weber D, Englund E (1992) Evaluation and comparison of spatial interpolators. Math Geol 24(4):381–391
100. Webster R, Oliver MA (2007) Geostatistics for environmental scientists. Wiley
101. Weisz R, Fleischer S, Smilowitz Z (1995) Map generation in high-value horticultural integrated pest management: appropriate interpolation methods for site-specific pest management of colorado potato beetle (Coleoptera: Chrysomelidae). J Econ Entomol 88(6):1650–1657
102. Wentz EA, Peuquet DJ, Anderson S (2010) An ensemble approach to space-time interpolation. Int J Geogr Inf Sci 24(9):1309–1325
103. Xie Y, Chen Tb, Lei M, Yang J, Guo Qj, Song B, Zhou Xy (2011) Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: accuracy and uncertainty analysis. Chemosphere 82(3):468–476
104. Yang D, Gu C, Dong Z, Jirutitijaroen P, Chen N, Walsh WM (2013) Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging. Renew Energy 60:235–245
105. Yasrebi J, Saffari M, Fathi H, Karimian N, Moazallahi M, Gazni R et al (2009) Evaluation and comparison of ordinary kriging and inverse distance weighting methods for prediction of spatial variability of some soil chemical parameters. Res J Biol Sci 4(1):93–102
106. Zhang GP (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175
107. Zhou F, Guo HC, Ho YS, Wu CZ (2007) Scientometric analysis of geostatistics using multivariate methods. Scientometrics 73(3):265–279
108. Zimmerman D, Pavlik C, Ruggles A, Armstrong MP (1999) An experimental comparison of ordinary and universal kriging and inverse distance weighting. Math Geol 31(4):375–390

Chapter 3
Spatial Semantic Kriging

Abstract The spatial *semantic kriging* (*SemK*) based spatial interpolation method is applied for the interpolation of meteorological parameters, aiming to enhance the accuracy in results. The *SemK* considers the semantic properties of the terrain, which is influential to the meteorological parameters and incorporates into the prediction process. One such property is the terrestrial land-use/land-cover (*LULC*) distribution. An ontology hierarchy is built with the available *LULC* classes to find the influence of each of the classes to the *land surface temperature* (*LST*). This interpolation process belongs to the family of *kriging* and extends the the *ordinary kriging* (*OK*) based spatial interpolation with *LULC* knowledge. The empirical experiments show that this auxiliary knowledge is highly significant to achieve more accuracy in prediction. The theoretical performance analysis is also carried out in this chapter to prove the efficiency of *SemK* over other existing methods.

3.1 Introduction

Prediction of spatial parameters with enhanced accuracy has attracted significant research interest in *remote sensing* and *GIS* research community. Further, the raw raster imagery, which is processed to generate different derived meteorological parameters, generally contain missing values in terms of line gaps (a line of missing pixels), cloud covers, etc. Prediction of the parameter values at those pixels is needed to generate continuous surfaces for further analysis. For estimating the parameter value with higher accuracy, modeling spatial autocorrelation plays an important role. It is defined as the dependency among the sampled locations with respect to a spatial parameter. There exist several variations in existing spatial interpolation methods based on how accurately or pragmatically the autocorrelation is modeled to achieve high prediction accuracy. As discussed in the literature survey (such as [10, 11]), the *kriging* is the most commonly used and well accepted geostatistical spatial interpolation method, which can model this dependency most efficiently. All the variants of *kriging* interpolators model the spatial autocorrelation as a function of *Euclidean* distances among the sampled locations.

S. Bhattacharjee et al., *Semantic Kriging for Spatio-temporal Prediction*, Studies in Computational Intelligence 839,
https://doi.org/10.1007/978-981-13-8664-0_3

For most of the meteorological parameters such as *LST*, *NDBI*, *NDVI*, *MSI*, *NDWI*, etc., the *LULC* of the sampled location is highly influential. Especially for the *land surface temperature*, the *LULC* of the terrain is of utmost importance. For example, say two sampled locations are at same *Euclidean* distance apart from the prediction point. When the former location is represented by a *industrial* area, the later is by a *waterbody*. In this scenario, the first sampled location may increase the *LST* value of the prediction point, while the later decreases it. Hence, there is a need for land–atmospheric interaction modeling for prediction by analyzing the terrestrial *LULC* distribution.

The spatial *SemK* approach incorporates the *LULC* knowledge into the prediction process. The "semantics" or the categorical *LULC* information of a sample point is assumed to be the terrestrial qualitative property [5], not a secondary covariate. Hence, in this work, no the multivariate *kriging* (such as *co-kriging*) methods are not suitable. This method extends the univariate *ordinary kriging* (*OK*) method with semantic information of the *RoI*. This research considers the *LULC* information of the terrain to be a contextual knowledge of the terrain. Hence, one of the major challenges is the proper and accurate quantification of this semantic knowledge. Once it is quantified, the *ordinary kriging* (*OK*) needs to be extended properly, without violating its statistical properties. The next section specifies the objectives of the newly proposed method *semantic kriging*.

3.2 Objectives of Semantic Kriging

This chapter establishes a new scheme of interpolation, namely, *semantic kriging*. Like all other univariate *kriging* interpolation methods, *SemK* aims to minimize the mean square error in prediction. It extends the existing *OK* interpolation process by integrating the semantics of the *LULC* classes with the same. The semantic knowledge is captured through an ontology-based concept hierarchy [5]. The proposed two metrics, *spatial importance* and *semantic similarity* are considered to model the spatio-semantic relations among the *LULC* classes for the interpolation process. The objectives of the chapter can be briefed as follows:

- to propose a unique and accurate approach to quantify the qualitative semantic *LULC* information for modeling the meteorological and terrestrial interaction.
- to propose new semantic metrics, which can be used to modify the *Euclidean* distance based *variance* between two sampled locations.
- to extend the *covariance*s and the weights assigned by *OK* with the proposed semantic metrics.
- mathematical formulation of the new weight matrices and other related variables of *SemK*.
- performance analysis and comparison of *SemK* with few existing methods using *LST* data.

- performance analysis of *SemK* for the establishment of its relation with the existing techniques and optimal interpolator and evaluate the effect of the granularity of the ontology on *SemK*.

3.3 SemK: Semantic Kriging

The *SemK* extends the *OK* method by combining the correlation and the semantics similarity between the *LULC* classes with the interpolation process, aiming to yield better predicted value. It extends the traditional *covariance* measure to higher dimension by blending the *LULC* distribution information. Similarity analysis among the *LULC* classes is accomplished in a way so that the semantically similar and the correlated *LULC* pairs (the representatives of the sampled or unsampled locations) will be assigned higher score compared to the distant ones. The flow diagram of *SemK* is depicted in Fig. 3.1, which can be described as follows:

- the satellite image of the derived meteorological parameter is considered as the input and the missing pixels are identified to carry out interpolation.
- an ontology of spatial *LULC* classes of the *RoI* is also considered as this input to the framework.
- in *SemK* process, the *Euclidean* distance based proximity is blended with semantic proximity of the terrain with two metrics: *spatial importance* and *semantic similarity*.
  - the *semantic similarity* is measured with respect to the ontological hop distances between the leaf *LULC* classes.
  - the *spatial importance* is evaluated by correlation analysis with sample data representing those *LULC* classes.
- these semantic metrics then modify the traditional *Euclidean* distance based proximity of the *ordinary kriging* (*OK*) method, resulting spatio-semantic proximity model.
- the weight vector of the *SemK* is evaluated, which is further utilized to evaluate the parameter value of a missing pixel.

Given the *RoI*, a *LULC* ontology is generated by considering all possible *LULC*s in that *RoI*. In the ontology, the *LULC*s are represented as concepts. These concepts are further organized as an ontology hierarchy with respect to a standard semantic relation. Examples of such relations include *meronym*, *hyponym*, *hypernym*, etc. [14]. According to the properties of hierarchical ontology, the semantically similar concepts are closer in the ontology hierarchy compared to the dissimilar ones. For example, Kolkata, WB, India is considered here as one of the *RoI*s for the case study to predict *LST*. Kolkata is a metropolitan city in eastern India with central coordinate: (22.567°N 88.367°E). It consists of different *LULC* classes such as *built-up, wastelands, cropland, waterbodies, forest, wetlands*, etc. The ontology hierarchy of Kolkata, consisting of these *LULC* concepts is depicted in Fig. 3.2. The hierarchy

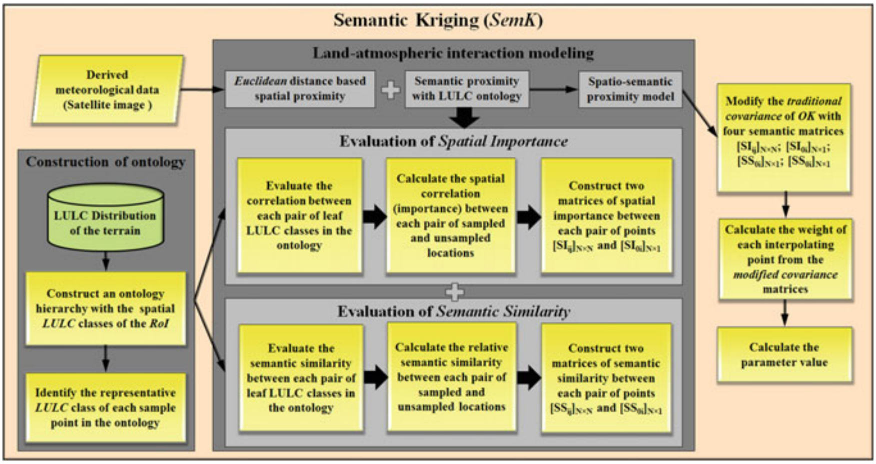

Semantic Kriging (*SemK*)
Land-atmospheric interaction modeling
Derived meteorological data (Satellite image)
Euclidean distance based spatial proximity
Semantic proximity with LULC ontology
Spatio-semantic proximity model
Construction of ontology
LULC Distribution of the terrain
Construct an ontology hierarchy with the spatial *LULC* classes of the *RoI*
Identify the representative *LULC* class of each sample point in the ontology
Evaluation of *Spatial Importance*
Evaluate the correlation between each pair of leaf LULC classes in the ontology
Calculate the spatial correlation (importance) between each pair of sampled and unsampled locations
Construct two matrices of spatial importance between each pair of points $[SI_{ij}]_{N\times N}$ and $[SI_{0i}]_{N\times 1}$
Evaluation of *Semantic Similarity*
Evaluate the semantic similarity between each pair of leaf LULC classes in the ontology
Calculate the relative semantic similarity between each pair of sampled and unsampled locations
Construct two matrices of semantic similarity between each pair of points $[SS_{ij}]_{N\times N}$ and $[SS_{0i}]_{N\times 1}$
Modify the *traditional covariance* of *OK* with four semantic matrices $[SI_{ij}]_{N\times N}$; $[SI_{0i}]_{N\times 1}$; $[SS_{0i}]_{N\times 1}$; $[SS_{0i}]_{N\times 1}$
Calculate the weight of each interpolating point from the *modified covariance* matrices
Calculate the parameter value

Fig. 3.1 *SemK* framework [1, 6]

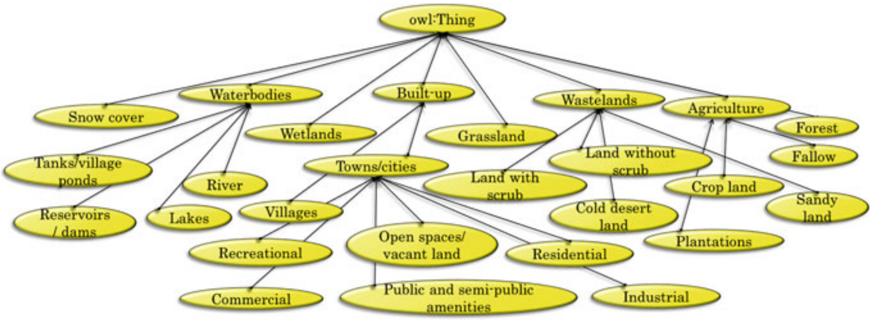


Fig. 3.2 Land-use/land-cover (*LULC*) ontology [2, 4–6, 8]

is created in accordance with the *LULC* classification, proposed by the Department of Science and Technology (DST), Government of India [13]. The *hyponym* relation is considered to construct the hierarchy.

Being region- and domain-specific, i.e., construction of an ontology hierarchy is dependent on the domain of interest and studied spatial region. Depending on the application domain and the *RoI*, the ontology varies in terms of the number and the type of concepts, semantic relation, permitted levels, etc. In this study, the ontology of *LULC* is adaptive, such that, new relevant concepts can be appended and old concepts can be discarded depending on the requirement of the application. From the given ontology in Fig. 3.2, it must be observed that the sampled and unsampled locations are always represented by one leaf concept. Hence, the prediction and every interpolating points are further mapped to the representative *leaf* concept of the hierarchy. For *SemK*, the mapping is mandatory to relate the appropriate *LULC* class of every prediction or interpolating points. Following this *LULC* mapping, the semantic/ ontological association among every pair of leaf *LULC* classes is estimated further. For this purpose, two metrics have been proposed: *spatial importance*, which is the correlation analysis among every pair of leaf *LULC* classes, and the *semantic similarity*, which is the analysis of their semantic distance in the ontology hierarchy. The processes of evaluating the metrics are presented in the Sects. 3.3.1 and 3.3.2. The hierarchical hop distance in the ontology is considered as the heuristic to evaluate the *semantic similarity*. And, for the *spatial importance* evaluation, the actual sample points are considered in the *RoI* representing different *LULC* classes to get their correlation score, related to the prediction parameter. These two semantic metrics extends the traditional (*Euclidean* distance based) *covariance* measurement process into spatio-semantic dimension. Therefore, though the weights of the interpolating points, assigned by *OK*, are based on *Euclidean* distance only, in the newly proposed *SemK*, it is the function of both *Euclidean* distance and the cumulative semantic score (both *semantic similarity* and *spatial importance*) of their representative *LULC* classes. Further, this spatio-semantic scores of the sampled locations are normalized to estimate the parameter value by *SemK*. Having more decision parameters compared to *OK*, makes the *SemK* process more informative than *OK*.

3.3.1 Semantic Similarity

The *semantic similarity* (*SS*) between two representative leaf *LULC* classes in the ontology is evaluated using modified *context resemblance* approach [12]. For this metric, the pair of points having higher semantic distance in the hierarchy will be assigned less score and vice-versa. It is proportional to the assigned weight. The score of the ith interpolating point x_i (where i ∈ 1 $\cdots$ N) with the prediction point x_0 is referred to as SS_{0i}. It is formulated as follows:

$$SS_{0i} = \frac{\frac{m_0}{|f_0|} + \frac{m_i}{|f_i|}}{2} \tag{3.1}$$

where f_i and f_0 are the corresponding representative *LULC* of x_i and x_0 respectively. $|f_i|$ and $|f_0|$ are the total number of concepts in the path of f_i and f_0 in the ontology, staring from the most general concept in the hierarchy, i.e., *owl:Thing*. The m_i and m_0 represent the number of concepts matching in the paths of f_i and f_0. This metric forms a [N × 1] vector, as it is evaluated for all the interpolating points with respect to the prediction point. It is given as $[\mathbf{SS}]_{0i}^T$ = $[SS_{01} SS_{02} \cdots SS_{0N}]$.

Being spatially related, the semantic dependency exists among the representative *LULC* classes, and thus between every pair of interpolating points. Therefore, the relative similarity score should also be measured between every pair of interpolating points. The similarity between ith and jth sampled locations, x_i and x_j (where i, j ∈ 1 $\cdots$ N) is referred to as SS_{ij} and can be represented as follows:

$$SS_{ij} = \frac{\frac{m_i}{|f_i|} + \frac{m_j}{|f_j|}}{2} \tag{3.2}$$

where f_i and f_j are the representative *LULC* of x_i and x_j respectively. The $|f_i|$, $|f_j|$ represent the total number of hops in the path of f_i and f_j respectively, staring from the root (*owl:Thing*) in the ontology. The m_i and m_j are the number of matching concepts in the paths of f_i, f_j. For all the interpolating points, this metric is measured between each pair of locations, which forms a symmetric matrix of dimension [N × N], denoted as $[\mathbf{SS}]_{ij}$.

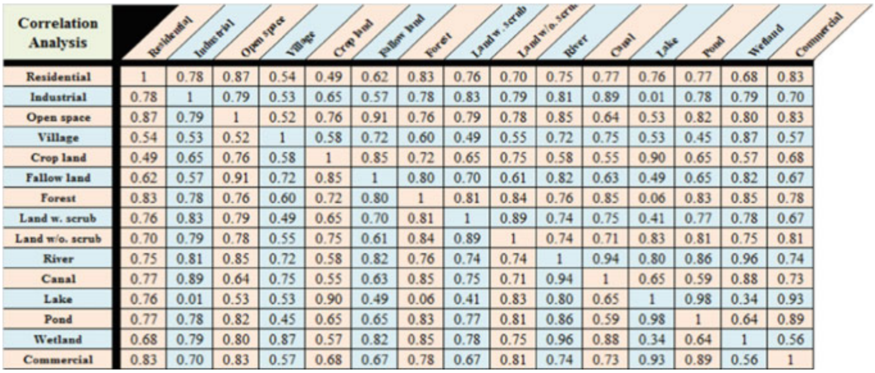
For example, the *semantic similarity* between the concepts *industrial* and *commercial* is given as $\frac{\frac{3}{4}+\frac{3}{4}}{2}$ = 0.75. Similarly, the *semantic similarity* between *commercial* and *plantations* is given as $\frac{\frac{1}{4}+\frac{1}{3}}{2}$ = 0.29. Finally these scores are scaled as difference scores by subtracting each of the SS_{ij} from its maximum possible value.

3.3.2 Spatial Importance

The *spatial importance* (*SI*) score between every pair of leaf *LULC* classes is measured by correlating their sample points with respect to the prediction parameter. In this regard, the *RoI* is splitted into some nonoverlapping subregions. Let *k* be a constant to define the number of paired sampled locations (other than *N* interpolating points) for this correlation analysis. Then, the spatial region is splitted into *k* zones (R_k) such that $\bigcup_{i=1}^{k} R_k = RoI$. Now, *k* pairs of sampled locations are selected from each of the *k* subregions, where the first location in every pair represents the first *LULC* considered for the correlation analysis and vice-versa. Each pair is selected in a way so that physically they represents the locations within a predefined *Euclidean* distance. The correlation is evaluated in terms of these *k* pairs of points. For example, to measure the correlation score between "*Industrial*" and "*River*" (refer Fig. 3.2) with respect to *land surface temperature* in Kolkata, WB, India, the *k* is defined as 50. Thus, 50 nonoverlapping subregions are first selected in the *RoI*. Then, 50 random (uniformly random) sampled locations are selected for the *LULC* class *Industrial* from each of the subregions. Next, a new set of *k* sample points are identified against each *k Industrial* points, which represents the *LULC* class *River*. Then a correlation score between "*Industrial*" and "*River*" is measured in terms of these fifty pairs of locations over the *RoI*. In this study, the correlation score between *Industrial* and *River* is measured as ***0.81***. For the given ontology hierarchy in Fig. 3.2 of the *RoI* Kolkata, WB, India, the correlation between some pairs of leaf concepts are evaluated and presented in Table 3.1. The correlation scores range between [−1, 1]. To further restrict the *SemK* method to deal with the negative mapping of the *covariance*s, these scores are normalized to any positive range. The *spatial importance* evaluation process exhibits the following characteristics:

- this process is dependent on the prediction parameter. For example, correlation score between *industrial* and *river* is ***0.81*** for the prediction of *land surface temperature*. However, for *moisture stress index*, it is ***0.22*** [*RoI* is Kolkata, WB, India].

Table 3.1 Correlation scores between pair of *LULC* classes in Kolkata, WB, India [2]



Correlation Analysis	Residential	Industrial	Open space	Village	Crop land	Fallow land	Forest	Land w. scrub	Land w/o. scrub	River	Canal	Lake	Pond	Wetland	Commercial
Residential	1	0.78	0.87	0.54	0.49	0.62	0.83	0.76	0.70	0.75	0.77	0.76	0.77	0.68	0.83
Industrial	0.78	1	0.79	0.53	0.65	0.57	0.78	0.83	0.79	0.81	0.89	0.01	0.78	0.79	0.70
Open space	0.87	0.79	1	0.52	0.76	0.91	0.76	0.79	0.78	0.85	0.64	0.53	0.82	0.80	0.83
Village	0.54	0.53	0.52	1	0.58	0.72	0.60	0.49	0.55	0.72	0.75	0.53	0.45	0.87	0.57
Crop land	0.49	0.65	0.76	0.58	1	0.85	0.72	0.65	0.75	0.58	0.55	0.90	0.65	0.57	0.68
Fallow land	0.62	0.57	0.91	0.72	0.85	1	0.80	0.70	0.61	0.82	0.63	0.49	0.65	0.82	0.67
Forest	0.83	0.78	0.76	0.60	0.72	0.80	1	0.81	0.84	0.76	0.85	0.06	0.83	0.85	0.78
Land w. scrub	0.76	0.83	0.79	0.49	0.65	0.70	0.81	1	0.89	0.74	0.75	0.41	0.77	0.78	0.67
Land w/o. scrub	0.70	0.79	0.78	0.55	0.75	0.61	0.84	0.89	1	0.74	0.71	0.83	0.81	0.75	0.81
River	0.75	0.81	0.85	0.72	0.58	0.82	0.76	0.74	0.74	1	0.94	0.80	0.86	0.96	0.74
Canal	0.77	0.89	0.64	0.75	0.55	0.63	0.85	0.75	0.71	0.94	1	0.65	0.59	0.88	0.73
Lake	0.76	0.01	0.53	0.53	0.90	0.49	0.06	0.41	0.83	0.80	0.65	1	0.98	0.34	0.93
Pond	0.77	0.78	0.82	0.45	0.65	0.65	0.83	0.77	0.81	0.86	0.59	0.98	1	0.64	0.89
Wetland	0.68	0.79	0.80	0.87	0.57	0.82	0.85	0.78	0.75	0.96	0.88	0.34	0.64	1	0.56
Commercial	0.83	0.70	0.83	0.57	0.68	0.67	0.78	0.67	0.81	0.74	0.73	0.93	0.89	0.56	1

- this process is a *a-priori* correlation analysis, such that, the correlation score between a pair of *LULC* classes is evaluated without estimating the impacts of nearby *LULC*s.
- correlation scores among every pair of *LULC* classes are considered to be the global scores for the whole *RoI*.

Once representative *LULC* classes of all the sampled and unsampled locations are identified in the ontology, the correlation between the representative *LULC* classes of each sampled location and the prediction location is to be evaluated further. These scores are referred to as *spatial importance*, as it represents the significance of one *LULC* class over other. Let the representative *LULC* classes of prediction point x_0 be f_0 and for the ith interpolating point x_i be f_i. Let the *spatial importance* for f_i with respect to f_0 (and vice-versa) be SI_{0i}. It is given as follows:

$$
\begin{aligned}
SI_{0i} &= Corr_{prediction\_parameter}(x_0, x_i) && (3.3)\\
&= Corr_{prediction\_parameter}(f_0, f_i) && (3.4)\\
&= \frac{\sum\limits_{m=1}^{k}(Z(f_{0_m}) - \overline{Z(f_0)})(Z(f_{i_m}) - \overline{Z(f_i)})}{\sqrt{\sum\limits_{m=1}^{k}(Z(f_{0_m}) - \overline{Z(f_0)})^2 \sum\limits_{m=1}^{k}(Z(f_{i_m}) - \overline{Z(f_i)})^2}} && (3.5)
\end{aligned}
$$

where $Z(f_{p_q})$ is the parameter value of qth sampled location, represented by the f_p *LULC* class, $Z(\bar{f}_p)$ is the mean of the parameter values over k sampled locations that are represented by the *LULC* f_p. Thus, being evaluated for every sampled locations with respect to the prediction point, it constructs a [N × 1] vector/ matrix, given as $[\mathbf{SI}]_{0i}^T = [SI_{01} SI_{02} \cdots SI_{0N}]$.

Following the spatial autocorrelation of the terrain, correlation exists among every pair of sampled locations as well. In a similar manner, the *spatial importance* SI_{ij} between any ith and jth sampled locations can be evaluated by measuring their representative *LULC* classes' correlation. Therefore, SI_{ij} is given as follows:

$$
\begin{aligned}
SI_{ij} &= Corr_{prediction\_parameter}(x_i, x_j) && (3.6)\\
&= Corr_{prediction\_parameter}(f_i, f_j) && (3.7)\\
&= \frac{\sum_{m=1}^{k}(Z(f_{i_m}) - \overline{Z(f_i)})(Z(f_{j_m}) - \overline{Z(f_j)})}{\sqrt{\sum_{m=1}^{k}(Z(f_{i_m}) - \overline{Z(f_i)})^2 \sum_{m=1}^{k}(Z(f_{j_m}) - \overline{Z(f_j)})^2}} && (3.8)
\end{aligned}
$$

The *spatial importance* score is measured for each pair of the N sampled locations the scores are scaled as difference scores by subtracting each of the SI_{ij} from its maximum possible value. It constructs a [N × N] symmetric matrix, given as $[\mathbf{SI}]_{ij}$.

These pair of *semantic similarity* and *spatial importance* matrices ($[\mathbf{SS}]_{ij[N\times N]}$, $[\mathbf{SI}]_{ij[N\times N]}$, $[\mathbf{SS}]_{0i[N\times 1]}$ and $[\mathbf{SI}]_{0i[N\times 1]}$) are further considered to modify the *semi-variance matrix* and the *distance matrix* ($[\mathbf{C}]_{ij[N\times N]}$ and $[\mathbf{D}]_{0i[N\times 1]}$) of *OK*. The theoretical error estimation of *SemK* and its evaluation procedure are described in Sect. 3.4.

3.4 Theoretical Error Analysis

To analyze the theoretical error of spatial *SemK*, the variables and constrains of this method are formalized in this section. Here, the relationship of *OK* and *SemK* is also established in terms of their variables. As *SemK* extends the exiting *OK* method with semantic metrics, the statistical properties of this base method remains unchanged in *SemK*. Let the random field (meteorological parameter, say *LST*) measure at the unsampled point x_0 be Z(x_0). The prediction is carried out considering the known sampled locations, given as $\mathbf{Z}^T$ = [Z(x_1) $\cdots$ Z(x_N)], where N represents the number of sampled locations, $Z(x_i)$ represents the parameter value at the sampled location x_i. In this monograph, the Z(x_i)s are symbolized as Z_i for simplicity. The weight vector of *OK*, i.e., $[\mathbf{W}^{OK}]^T$ is given as $[\mathbf{W}^{OK}]^T = [w_1^{OK}\ w_2^{OK}\ \cdots\ w_N^{OK}]$, where w_i^{OK} represents the assigned weight to the ith sampled location by *OK*. Let's consider that $\hat{Z}_0$ is the estimated/predicted parameter value at the prediction location x_0.

For *ordinary kriging*, the following holds:

$$
\hat{Z}_0 = \sum_{i=1}^{N} w_i^{OK} Z_i = [\mathbf{W}]^T\mathbf{Z};\ [\mathbf{W}]^T\mathbf{1} = 1, or \sum_{i=1}^{N} w_i^{OK} = 1.
$$

As *ordinary kriging* follows the concept of mean square error minimization in prediction. Therefore, it aims to optimize the weight vector $\mathbf{W}^{OK}$ so that the estimation variance $\sigma_{OK}^2 = E([Z_0 - \hat{Z}_0]^2)$ is minimum. Let's assume the two traditional matrices, the *semivariance matrix* ($[\mathbf{C}]_{ij[N\times N]}$) and the *distance matrix* of *OK* ($[\mathbf{D}]_{0i[N\times 1]}$) is expressed as follows:

$$\mathbf{C} = \begin{bmatrix} Var(Z_1) & Cov(Z_1, Z_2) & \cdots & Cov(Z_1, Z_N) \\ Cov(Z_2, Z_1) & Var(Z_2) & \cdots & Cov(Z_2, Z_N) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Z_N, Z_1) & Cov(Z_N, Z_2) & \cdots & Var(Z_N) \end{bmatrix} \tag{3.9}$$

$$\mathbf{D} = \begin{bmatrix} Cov(Z_0, Z_1) \\ Cov(Z_0, Z_2) \\ \vdots \\ Cov(Z_0, Z_N) \end{bmatrix} \tag{3.10}$$

where $Cov(Z_i, Z_j)$ represents the *covariance* score between Z(x_i) and Z(x_j) and the variance of Z(x_i) (or, self-covariance) is expressed as $Var(Z_i)$. The weight vector of *ordinary kriging*, $\mathbf{W}^{OK}$ is evaluated by the *semivariance* between interpolating and interpolation points. Here, these *semivariance* relations are represented by the *semivariance matrix* and the *distance matrix* ($[\mathbf{C}]_{ij\,[N\times N]}$ and $[\mathbf{D}]_{0i\,[N\times 1]}$).

In *semantic kriging*, the *covariance* between any pair of locations in the terrain is considered to be influenced by the local uncertainties, which includes the effect of the underlying spatial *LULC* classes. Therefore, in *SemK*, the *covariance* between locations are extended to semantic dimension by modeling this terrestrial property. In spatial *SemK*, the semantic and the spatial correlation among the spatial *LULC* classes are captured by four matrix components $[\mathbf{SS}]_{ij\,[N\times N]}$, $[\mathbf{SS}]_{0i\,[N\times 1]}$, $[\mathbf{SI}]_{ij\,[N\times N]}$, and $[\mathbf{SI}]_{0i\,[N\times 1]}$. The *covariance* models the variance of the random field values, both the proposed metrics, the *SS* and the *SI* are inversely proportional to the *covariance* measures among the sampled locations. However, in *OK*, the traditional *semivariance* is proportional to the *covariance* measure. Therefore, in *SemK*, the newly proposed *covariance* between any ith and jth sampled location is given as follows: $\frac{C_{ij}}{SI_{ij} * SS_{ij}}$. The physical significance of the semantic *covariance* of *SemK* is given as follows: being at the same distance with respect to the prediction location, the *covariance* between two sampled location might be different, depending on the semantic properties (*SS* and *SI*) of their representative *LULC* classes. At the same distance d, the *covariance* between two locations increases if *SS* and *SI* are less and vice-versa. Hence, the new *semivariance matrix* and the *distance matrix* ($\mathbf{C}^{SemK}$ and $\mathbf{D}^{SemK}$) of *SemK* are expressed as follows:

$$[\mathbf{C}]_{ij}^{SemK}{}_{[N\times N]} = \frac{[\mathbf{C}]_{ij\,[N\times N]}}{\left([\mathbf{SI}]_{ij\,[N\times N]} \circ [\mathbf{SS}]_{ij\,[N\times N]}\right)} \tag{3.11}$$

$$[\mathbf{D}]_{0i}^{SemK}{}_{[N\times 1]} = \frac{[\mathbf{D}]_{0i\,[N\times 1]}}{\left([\mathbf{SI}]_{0i\,[N\times 1]} \circ [\mathbf{SS}]_{0i\,[N\times 1]}\right)} \tag{3.12}$$

where "— · — · —" and "∘" denote the *Hadamard division* and the *Hadamard product* between matrices, respectively. The $\mathbf{W}^{SemK}$ denotes the *SemK* weight matrix of dimension [N × 1]. The mean square error of *SemK* at x_0 is given as σ^2_{SemK}. Being a variant of *ordinary kriging*, the *SemK* assumes that the *mean* of the random field is constant over the study region. It implies the following equality: $E(\sigma^2_{SemK}) = 0 \Rightarrow \mathbf{1}^T\mathbf{W}^{SemK} = 1$.

Hence, the mean square error for *SemK* (σ^2_{SemK}) is defined as follows:

$$\begin{aligned}\sigma^2_{SemK} &= Var([[\mathbf{W}^{SemK}]^T - 1] \times [Z(x_1)\cdots Z(x_N)Z(x_0)]^T)\\ &= [[\mathbf{W}^{SemK}]^T - 1] \times [Var([Z(x_1)\cdots Z(x_N)Z(x_0)]^T)] \times [[\mathbf{W}^{SemK}]^T - 1]^T\\ &= C^{SemK}_{00} + [\mathbf{W}^{SemK}]^T\mathbf{C}^{SemK}\mathbf{W}^{SemK} - 2[\mathbf{W}^{SemK}]^T\mathbf{D}^{SemK} \qquad (3.13)\end{aligned}$$

where $C^{SemK}_{00} = \frac{C_{00}}{(SI_{00} * SS_{00})}$, C_{00} is $Cov\{Z_0, Z_0\}$, SI_{00} and SS_{00} are the *spatial importance* and *semantic similarity* between (f_0, f_0) respectively. Following the notion of least-square regression approach, *SemK* aims to minimize the mean square error σ^2_{SemK} by minimizing the equation as follows:

$$C^{SemK}_{00} + [\mathbf{W}^{SemK}]^T\mathbf{C}^{SemK}\mathbf{W}^{SemK} - 2[\mathbf{W}^{SemK}]^T\mathbf{D}^{SemK}; \ni [\mathbf{W}^{SemK}]^T\mathbf{1} = 1 \quad (3.14)$$

To optimize the Eq. 3.14 without constraints, a *Lagrange multiplier* is considered in the error expression, by converting a constrained optimization problem into a corresponding unconstrained problem. Let's assume the *Lagrange multiplier* be $-2\lambda_{SemK}$ which makes the Eq. 3.14 unconstrained. If K is the unconstrained error for *SemK*, it is expressed as follows:

$$K = C^{SemK}_{00} + [\mathbf{W}^{SemK}]^T\mathbf{C}^{SemK}\mathbf{W}^{SemK} - 2[\mathbf{W}^{SemK}]^T\mathbf{D}^{SemK} + 2\lambda^{SemK}([\mathbf{W}^{SemK}]^T\mathbf{1} - 1) \quad (3.15)$$

Like any other methods in *kriging* family, the *SemK* can also be mapped to a minimization problem. For minimizing the variance of error, the partial first-order derivative of Eq. 3.15 with respect to the unknowns, i.e., $\mathbf{W}^{SemK}$ and λ^{SemK} is expressed as follows:

$$\frac{\delta K}{\delta \mathbf{W}^{SemK}} = (\mathbf{C}^{SemK} + [\mathbf{C}^{SemK}]^T)\mathbf{W}^{SemK} - 2\mathbf{D}^{SemK} + 2\lambda^{SemK}\mathbf{1} \quad (3.16)$$

$$\frac{\delta K}{\delta \mathbf{W}^{SemK}} = 2\mathbf{C}^{SemK}\mathbf{W}^{SemK} - 2\mathbf{D}^{SemK} + 2\lambda^{SemK}\mathbf{1}(\because [\mathbf{C}^{SemK}]^T = \mathbf{C}^{SemK}) \quad (3.17)$$

$$\frac{\delta K}{\delta \lambda^{SemK}} = 2[\mathbf{W}^{SemK}]^T\mathbf{1} - 2 \quad (3.18)$$

According to the properties of the minimization problem, the partial first-order derivative can be set to zero to get the value at minima (with respect to the $\mathbf{W}^{SemK}$ and λ^{SemK}). Thus, setting $\frac{\delta K}{\delta \mathbf{W}^{SemK}} = 0$ and $\frac{\delta K}{\delta \lambda^{SemK}} = 0$, the following can be obtained:

$$2\mathbf{C}^{SemK}\mathbf{W}^{SemK} - 2\mathbf{D}^{SemK} + 2\lambda^{SemK}\mathbf{1} = 0 \tag{3.19}$$

$$2[\mathbf{W}^{SemK}]^T\mathbf{1} - 2 = 0 \tag{3.20}$$

From Eqs. 3.19 and 3.20, the following can be obtained:

$$\mathbf{C}^{SemK}\mathbf{W}^{SemK} + \lambda^{SemK}\mathbf{1} = \mathbf{D}^{SemK} \tag{3.21}$$

$$[\mathbf{W}^{SemK}]^T\mathbf{1} = 1 \tag{3.22}$$

From Eq. 3.21, the following can be derived:

$$\mathbf{C}^{SemK}\mathbf{W}^{SemK} + \lambda^{SemK}\mathbf{1} = \mathbf{D}^{SemK} \tag{3.23}$$

$$\Rightarrow \mathbf{W}^{SemK} = [\mathbf{C}^{SemK}]^{-1}[\mathbf{D}^{SemK} - \lambda^{SemK}\mathbf{1}] \tag{3.24}$$

Substituting $[\mathbf{C}^{SemK}]_{ij} = \overset{[\mathbf{C}]_{ij}}{\underset{\left([\mathbf{SI}]_{ij}\circ[\mathbf{SS}]_{ij}\right)}{- \cdot - \cdot -}}$ and $[\mathbf{D}^{SemK}]_{0i} = \overset{[\mathbf{D}]_{0i}}{\underset{\left([\mathbf{SI}]_{0i}\circ[\mathbf{SS}]_{0i}\right)}{- \cdot - \cdot -}}$ in Eq. 3.24, the following can be derived:

$$\mathbf{W}^{SemK} = \left[\overset{[\mathbf{C}]_{ij}}{\underset{\left([\mathbf{SI}]_{ij}\circ[\mathbf{SS}]_{ij}\right)}{- \cdot - \cdot -}}\right]^{-1}\left[\left[\overset{[\mathbf{D}]_{0i}}{\underset{\left([\mathbf{SI}]_{0i}\circ[\mathbf{SS}]_{0i}\right)}{- \cdot - \cdot -}}\right] - \lambda^{SemK}\mathbf{1}\right] \tag{3.25}$$

In Eq. 3.23, multiplying both the sides of the equality by $[\mathbf{C}^{SemK}]^{-1}$ and $[\mathbf{1}]^T$ respectively, the following can be obtained:

$$\begin{aligned}
\mathbf{C}^{SemK}\mathbf{W}^{SemK} + \lambda^{SemK}\mathbf{1} &= \mathbf{D}^{SemK} \\
\mathbf{W}^{SemK} + \lambda^{SemK}[\mathbf{C}^{SemK}]^{-1}\mathbf{1} &= [\mathbf{C}^{SemK}]^{-1}\mathbf{D}^{SemK} \\
[\mathbf{1}]^T\mathbf{W}^{SemK} + \lambda^{SemK}[\mathbf{1}]^T[\mathbf{C}^{SemK}]^{-1}\mathbf{1} &= [\mathbf{1}]^T[\mathbf{C}^{SemK}]^{-1}\mathbf{D}^{SemK} \\
1 + \lambda^{SemK}[\mathbf{1}]^T[\mathbf{C}^{SemK}]^{-1}\mathbf{1} &= [\mathbf{1}]^T[\mathbf{C}^{SemK}]^{-1}\mathbf{D}^{SemK}
\end{aligned}$$

[As, $\mathbf{1}^T\mathbf{W}^{SemK} = 1$]

$$\Rightarrow \lambda^{SemK} = \frac{\mathbf{1}^T[\mathbf{C}^{SemK}]^{-1}\mathbf{D}^{SemK} - 1}{\mathbf{1}^T[\mathbf{C}^{SemK}]^{-1}\mathbf{1}} \tag{3.26}$$

Again substituting $[\mathbf{C}]_{ij}^{SemK} = \overset{[\mathbf{C}]_{ij}}{\underset{\left([\mathbf{SI}]_{ij}\circ[\mathbf{SS}]_{ij}\right)}{- \cdot - \cdot -}}$ and $[\mathbf{D}]_{0i}^{SemK} = \overset{[\mathbf{D}]_{0i}}{\underset{\left([\mathbf{SI}]_{0i}\circ[\mathbf{SS}]_{0i}\right)}{- \cdot - \cdot -}}$ in Eq. 3.26, the following can be deduced:

$$\lambda^{SemK} = \frac{\left[\mathbf{1}^T \begin{bmatrix} [\mathbf{C}]_{ij} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}]_{ij} \circ [\mathbf{SS}]_{ij}\right) \end{bmatrix}^{-1} \begin{bmatrix} [\mathbf{D}]_{0i} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}]_{0i} \circ [\mathbf{SS}]_{0i}\right) \end{bmatrix}\right] - 1}{\mathbf{1}^T \begin{bmatrix} [\mathbf{C}]_{ij} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}]_{ij} \circ [\mathbf{SS}]_{ij}\right) \end{bmatrix}^{-1} \mathbf{1}} \tag{3.27}$$

The $\mathbf{W}^{SemK}$ and λ^{SemK} can be calculated from Eqs. 3.25 and 3.27, where $\mathbf{W}^{SemK}$ is the [N × 1] dimensional weight matrix (or N dimensional vector) of *semantic kriging*, given as: $[\mathbf{W}^{SemK}]^T = [w_1^{SemK} w_2^{SemK} \cdots w_N^{SemK}]$. Once it is derived from the real sample points, it can be further normalized in [0,1] to satisfy the constraint $[\mathbf{W}^{SemK}]^T \mathbf{1} = 1$ of Eq. 3.22. In the derivation process of different parameters, the λ^{SemK} should be evaluated first, because it is not dependent on the term $\mathbf{W}^{SemK}$. The predicted parameter value is obtained through the following equation: $\hat{Z}(x_0) = \sum_{i=1}^{N} w_i^{SemK} Z(x_i)$. Here, w_i^{SemK} represents the assigned weight to the sampled location x_i by *SemK*. The minimum variance of *SemK*'s mean square is the following:

$$\sigma^2_{SemK} = C_{00}^{SemK} + [\mathbf{W}^{SemK}]^T \mathbf{D}^{SemK} - \lambda^{SemK} \tag{3.28}$$

$$= \frac{C_{00}}{(SI_{00} * SS_{00})} + [\mathbf{W}^{SemK}]^T \begin{bmatrix} [\mathbf{D}]_{0i} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}]_{0i} \circ [\mathbf{SS}]_{0i}\right) \end{bmatrix} - \lambda^{SemK} \tag{3.29}$$

As σ^2_{SemK} is the error variance, its value should be preferably closer to zero (as small as possible). Therefore, to prove the betterment in *SemK* prediction process, the following relation should be satisfied for any given surface: $\sigma^2_{OK} > \sigma^2_{SemK}$. Further, due to incorporating secondary terrestrial information, the *SemK* has higher information content than *OK* and other interpolation approaches. Sections 3.5 and 3.6 present the comparison of *semantic kriging* with other methods considering two factors: information content and prediction accuracy, respectively.

3.5 Information Content

To assess the performance of *semantic kriging* in terms of its information content is described in this section [7]. For comparison, the information content of *ordinary kriging* is also analyzed, where *OK* is the representative of other geostatistical univariate methods. Let's consider an example in Table 3.2, where each row/tuple represents the specification of a sampled location with respect to three supporting attributes. Also, there is one derived class label attribute. There are six interpolating points ($N = 6$) with supporting attributes as: *Euclidean* distance ($\mathbf{A_d}$), *semantic similarity* ($\mathbf{A_{SS}}$) and *spatial importance* ($\mathbf{A_{SI}}$). The class labell attribute id is the derived weight by the interpolation method denoted as *Assigned Weight*. All the supporting

Table 3.2 Example scenario for information content assessment of *SemK* [7]

Sample point	A_d	A_{SS}	A_{SI}	Assigned weight
x_1	d_{01}	SS_{01}	SI_{01}	w_1^{SemK}
x_2	d_{01}	SS_{01}	SI_{02}	w_2^{SemK}
x_3	d_{02}	SS_{02}	SI_{03}	w_3^{SemK}
x_4	d_{03}	SS_{03}	SI_{04}	w_4^{SemK}
x_5	d_{03}	SS_{03}	SI_{04}	w_5^{SemK}
x_6	d_{04}	SS_{04}	SI_{05}	w_6^{SemK}

attributes are measured with respect to the unsampled location or the interpolation point.

In this scenario, the class label attribute has six unique values for six sampled locations. As *OK* assigns the weight to the sampled locations considering the *Euclidean* distance only, thus for the given example in Table 3.2, $w_1^{OK} = w_2^{OK}$ and $w_4^{OK} = w_5^{OK}$. Because, for pairs of sampled locations 1, 2 and 4, 5, the supporting attribute *Euclidean* distance ($\mathbf{A_d}$) are the same values, d_{01} and d_{03}, respectively.

Let's consider that the number of unique class labels for *ordinary kriging* is m^{OK}, which is referred to as C_i (i = 1 $\cdots$ m^{OK}). Thus, every distinct class label C_i is associated to a unique weight value, assigned to the sampled locations. According to Table 3.2, $m^{OK} = 4$, $C_1 = w_1^{SemK} = w_2^{SemK}$, $C_2 = w_3^{SemK}$, $C_3 = w_4^{SemK} = w_5^{SemK}$, $C_4 = w_6^{SemK}$. Let C_i^{OK} be the set of sampled locations that correspond to the class (of assigned weight) C_i. Therefore, $\sum_{i=1}^{m^{OK}} C_i^{OK}$ is equal to the number of sample points N. If p_i^{OK} denotes the probability that the weight of an interpolating point corresponds to class C_i, then $p_i^{OK} = \frac{C_i^{OK}}{N}$. Then the information content [9] for *OK*, represented as I^{OK} is expressed as follows:

$$I^{OK} = -\sum_{i=1}^{m^{OK}} p_i^{OK} log_2(p_i^{OK}) \tag{3.30}$$

According to Eq. 3.30 then, for the example scenario in Table 3.2, the information content for *OK* can be evaluated as follows:

$$\begin{aligned} I^{OK} &= -\frac{2}{6}log_2(\frac{2}{6}) - \frac{1}{6}log_2(\frac{1}{6}) - \frac{2}{6}log_2(\frac{2}{6}) - \frac{1}{6}log_2(\frac{1}{6}) \\ &= 1.92 \end{aligned} \tag{3.31}$$

On the other hand, for *SemK*, though $w_4^{SemK} = w_5^{SemK}$ (as they match in all the supporting attributes), $w_1^{SemK} \neq w_2^{SemK}$. The reason is, though first and second tuple or the sampled location have similar *Euclidean* distance as d_{01}, their semantic supporting attribute $\mathbf{A_{SI}}$ vary, i.e., SI_{01} and SI_{02}, respectively. This extra knowledge is exclusively captured in the prediction process by *SemK*. Considering more number of supporting attributes, the number of unique class labels reported by *SemK* (m^{SemK})

for a given scenario is always greater than or equal to that of *OK*. Thus, the following is true: $m^{SemK} \geq m^{OK}$. For the given example in Table 3.2, the unique number of class labels evaluated by *SemK* is 5, but for *OK* it is 4. Correspondingly, $m^{SemK} \geq m^{OK} \Rightarrow p_i^{OK} \geq p_i^{SemK}$. The general estimation equation of the expected information content for *SemK* (I^{SemK}) is presented as follows:

$$I^{SemK} = -\sum_{i=1}^{m^{SemK}} p_i^{SemK} log_2(p_i^{SemK}) \tag{3.32}$$

And for the example in Table 3.2, the information content in *SemK* is estimated as follows:

$$\begin{aligned} I^{SemK} &= -\frac{1}{6}log_2(\frac{1}{6}) - \frac{1}{6}log_2(\frac{1}{6}) - \frac{1}{6}log_2(\frac{1}{6}) - \frac{2}{6}log_2(\frac{2}{6}) - \frac{1}{6}log_2(\frac{1}{6}) \\ &= 2.25 \end{aligned} \tag{3.33}$$

Therefore, for the given example scenario in Table 3.2, $I^{SemK} \geq I^{OK}$, i.e., information content of *SemK* is always greater than or equal to *ordinary kriging*. Here, as *OK* is considered as the representative of all the univariate spatial interpolation methods. Thus, in general, the information content for most of the other existing univariate interpolation methods are always less than or equal to *SemK*.

3.6 Empirical Proof for SemK

Empirical experimentation is performed with *land surface temperature* (*LST*) data captured by the Landsat-7 ETM+ satellite imagery, by United States Geological Survey (USGS). For the empirical evaluation of the spatial *SemK* and comparing it with other existing interpolation methods, the *LST*s of two cities have been considered: Kolkata, WB, India and Dallas, TX, USA. Five spatial zones from each of these cities have been considered as the region of interest (*RoI*) in this chapter. These zones are depicted in Figs. 3.3 and 3.4. For this study, the satellite image of the year 2015 (within the range of mid-October–mid-November) has been considered. The performance of *SemK*, in terms of prediction accuracy, is compared with other univariate interpolation methods. In this work, four popular interpolation techniques, considered for the empirical comparison, are as follows:

- *nearest neighbors* (*NN*)
- *inverse distance weighting* (*IDW*)
- *universal kriging* (*UK*)
- *ordinary kriging* (*OK*)

For this particular empirical evaluation, the generic experimental specifications are given as follows:

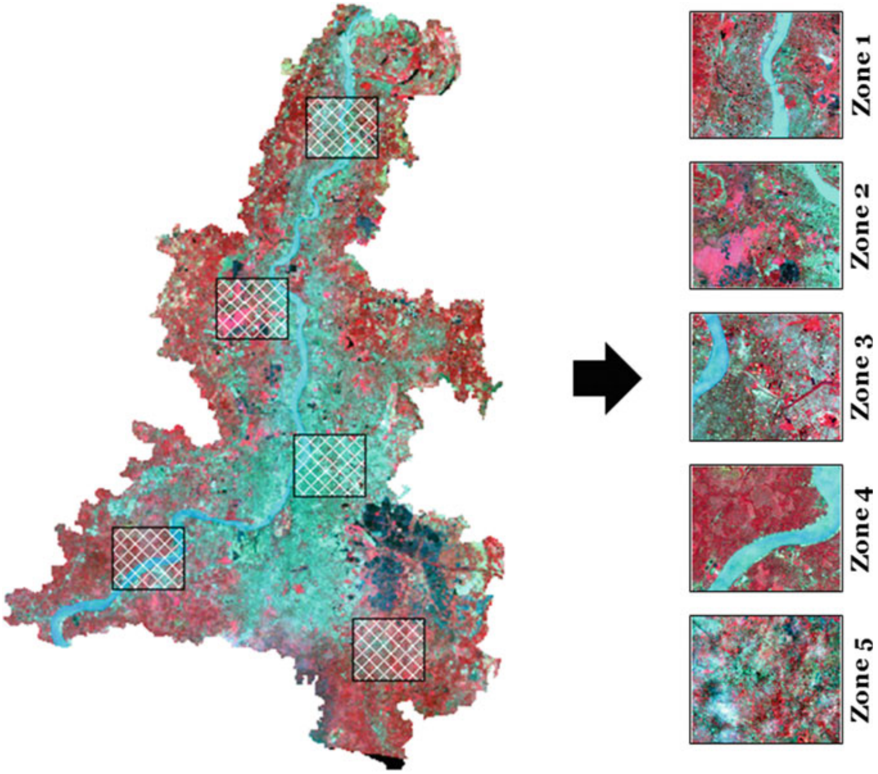


Fig. 3.3 Selected spatial zones of Kolkata, WB, India for *SemK*

- around 500 random sampled locations (uniformly random) are considered for modeling the experimental *semivariogram* with lag distance $h = 5$ km.
- a fixed search radius of 1 km is considered against each unsampled prediction point to select the interpolating points.
- around 20 interpolating points are selected randomly against each unsampled prediction point, within a predefined radius.
- the *exponential semivariogram* model is considered for both *OK* and *SemK*.
- the *linear semivariogram* model (with linear drift) is considered for *UK*.

The performance of the considered approaches, including *SemK*, is evaluated by two standard error metrics: *mean absolute error* (*MAE*) and *root mean square error* (*RMSE*). The graphical representations of the error measures by different methods are depicted in Figs. 3.5 and 3.6. It has been observed that the *SemK* outperforms the considered approaches, mainly the *ordinary kriging*, in terms of prediction accuracy.

To validate the performance of *SemK* to generate the mapping surfaces, the predicted imagery of the selected zones of Kolkata, WB, India and Dallas, TX, USA are depicted in Tables 3.7 and 3.8. The bounding box (BB) of a zone is specified in

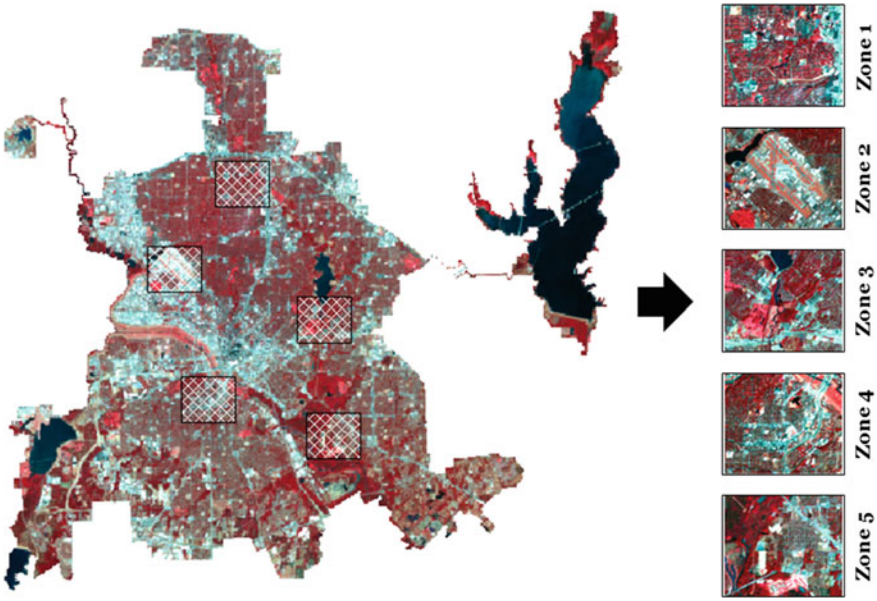
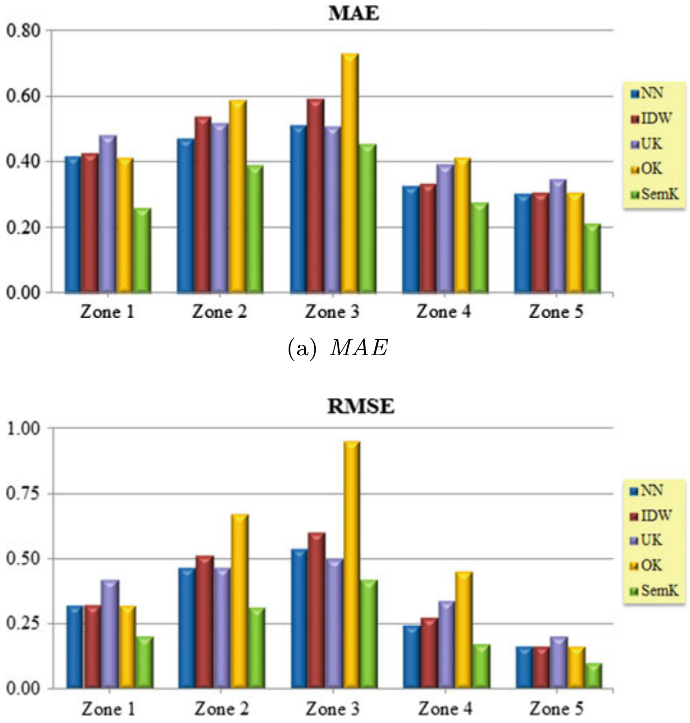


Fig. 3.4 Selected spatial zones of Dallas, TX, USA for *SemK*

the table itself as [lower-left upper-right] corner. The actual *land surface temperature* imagery and the predicted imagery by different methods (*NN*, *IDW*, *UK*, *OK*, and *SemK*) are tabularized respectively. For each of the predicted image, the error surfaces are also produced in gray scale where black pixel represents higher error and white represents lower error. Corresponding *peak signal-to-noise* ratio is also reported, which is evaluated against the actual surface of the respective zone.

3.6.1 Discussions on Empirical Proof

It is evident from Figs. 3.5 and 3.6 that the *SemK* outperforms the *ordinary kriging* and other popular interpolation methods by incorporating the terrestrial semantic *LULC* knowledge into the prediction process. It may also be observed from Tables 3.7 and 3.8 that the imagery (of the *LST* distribution of the terrain), that are predicted by *SemK* are more accurate compared to that of produced by *OK* and other methods. The *SemK* reports higher *peak signal-to-noise ratio* than *NN*, *IDW*, and *OK* (≈2–8 dB for Kolkata, WB, India and ≈3–10 dB for Dallas, TX, USA).



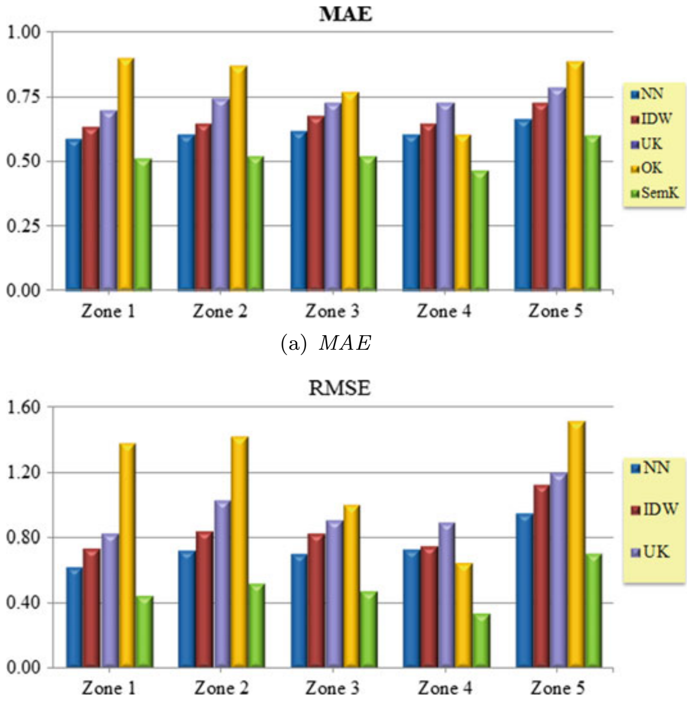
(a) *MAE*

(b) *RMSE*

Fig. 3.5 Comparison study with error graph for *SemK* (Region: Kolkata)

3.7 Theoretical Performance Evaluation of SemK

This section presents a theoretical performance analysis of *semantic kriging* in order to prove its efficacy for modeling the semantic *LULC* knowledge for spatial interpolation, verify its functionality, and establish its formal relationship with *ordinary kriging* method. It basically relates different modified parameters of *SemK* with the basic parameters of *OK*. This analysis theoretically investigates the capability of *SemK* in incorporating the domain knowledge into the interpolation process. At the same time, this study also validates the impact of the *LULC* ontology and its structure, granularity, to establish the benefits of using *SemK*. As the notion of *ordinary kriging* is considered as the building block for *SemK*, this theoretical performance evaluation has been compared with *OK*. However, the similar analysis can be done considering other univariate *kriging* approaches as well. The *Euclidean* vector analysis approach is performed for the formal proofs, which are presented further as four *lema*s and a *proposition*. Each of these proofs exhibit some significant characteristics of *semantic kriging*. The *SemK* process assumes that the sampled locations represented by the spatially correlated and semantically similar *LULC* classes should be having



(a) *MAE*

(b) *RMSE*

Fig. 3.6 Comparison study with error graph for *SemK* (Region: Dallas)

more influence to the unsampled prediction point, and thus more weight should be assigned, compared to the loosely correlated and less similar sample points. All the proofs rely on this assumption. The idea and the outlines of the *lema*s and *proposition* are stated as follows:

- ***Lemma*** 3.1: The semantically similar and spatially correlated sampled locations are assigned more weight by the *semantic kriging* process, compared to other points.
- ***Lemma*** 3.2: The angular difference between the *OK*'s and the *SemK*'s weight vectors represents the extra amount of domain knowledge (in terms of *LULC* distribution) captured by *SemK* over *OK*. That is, the *LULC* knowledge is correctly captured and modeled by *SemK*.
- ***Lemma*** 3.3: The optimal weight vector is closer to the one that is produced by *SemK*, compared to the weight vector of *OK*. Thus, *SemK* outperforms *OK*.
- ***Lemma*** 3.4: More generalized ontology will eventually converge *SemK* to *OK*.
- ***Proposition*** 3.1: The misclassified sample points in the ontology can be identified by a preprocessing of *SemK* method.

Table 3.3 An example scenario for the theoretical analysis of *SemK* [3]

Sample point	Representative *LULC* class	A_d	A_{SI}	A_{SS}	Assigned weight
x_0	f_0	0	1	1	–
x_i	f_i	d_{0i}	SI_{0i}	SS_{0i}	w_i^{SemK}
x_j	f_j	d_{0j}	SI_{0j}	SS_{0j}	w_j^{SemK}
x_k	f_k	d_{0k}	SI_{0k}	SS_{0k}	w_k^{SemK}

To further proceed with the proofs, an example scenario of interpolation process is considered in Table 3.3. The scenario consists of three sampled interpolating points x_i, x_j, and x_k against one prediction point x_0. Similar to Table 3.2, three supporting attributes are: *Euclidean* distance (A_d), *spatial importance* (A_{SI}), and *semantic similarity* (A_{SS}), which are measured with reference to the interpolation point x_0. The *assigned weight* (w^{SemK}) is the class label attribute. Given the supporting attributes, the class label attribute is evaluated by *SemK* for each of the interpolating points.

Lemma 3.1 *Between a pair of sampled locations at the same Euclidean distance from the unsampled prediction location, SemK assigns more weight to the location which is represented with more similar LULC class to the representative LULC of the prediction point, compared to the other one.*

Proof From the Table 3.3, let's consider a pair of interpolating points x_i and x_j and let their *Euclidean* distances from the unsampled location be same, i.e., $d_{0i} = d_{0j}$. Let the representative *LULC* class of x_i, i.e., f_i be more similar and correlated with f_0 than f_j, the representative *LULC* class of x_j. Now, according to the Tobler's law of spatial proximity [15] and the properties of hierarchical ontology property, the following holds: $(SI_{0i} * SS_{0i}) > (SI_{0j} * SS_{0j}) \Rightarrow SIS_{0i} > SIS_{0j}$, where $(SI_{mn} * SS_{mn})$ is referred to as SIS_{mn}. Now, it is needed to be proved: $w_i^{SemK} > w_j^{SemK}$. According to [7], the *covariance matrix* and the *distance matrix* of *SemK* ($\mathbf{C}^{SemK}$ and $\mathbf{D}^{SemK}$) are given as follows:

$$\mathbf{C}^{SemK} = \begin{bmatrix} \frac{\gamma(d_{ii})}{1*1} & \frac{\gamma(d_{ij})}{SIS_{ij}} \\ \frac{\gamma(d_{ji})}{SIS_{ji}} & \frac{\gamma(d_{jj})}{1*1} \end{bmatrix} \qquad \mathbf{D}^{SemK} = \begin{bmatrix} \frac{\gamma(d_{0i})}{SIS_{0i}} \\ \frac{\gamma(d_{0j})}{SIS_{0j}} \end{bmatrix} \tag{3.34}$$

Now, the weight matrix of *SemK*, $\mathbf{W}^{SemK}$ is given as follows: $[\mathbf{C}^{SemK}]^{-1}$ $[\mathbf{D}^{SemK} - \lambda^{SemK}\mathbf{1}]$. Considering the normalization constraint of *SemK* ($\mathbf{1}^T\mathbf{W}^{SemK} = 1$), let the expression $[\mathbf{D}^{SemK} - \lambda^{SemK}\mathbf{1}]$ be referred to as $\mathbb{D}$, where $\mathbb{D}_{0i} = (\frac{\gamma(d_{0i})}{SIS_{0i}} - \lambda^{SemK})$. Hence, $\mathbf{W}^{SemK}$ is modified as $[\mathbf{C}^{SemK}]^{-1}\mathbb{D}$ and is given as follows:

Table 3.4 Empirical study of *Lemma* 3.1 [3]

Point type	Point no.	Representative *LULC* class	Similarity score	Assigned weight
Prediction point	—	Commercial	—	—
Interpolating points	1	Land with scrub	0.194	0.128
	2	Industrial	0.525	0.357
	3	Residential	0.623	0.515

$$\mathbf{W}^{SemK} = \begin{bmatrix} w_i^{SemK} \\ \\ w_j^{SemK} \end{bmatrix} = \frac{1}{K}\begin{bmatrix} -\gamma(d_{jj}) & \frac{\gamma(d_{ij})}{SIS_{ij}} \\ \frac{\gamma(d_{ji})}{SIS_{ji}} & -\gamma(d_{ii}) \end{bmatrix}\begin{bmatrix} \mathbb{D}_{0i} \\ \\ \mathbb{D}_{0j} \end{bmatrix} \tag{3.35}$$

$$= \frac{1}{K}\begin{bmatrix} \frac{\gamma(d_{ij})*\mathbb{D}_{0j}}{SIS_{ij}} \\ \frac{\gamma(d_{ij})*\mathbb{D}_{0i}}{SIS_{ij}} \end{bmatrix} \tag{3.36}$$

where $K = -\gamma(d_{ii})\gamma(d_{jj}) + \frac{\gamma(d_{ij})^2}{(SI_{ij}*SS_{ij})^2}$ and d_{mn} represent *Euclidean* distance between point x_m and x_n. Now, from the notion of *ordinary kriging*, $\gamma(d_{ii}) = \gamma(d_{jj}) = 0$ (as *self-covariance* in the same location is 0) and $d_{ij} = d_{ji}$ (as *Euclidean* distance) is omnidirectional). From the definition of *SemK*, $SI_{ij} = SI_{ji}$ and $SS_{ij} = SS_{ji} \Rightarrow SIS_{ij} = SIS_{ji}$. Therefore, K is modified as $\frac{\gamma(d_{ij})^2}{(SIS_{ij})^2}$. Hence, from the definition of $\mathbb{D}$, the following inequality holds: $w_i^{SemK} > w_j^{SemK}$. This concludes the proof.

For the empirical proof of ***Lemma*** 3.1, a real scenario is considered with three interpolating points, which are at the same distance, but in different directions from the prediction point. The representative *LULC* classes are also specified in the table. As mentioned in the description of *SemK*, the measured semantic metrics are specified in (0, 1]. It may be observed from the Table 3.4 that the assigned weight by *SemK* to the sampled locations increases with the increment of their semantic score. The result proves the claim of ***Lemma*** 3.1.

Lemma 3.2 *The additional semantic knowledge captured by SemK for the sampled and unsampled locations with the semantic similarity and spatial importance metric is correctly modeled as the semantic weight assigned by semantic kriging. This auxiliary domain knowledge that is exclusively captured by SemK (not by OK), and can be characterized as the angular difference between SemK's and OK's weight vectors.*

Proof This *lema* analyzes the amount of semantic change that is captured in the weight vector of *SemK* over *OK*. Here, it is needed to prove that the angular difference between the weight vectors of *OK* and *SemK*, $\overrightarrow{\mathbf{W}^{OK}}$ and $\overrightarrow{\mathbf{W}^{SemK}}$, is equal to the amount of extra semantic knowledge captured by *SemK*, compared to *OK*, i.e., the angular difference between the domain knowledge captured in both the methods. Thus, the formal statement of the *lema* is given as follows:

$$\theta_{(\overrightarrow{\mathbf{Change}},(1-\lambda^{OK})\overrightarrow{\mathbf{1}})} = \theta_{(\overrightarrow{\mathbf{W}^{OK}_{sem}},\overrightarrow{\mathbf{W}^{SemK}_{sem}})}$$

where λ^{OK} is the *Lagrange multiplier* of *OK*. For N interpolating points, $\overrightarrow{\mathbf{Change}}$ is a *Euclidean* vector given as $[Change_1 Change_2 \cdots Change_N]^T$, where $Change_i$ represents the semantic knowledge of the ith interpolating point captured by *SemK* over *OK*. Therefore, $Change_i$ is modeled with respect to the SI and the SS metric with respect to the *LULC* classes. It is expressed as follows:

$$Change_i = \overrightarrow{\mathbf{SIS_{iN}}} \circledast (\overrightarrow{\mathbf{SIS_{0N}}} - \overrightarrow{\lambda^{SemK}\mathbf{1}})$$

Formally, $\overrightarrow{\mathbf{SIS_{iN}}}$ and $\overrightarrow{\mathbf{SIS_{0N}}}$ represent the ith rows of two matrices, $[([\mathbf{SI_{ij}}]_{\mathbf{N\times N}} \circ [\mathbf{SS_{ij}}]_{\mathbf{N\times N}})^{-H}]^{-1}$ and $([\mathbf{SI_{0i}}]_{\mathbf{N\times 1}} \circ [\mathbf{SS_{0i}}]_{\mathbf{N\times 1}})^{-H}$, respectively. The $\circledast$ and the $-H$ represent the *dot product* and the *Hadamard inverse* between matrices. For sampled locations in Table 3.3, x_i and x_j, the $\overrightarrow{\mathbf{Change}}$ matrix can be expressed as follows:

$$\mathbf{Change} = \begin{bmatrix} Change_i \\ Change_j \end{bmatrix} \tag{3.37}$$

$$= [[[\mathbf{SI_{ij}}] \circ [\mathbf{SS_{ij}}]]^{-H}]^{-1}_{2\times 2}[[[\mathbf{SI_{0i}}] \circ [\mathbf{SS_{0i}}]]^{-H}_{2\times 1} - [\lambda^{SemK}\mathbf{1}]_{2\times 1}] \tag{3.38}$$

$$= \frac{1}{K_{change}} \begin{bmatrix} \frac{1}{SIS_{ij} * \mathbb{SIS}_{0j}} \\ \frac{1}{SIS_{ij} * \mathbb{SIS}_{0i}} \end{bmatrix} \tag{3.39}$$

The $\overrightarrow{\mathbf{1}}$ denotes *OK*'s semantic vector, the $\mathbb{SIS}^{-H}$ can be expressed as $(\mathbf{SIS}^{-H} - \lambda^{SemK}\mathbf{1})$. While quantifying the semantic knowledge, d_{0i} is exactly equal to d_{0j}. With respect to these constraints and after normalizing the *distance matrices* of both *OK* and *SemK*, the weight matrices of *OK* and *SemK* ($\mathbf{W}^{OK}$ and $\mathbf{W}^{SemK}$) are expressed as follows:

$$\mathbf{W}^{OK} = \frac{1}{K} \begin{bmatrix} \gamma(d_{ij}) * (1 - \lambda^{OK}) \\ \gamma(d_{ij}) * (1 - \lambda^{OK}) \end{bmatrix} \tag{3.40}$$

$$\mathbf{W}^{SemK} = \frac{1}{K'} \begin{bmatrix} \frac{\gamma(d_{ij})}{SIS_{ij}} * (\frac{1}{SIS_{0j}} - \lambda^{SemK}) \\ \frac{\gamma(d_{ij})}{SIS_{ij}} * (\frac{1}{SIS_{0i}} - \lambda^{SemK}) \end{bmatrix} \tag{3.41}$$

where K' is a constant for *SemK*. With respect to the semantic knowledge of the surrounding spatial *LULC* classes, the angular difference between the *LULC* knowledge captured by *SemK* over *OK* is expressed as follows:

$$\theta_{(\overrightarrow{\mathbf{Change}},1-\lambda^{OK}\overrightarrow{\mathbf{1}})} = Cos^{-1}\left(\frac{1}{\sqrt{2}} \frac{\mathbb{SIS}_{0i} + \mathbb{SIS}_{0j}}{\sqrt{\mathbb{SIS}^2_{0i} + \mathbb{SIS}^2_{0j}}}\right) \tag{3.42}$$

Table 3.5 Empirical study of *Lemma* 3.2 [3]

Scenario	Point no.	Representative *LULC* class	Similarity score	Assigned weight
1	1	Land with scrub	0.194	0.128
	2	Industrial	0.525	0.357
	3	Residential	0.623	0.515
	Semantic angular difference		Weight vectors' angular difference	
	25.48°		25.48°	
2	1	Cropland	0.570	0.629
	2	Residential	0.180	0.073
	3	Wetlands	0.344	0.297
	Semantic angular difference		Weight vectors' angular difference	
	34.40°		34.40°	
3	1	River	0.218	0.120
	2	Commercial	0.623	0.657
	3	Forest	0.315	0.223
	Semantic angular difference		Weight vectors' angular difference	
	34.88°		34.88°	

As, for the semantic knowledge, $d_{0i} = d_{0j}$ (refer Table 3.3), the angular difference between $\overrightarrow{\mathbf{W}_{sem}^{OK}}$ and $\overrightarrow{\mathbf{W}_{sem}^{SemK}}$ is given as $\theta_{(\overrightarrow{\mathbf{W}_{sem}^{OK}}, \overrightarrow{\mathbf{W}_{sem}^{SemK}})}$. Now, $Cos_{(\overrightarrow{\mathbf{W}_{sem}^{OK}}, \overrightarrow{\mathbf{W}_{sem}^{SemK}})} \theta$ is given as follows:

$$= \frac{(1-\lambda^{OK})(\frac{1}{SIS_{0j}} - \lambda^{SemK}) + (1-\lambda^{OK})(\frac{1}{SIS_{0i}} - \lambda^{SemK})}{\sqrt{(1-\lambda^{OK})^2 + (1-\lambda^{OK})^2}\sqrt{\left(\frac{1}{SIS_{0j}} - \lambda^{SemK}\right)^2 + \left(\frac{1}{SIS_{0i}} - \lambda^{SemK}\right)^2}} \tag{3.43}$$

$$= \frac{1}{\sqrt{2}}\left(\frac{\mathbb{SIS}_{0i} + \mathbb{SIS}_{0j}}{\sqrt{\mathbb{SIS}_{0i}^2 + \mathbb{SIS}_{0j}^2}}\right) \tag{3.44}$$

which implies $\theta_{(\overrightarrow{\mathbf{W}_{sem}^{OK}}, \overrightarrow{\mathbf{W}_{sem}^{SemK}})} = Cos^{-1}\frac{1}{\sqrt{2}}\left(\frac{\mathbb{SIS}_{0i}+\mathbb{SIS}_{0j}}{\sqrt{\mathbb{SIS}_{0i}^2+\mathbb{SIS}_{0j}^2}}\right)$.

Hence, $\theta_{(\overrightarrow{\mathbf{Change}}, (1-\lambda^{OK})\vec{\mathbf{1}})} = \theta_{(\overrightarrow{\mathbf{W}_{sem}^{OK}}, \overrightarrow{\mathbf{W}_{sem}^{SemK}})}$. This concludes the proof.

For empirical validation of this *lema*, three real scenarios have been considered, which satisfies the specifications given in Table 3.3. The weight vectors of both *OK* and *SemK* and the semantic vectors are compared for each scenario in Table 3.5. From the table, it is observed that the angle between the weight vectors of *OK* and *SemK* is equal to the angle between the semantic knowledge captured, i.e., the semantic knowledge vectors.

Lemma 3.3 *The optimal weight vector for an interpolation scenario is closer to weight vector produced by SemK compared to that of OK.*

Proof To prove the *lemma*, an alternate proof can be conducted which validates that the angle between the optimal weight vector of an interpolation scenario and the weight vector produced by *SemK* is smaller compared to the angle between optimal and *OK*'s weight vector. Formally, it can be expressed as $\theta_{(\overrightarrow{\mathbf{W}^{OK}},\overrightarrow{\mathbf{W}^{OPT}})} > \theta_{(\overrightarrow{\mathbf{W}^{SemK}},\overrightarrow{\mathbf{W}^{OPT}})}$.

Let us assume for any two interpolating points x_i and x_j, the optimal weight vector is $[w_i^{OPT} w_j^{OPT}]^T$. Let's consider $\mathbf{U}$ be that extra domain knowledge by incorporation of which we can get the optimal solution, i.e., the optimal weight vector. For ***Lemma*** 3.1, the following inequality holds: if $(w_i^{SemK} - w_j^{SemK}) \geq 0$, then $(w_i^{OPT} - w_j^{OPT}) \geq (w_i^{SemK} - w_j^{SemK}) \geq 0$ and vice-versa. The physical significance of this inequality can be stated as, similar to the semantic knowledge (in terms of *semantic similarity* and *spatial importance*), the knowledge $\mathbf{U}$ is also inversely proportional to the traditional *covariance* which is based on *Euclidean* distance. The $\mathbf{U}$ also ranges between positive real values of (0, 1]. Now, considering ***Lemma*** 3.2, $Cos_{(\overrightarrow{\mathbf{W}^{OK}},\overrightarrow{\mathbf{W}^{OPT}})}\theta$ is expressed as follows:

$$\frac{(\frac{1}{SIS_{0j}U_{0j}} - \lambda^{OPT}) + (\frac{1}{SIS_{0i}U_{0i}} - \lambda^{OPT})}{\sqrt{2}\sqrt{\left(\frac{1}{SIS_{0j}U_{0j}} - \lambda^{OPT}\right)^2 + \left(\frac{1}{SIS_{0i}U_{0i}} - \lambda^{OPT}\right)^2}}$$

Similarly, $Cos_{(\overrightarrow{\mathbf{W}^{SemK}},\overrightarrow{\mathbf{W}^{OPT}})}\theta$ is expressed as $\frac{(\frac{1}{U_{0j}}-\lambda^{OPT})+(\frac{1}{U_{0i}}-\lambda^{OPT})}{\sqrt{2}\sqrt{(\frac{1}{U_{0j}}-\lambda^{OPT})^2+(\frac{1}{U_{0i}}-\lambda^{OPT})^2}}$. As, the semantic score considering any two sampled locations, SIS_{ij} is inversely proportional to their *Euclidean* distance based traditional *covariance* measure and ranges between (0, 1], thus the following is true:

$$Cos_{(\overrightarrow{\mathbf{W}^{OK}},\overrightarrow{\mathbf{W}^{OPT}})}\theta < Cos_{(\overrightarrow{\mathbf{W}^{SemK}},\overrightarrow{\mathbf{W}^{OPT}})}\theta \Rightarrow \theta_{(\overrightarrow{\mathbf{W}^{OK}},\overrightarrow{\mathbf{W}^{OPT}})} > \theta_{(\overrightarrow{\mathbf{W}^{SemK}},\overrightarrow{\mathbf{W}^{OPT}})}$$

Hence, it is proved that the optimal weight vector for an interpolation scenario is closer to weight vector produced by SemK compared to that of OK.

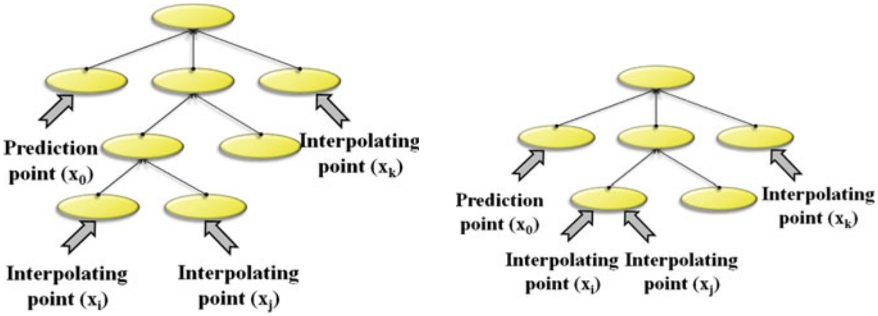
The interpolation scenario given in Table 3.3 is further considered for empirically analyzing ***Lemma*** 3.3. An additional terrestrial knowledge, "elevation of the earth surface", which is also significant for the interpolation accuracy beside the semantic knowledge has been chosen for the improvement of prediction. These three knowledge, the *LULC*-based *semantic property*, the traditional *Euclidean* distance, and the newly considered *elevation* are considered to be sufficient information to produce the optimal weight vector for this interpolation scenario. The angles between optimal weight vector with the weight vectors produced by *OK* and *SemK*, respectively, are tabularized in Table 3.6. The result supports the claim of ***Lemma*** 3.3.

Lemma 3.4 *More generalized ontology will eventually converge SemK to OK.*

Proof Let us again consider three sampled locations as given example scenario in Table 3.3. Further, let there be two ontologies, as given in Fig. 3.7, one of which is more general (refer Fig. 3.7b) and the other one is more specific as one leaf *LULC* class was previously split into two specialized classes (refer Fig. 3.7a). The positions

Table 3.6 Empirical study of *Lemma* 3.3 [3]

Scenario	Angle with the optimal weight vector	
	OK	*SemK*
1	49.95°	29.90°
2	34.40°	5.45°
3	29.09°	10.70°


Prediction point (x$_0$)
Interpolating point (x$_k$)
Interpolating point (x$_i$)
Interpolating point (x$_j$)
Prediction point (x$_0$)
Interpolating point (x$_k$)
Interpolating point (x$_i$)
Interpolating point (x$_j$)

(a) Position before modification (b) Position after modification

Fig. 3.7 Position of three sampled locations in the ontologies before and after modification [3]

of the representative *LULC* classes of the sample points are identified in both the ontologies.

In Fig. 3.7a, i.e., for the initial ontology, the sampled locations x_i and x_j are represented by two different specialized *LULC* classes in the hierarchy. After modification, i.e., in the ontology of Fig. 3.7b, let those two specialized classes are converged to the same parent class. Therefore, now both the sampled locations, x_i and x_j are represented by the same parent *LULC* class in the hierarchy. According to the *SemK* process and with the newly modified ontology (*SemK_mod*) in Fig. 3.7b, the semantic metrics for both x_i and x_j (with reference to the prediction point x_0) have been changed to $SI_{0i}^{SemK\_mod}$, $SS_{0i}^{SemK\_mod}$ and $SI_{0j}^{SemK\_mod}$, $SS_{0j}^{SemK\_mod}$ respectively. However, the semantic properties for the third interpolating point x_k do not change.

Now, according to the hierarchical ontology property, it must be noted that more general classes are assigned higher *semantic similarity* score with respect to others, compared to a specialized *LULC* class in the same path of an hierarchy [7]. And in order to evaluate the *spatial importance* of the parent class, let's assume m number of specialized classes are converged to their parent *LULC* class. Further, let k be the predefined number of sampled location that have been chosen for each of those m specialized/child classes. After modification in the ontology, now the parent class consists of total $(m * k)$ sampled locations. Among these, the first k sampled locations, which are closer to each other within a predefined radius, are considered further correlation stud. According to the Tobler's law of spatial proximity [15], this modified correlation scores of the parent classes are always higher compared to their specialized classes. Therefore, for any $p = i, j$ and $q = 1, k$, the following is true:

Table 3.7 Comparison study for *SemK* (Region: Kolkata, WB, India)

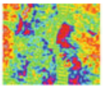
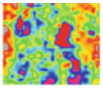
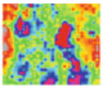
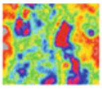
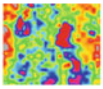
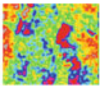
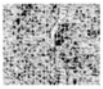
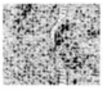
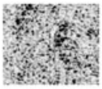
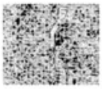
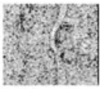
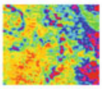
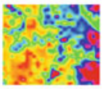
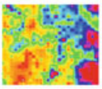
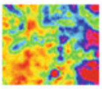
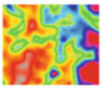
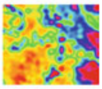
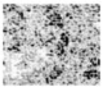
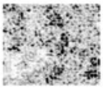
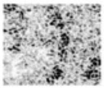
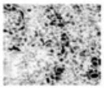
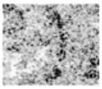
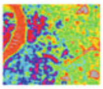
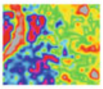
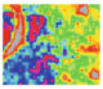
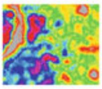
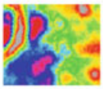
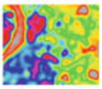
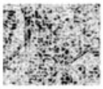
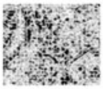
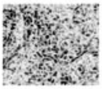
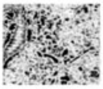
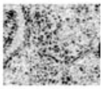
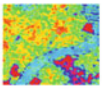
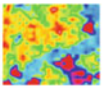
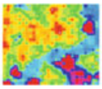
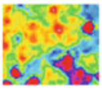
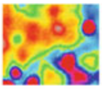
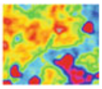
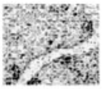
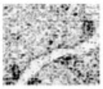
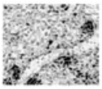
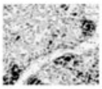
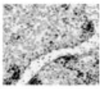
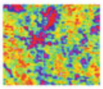
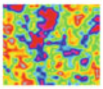
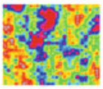
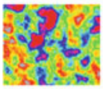
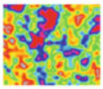
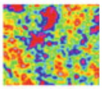
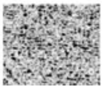
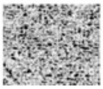
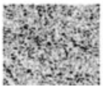
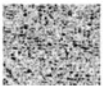
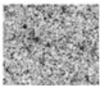
Zone	Actual image	Predicted image				
		NN	IDW	UK	OK	SemK
Zone 1	BB: [(88°21′56.75″E 22°53′3.338″N); (88°26′11.684″E 22°56′20.576″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	41.11dB	40.27dB	38.74dB	41.11dB	45.06dB
Zone 2	BB: [(88°16′37.345″E 22°43′16.086″N); (88°20′51.518″E 22°46′33.581″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	38.24dB	37.34dB	38.24dB	35.04dB	41.71dB
Zone 3	BB: [(88°21′3.362″E 22°34′43.896″N); (88°25′17.437″E 22°38′1.121″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	37.10dB	36.10dB	37.76dB	32.13dB	39.29dB
Zone 4	BB: [(88°10′24.797″E 22°29′45.122″N); (88°14′38.046″E 22°33′3.055″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	43.63dB	42.66dB	40.84dB	38.30dB	46.67dB
Zone 5	BB: [(88°24′21.572″E 22°24′41.309″N); (88°28′35.399″E 22°27′58.463″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	45.78dB	45.62dB	43.91dB	45.75dB	50.08dB

Table 3.8 Comparison study for *SemK* (Region: Dallas, TX, USA)

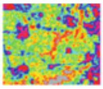
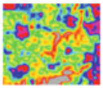
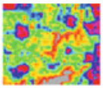
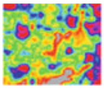
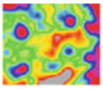
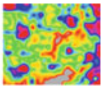
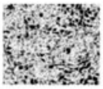
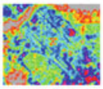
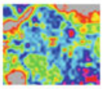
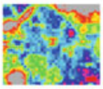
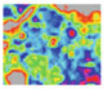
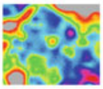
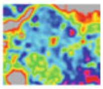
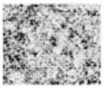
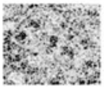
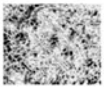
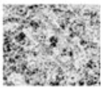
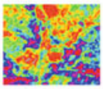
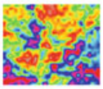
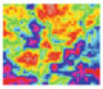
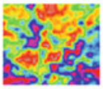
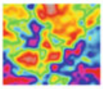
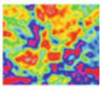
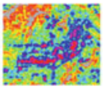
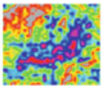
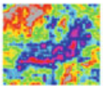
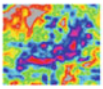
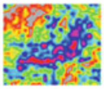
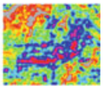
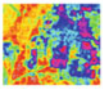
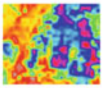
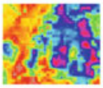
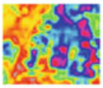
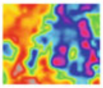
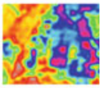
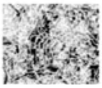
Zone	Actual image	Predicted image				
		NN	IDW	UK	OK	SemK
Zone 1	BB: [(96°48′52.626″W 32°53′19.252″N); (96°46′4.031″W 32°55′15.003″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	35.27dB	33.83dB	32.83dB	28.38dB	38.16dB
Zone 2	BB: [(96°52′26.827″W 32°49′41.121″N); (96°49′38.61″W 32°51′36.489″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	34.59dB	33.33dB	31.55dB	28.78dB	37.44dB
Zone 3	BB: [(96°44′48.861″W 32°47′22.359″N); (96°42′0.958″W 32°49′17.557″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	34.37dB	32.93dB	32.18dB	31.30dB	37.84dB
Zone 4	BB: [(96°50′50.616″W 32°44′1.516″N); (96°48′2.552″W 32°45′56.856″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	34.10dB	33.81dB	32.36dB	35.08dB	40.82dB
Zone 5	BB: [(96°44′27.666″W 32°42′20.309″N); (96°41′39.631″W 32°44′15.74″N)]					
						
	Error surfaces (In gray scale) High Low					
	PSNR	32.03dB	30.53dB	29.97dB	27.94dB	34.65dB

Table 3.9 Empirical study of *Lemma* 3.4 [3]

Scenario	Angle with the weight vector of *OK*	
	SemK with initial ontology	*SemK* with modified ontology
1	25.48°	24.19°
2	39.68°	30.22°
3	34.88°	29.47°

$$1 \geq SIS_{pq}^{SemK\_mod} > SIS_{pq} > 0$$

Therefore, for both x_i and x_j, $w_i^{SemK} < w_i^{SemK\_mod} < w_i^{OK}$ and $w_j^{SemK} < w_j^{SemK\_mod} < w_j^{OK}$. Also, with respect to ***Lemma*** 3.2, the following inequality can be proved: $\theta_{(\overrightarrow{\mathbf{W}^{OK}},\overrightarrow{\mathbf{W}^{SemK}})} < \theta_{(\overrightarrow{\mathbf{W}^{OK}},\overrightarrow{\mathbf{W}^{SemK}})}$. This concludes the proof.

For the empirical validation of ***Lemma*** 3.4, the *LULC* ontology shown in Fig. 3.2 is revised by converging all the *level-4* spatial *LULC* classes to their corresponding *level-3* parent classes. The change in the performance of *SemK* is captured with both the initial and altered ontologies, as decribed in ***Lemma*** 3.4 and presented in Table 3.9. The results also satisfy the claim of ***Lemma*** 3.4 and prove that more generalized ontology will eventually converge *SemK* to *OK*. *SemK* to *OK*.

Proposition 3.1 *The misclassified sample points, that are represented by wrong* LULC *classes in the ontology, can be identified by a preprocessing of* SemK *method.*

Proof In case of any misclassification of the sampled locations in the hierarchical ontology, the error eventually propagates to the *SemK* process as well and generates erroneous semantic *covariance*s, and thus erroneous prediction results. For example, if any sampled location is erroneously represented by a less similar *LULC* class with respect to the unsampled prediction location, The assigned weight by *SemK* is lesser than the actual weight that should be assigned and vice-versa.

Through some preprocessing steps of the actual *SemK* process, this method is capable to identify these misclassifications of the *LULC* classes in the hierarchy. To carry out the same, few dummy sample points are introduced, each of which represents one unique leaf *LULC* class of the ontology hierarchy. It is assumed that all these points are equidistant from the prediction location. Hence, according to ***Lemma*** 3.2, the assigned weight to these dummy points by *SemK* should be based on the semantic scores only, i.e., in terms of *semantic similarity* and *spatial importance*. Now, let us consider two interpolating points x_i and x_j from the given scenario in Table 3.3. Then the the ith dummy point ($dummy_i$) will be assigned the semantic weight given as follows: $w_{f_{dummy_i}}^{SemK} = \left(\frac{1}{SIS_{0j}} - \lambda^{SemK}\right)$, where f_{dummy_i} represents the *LULC* class of the ith dummy point. If the ith sampled location x_i is represented by the *LULC* class f_{dummy_i} in reality, the normalized w_i^{SemK} is expressed as follows: $w^{SemK} = \left(\frac{\gamma(d_{0j})}{SIS_{0j}} - \lambda^{SemK}\right)$. However, erroneously if x_i is misclassified in the ontology and misrepresented by any *LULC*: f_{dummy_k}, then the w_i^{SemK} is expressed as follows:

$$w_i^{SemK} = \left(\frac{\gamma(d_{0j})}{SIS_{0k}} - \lambda^{SemK}\right) \tag{3.45}$$

Therefore, is can be observed that the assigned weight by *SemK* to the ith sampled location x_i is misinterpreted here. Thus, it is identified that the sampled location x_i is misclassified as *LULC* class f_{dummy_k} instead of f_{dummy_i} in the ontology. This concludes the proof.

3.8 Further Discussions

The prediction of spatial parameters is a significant research problem in the field of *geographic information system*. For the satellite imagery, the data values are often missing in some locations due to faulty sensors. Prediction in the missing location is an indispensable data staging process, which is highly required in *remote sensing* problems. However, for the meteorological parameters which are influenced by the terrain dynamics, the land–atmospheric interaction modeling is important, but still unpragmatic. It should be modeled efficiently to incorporate the terrestrial proximity into the interpolation process for better accuracy. The spatial *SemK* interpolation method extends a popular univariate regression-based interpolation technique *ordinary kriging*. The *SemK* considers the fact that there exist additional knowledge in the terrain, which influence most of the meteorological parameters. This work identifies one such knowledge for the parameter *land surface temperature*, i.e., *LULC* distribution in the study region. In *SemK*, the spatial autocorrelation is modeled considering both *Euclidean* distances and the semantic *LULC* properties between the sampled locations. Empirical experimentation shows that the *SemK* method yields better interpolation accuracy than the popular interpolation methods, primarily *ordinary kriging*. Analysis of information content in *SemK* and the theoretical performance evaluation by *Euclidean* vector analysis approach show the efficacy of *SemK* over others. The major contributions of the spatial *SemK* method can be stated as follows:

- modeling land–atmospheric interaction by semantically analyzing the *LULC* classes of the *RoI* and representing the formally by building a spatial *LULC* ontology.
- proposing two semantic parameters: the *semantic similarity* and the *spatial importance* to extend the traditional spatial autocorrelation measure.
- enrichment of the spatial interpolation process by amalgamating the terrestrial *LULC* knowledge into the *OK* method.
- empirical study with *land surface temperature* data to evaluate the performance of *SemK* with the existing interpolators.
- theoretical performance evaluation of *SemK* to establish its relations with existing techniques, evaluate the effect of the granularity of the ontology in *SemK*, etc.

To the best of our knowledge, the *SemK* is a preliminary attempt to quantify the terrestrial semantic *LULC* knowledge using ontology. This framework can be used for any semantic knowledge quantification, which is a qualitative knowledge of the domain of discourse. The *SemK* evaluates the *semantic similarity* and correlation among the *LULC* classes in the terrain. However, it might be noted that this correlation evaluation process can be regarded as an a-priori approach where the influence of surrounding *LULC* classes are not taken into account. Hence, to formulate the *SemK* process more pragmatically, this basic model can be improved further by extending this process to an a-posterior correlation analysis. This extension may improve the prediction accuracy further, resulting in a new pragmatic interpolation process. This improvement process is presented in the next chapter.

References

1. Bhattacharjee S (2015) Prediction of meteorological parameters: a semantic kriging approach. In: 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2015) PhD Symposium, ACM, p 1
2. Bhattacharjee S (2016) Semantic kriging: a semantically enhanced approach for spatial interpolation. PhD thesis, Indian Institute of Technology (IIT) Kharagpur, India
3. Bhattacharjee S, Ghosh SK (2015a) Performance evaluation of semantic kriging: A Euclidean vector analysis approach. IEEE Geosci Remote Sens Lett 12(6):1185–1189
4. Bhattacharjee S, Ghosh SK (2015b) Spatio-temporal change modeling of LULC: a semantic kriging approach. ISPRS Ann Photogramm, Remote Sens Spat Inf Sci 1:177–184
5. Bhattacharjee S, Ghosh SK (2016) Measuring semantic similarity between land-cover classes for spatial analysis: an ontology hierarchy exploration approach. Innov Syst Softw Eng 12(3):193–200
6. Bhattacharjee S, Ghosh SK (2017) Semantic kriging. In: Encyclopedia of GIS, vol 2. Springer International Publishing, pp 1868–1879
7. Bhattacharjee S, Mitra P, Ghosh SK (2014) Spatial interpolation to predict missing attributes in GIS using semantic kriging. IEEE Trans Geosci Remote Sens 52(8):4771–4780
8. Bhattacharjee S, Das M, Ghosh SK, Shekhar S (2016) Prediction of meteorological parameters: an a-posteriori probabilistic semantic kriging approach. In: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, p 38
9. Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier
10. Li J (2008) A review of spatial interpolation methods for environmental scientists. Record (Australia. Geoscience Australia), Geoscience Australia
11. Li J, Heap AD (2011) A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol Inform 6(3):228–241
12. Manning CD, Raghavan P, Schütze H, et al (2008) Introduction to information retrieval, vol 1. Cambridge university press, Cambridge
13. Mendiratta N, Kumar RS, Rao KS (2008) Standards for bio-geo database. Tech Rep 1 natural resources data management system (NRDMS) division. New Delhi, India
14. Miller GA (1995) WordNet: a lexical database for english. Commun ACM 38(11):39–41
15. Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Econ Geogr 46:234–240

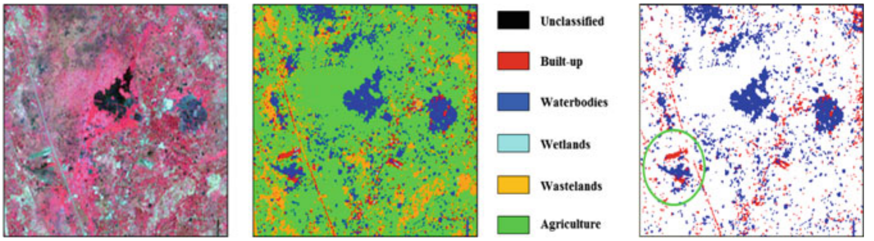
Chapter 4
Fuzzy Bayesian Semantic Kriging

Abstract The spatial *semantic kriging* (*SemK*) based interpolation approach is an attempt to amalgamate semantic knowledge into the prediction process. It considers land-use/land-cover (*LULC*) information for the land–atmospheric interaction modeling to achieve better prediction outcome. However, the correlation study between every pair of *LULC* classes in *SemK* is a-priori, which is not a pragmatic approach. In this a-priori process, the influences of other nearby *LULC* classes is ignored in the interpolation process. This chapter establishes a modification of spatial *SemK* by extending this process with an a-posterior probability-based correlation analysis among different *LULC* classes. The fuzzy *Bayesian network* principle is utilized here to carry out the probabilistic analysis. The empirical evaluations with real *land surface temperature* data shows the need for probability-based correlation analysis in *SemK* by achieving more prediction accuracy.

4.1 Introduction

It is evident from the existing literature that the geostatistical analysis for meteorological parameters is highly recommended for their prediction as it models spatial autocorrelation more accurately and pragmatically, thereby minimizing the error in prediction. However, the theoretical (refer Sect. 3.4) and empirical analysis (refer Sect. 3.6) of *SemK* show that other than geostatistical analysis, the interdependencies between the meteorological and terrestrial (e.g., land-use/land-cover (*LULC*), elevation, etc.) factors play a crucial role for spatial autocorrelation (proximity) estimation. Thus, enhancement of the accuracy of meteorological parameter's interpolation demands their semantic modeling and the assessment of the relation among different factors. To address the drawback, an unique spatial interpolation method is established, named as *semantic kriging* (*SemK*) [2, 3]. It models the semantic land-use/land-cover (*LULC*) distribution information for the land–atmospheric interaction assessment and integrates it into the existing techniques to make it more adequate

S. Bhattacharjee et al., *Semantic Kriging for Spatio-temporal Prediction*, Studies in Computational Intelligence 839,
https://doi.org/10.1007/978-981-13-8664-0_4



(a) *RoI* (b) Classified *RoI* (c) Legends (d) A-priori correlation

Fig. 4.1 Shortcomings of *SemK*

and pragmatic for enhancing the prediction accuracy. However, from the assumption of the *spatial importance* evaluation process in *SemK*, it might be noted that the correlation analysis process is a-priori in nature. Thus, *SemK* doesn't consider the impact of the neighboring *LULC* classes on each other.

For example, refer Fig. 4.1a, which shows a terrain map, considered as the region of interest (*RoI*). The same terrain that is processed by supervised classification technique is shown in Fig. 4.1b, with the legend information in Fig. 4.1c. Now, if any spatial analysis requires to evaluate the correlation between "*built-up*" and "*waterbodies*", the a-priori analysis considers the terrain as depicted in Fig. 4.1d. However, it is evident that in the given terrain, the *LULC* class "*agriculture*" is dominating and its effect on both "*built-up*" and "*waterbodies*" is very much profound. A portion of the terrain is identified with green circle (○) in Fig. 4.1d, where all other *LULC* classes are present to influence both "*built-up*" and "*waterbodies*". Thus, ignoring the impact of the neighboring *LULC* classes will generate some erroneous correlation value in a-priori analysis.

This chapter establishes a revised *SemK*, which is named as probabilistic *semantic kriging* or *fuzzy Bayesian semantic kriging* (*FB-SemK*) [4]. It overcomes the shortcomings of spatial *SemK* technique by modeling the mutual impact of the terrestrial *LULC* classes for the evaluation of spatial autocorrelation. The *SemK*'s a-priori correlation evaluation process is extended by the proposed *FB-SemK*. For example, for the given ontology in Fig. 3.2, the *spatial importance* of *industrial* and *river* is supposed to be evaluated. For the a-priori correlation analysis, the effect of other nearby *LULC* classes, such as, *residential*, *lakes*, *grassland*, etc., are not considered while choosing the sampled locations for both *industrial* and *river*. However, in *FB-SemK*, the a-posterior probability-based correlation analysis considers the effect of other nearby *LULC* classes for more accurate estimation of spatial autocorrelation and parameter value. As this monograph focuses on the analysis of meteorological parameters, which are continuous in nature, discretization of these parameters are obvious. The fuzzy analysis helps to take care of the imprecision and the uncertainties introduced due to discretization of the continuous variables.

Motivating Example

Let us assume a scenario to understand the flow of the proposed *FB-SemK* approach, in contrast to *OK* and *SemK*. Here, an *RoI* with 9 pixels (refer Fig. 4.2a) is considered, where parameter value at pixel no. 5 is missing. It is the unsampled locations and is supposed to be predicted using *FB-SemK* method. This pixel is represented with a ★. The eight other pixels (pixels: 1, 2, 3, 4, 6, 7, 8, 9) are the sampled locations or the interpolating points within one hop distance from the prediction point (pixel: 5). The parameter values (*LST* in °C) at the sampled locations are specified in Fig. 4.2b. Let us also assume that there are five types of *LULC* classes in the terrain: f_A: 🟨, f_B: 🟦, f_C: 🟥, f_D: 🟩, f_E: ⬜, for which the correlation between *LULC* classes are supposed to be evaluated.

In *SemK*, for the a-priori correlation evaluation between a pair of *LULC* classes in *spatial importance* analysis, say f_A and f_B (🟨 and 🟦), k number of pixels are chosen for each of classes in the pair, without considering other classes (🟥, 🟩 and ⬜). It is depicted in Fig. 4.2c. However, for the a-posterior correlation analysis, the effect other other *LULC* classes are also considered. The effect of 🟥, 🟩 and ⬜ classes on 🟨 and 🟦 are evaluated individually first. The principles of a fuzzy *Bayesian network* is utilized here to carry out the probabilistic analysis. Through this study, each of the pixels values of both the classes f_A and f_B (🟡 and 🔵) get updated as ◐ and ◐. The correlation between *LULC* classes f_A and f_B are carried out with the influenced values of the pixels. This process is depicted in Fig. 4.2d.

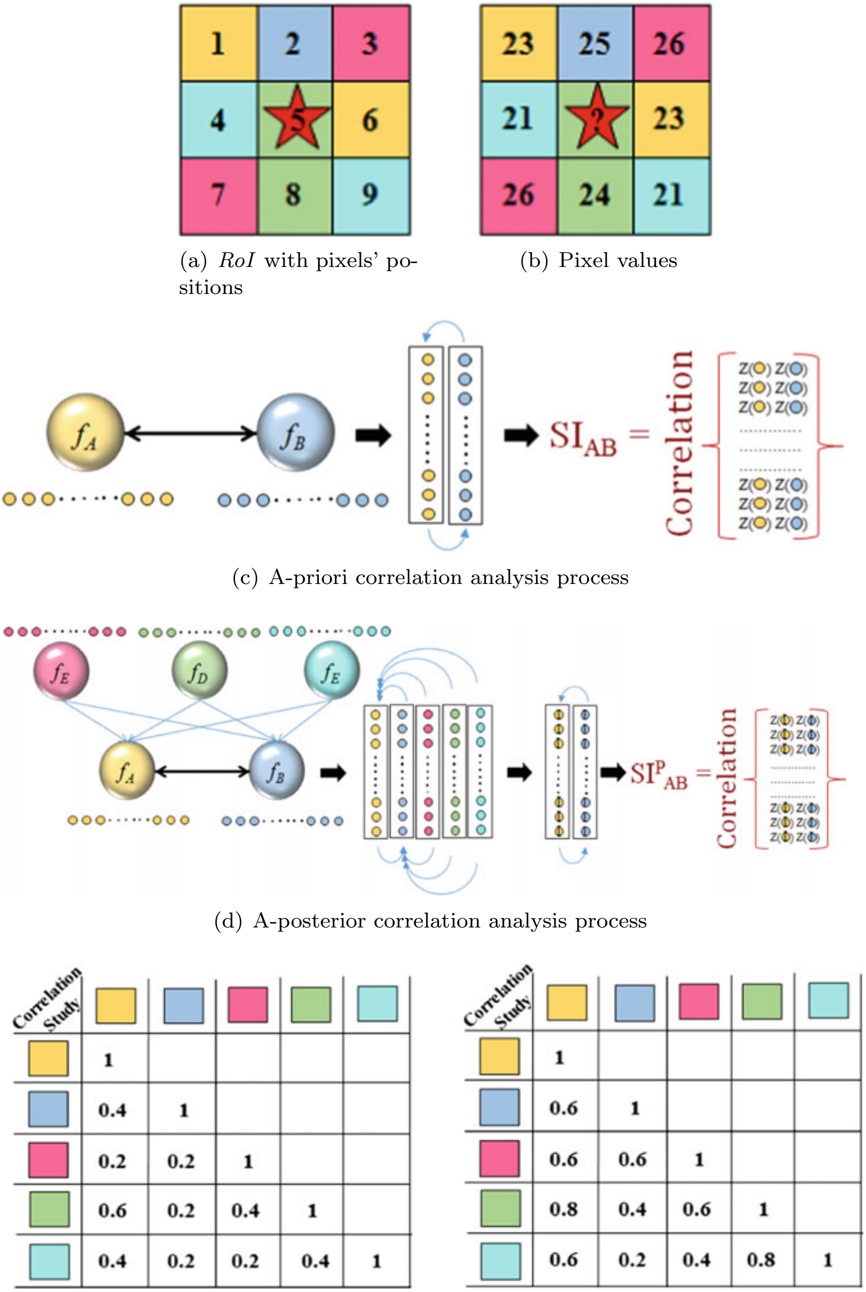
Now, consider a toy example of correlation evaluated by *SemK* process that is depicted in Fig. 4.2e. However, these values are likely to get updated for a-posterior probability-based correlation analysis in *FB-SemK* (refer Fig. 4.2f). Now, for both *SemK* and *FB-SemK*, the parameter value at pixel ★ is a function of known pixels: 1 2 3 4 6 7 8 9. Therefore, the normalized weight assigned to the sampled locations by ordinary kriging (*OK*), considering the *Euclidean* distance based proximity only, would be as follows:

$$\mathbf{W}^{OK} = 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125$$

However, as the *spatial importance* (correlation) between the *LULC* classes are not equal, weight assigned by the *SemK* would be different for pixels with different *LULC* classes. Hence, the normalized weight assigned to the sampled locations by the *semantic kriging* and *fuzzy Bayesian semantic kriging* would be as follows:

$$\mathbf{W}^{SemK} = 0.15, 0.05, 0.1, 0.1, 0.15, 0.1, 0.25, 0.1$$
$$\mathbf{W}^{FB\text{-}SemK} = 0.11, 0.21, 0.14, 0.11, 0.11, 0.14, 0.08, 0.11$$

Therefore, the parameter value at the pixel 6 (★) predicted by *OK*, *SemK* and *FB-SemK* will be as follows:



1	2	3
4	★5	6
7	8	9

(a) *RoI* with pixels' positions

23	25	26
21	★?	23
26	24	21

(b) Pixel values

(c) A-priori correlation analysis process

(d) A-posterior correlation analysis process

Correlation Study	🟨	🟦	🟥	🟩	⬜
🟨	1				
🟦	0.4	1			
🟥	0.2	0.2	1		
🟩	0.6	0.2	0.4	1	
⬜	0.4	0.2	0.2	0.4	1

(e) A-priori correlations

Correlation Study	🟨	🟦	🟥	🟩	⬜
🟨	1				
🟦	0.6	1			
🟥	0.6	0.6	1		
🟩	0.8	0.4	0.6	1	
⬜	0.6	0.2	0.4	0.8	1

(f) A-posterior correlations

Fig. 4.2 Example scenario for *FB-SemK*

$$\begin{aligned}
\hat{\mathbf{Z}}^{OK}(\bigstar) &= 0.125 * 23 + 0.125 * 25 + 0.125 * 26 + 0.125 * 21 + 0.125 * 23 \\
&\quad + 0.125 * 26 + 0.125 * 24 + 0.125 * 21 = 23.63 \\
\hat{\mathbf{Z}}^{SemK}(\bigstar) &= 0.15 * 23 + 0.05 * 25 + 0.1 * 26 + 0.1 * 21 + 0.15 * 23 \\
&\quad + 0.1 * 26 + 0.25 * 24 + 0.1 * 21 = 23.55 \\
\hat{\mathbf{Z}}^{FB\text{-}SemK}(\bigstar) &= 0.11 * 23 + 0.21 * 25 + 0.14 * 26 + 0.11 * 21 + 0.11 * 23 \\
&\quad + 0.14 * 26 + 0.08 * 24 + 0.11 * 21 = 23.93
\end{aligned}$$

4.2 Objectives of Fuzzy Bayesian Semantic Kriging

The spatial *SemK* approach [2, 3], discussed in Chap. 3, is capable of incorporating semantic domain knowledge in the interpolation technique. However, the land–atmospheric interaction modeling technique needs further improvement. One crucial stage involved in the *SemK* process is the evaluation of the *spatial importance* between every pair of spatial *LULC* classes. This aims to evaluate the correlation among *LULC* classes in terms of the prediction parameter. Therefore, this spatial correlation analysis can be referred to as a study to check how one *LULC* is semantically influenced by the another class. In *semantic kriging*, the relative *correlation* between two *LULC*s, say f_i and f_j, the impact of other neighboring *LULC*s f_p ($p \in 1 \cdots |F|$, $p \neq i, j$; $|F|$ represents the total number of leaf *LULC*s in the ontology) on both f_i and f_j is not considered. Thus, this a-priori land–atmospheric interaction modeling approach is irrational, which results in unpragmatic *spatial importance* measures, indirectly affecting prediction accuracy.

This chapter proposes a revised *SemK* that establishes a fuzzy *Bayesian network* based probabilistic correlation analysis among different pairs of *LULC* classes. This probabilistic a-posterior estimation of inter-*LULC* correlation facilitates proper estimation of *spatial importance*, consequently outperforming the *SemK* approach by improving the prediction performance. Therefore, the major objectives of the proposed approach is stated below:

- to propose a *fuzzy Bayesian semantic kriging* (*FB-SemK*) framework for interpolating meteorological parameters.
- to model a probabilistic a-posterior correlation study model for the terrestrial *LULC* classes in the *RoI*.
- extending *SemK*'s *spatial importance* metric with the conditional probability analysis.
- to validate the performance of the *FB-SemK* approach and compare it with the existing interpolation techniques, including *SemK*.

4.3 FB-SemK: Fuzzy Bayesian Semantic Kriging

This probabilistic extension of *semantic kriging* approach, referred to as *FB-SemK*, applies the principle of *Bayesian network* to model the probabilistic correlation analysis, which improves the prediction accuracy of *SemK*. The *FB-SemK* replaces the a-priori *spatial importance* evaluation of *SemK* with a-posterior correlation estimation by considering the mutual impact of the terrestrial *LULC* classes. The probabilistic analysis in *FB-SemK* is done considering discrete fuzzy *Bayesian network* based learning and inference generation approach. Here, the impact of one *LULC* over others is instinctively captured by the *Bayesian network*'s causal dependency graph and the incorporated fuzzy logic helps to handle the uncertainties and imprecision present in the data. This section describes the *FB-SemK* method, preceded by a description of the *Bayesian network* principle and its fuzzy extension in the following subsections.

Figure 4.3 presents the *FB-SemK* interpolation framework. The probabilistic *spatial importance* calculation has a major component of probabilistic a-posterior correlation evaluation among different pairs of leaf *LULC*s in the hierarchical ontology. In this component, a *Bayesian network* is established with all the available terrestrial *LULC*s classes. The parameter values of the sampled locations representing a *LULC* class get updated by evaluating the impact of other classes with the probabilistic fuzzy analysis of the Bayesian network. The actual correlation score is evaluated

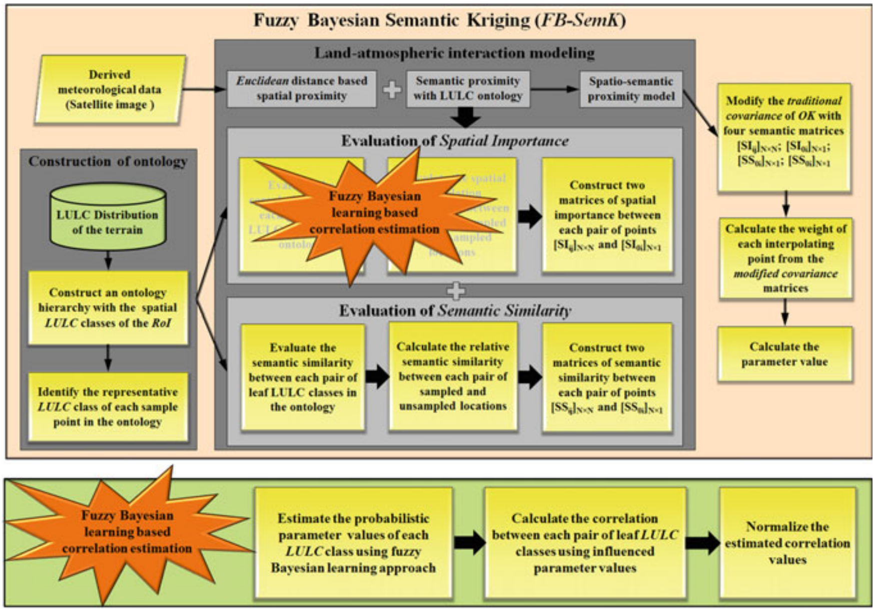

Fuzzy Bayesian Semantic Kriging (FB-SemK)
Land-atmospheric interaction modeling
Derived meteorological data (Satellite image)
Euclidean distance based spatial proximity
Semantic proximity with LULC ontology
Spatio-semantic proximity model
Evaluation of Spatial Importance
Fuzzy Bayesian learning based correlation estimation
Construct two matrices of spatial importance between each pair of points $[SI_{ij}]_{N\times N}$ and $[SI_{0i}]_{N\times 1}$
Construction of ontology
LULC Distribution of the terrain
Construct an ontology hierarchy with the spatial LULC classes of the RoI
Identify the representative LULC class of each sample point in the ontology
Evaluation of Semantic Similarity
Evaluate the semantic similarity between each pair of leaf LULC classes in the ontology
Calculate the relative semantic similarity between each pair of sampled and unsampled locations
Construct two matrices of semantic similarity between each pair of points $[SS_{ij}]_{N\times N}$ and $[SS_{0i}]_{N\times 1}$
Modify the traditional covariance of OK with four semantic matrices $[SI_{ij}]_{N\times N}$; $[SI_{0i}]_{N\times 1}$; $[SS_{0i}]_{N\times 1}$; $[SS_{0i}]_{N\times 1}$
Calculate the weight of each interpolating point from the modified covariance matrices
Calculate the parameter value
Fuzzy Bayesian learning based correlation estimation
Estimate the probabilistic parameter values of each LULC class using fuzzy Bayesian learning approach
Calculate the correlation between each pair of leaf LULC classes using influenced parameter values
Normalize the estimated correlation values

Fig. 4.3 *FB-SemK* framework [4]

using the parameter values resulted from the conditional probability analysis. From Fig. 4.3, it is observed that most of the other components are same as the previously proposed approach *SemK* (refer Fig. 3.1). However, a new module is introduced in *FB-SemK* framework for the evaluation of *spatial importance*, indicated as *fuzzy Bayesian learning based correlation estimation* module. The steps associated with this component are shown at the bottom of the base process *SemK*.

4.3.1 Bayesian Network and Its Fuzzy Extension

The *Bayesian network* (*BN*), which is also referred to as Bayes network and belief network, is a directed acyclic graph (*DAG*) with the characteristics specified below:

- *BN* represents a probabilistic graph (*DAG*) with a number of random variables, where their conditional dependencies are represented as the edges between the nodes.
- The random variables in the network are from the domain of interest, $X = \{X_1, X_2, \ldots, X_i\}$, which are observable quantities or latent variables, sometimes unknown parameters or hypotheses as well.
- the directed edges between a pair of nodes (or links), say $X_i \rightarrow X_j$, represents direct dependency between the variables. Here, the variable X_i is referred to as the parent of the variable X_j.
- there might be nodes which are isolated and not connected to other nodes. These variables are conditionally independent of other variables in the network.
- each node X in the *BN* is associated with a conditional probability distribution as $P\left(X_i \mid Parents\left(X_i\right)\right)$. It is the measure of the impact of the parents on the child node X.

Bayesian network assumes that a node X_i and its parents $Parents(X_i)$, X_i is conditionally independent of all its non-descendant nodes $ND(X_i)$. It can be formally represented as

$$P\left(X_i|Parents\left(X_i\right), ND\left(X_i\right)\right) = P\left(X_i|Parents\left(X_i\right)\right) \tag{4.1}$$

Further, in the directed acyclic graph of a *BN*, the dependencies among the variables can be evaluated by a joint probability density function (*PDF*). The random variable range can be factorized as a product of conditional/marginal probability distributions. The notion can be expressed as follows:

$$P\left(x_1, x_2, \ldots, x_i, \ldots, x_n\right) = \prod_{i=1}^{n} P\left(x_i \mid parents\left(X_i\right)\right) \tag{4.2}$$

where x_i and $parents\,(x_i)$ represent a specific value for both X_i and its parent $Parents(X_i)$, respectively. This probabilistic analysis facilitates us to understand the dependencies among the random variables of a domain of interest.

4.3.2 *Fuzzy Bayesian Network (FBN)*

Fuzzy probability can be referred to as an extension of the notion of simple probability [1]. It utilizes the hypothesis of fuzzy set theory to check the probability of a variable to be in a set. On the other hand, the simple probability theory is based on the subjective belief to check the probability of a variable to be in a set. The simple probability theory cannot manage the uncertainties and the imprecision present in datasets. However, the fuzzy probability theory is able to handle probabilistic and non-probabilistic uncertainties. In literature, several works have been proposed on *Bayesian network* with fuzzy extension [5, 7–9]. Among all these works, the most relevant and fundamental is the "Fuzzy *Bayesian network*s or *FBN*s", proposed by Tang and Lui [9]. The *FBN* can be referred to as a hybrid inference system to incorporate fuzzy logic into the *Bayesian network*. It can be used effectively and efficiently to solve many practical problems, which are difficult to be expressed by accurate mathematics.

Let the two sets of events be $X = \{X_1, X_2, \ldots, X_m\}$ and $Y = \{Y_1, Y_2, \ldots, Y_n\}$. The corresponding fuzzy events of X and Y be $\tilde{X}$ and $\tilde{Y}$. Thus, with reference to *FBN*, the fuzzy conditional probability of $\tilde{Y}$ given $\tilde{X}$ is estimated as follows:

$$P\left(\tilde{Y}|\tilde{X}\right) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} \mu_{\tilde{Y}}\left(Y_j\right) \cdot \mu_{\tilde{X}}\left(X_i\right) \cdot P\left(X_i|Y_j\right) \cdot P\left(Y_j\right)}{P\left(\tilde{X}\right)} \tag{4.3}$$

$$= \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} \mu_{\tilde{Y}}\left(Y_j\right) \cdot \mu_{\tilde{X}}\left(X_i\right) \cdot P\left(Y_j|X_i\right) \cdot P\left(X_i\right)}{P\left(\tilde{X}\right)} \tag{4.4}$$

where $\mu_{\tilde{X}}\left(X_i\right)$ is the membership value of the fuzzy event $\tilde{X}$ in the event X_i, $\mu_{\tilde{Y}}\left(Y_j\right)$ is the membership value of the fuzzy event $\tilde{Y}$ in the event Y_j, and the $P\left(\tilde{X}\right)$ represents event $\tilde{X}$'s fuzzy marginal probability. $P\left(\tilde{X}\right)$ is defined as follows:

$$P\left(\tilde{X}\right) = \sum_{i=1}^{m} \mu_{\tilde{X}}\left(X_i\right) \cdot P\left(X_i\right) \tag{4.5}$$

The principles of *FBN* are used in the proposed *FB-SemK* to calculate the probabilistic *spatial importance* between every pair of sampled and unsampled locations.

4.3.3 FB-SemK and Its Working Principles

In this section, the proposed *FB-SemK* is presented in details, with the description of its working principles. It follows the notion of basic *semantic kriging* approach, however, the *spatial importance* metric evaluation process of *SemK* is extended with the probabilistic analysis using *FBN*. The probabilistic *spatial importance* evaluation process has two components: a-posterior correlation study between every pair of leaf *LULC*s in the ontology and the second is the evaluation of the actual spatial *importance* scores and the matrices. Among these two components, the first one improves the basic *SemK* process with more pragmatic semantic analysis of *LULC* classes. First, a *Bayesian network* or a *DAG* is established considering the available terrestrial *LULC* classes to model their impact on each other.

Now, a probabilistic fuzzy Bayesian analysis is carried out with this *LULC DAG* to update the parameter value of each *LULC* class by considering the impact of others. The actual correlation between every pair of *LULC* classes is evaluated with these updated parameter values of the sampled locations. The overall process flow of the *FB-SemK* approach is described in Algorithm 2 [4]. The evaluation process of the *spatial importance* metric of *FB-SemK* is presented in Sect. 4.3.3.1.

4.3.3.1 Spatial Importance Evaluation by FB-SemK

Similar to the evaluation process in *SemK*, the *spatial importance* evaluation by *FB-SemK* also measures the score for every pair of leaf *LULC* classes in the hierarchical ontology, in terms of the prediction parameter. This correlation follows the below-mentioned assumptions:

- this correlation analysis is highly dependent on the prediction parameter.
- a correlation measure between a pair of *LULC* classes is a global correlation score representative for the whole study region.
- this analysis overcomes the shortcoming of basic *SemK*'s a-priori correlation study.

As evaluated in *SemK* process, first the entire *RoI* is split into k nonoverlapping random zones (R_k) which satisfies the following criteria: $\bigcup^k{}_{i=1} R_k = RoI$. Then, k pairs of points are sampled for both the considered *LULC* classes from those k zones. Up to this, the process conforms to the standard *SemK* process. However, here in *FB-SemK*, the correlation scores are not directly calculated from the random field values of the chosen k pair of points. The values of these sample points are further revised by considering the other neighboring *LULC*s' impact. To evaluate these revised values, again k set of sample points are collected for each of the $|F| - 2$ number of the other *LULC* classes, where $|F|$ represents the total number of available leaf *LULC*s. Each of the samples in each set of k samples are considered from every nonoverlapping random zones. Therefore, now a total number of $|F|$ (2 *LULC* classes considered for the correlation analysis +($|F| - 2$) number of neighboring *LULC* classes) sample points are collected from k zones. This preprocessing generates a

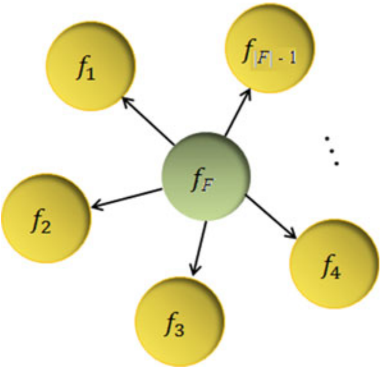


Fig. 4.4 *BN* with *LULC* classes [4]

matrix of sample points of dimension $k \times |F|$. These points satisfy the criteria of being within a predefined distance d in the corresponding zone. Next, a pair of *LULC* classes is considered and the random field value of all the sample points in that pair is inferred separately with respect to the impact of other $|F| - 2$ *LULC* classes in the *RoI*. For this analysis, the purpose, the principle of fuzzy *Bayesian network* (*FBN*), proposed by Tang and Liu [9] is considered. A *DAG* of causal dependency among the *LULC* classes is constructed using the principle of *Bayesian network*, which captures the influence of one *LULC* class on others. One example *Bayesian network* structure is presented in Fig. 4.4. It depicts the impact of the $|F|^{th}$ *LULC* class $f_{|F|}$ on the others f_is, where $f_i \in \{f_1, f_2, f_3, \ldots, f_{|F|-1}\}$. Analyzing this fuzzy *Bayesian network*, once the conditional probabilities are evaluated for all the sample points in the $k \times |F|$-dimensional matrix by considering every *LULC* pair separately, a Pearson correlation evaluation is performed by considering the revised random field value of the sample points. This process is formalized in *Probabilistic correlation calculation* algorithm (refer Algorithm 1 [4]).

The idea of *FB-SemK*'s correlation evaluation process is shown in Fig. 4.5. The a-posterior correlation evaluation between two *LULC* classes, f_i and f_j, is shown,

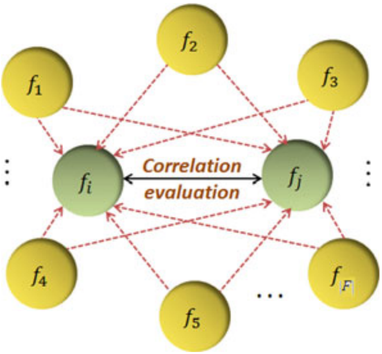


Fig. 4.5 Correlation study between f_i and f_j [4]

however, the impact of other neighboring *LULC*s $f_p \in \{f_1, f_2, \ldots, f_{|F|}\}$, $(p \neq i, j)$, is also considered. First, using learning and inference generation principles of *FBN*, the conditional probability of the considered *LULC* classes, f_i and f_j, are inferred from the sample points representing rest of the *LULC* classes $f_{|F|}$. The correlation or *spatial importance* analysis takes place between the revised samples or the conditional probability-based samples values of both f_i and f_j. In Fig. 4.5, the red dashed arrows represent the impact of the rest of the neighboring *LULC*s f_p $(p \neq i, j)$ on the participating *LULC*s f_i and f_j. The actual correlation between f_i and f_j is represented by the black solid arrow.

As described in *SemK*, once representative *LULC*s of the prediction and all the sampled locations are linked in the hierarchical ontology, each of the sampled locations is associated with an *importance* score, which is the estimated correlation value with respect to the prediction point. Therefore, the relative probabilistic *importance* of the i^{th} sampled location (SI_{0i}^{prob}), taking the prediction point (x_0) as the reference, is given as follows:

$$\begin{aligned} SI_{0i}^{prob} &= Corr_{prediction\_parameter}(x_0, x_i) && (4.6)\\ &= Corr_{prediction\_parameter}(f_0, f_i) && (4.7)\\ &= \frac{\sum\limits_{m=1}^{k} (\mathbb{Z}(f_{0_m}) - \overline{\mathbb{Z}(f_0)})(\mathbb{Z}(f_{i_m}) - \overline{\mathbb{Z}(f_i)})}{\sqrt{\sum\limits_{m=1}^{k} (\mathbb{Z}(f_{0_m}) - \overline{\mathbb{Z}(f_0)})^2 \sum\limits_{m=1}^{k} (\mathbb{Z}(f_{i_m}) - \overline{\mathbb{Z}(f_i)})^2}} && (4.8) \end{aligned}$$

where $\mathbb{Z}(f_{p_q})$ denotes the revised value of the q^{th} sample point(i.e., selected from the q^{th} zone) with representative *LULC* as f_p using conditional probability analysis. The $\mathbb{Z}(\overline{f_p})$ denotes the mean of these revised values of the f_p *LULC* class with k sample points. Now, similar to the basic *SemK* process, the *FB-SemK* also forms a [N × 1] vector, given as $[\mathbf{SI}^{prob}]_{0i}^{T} = [SI_{01}^{prob} SI_{02}^{prob} \cdots SI_{0N}^{prob}]$, which is the importance matrix considering all of the sampled location with respect to the unsampled prediction location. Similarly, as there exists spatial autocorrelation in the terrain, each pair of the interpolating points are also correlated to each other. Their mutual *spatial importance* can be measured which is their representative *LULC* classes' correlation. Therefore, the relative probabilistic importance score between i^{th} and j^{th} sampled locations is given as SI_{ij}^{prob} and is derived as follows:

$$\begin{aligned} SI_{ij}^{prob} &= Corr_{prediction\_parameter}(x_i, x_j) && (4.9)\\ &= Corr_{prediction\_parameter}(f_i, f_j) && (4.10)\\ &= \frac{\sum\limits_{m=1}^{k} (\mathbb{Z}(f_{i_m}) - \overline{\mathbb{Z}(f_i)})(\mathbb{Z}(f_{j_m}) - \overline{\mathbb{Z}(f_j)})}{\sqrt{\sum\limits_{m=1}^{k} (\mathbb{Z}(f_{i_m}) - \overline{\mathbb{Z}(f_i)})^2 \sum\limits_{m=1}^{k} (\mathbb{Z}(f_{j_m}) - \overline{\mathbb{Z}(f_j)})^2}} && (4.11) \end{aligned}$$

Here, a symmetric matrix of dimension [N × N] is formed for N sampled locations and is referred as $[\mathbf{SI}^{prob}]^T_{ij}$.

Example of A-Posterior Correlation Analysis

Let us consider an interpolation scenario to understand the evaluation process of correlation between a pair of leaf *LULC* classes in the ontology. Let f_i and f_j the considered pair of *LULC* classes and the interpolation is to be carried out in Kolkata, WB, India (*RoI*), considering the prediction parameter *LST*. Though according to the ontology in Fig. 3.2, there are 22-leaf *LULC* classes, however, for this example we have chosen 15 out of them, i.e., $|F| = 15$. Further. let $k = 30$, thus the whole *RoI* is divided into 30 nonoverlapping random zones. According to the description of *FB-SemK*'s correlation analysis process, 30 pair of random sampled locations are chosen from each of these zones where the former 30 points correspond to the *LULC* f_i and the later 30 corresponds to f_j. Now, for the conditional probability evaluation of f_i and f_j's sample points considering other $(|F| - 2) = 13$ *LULC* classes with respect to*LST*, 30 points against each of the $|F| - 2$ classes are selected from the random zones. Now, as these additional locations are the observed values of the neighboring 13 *LULC* classes, the revised *LST* values of the 30 pairs of f_i and f_j are inferred from these locations using the *FBN* principles. Now, considering these revised values of f_i and f_j, the correlation between these *LULC* classes is measured further using Eq. 4.11, following the same process as described in *SemK*. These correlation scores are then normalized to a positive range from its original values, ranging between $[-1, 1]$.

As the ontology structure is unchanged with respect to time and other related factors, the *FB-SemK* follows the same evaluation process for the metric *semantic similarity*, as proposed in basic *SemK* [3]. Therefore, for this metric, *FB-SemK* generates the similar $[\mathbf{SS}]_{0i}$ and $[\mathbf{SS}]_{ij}$ matrices for any leaf *LULC* classes f_i and f_j in the ontology (described in Sect. 3.3.1). Hence, with respect to these four matrices ($[\mathbf{SI}^{prob}]_{0i}$, $[\mathbf{SI}^{prob}]_{ij}$, $[\mathbf{SS}]_{0i}$ and $[\mathbf{SS}]_{ij}$), the weight vector of *FB-SemK* ($\mathbf{W}^{FB\text{-}SemK}$) and its *Lagrange multiplier* ($\lambda^{FB\text{-}SemK}$) are given as follows, where $\mathbf{C}$ represents the traditional *semivariance* matrix and $\mathbf{D}$ is the traditional *distance matrix* of *ordinary kriging*.

$$\mathbf{W}^{FB-SemK} = \begin{bmatrix} [\mathbf{C}]_{ij} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}^{prob}]_{ij} \circ [\mathbf{SS}]_{ij}\right) \end{bmatrix}^{-1} \left[\begin{bmatrix} [\mathbf{D}]_{0i} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}^{prob}]_{0i} \circ [\mathbf{SS}]_{0i}\right) \end{bmatrix} - \lambda^{FB-SemK}\mathbf{1}\right] \tag{4.12}$$

$$\lambda^{FB-SemK} = \frac{\left[[\mathbf{1}]^T \begin{bmatrix} [\mathbf{C}]_{ij} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}^{prob}]_{ij} \circ [\mathbf{SS}]_{ij}\right) \end{bmatrix}^{-1} \left[\begin{bmatrix} [\mathbf{D}]_{0i} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}^{prob}]_{0i} \circ [\mathbf{SS}]_{0i}\right) \end{bmatrix}\right]\right] - 1}{[\mathbf{1}]^T \begin{bmatrix} [\mathbf{C}]_{ij} \\ - \cdot - \cdot - \\ \left([\mathbf{SI}^{prob}]_{ij} \circ [\mathbf{SS}]_{ij}\right) \end{bmatrix}^{-1} \mathbf{1}} \tag{4.13}$$

Algorithm 1: Probabilistic correlation analysis (R_k, F)

```
Input: $R_k$ = Set of random nonoverlapping zones in the RoI, where $\bigcup_{i=1}^{k} R_k = RoI$; $F$ = Set
       of leaf LULC classes $\{f_1, f_2, \ldots, f_{|F|}\}$ in the ontology; $|F|$ = Total number of leaf
       LULC classes in the ontology
Output: A-posterior probabilistic correlation scores between every pair of LULC
        $(f_i, f_j) \in F$
foreach $R_i \in R_k$ $(i = 1, 2, \ldots, k)$ do
    Apply FBN learning rules to capture the mutual impact among LULC classes
    foreach pair of leaf LULC $(f_p, f_q) \in F$ $(p, q = 1, \ldots, |F|)$ do
        Infer the conditional probability for $f_p$ $(\mathbb{Z}(f_{p_i}))$ considering $\forall f_y \in$ F (y $\neq$ p) using
        FBN
        Infer the conditional probability for $f_q$ $(\mathbb{Z}(f_{q_i}))$ considering $\forall f_y \in$ F (y $\neq$ q) using
        FBN
    end
end
foreach pair of leaf LULC $(f_i, f_j) \in F$ $(i, j = 1, \ldots, |F|)$ do
    Probabilistic_correlation$(f_i, f_j)$
    $= \frac{\sum_{m=1}^{k} (\mathbb{Z}(f_{i_m}) - \overline{\mathbb{Z}(f_i)})(\mathbb{Z}(f_{j_m}) - \overline{\mathbb{Z}(f_j)})}{\sqrt{\sum_{m=1}^{k} (\mathbb{Z}(f_{i_m}) - \overline{\mathbb{Z}(f_i)})^2 \sum_{m=1}^{k} (\mathbb{Z}(f_{jm}) - \overline{\mathbb{Z}(f_j)})^2}}$
end
```

4.4 Empirical Proof for FB-SemK

Empirical experiment is performed using *land surface temperature* data. As presented in *semantic kriging*, the similar study is executed in two spatial regions: Kolkata, WB, India and Dallas, TX, USA. Here, five different zones (in contrast to the zones considered in Chap. 3) from each of the region. These zones are depicted in Fig. 4.6 and in Fig. 4.7 respectively. For the empirical analysis of *FB-SemK*, the same satellite image of the year 2015 (within the range of mid-October– mid-November) has been considered. The same experimental specifications have been considered for all the methods as mentioned in Sect. 3.6.

While measuring the probabilistic *spatial importance* of *FB-SemK*, the mutual impact of among the *LULC* classes have been estimated using *FBN* learning and inference generation principles. Generally, the *Bayesian network* deals with the discrete values of the considered parameter. However, most of the meteorological parameters (here, *LST*) are originally continuous variables. Thus, the discretization of the parameters is required to apply *FBN* principles. To discretize the range of *LST* values in the region Kolkata, it is further classified into seven groups (refer Table 4.1) and further have been converted into fuzzy variables, which helps in dealing with the imprecision induced in the data due to its discretization. The appropriate fuzzy membership functions need to be considered for respective meteorological parameters.

Algorithm 2: *FB-SemK procedure*

```
Input: $N$ number of sampled locations Interpolating points X = $[x_1, x_2, \ldots, x_N]$
Unsampled prediction point $x_0$
$M$ pairs of samples to plot semivariogram
Random field/ parameter values at every samples
$R_k$ = Set of random nonoverlapping zones in the RoI, where $\bigcup_{i=1}^{k} R_k = RoI$
$F$ = Set of leaf LULC classes $\{f_1, f_2, \ldots, f_{|F|}\}$ in the ontology

Output: Estimated parameter value $\hat{Z}(x_0)$ at the unsampled prediction location $x_0$

foreach $(x_i, x_i + h) \in X$ do
    $\gamma(h) = \frac{\sum_{i=1}^{M} [Z(x_i) - Z(x_i + h)]^2}{2M}$
    Plot ($\gamma(h)$ vs. $h$)
    Evaluate the semivariogram parameters: $C_0, C_1, R$
end
Call Probabilistic correlation analysis ($R_k$, F)
foreach $(x_i, x_i + h) \in X$ do
    Set $\gamma(h) = C_0 + C_1(1 - e^{\frac{3h}{R}})$
    $SS_{ij} = \frac{\frac{m_i}{|f_i|} + \frac{m_j}{|f_j|}}{2}$
    $SI_{ij}^{prob} = Probabilistic\_correlation(f_i, f_j)$
    Set $C_{ij}^{FB\text{-}SemK} = \frac{C_{ij}}{SI_{ij}^{prob} * SS_{ij}} = \frac{\gamma(h)}{SI_{ij}^{prob} * SS_{ij}}$ [Considering $(x_i + h)$ as $x_j$]
end
Generate two [N × N] matrices: $[\mathbf{SI}]_{ij}^{prob}$, $[\mathbf{SS}]_{ij}$
Generate two [N × 1] matrices: $[\mathbf{SI}]_{0i}^{prob}$, $[\mathbf{SS}]_{0i}$
Determine $\mathbf{C}^{FB\text{-}SemK} = \frac{[\mathbf{C}]_{ij}}{([\mathbf{SI}]_{ij}^{prob} \circ [\mathbf{SS}]_{ij})}$
Determine $\mathbf{D}^{FB\text{-}SemK} = \frac{[\mathbf{D}]_{0i}}{([\mathbf{SI}]_{0i}^{prob} \circ [\mathbf{SS}]_{0i})}$
Determine $\lambda^{FB\text{-}SemK} = \frac{\left[ [\mathbf{1}]^T \left[ \frac{[\mathbf{C}]_{ij}}{\left([\mathbf{SI}]_{ij}^{prob} \circ [\mathbf{SS}]_{ij}\right)} \right]^{-1} \left[ \frac{[\mathbf{D}]_{0i}}{\left([\mathbf{SI}]_{0i}^{prob} \circ [\mathbf{SS}]_{0i}\right)} \right] \right] - 1}{[\mathbf{1}]^T \left[ \frac{[\mathbf{C}]_{ij}}{\left([\mathbf{SI}]_{ij}^{prob} \circ [\mathbf{SS}]_{ij}\right)} \right]^{-1} \mathbf{1}}$
Determine $\mathbf{W}^{FB\text{-}SemK} =$
$\left[ \frac{[\mathbf{C}]_{ij}}{\left([\mathbf{SI}]_{ij}^{prob} \circ [\mathbf{SS}]_{ij}\right)} \right]^{-1} \left[ \left[ \frac{[\mathbf{D}]_{0i}}{\left([\mathbf{SI}]_{0i}^{prob} \circ [\mathbf{SS}]_{0i}\right)} \right] - \lambda^{FB\text{-}SemK} \mathbf{1} \right]$
```

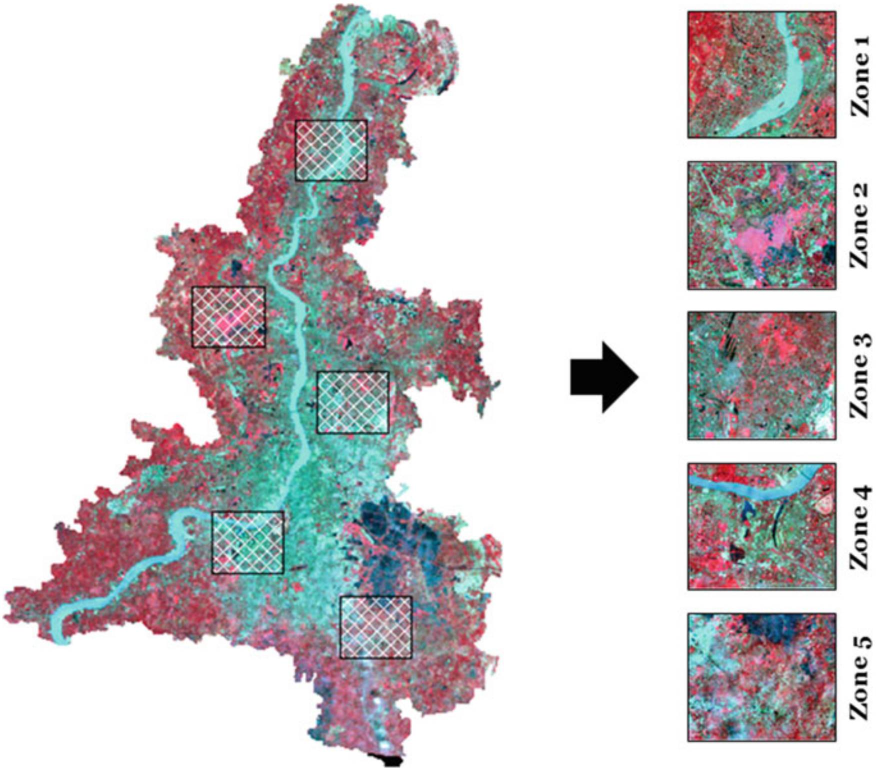


Fig. 4.6 Selected spatial zones of Kolkata, WB, India for *FB-SemK*

For this study, the trapezoidal membership function has been chosen to fuzzify the parameter, which is depicted in Fig. 4.8. For the actual correlation analysis process in *FB-SemK*, each of the *RoI*s is subdivided into sixteen zones (k = 16). This discretization process and finally applying the *FBN* principle is performed for every zone in the *RoI*, which desists the generalization of the local variability.

The performance of *FB-SemK* is compared with the same univariate spatial interpolation methods that are considered in Chap. 3, along with *SemK* itself. The *semantic kriging* is also considered here to check whether the a-posterior correlation analysis actually improves the performance of *SemK*. Hence, four existing popular interpolation techniques, considered for the comparative study along with *SemK*, are specified below.

- *nearest neighbors* (*NN*)
- *inverse distance weighting* (*IDW*)
- *universal kriging* (*UK*)
- *ordinary kriging* (*OK*)
- *semantic kriging* (*SemK*)

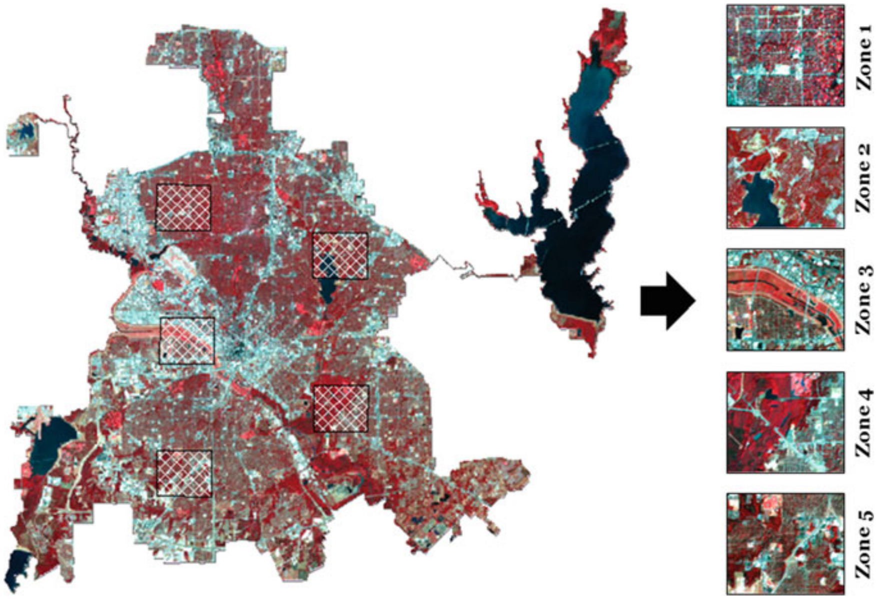


Fig. 4.7 Selected spatial zones of Dallas, TX, USA for *FB-SemK*

Table 4.1 Discretized values for *LST* [4]

Sub-range	*LST* values
1	R1: <20 °C
2	R2: 20–25 °C
3	R3: 25–30 °C
4	R4: 30–35 °C
5	R5: 35–40 °C
6	R6: 40–45 °C
7	R7: >45 °C

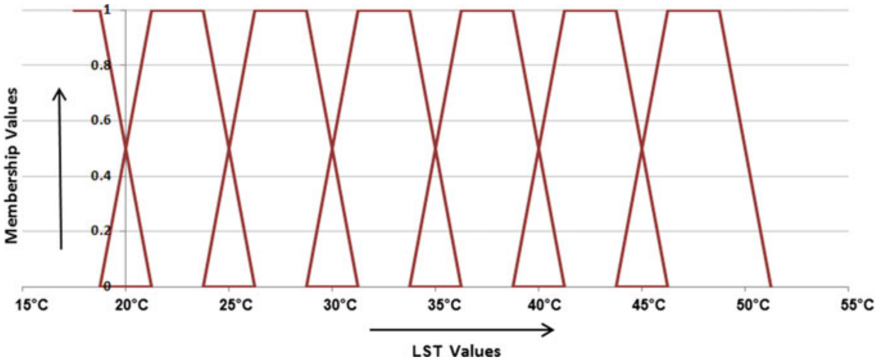
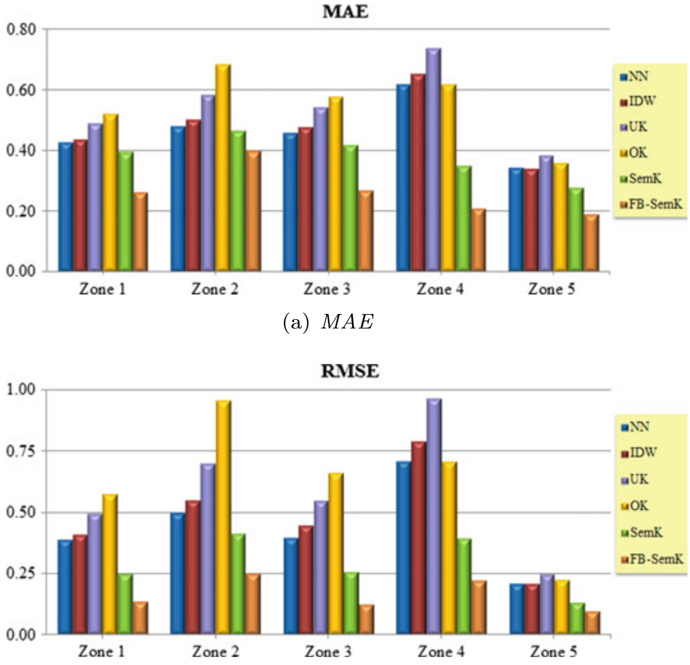


Fig. 4.8 Fuzzy membership function for *LST* [4]



(a) *MAE*

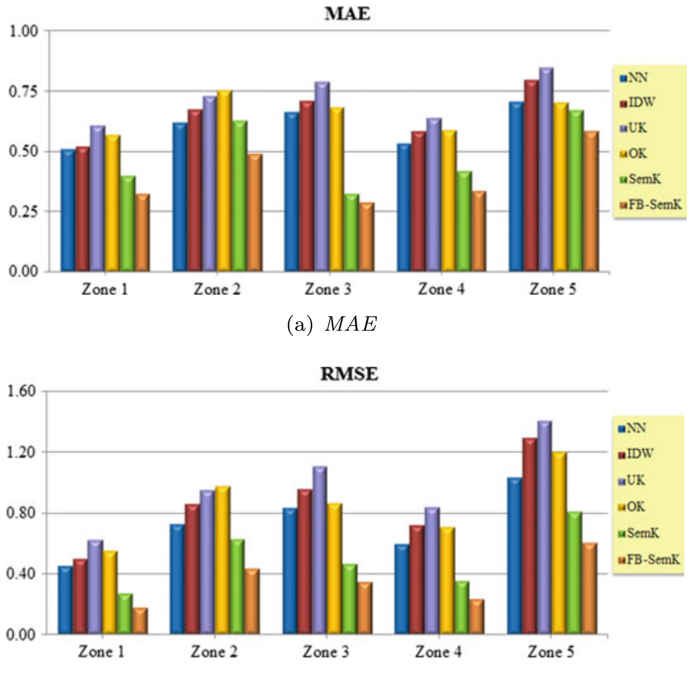
(b) *RMSE*

Fig. 4.9 Comparison study with error graph for *FB-SemK* (Region: Kolkata, WB, India)

Two standard error metrics: *mean absolute error* (*MAE*) and *root mean square error* (*RMSE*) [6] are evaluated to check for discrepancies produced by different methods between the actual and the measured *LST* at some point locations. The graphical representations of error are depicted through Fig. 4.9 and Fig. 4.10 respectively.

4.4.1 Discussions on Empirical Proof

The performance of *FB-SemK* is also evaluated in terms of generating the predicted surfaces for the selected zones in both the cities: Kolkata, WB, India and Dallas, TX, USA. The predicted surfaces along with the actual one are depicted in Tables 3.8 and 3.9. Beside *FB-SemK*, here the predicted surfaces are generated from five other spatial interpolation methods: *NN*, *IDW*, *UK*, *OK* and *SemK* as well. As mentioned in Chap. 3, the error surface (in gray scale) and the *peak signal-to-noise* ratio are also reported against each predicted surface with respect to the actual surface. As can be observed from the figures and the table reported for this empirical analysis, both the methods, the *SemK* and the *FB-SemK*, generates better mapping surfaces



(a) *MAE*

(b) *RMSE*

Fig. 4.10 Comparison study with error graph for *FB-SemK* (Region: Dallas, TX, USA)

of *LST* compared to others by considering the terrestrial *LULC* information for the prediction. Further, the *FB-SemK* generates the most accurate surface of *LST* by performing a-posterior correlation analysis among the *LULC* classes. The *FB-SemK* reports the highest *PSNR* compared to other methods (≈6–13 dB for Kolkata, WB, India and ≈5–11 dB for Dallas, TX, USA) and approximately 3–6 dB (for Kolkata, WB, India) and ≈3–4 dB (for Dallas, TX, USA) higher than *SemK* as well (Tables 4.2 and 4.3).

This chapter has also considered the zones from the basic *SemK*'s empirical study (i.e., from Chap. 3). Ten spatial zones depicted in Figs. 3.3 and 3.4 have also gone through the empirical analysis of *FB-SemK*, to check whether the performance of those zones' are also improved by *FB-SemK*. The Table 4.4 presents the empirical analysis for the same. In this table, the predicted imagery for *SemK* and *FB-SemK* for both the regions have been reported for the comparison. The corresponding error surfaces for each of the predicted imagery are also reported in the table along with the *PSNR*. For these zones also, *FB-SemK* yields higher *PSNR* over *SemK* (≈2–5 dB for Kolkata, WB, India and ≈3–4 dB for Dallas, TX, USA).

It must be noted that the performance of an interpolation method depends on the application and is highly influenced by the surrounding spatial variability. This notion is applicable for *SemK* and *FB-SemK* as well. The performance of these two methods

Table 4.2 Comparison study for *FB-SemK* (Region: Kolkata, WB, India) [4]

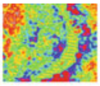
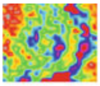
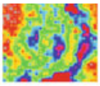
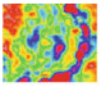
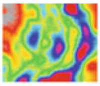
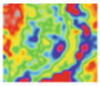
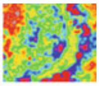
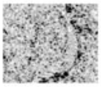
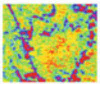
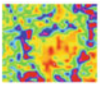
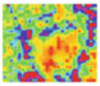
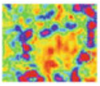
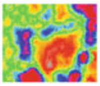
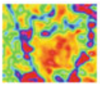
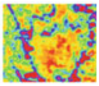
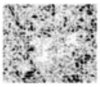
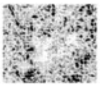
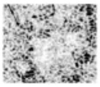
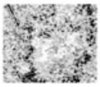
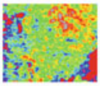
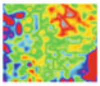
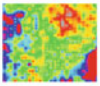
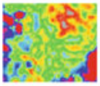
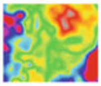
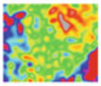
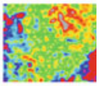
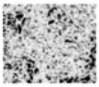
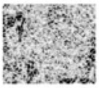
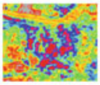
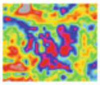
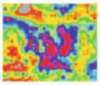
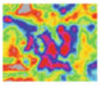
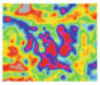
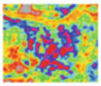
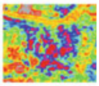
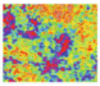
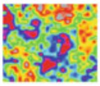
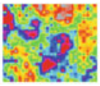
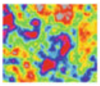
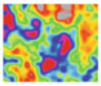
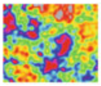
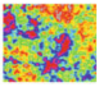
Zone	Actual image	Predicted image					
		NN	IDW	UK	OK	SemK	FB-SemK
Zone 1	BB: [(88°21′23.431″E 22°51′48.356″N); (88°25′38.059″E 22°55′5.136″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	39.73dB	39.24dB	37.58dB	36.30dB	43.50dB	48.72dB
Zone 2	BB: [(88°15′15.139″E 22°42′45.687″N); (88°19′29.709″E 22°46′4.027″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	37.59dB	36.78dB	34.64dB	31.94dB	39.23dB	43.65dB
Zone 3	BB: [(88°22′32.01″E 22°38′4.089″N); (88°26′45.357″E 22°41′21.274″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	39.49dB	38.50dB	36.71dB	35.08dB	43.22dB	49.47dB
Zone 4	BB: [(88°16′16.967″E 22°30′27.727″N); (88°20′31.233″E 22°33′45.087″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	34.42dB	33.49dB	31.74dB	34.43dB	39.53dB	44.52dB
Zone 5	BB: [(88°23′44.459″E 22°25′46.988″N); (88°27′57.874″E 22°29′4.16″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	43.55dB	43.52dB	42.03dB	42.92dB	47.43dB	50.05dB

Table 4.3 Comparison study for *FB-SemK* (Region: Dallas, TX, USA)

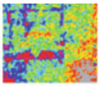
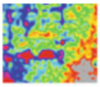
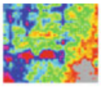
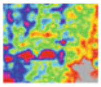
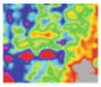
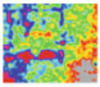
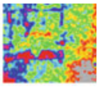
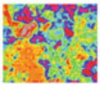
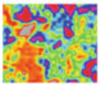
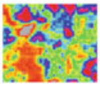
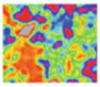
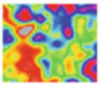
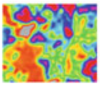
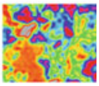
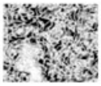
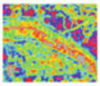
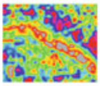
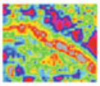
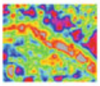
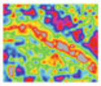
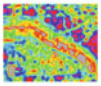
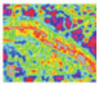
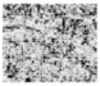
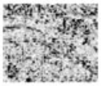
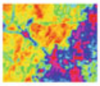
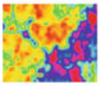
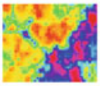
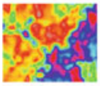
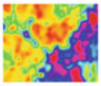
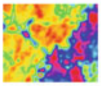
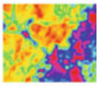
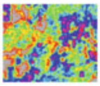
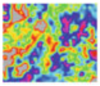
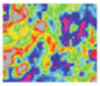
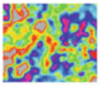
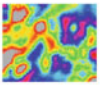
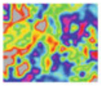
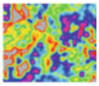
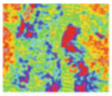
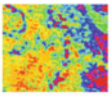
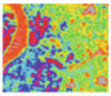
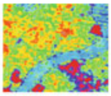
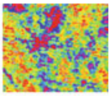
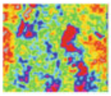
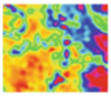
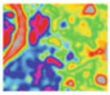
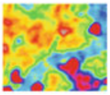
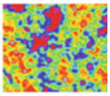
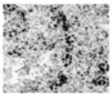
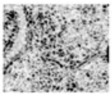
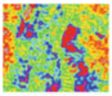
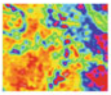
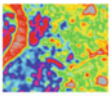
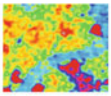
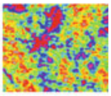
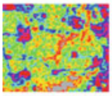
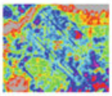
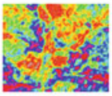
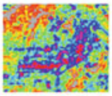
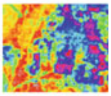
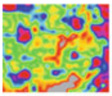
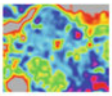
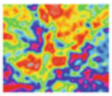
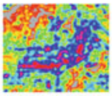
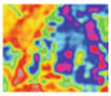
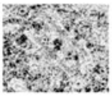
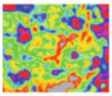
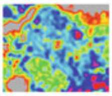
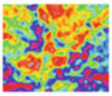
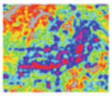
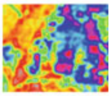
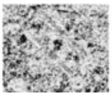
Zone	Actual image	Predicted image					
		NN	**IDW**	**UK**	**OK**	**SemK**	**FB-SemK**
Zone 1	BB: [(96°52′6.416″W 32°52′14.903″N); (96°49′17.242″W 32°54′10.726″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	37.73dB	36.99dB	35.04dB	36.08dB	42.12dB	45.69dB
Zone 2	BB: [(96°44′10.49″W 32°50′4.8″N); (96°41′21.511″W 32°52′0.445″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	33.82dB	32.38dB	31.54dB	31.29dB	35.15dB	38.32dB
Zone 3	BB: [(96°52′1.329″W 32°46′32.721″N); (96°49′12.051″W 32°48′28.785″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	34.86dB	33.62dB	32.39dB	34.50dB	39.85dB	42.41dB
Zone 4	BB: [(96°44′17.283″W 32°43′31.219″N); (96°41′28.232″W 32°45′27.113″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	35.83dB	34.21dB	32.86dB	34.37dB	40.33dB	43.92dB
Zone 5	BB: [(96°52′16.36″W 32°40′51.843″N); (96°49′28.131″W 32°42′47.697″N)]						
							
	Error surfaces (In gray scale) High Low						
	PSNR	32.71dB	30.75dB	30.05dB	31.42dB	34.81dB	37.39dB

Table 4.4 Comparison study for *FB-SemK* with *SemK*

Region	Method	Type	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Kolkata							
	SemK	Predicted					
		Error					
		PSNR	45.06dB	41.71dB	39.29dB	46.67dB	50.08dB
	FB-SemK	Predicted					
		Error					
		PSNR	47.92dB	45.36dB	43.79dB	51.01dB	54.02dB
Dallas							
	SemK	Predicted					
		Error					
		PSNR	38.16dB	37.44dB	37.84dB	40.82dB	34.65dB
	FB-SemK	Predicted					
		Error					
		PSNR	41.33dB	41.76dB	40.65dB	44.50dB	38.13dB

is dependent on the amount of semantic variability or the entropy in terms of *LULC* distribution of the *RoI*. Therefore, the performance of both *SemK* and *FB-SemK* vary in different study regions. The structure or the granularity of the ontology hierarchy impacts the prediction accuracy reported by *SemK* and *FB-SemK* significantly. The *SemK* and *FB-SemK* have more number of decision variables compared to *ordinary kriging* and other univariate interpolation approaches. However, to reduce the chance of over-fitting [10], the number of sampled locations considered in both *SemK* and *FB-SemK* is always assured to be ten times more than their number of independent factors.

4.5 Further Discussions

Spatial analysis for meteorological parameters, which are nearby to the earth surface often require the land–atmospheric interaction analysis, by modeling the spatial variability in terms of terrestrial *LULC* distribution. The basic *SemK* approach models these *LULC* information and incorporates this semantic knowledge into the interpolation for better accuracy. The *FB-SemK* improves *SemK* by revising its a-priori correlation analysis by probabilistic a-posterior correlation analysis among the leaf *LULC* classes in the ontology. It considers the mutual impact of *LULC*s on each other. Here, a probabilistic analysis of correlation is based on fuzzy *Bayesian network* (*FBN*) learning and inference generation principle. A *DAG* in the *Bayesian network* is utilized to properly capture the inter-*LULC* influences and the incorporated fuzziness deals with the uncertainties and the imprecision of the datasets that occurred due to the discretionary of the continuous meteorological parameters. The empirical studies exhibit that the *FB-SemK* improves the interpolation of *LST* by outperforming the existing popular techniques and also the *SemK*. Therefore, the contributions of *FB-SemK* are stated below:

- revising *semantic kriging*'s a-priori correlation study to a probabilistic a-posterior correlation study by modeling land–atmospheric interaction, i.e., the terrestrial *LULC* information for the meteorological parameters.
- utilizing a fuzzy extension of *Bayesian network* (*FBN*) of *LULC* classes to deal with the imprecision and uncertainty present in the data.
- improving the *SemK* process and thus enhancing the prediction accuracy by developing a probabilistic spatial importance evaluation algorithm.
- experimenting with *land surface temperature* data for validating the performance of the proposed *FB-SemK* compared to others.

It might be noted that both *SemK* and *FB-SemK* are capable of incorporating the terrestrial *LULC* knowledge into the prediction process. To achieve enhanced accuracy, both the methods, *SemK* and *FB-SemK*, can be considered based on the spatial application. However, there is a trade-off between both the methods. Though *FB-SemK* is more pragmatic interpolation process over *SemK*, however, the complexity in terms of number of parameters and the amount of processing required is higher in this process. Hence, with lower processing requirement, *SemK* is an efficient choice with negotiable amount of compromised accuracy. The proposed framework may facilitate the incorporation of other domain-specific knowledge and testing other fuzzy membership functions for achieving better interpolation accuracy.

It is evident that *FB-SemK* is an extension of basic spatial *SemK* process to make the spatial interpolation process more pragmatic one to achieve better accuracy. However, nowadays, most of the spatial data infrastructure contains time-series data. Past time-series meteorological parameters' information can be utilized for spatio-temporal prediction or forecasting of the parameters in the future. Even the spatio-temporal forecasting framework can be utilized for the forecasting of other spatial events. This spatio-temporal framework may facilitate modeling the applications, such as, change modeling, urban planning, natural resource management, etc. Therefore, a spatio-temporal interpolation framework, which can analyze past time-series data for prediction and forecasting, can be developed further by extending the basic *SemK* or *FB-SemK* framework. The subsequent chapters present these extensions for different spatio-temporal applications.

References

1. Beer M (2010) A summary on fuzzy probability theory. In: IEEE International conference on granular computing (GrC), IEEE, pp 5–6
2. Bhattacharjee S, Ghosh SK (2015) Performance evaluation of semantic kriging: a Euclidean vector analysis approach. IEEE Geosci Remote Sens Lett 12(6):1185–1189
3. Bhattacharjee S, Mitra P, Ghosh SK (2014) Spatial interpolation to predict missing attributes in GIS using semantic kriging. IEEE Trans Geosci Remote Sens 52(8):4771–4780
4. Bhattacharjee S, Das M, Ghosh SK, Shekhar S (2016) Prediction of meteorological parameters: an a-posteriori probabilistic semantic kriging approach. In: Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems, ACM, p 38
5. Ferreira L, Borenstein D (2012) A fuzzy-Bayesian model for supplier selection. Expert Syst Appl 39(9):7834–7844
6. Li J (2008) A review of spatial interpolation methods for environmental scientists. Record. Geoscience Australia, Australia
7. Li PC, Chen GH, Dai LC, Zhang L (2012) A fuzzy Bayesian network approach to improve the quantification of organizational influences in HRA frameworks. Saf Sci 50(7):1569–1583
8. Penz CA, Flesch CA, Nassar SM, Flesch RC, De Oliveira MA (2012) Fuzzy-Bayesian network for refrigeration compressor performance prediction and test time reduction. Expert Syst Appl 39(4):4268–4273
9. Tang H, Liu S (2007) Basic theory of fuzzy Bayesian networks and its application in machinery fault diagnosis. In: Fourth international conference on fuzzy systems and knowledge discovery, vol 4. IEEE, pp 132–137
10. Wilcox A, Hripcsak G (1999) Classification algorithms applied to narrative reports. In: Proceedings of the AMIA symposium, American Medical Informatics Association, p 455

Chapter 5
Spatio-Temporal Reverse Semantic Kriging

Abstract Spatio-temporal prediction and forecasting of the terrestrial land-use/land-cover (*LULC*) distribution of a *RoI* facilitates their management, city planning, mitigation of adverse climate change, etc. However, the spatio-temporal change in *LULC* distribution is not a trivial phenomena to be modeled as it often shows nonlinear behavior in the presence of different factors such as human impact to the ecosystem (e.g., anthropogenic activities, urbanization), change in meteorological parameters, etc. Therefore, for the prediction and forecasting of *LULC* distribution, incorporation of the meteorological knowledge into the prediction process may facilitate us to develop an advanced model. This work aims to model the behavioral change of interannual *LULC* pattern of a region by analyzing different related meteorological parameters. This study also attempts to forecast the future *LULC* distribution using the basic idea of *semantic kriging*. Here, *SemK* approach is extended for the spatio-temporal analysis and a new variant is proposed, which is referred to as *ST-RevSemK*. It captures the semantic relationships among different related meteorological parameters and use this relation to forecast the semantic terrestrial distribution pattern. From the empirical performance evaluation of this framework, it is found that the spatio-temporal modeling of meteorological parameters facilitates improved prediction of *LULC* pattern.

5.1 Introduction

Prediction and forecasting of the meteorological and terrestrial patterns with high accuracy is one of the major challenges in the field of environmental changes modeling. In this regard, modeling the *LULC* change and forecasting their future pattern helps to improve the urbanization, natural resource planning, different socioeconomic activities, etc., This modeling is important because the abovementioned factors are the significant driving force for different environmental threats such as *drought*, *flood*, *urban heat island*, etc. From the existing literature and the previous empirical evidence, it is evident that the meteorological parameters such as *land surface temperature*, *moisture stress index*, *vegetation index*, etc., are influenced by the earth surface dynamics in terms of *LULC* and are highly correlated with its pattern of

S. Bhattacharjee et al., *Semantic Kriging for Spatio-temporal Prediction*, Studies in Computational Intelligence 839,
https://doi.org/10.1007/978-981-13-8664-0_5

distribution [5, 13]. Thus, these two factors, the meteorological (the parameters) and the terrestrial (the *LULC*), are interdependent, which influence the environmental changes together. As a result, the proposed methods, the basic *SemK*, *FB-SemK* and the spatio-temporal *SemK* incorporates this knowledge with their semantic analysis for the prediction with higher accuracy. This chapter focuses on the spatio-temporal forecasting of *LULC* pattern of a given *RoI* by modeling the behavioral pattern of different meteorological parameters. Hence, this work can be regarded as an application of the proposed *SemK* approach, where *SemK* and its variant have been applied for the change pattern analysis of *LULC* and its forecasting through urban landscape modeling.

Environmental scientists have already declared that the twenty-first century is going to be an era of environmental pattern recognition, weather/climate prediction, using some well-defined mathematical equations [20]. However, the prediction and forecasting by handling all the surrounding and influential uncertainties with high accuracy is difficult to achieve. One of the major reasons behind this unsatisfactory level of accuracy is that, in most of the cases other critical covariates' information are not considered which are highly influential to each other. One best possible way to handle the uncertainties is to incorporate different influential secondary knowledge for the prediction of one parameter. In this regard, the primary parameter should exhibit high correlation with the secondary parameters, such that they influence each other significantly. Therefore, it is obvious to model any environmental change phenomena considering a multivariate scenario for more pragmatic estimation compared to a univariate one.

According to Tobler's law, the spatial parameters are highly correlated among themselves in 2D space [10]. Therefore, in a multivariate scenario, it is a challenging task to select the best group of auxiliary parameters and prioritize them with respect to their degree of influence to the primary parameter. These best possible secondary parameters facilitate us to achieve better results for the actual analysis. Therefore, one of the major preprocessing steps involved in any multivariate spatial analysis is to check and extract the causal dependencies among the meteorological parameters to evaluate their degree of influence to each other.

As the spatio-temporal alteration of *LULC* follows nonlinear behavior in the presence of different dynamic factors such as human impact to the ecosystem (e.g., anthropogenic activities, urbanization), change in meteorological parameters, etc., thus a proper prediction and forecasting model of *LULC* should study historic data and model the past trend of different parameters. Therefore, for these types of data-driven approaches to analyze the past pattern of different environmental parameters, the *kriging* [14] estimators are reported to be the most popularly used and suitable approaches in the literature to handle spatial uncertainties. The example of multivariate geostatistical *kriging* estimators are *co-kriging*, *kriging with external drift*, *regression kriging*, etc.

However, for meteorological and terrestrial applications in the literature, most of these univariate and multivariate methods have not considered modeling terrestrial *LULC* knowledge. In *semantic kriging* [5] (and its variants, *FB-SemK* and spatio-temporal *SemK*), the *LULC* distribution of the terrain is amalgamated within the prediction equations of the meteorological parameters, mainly *LST* for better estimation. In this chapter, a spatio-temporal forecasting framework is proposed by considering *LULC* distribution of the terrain as the primary parameter to be predicted. This framework is named as *spatio-temporal reverse semantic kriging* (*ST-RevSemK*) [3]. Here, the related meteorological parameters are used as the auxiliary or the secondary parameters to support the prediction of *LULC*. Therefore, *ST-RevSemK* can be considered as a multivariate *kriging* approach. The variances between secondary parameters with the primary parameter is modeled by *semivariogram*, *cross-semivariogram* models [6]. For empirical evaluation, three correlated parameters are considered, namely, *LST*, *NDVI*, and *MSI* as the secondary information to forecast the primary (*LULC*). From the empirical analysis with spatio-temporal meteorological and terrestrial *LULC* data, it is observed that the incorporation of semantic knowledge into the spatio-temporal modeling prediction of *LULC* facilitate more precise estimation.

5.2 Objectives of Spatio-Temporal Reverse Semantic Kriging

Spatio-temporal change modeling of *LULC* distribution of a region is required to forecast their future pattern which facilitates different types of environmental planning. Being influenced by terrestrial dynamics, the temporal changes in different meteorological parameters rely on the properties of *LULC*. Therefore, a reverse mapping of meteorological information for the spatio-temporal prediction of *LULC* helps to develop better prediction model. The interannual variation of different meteorological parameters and the *LULC* pattern have been analyzed in this work to forecast the future *LULC* spread. A revised version of spatio-temporal *semantic kriging* is developed and is named as *ST-RevSemK*. It captures the semantic associations between different parameters and the terrestrial *LULC* distribution. The general objectives of *ST-RevSemK* is given as follows:

- executing causality test in different spatial locations to rank the selected parameters or a group of secondary parameters with respect to the amount of impact on the prediction parameter.
- extending the basic *semantic kriging* framework for the multivariate prediction scenario to consider auxiliary information.
- developing a reverse spatio-temporal *SemK* approach, i.e., *ST-RevSemK*, for quantifying the spatio-temporal change in *LULC* pattern in terms of related meteorological parameters.
- developing the spatio-temporal *semivariogram*s and *cross-semivariogram*s model to analyze the correlation among different input variables.
- mathematical formalization of the proposed *ST-RevSemK* framework to forecast *LULC* pattern in future.
- conducting empirical experiment with meteorological and terrestrial data to validate the efficacy of the *ST-RevSemK* framework.

5.3 ST-RevSemK: Spatio-Temporal Reverse Semantic Kriging Framework

The proposed *spatio-temporal reverse semantic kriging* (*ST-RevSemK*) framework can be considered as the reverse and multivariate approach of basic *SemK* in spatio-temporal domain. A comprehensive process flow diagram of *ST-RevSemK* has been shown in Fig. 5.1. This model considers the spatio-temporal data of meteorological parameters and terrestrial *LULC* as the input. Besides the quantification of the terrestrial knowledge, by following the basic *SemK* principle (refer Chap. 3 and [5]), the spatio-temporal interrelationships between different combinations of meteorological parameters and the *LULC*s are analyzed further. Here, for spatio-temporal analysis, a separable spatio-temporal *SemK* approach [4, 24] is chosen to be extended in multivariate scenario, taking *LULC* as the primary parameter for prediction. Considering three- dimensional spatio-semantic and temporal-semantic *semivariogram*s, the *ST-RevSemK* framework forecasts the future *LULC* distribution of the chosen *RoI*.

A major preprocessing step in the *ST-RevSemK* approach is the *CTF* component. It is basically a causality testing framework (*CTF*), which takes multiple meteorological parameters (primary or derived) as input. From the pool of the parameters, this module checks whether the input secondary parameters actually influence the primary parameter and further selects those parameters, which are causal to the primary parameter to be predicted (here, it is *LULC*), by pruning the rest. Here, a *Granger causality* (*GC*) testing approach [2] has been adopted, which is a data preprocessing task for the dependency analysis between meteorological and terrestrial

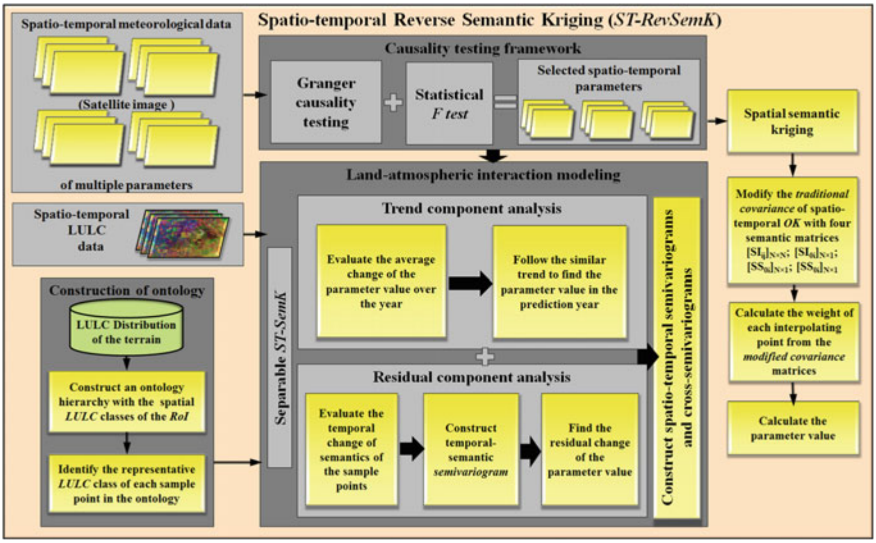

Spatio-temporal meteorological data
(Satellite image)
of multiple parameters
Spatio-temporal LULC data
Construction of ontology
LULC Distribution of the terrain
Construct an ontology hierarchy with the spatial LULC classes of the RoI
Identify the representative LULC class of each sample point in the ontology
Spatio-temporal Reverse Semantic Kriging (ST-RevSemK)
Causality testing framework
Granger causality testing
+
Statistical F test
=
Selected spatio-temporal parameters
Land-atmospheric interaction modeling
Separable ST-SemK
Trend component analysis
Evaluate the average change of the parameter value over the year
Follow the similar trend to find the parameter value in the prediction year
+
Residual component analysis
Evaluate the temporal change of semantics of the sample points
Construct temporal-semantic semivariogram
Find the residual change of the parameter value
Construct spatio-temporal semivariograms and cross-semivariograms
Spatial semantic kriging
Modify the traditional covariance of spatio-temporal OK with four semantic matrices [SI_ij]_N×N; [SI_0i]_N×1; [SS_0i]_N×1; [SS_0i]_N×1
Calculate the weight of each interpolating point from the modified covariance matrices
Calculate the parameter value

Fig. 5.1 *ST-RevSemK* framework [3]

parameters. The details of this causality testing framework and modeling the *ST-RevSemK* prediction equations are discussed in the following subsections.

5.3.1 CTF: Causality Testing Framework

The *CTF* module investigates the causal linkages between different input meteorological parameters that has been considered to forecast the primary parameter, *LULC*. In this regard, the causal dependencies among the primary parameter, i.e., the *LULC* classes and the *LST*, *NDVI* and *MSI* (as secondary parameters) [7, 9] are investigated further. For the causality analysis, the CTF considers a *Granger causality* (*GC*) [11, 22] testing approach for the causal relationship extraction among different parameters. As per existing the literature, the *Granger causality* can be referred to as a data-driven framework, which was primarily developed for the econometrics applications by the British economist *Clive Granger*. Further, many studies have extended this approach in different fields to analyze the causal dependencies among different stochastic variables. This approach aims to minimize the error in prediction and forecasting. For a bivariate scenario, the causality hypothesis, defined by Granger in [11] is given as follows: "if some other series y_t contains information in past terms that helps in the prediction of x_t and if this information is contained in no other series used in the predictor, then y_t is said to cause x_t." Further, this bivariate hypothesis has been extended in multivariate approaches as well to find the most influential and statistically significant combination of the secondary variables for the primary variable, which can also be prioritized accordingly as per their influence quotient.

For the application of *Granger causality* testing approach in the area of meteorological analysis, a significant amount of scientific investigations have been reported in the literature. Attanasio et al. [1] have outlined a review report on the *Granger causality* testing for reasoning the global warming phenomena. Salvucci et al. [18] have extracted the causal relationship between two meteorological parameters, the *soil moisture* and the *precipitation* in the region Illinois, USA. Similarly, Lozano et al. [16] have described a *GC*-based spatio-temporal data analysis approach for the causal analysis of climate change. Considering own datasets, the authors have reported that the presence of CO_2 and other greenhouse gases significantly influence the *temperature* change in the environment. Sfetsos and Vlachogiannis [19] have considered applied *GC* testing approach to check the causal dependencies between PM_{10} concentrations and its daily exceedances. Smirnov and Mokhov [21] have developed a long-term *GC* testing principle. They have reported that the CO_2 concentration is causal to the rise in *temperature* during the past decades. Kodra et al. [15] have developed a reverse cumulative *GC* testing method to extract the causal dependency between CO_2 and globally averaged *LST*. On the other hand, Dutta et al. [8] have developed an *online* feature extraction approach using *GC* testing principle for predicting *rainfall* with using neural network and reported influential features from the historic *rainfall* data to forecast it in future.

For a multivariate scenario, a group of variables often exhibits more impact or the influence toward another variable or the primary parameters compared to influencing it individually. Formally, it can be described as the group of parameters are not statistically significant individually, however, shows more significant dependency when grouped with others. Therefore, for this multivariate analysis, a bottom-up *GC* testing approach in *CTF* is considered, where the causality is first checked with individual secondary parameters and gradually making groups of them in different combinations of variables. The final group consists of all the secondary parameters as one group and all the groups are annotated and prioritized in terms of their influence measure for the primary parameter. For the *GC* testing, we have adopted the principle of statistical *F-test*. Some additional spatio-temporal properties can also be verified through this testing. For example, the due to the spatial autocorrelation property of the terrain, the nearby study locations should report similar dependency compared to the distant locations. Similarly, it can also be checked how the interannual change of the considered parameters has affected their inter-causal dependencies in different spatial locations.

5.3.1.1 Granger Causality Testing Principles

The basic principle of *Granger causality* follows the idea of one-way causality in temporal domain. That is, the future value always gets influenced by the past values but the opposite is not possible. Similarly, the *GC* testing principle states that, if the change of one variable X causes the change is another variable Y, then X can be included as an independent factor for the analysis of Y, which eventually increase the accuracy of the analysis of Y. Therefore, the following hypotheses can be stated based on the above notion:

- Y is caused by X indicates that the past values of both X and Y influences the future value of Y.
- X is caused by Y indicates that the past values of both X and Y influences the future value of X.
- X is caused by Y and Y is caused by X both indicate that past values of both X and Y influences the future value of both X and Y.
- X is not caused by Y and Y is not caused by X both indicate that X and Y are statistically independent from each other.

In univariate scenario, the autoregressive representation of Y (order N) can be given as follows:

$$y_t = \theta_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_N y_{t-N} + \epsilon_t \tag{5.1}$$

$$= \theta_0 + \sum_{i=1}^{N} a_i y_{t-i} + \epsilon_t \tag{5.2}$$

where y_i represents the value of Y at the ith time from its past time-series data, ϵ_t represents the residual value at time t, θ_0 and a_i are the constants. In the bivariate scenario, the similar autoregressive representation of Y, considering the lagged values of X, can be given as follows:

$$\begin{aligned} y_t &= \phi_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_N y_{t-N} + b_1 x_{t-1} + \\ &\quad b_2 x_{t-2} + \cdots + b_N x_{t-N} + \xi_t \qquad (5.3) \\ &= \phi_0 + \sum_{i=1}^{N} a_i y_{t-i} + \sum_{i=1}^{N} b_i x_{t-i} + \xi_t \qquad (5.4) \end{aligned}$$

where x_i represents the value of X at the ith time from its past time-series data, ξ_t represents the residual value at time t, ϕ_0 and b_i are the constants. In this bivariate scenario, if X *Granger* causes Y, the b can be considered as a N dimensional nonzero vector. Here, the bivariate autoregressive representation of Y produces higher accuracy for the analysis of Y compared to its univariate autoregression. As *GC* is verified through some statistical hypothesis tests, this work has adopted the statistical *F-test* for the testing of null hypothesis.

5.3.1.2 Statistical F-Test

In this work, the statistical *F-test* [17] is utilized to check whether a set of secondary parameters, individually or as a group, influence the primary parameter or not. The statistical *F-test* assumes the test statistic to have a *F-distribution* under null hypothesis. So, for this work, the null hypothesis of *GC* test is given as follows: *for the univariate autoregressive representation of* Y, $y_j \in Y$ *is not caused by* $y_i \in Y$. It must be noted that this hypothesis is a variant of actual *GC* hypothesis. It may also be explained by the scenario when every coefficient a_i is zero in vector a. The alternative hypothesis can be stated as for at least one i in vector a, $a_i \neq 0$.

To extend this method for the bivariate regression analysis, the the *F statistic* is evaluated between two models using the *F-test*. Let us assume that the number of data points available for the parameter estimation of both the models be n. Again, assuming the number of parameters for the first model be p_1 and for the second model be p_2 ($p_2 > p_1$), the corresponding *F statistic* is expressed as follows:

$$F = \frac{\left(\frac{\text{SSE}_1 - \text{SSE}_2}{p_2 - p_1}\right)}{\left(\frac{\text{SSE}_2}{n - p_2}\right)} \qquad (5.5)$$

where SSE_i represents the sum of squares of the residual of the ith model which can be expressed as follows: $SSE_i = \sum(y_i - \bar{y}_i)^2$, the $\bar{y}_i$ represents the mean of the series Y. In general, the group having more parameters always better fit the data compared to the group with lesser parameters. However, the null hypothesis violates this statement as the former model does not generate any better under null hypothesis. In F-distribution, if the *F statistic* is larger than the critical value for a given significance level s (usually 0.05), then the null hypothesis is rejected. The critical value can be formalized as $F_{crit}(m_1, m_2)$. Here, the between-group degrees of freedom is represented as m_1 and the within-group degrees of freedom is represented as m_2. Considering the acceptance of the null hypothesis, the p value represents the probability to obtain the *F statistic* at least as extreme as the observed one. For rejecting the null hypothesis, p must be less than s.

5.3.1.3 Causality Testing in ST-RevSemK

In the multivariate analysis of *ST-RevSemK*, the secondary parameters are needed to be ranked to according to their impact of the primary parameter. It also gives an insight to understand whether the secondary parameters are influencing the primary one as a group or individually. In this work, the group of secondary meteorological parameters are *LST*, *NDVI* and *MSI*, which are parameters that are ranked considering a bottom-up approach to forecast the primary parameters, terrestrial *LULC*. It is obvious from the rejection assumption of the null hypothesis in *F-test* is that at least one coefficient from each of a and b vector should be nonzero. Thus, the model is said to be statistically significant to forecast Y. To choose the best model for the forecasting of Y considering every possible combination of the secondary parameters, this work develops an exhaustive hierarchical approach to group them. At the bottom level, each of the secondary parameters are gone through *GC* testing with respect to the terrestrial *LULC*. The corresponding *F statistic* are measured and the parameters are ranked accordingly in terms of their *F statistic*. The model with higher *F statistic* are more statistically significant, and thus assigned lower rank compred to the others. This work assumes that a secondary parameter may not seem to be significant individually but shows higher impact for the secondary parameter when grouped together with other parameters. Therefore, even if the null hypothesis is accepted (with lower *F statistic*) for a individual parameter, it is further analyzed by grouping it with other secondary parameters. At the second level, the parameters are paired in every possible combination and the multivariate analysis is carried out for every grouping. This grouping is continued uptil level n (n represents the number of secondary parameters), in which all the secondary parameters are clustered in a single group.

The abovementioned bottom-up approach to measure the *F statistic* for the group of the secondary parameters is presented formally in Algorithm 3 [2]. Each of the combinations are tagged with their corresponding *F statistic* value. With respect to the physical significance of the *F statistic* metric, the combination of secondary parameters with higher *F statistic* value influences the primary parameter more. Therefore, for this multivariate scenario, the combination with the highest *F statistic* value or a lowest rank is the best model to forecast the the terrestrial *LULC*, given the null hypothesis is rejected for that model.

For the causality testing in *ST-RevSemK* among the four parameters (*LST*, *NDVI*, *MSI*, and *LULC*), the empirical analysis has been carried out for five spatial zones as considered in Figs. 3.3 and 3.4 for Kolkata, WB, India and Dallas, TX, USA, respectively (refer Chap. 3). The statistical *F-Test* results for each of the five zones for both Kolkata, WB, India and Dallas, TX, USA are specified in Table 5.1 and Table 5.2, respectively. An analytic software by IBM, IBM SPSS[1] Statistics 15.0 is used here for the bivariate and multivariate analysis of this work. For the bivariate scenario, the *One-way Anova* test of IBM SPSS is performed for individual secondary parameter. For the combination with three and more parameters, the multivariate linear regression of IBM SPSS is performed to evaluated their *F statistic*s.

Let <G> be a group or the combination of secondary parameters. Thus the null hypothesis can be stated as "The *LULC* is not *Granger* caused by G". In Tables 5.1 and 5.2, the *F statistic*s and the corresponding critical values of the *F-test* are tabularized for each of the combinations of the parameters, considering five zones from each of the two *RoI*s, Kolkata and Dallas. A ✓ in the "Reject?" column indicates to reject the null hypothesis, whereas a × mark indicates to accept it. The results in the Table indicate that the group that considers the combination of all secondary parameters together, i.e., <*land surface temperature*, *normalized difference vegetation index*, *moisture stress index*> is the most significant model for the prediction year 2015. It has also been found that even if the primary parameter is not *Granger* caused by the individual secondary parameter, but cased by them in a single group. This supports the notion of not rejecting the insignificant secondary parameters when tested individually.

5.3.2 Modeling ST-RevSemK

As *ST-RevSemK* deals with multivariate scenario, it assumes that for any ith sampled location from N sample points at a past time instance can be represented as $(x_i, \{Z_{1_i}, Z_{2_i}, \ldots, Z_{s_i}\}, f_i, t_{a_i})$. That is the sampled location x_i is measured at a past year t_{a_i}. The number of meteorological parameters considered is s and are represented as $Z_{1_i}, Z_{2_i}, \ldots, Z_{s_i}$ at a sample point x_i and its corresponding *LULC*, f_i is also known. Here, for our analysis, three meteorological parameters are considered, *LST*, *NDVI* and *MSI*, thus $s = 3$. For the forecasting of f_0 at the prediction point x_0, only past *LULC* information is not a pragmatic approach to follow. Modeling the other related meteorological information may enhance the accuracy in estimation. The proposed approach utilizes the notion of spatio-temporal *semantic kriging* [4] and extends it for multivariate approach, where three secondary parameters are chosen to predict the terrestrial *LULC*. In any multivariate approach it is assumed that the auxiliary parameters are highly correlated with the primary and the primary parmater can be predicted jointly considering a BLUE (best linear unbiased estimator) [12]. From our causality analysis, it is observed that all three secondary parameters together

[1] http://www-01.ibm.com/software/in/analytics/spss/; Accessed on August 2014.

Table 5.1 F-test for *ST-RevSemK* (Region: Kolkata, WB, India)

Zone	Level	Parameter groups	*F statistic*	*p* value	Critical value	Reject?	Rank	Level rank
Zone 1	BB: [(88°21′56.75″E 22°53′3.338″N); (88°26′11.684″E 22°56′20.576″N)]							
	Level 1	<LST>	5.929	0.001	1.82	✓	6	3
		<NDVI>	5.867	0.001		✓	7	
		<MSI>	5.980	0.001		✓	5	
	Level 2	<LST, NDVI>	11.704	0.001	1.48	✓	3	2
		<LST, MSI>	12.406	0.003		✓	2	
		<NDVI, MSI>	11.691	0.002		✓	4	
	Level 3	<LST, NDVI, MSI>	13.634	0.002	1.62	✓	1	1
Zone 2	BB: [(88°16′37.345″E 22°43′16.086″N); (88°20′51.518″E 22°46′33.581″N)]							
	Level 1	<LST>	7.369	0.001	1.95	✓	7	3
		<NDVI>	7.630	0.001		✓	6	
		<MSI>	8.446	0.001		✓	5	
	Level 2	<LST, NDVI>	10.562	0.003	1.62	✓	4	2
		<LST, MSI>	11.184	0.001		✓	2	
		<NDVI, MSI>	10.679	0.002		✓	3	
	Level 3	<LST, NDVI, MSI>	13.577	0.002	1.77	✓	1	1
Zone 3	BB: [(88°21′3.362″E 22°34′43.896″N); (88°25′17.437″E 22°38′1.121″N)]							

(continued)

Table 5.1 (continued)

Zone	Level	Parameter groups	*F statistic*	*p* value	Critical value	Reject?	Rank	Level rank
	Level 1	<LST>	7.140	0.001	1.80	✓	6	3
		<NDVI>	6.556	0.001		✓	7	
		<MSI>	7.316	0.001		✓	5	
	Level 2	<LST, NDVI>	11.928	0.002	1.47	✓	4	2
		<LST, MSI>	12.017	0.003		✓	3	
		<NDVI, MSI>	12.084	0.001		✓	2	
	Level 3	<LST, NDVI, MSI>	15.002	0.001	1.53	✓	1	1
Zone 4	BB: [(88°10′24.797″E 22°29′45.122″N); (88°14′38.046″E 22°33′3.055″N)]							
	Level 1	<LST>	8.248	0.001	1.55	✓	6	3
		<NDVI>	7.600	0.001		✓	7	
		<MSI>	8.282	0.001		✓	5	
	Level 2	<LST, NDVI>	11.511	0.004	1.43	✓	2	2
		<LST, MSI>	11.197	0.002		✓	4	
		<NDVI, MSI>	11.301	0.001		✓	3	
	Level 3	<LST, NDVI, MSI>	14.535	0.003	1.59	✓	1	1
Zone 5	BB: [(88°24′21.572″E 22°24′41.309″N); (88°28′35.399″E 22°27′58.463″N)]							
	Level 1	<LST>	7.291	0.005	1.68	✓	5	3
		<NDVI>	6.444	0.001		✓	7	
		<MSI>	6.980	0.001		✓	6	
	Level 2	<LST, NDVI>	12.102	0.001	1.50	✓	3	2
		<LST, MSI>	12.466	0.002		✓	2	
		<NDVI, MSI>	11.946	0.004		✓	4	
	Level 3	<LST, NDVI, MSI>	14.949	0.001	1.80	✓	1	1

Table 5.2 F-test for *ST-RevSemK* (Region: Dallas, TX, USA)

Zone	Level	Parameter groups	*F statistic*	*p* value	Critical value	Reject?	Rank	Level rank
Zone 1	BB: [(96°48′52.626″W 32°53′19.252″N); (96°46′4.031″W 32°55′15.003″N)]							
	Level 1	<LST>	4.165	0.001	1.54	✓	5	3
		<NDVI>	3.892	0.002		✓	7	
		<MSI>	4.059	0.001		✓	6	
	Level 2	<LST, NDVI>	9.613	0.001	2.04	✓	4	2
		<LST, MSI>	9.926	0.001		✓	3	
		<NDVI, MSI>	10.474	0.002		✓	2	
	Level 3	<LST, NDVI, MSI>	12.898	0.003	1.82	✓	1	1
Zone 2	BB: [(96°52′26.827″W 32°49′41.121″N); (96°49′38.61″W 32°51′36.489″N)]							
	Level 1	<LST>	5.233	0.001	1.80	✓	6	3
		<NDVI>	5.176	0.001		✓	7	
		<MSI>	5.861	0.001		✓	5	
	Level 2	<LST, NDVI>	10.290	0.001	1.99	✓	2	2
		<LST, MSI>	9.885	0.002		✓	4	
		<NDVI, MSI>	10.141	0.003		✓	3	
	Level 3	<LST, NDVI, MSI>	13.699	0.001	1.80	✓	1	1
Zone 3	BB: [(96°44′48.861″W 32°47′22.359″N); (96°42′0.958″W 32°49′17.557″N)]							
	Level 1	<LST>	5.550	0.001	1.95	✓	5	3
		<NDVI>	4.866	0.001		✓	7	
		<MSI>	5.131	0.001		✓	6	
	Level 2	<LST, NDVI>	8.771	0.003	1.53	✓	4	2
		<LST, MSI>	9.049	0.002		✓	3	
		<NDVI, MSI>	9.176	0.001		✓	2	
	Level 3	<LST, NDVI, MSI>	13.130	0.003	1.53	✓	1	1

(continued)

Table 5.2 (continued)

Zone	Level	Parameter groups	*F statistic*	*p* value	Critical value	Reject?	Rank	Level rank
Zone 4	BB: [(96°50′50.616″W 32°44′1.516″N); (96°48′2.552″W 32°45′56.856″N)]							
	Level 1	<LST>	6.737	0.001	1.99	✓	7	3
		<NDVI>	6.796	0.002		✓	6	
		<MSI>	7.259	0.001		✓	5	
	Level 2	<LST, NDVI>	10.631	0.001	1.77	✓	2	2
		<LST, MSI>	10.604	0.002		✓	3	
		<NDVI, MSI>	10.443	0.004		✓	4	
	Level 3	<LST, NDVI, MSI>	12.901	0.003	2.04	✓	1	1
Zone 5	BB: [(96°44′27.666″W 32°42′20.309″N); (96°41′39.631″W 32°44′15.74″N)]							
	Level 1	<LST>	5.525	0.001	1.82	✓	7	3
		<NDVI>	6.061	0.001		✓	6	
		<MSI>	6.408	0.001		✓	5	
	Level 2	<LST, NDVI>	9.886	0.002	1.73	✓	4	2
		<LST, MSI>	10.008	0.001		✓	3	
		<NDVI, MSI>	10.078	0.001		✓	2	
	Level 3	<LST, NDVI, MSI>	13.262	0.003	1.75	✓	1	1

Algorithm 3: Bottom-up *F statistic* evaluation

```
Input: Primary parameter $\{v_0\}$;
Secondary parameters $\{\{v_1\}, \{v_2\}, \cdots, \{v_n\}\}$
Output: F statistic
$Model_1 = \{v_0\}$;
Hierarchy $A = \phi$;
$G = \{\{v_1\}, \{v_2\}, \cdots, \{v_n\}\}$
foreach $i \in 1$ to $n$ do
    $A = A \cup G$
    $G = G \times v_i$
end
foreach $Model_i$ in $A$ do
    $Fstatistic_{(Model_1, Model_i)} = \frac{\left(\frac{SSE_{group1} - SSE_{group2}}{p_2 - p_1}\right)}{\left(\frac{SSE_{group2}}{n - p_2}\right)}$
    if $Fstatistic_{(Model_1, Model_i)} < F_{crit}(m_1, m_2)$ then
        accept null hypothesis with $Model_i$
        discard $Model_i$
    end
    else
        reject null hypothesis with $Model_i$
        annotate $Model_i$ with $Fstatistic_{(Model_1, Model_i)}$
    end
end
```

influence *LULC* (refer Tables 5.1 and 5.2). In univariate approach, the estimation by *spatio-temporal reverse semantic kriging* can be formalized as follows:

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_i^{ST-RevSemK} Z(x_i) \tag{5.6}$$

where $\hat{Z}(x_0)$ represents the predicted primary parameter value, (x_0) is the unsampled location to be predicted, N is the number of sampled locations, x_i represent the primary parameter value at the ith sampled location and $w_i^{ST-RevSemK}$ represents the assigned weight to the same location. For multivariate *spatio-temporal reverse semantic kriging*, as three parameters [*LST* (Z_1), *NDVI* (Z_2) and *MSI* (Z_3)] are chosen as the auxiliary parameter for the prediction of *LULC*, the corresponding estimation equations can be expressed as follows:

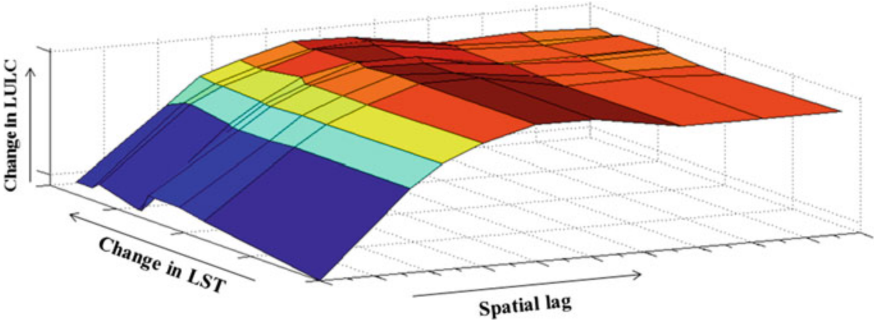


Fig. 5.2 Spatio-semantic *semivariogram* (Parameter: *LST*) [3]

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_{1_i}^{ST-RevSemK} Z(x_i) \text{ such that: } \hat{Z}_1(x_0) = \sum_{i=1}^{N} w_{1_i}^{ST-RevSemK} Z_{1_i} \quad (5.7)$$

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_{2_i}^{ST-RevSemK} Z(x_i) \text{ such that: } \hat{Z}_2(x_0) = \sum_{i=1}^{N} w_{2_i}^{ST-RevSemK} Z_{2_i} \quad (5.8)$$

$$\hat{Z}(x_0) = \sum_{i=1}^{N} w_{3_i}^{ST-RevSemK} Z(x_i) \text{ such that: } \hat{Z}_3(x_0) = \sum_{i=1}^{N} w_{3_i}^{ST-RevSemK} Z_{3_i} \quad (5.9)$$

Being a variant of the basic *semantic kriging* process, these estimation equations adheres to the following conditions: $\sum_{i=1}^{N} w_{1_i}^{ST-RevSemK} = \sum_{i=1}^{N} w_{2_i}^{ST-RevSemK} = \sum_{i=1}^{N} w_{3_i}^{ST-RevSemK} = 1$ (as unbiased estimator). The $w_{s_i}^{ST-RevSemK}$s are the weights assigned to the interpolating points, calculated from the *semivariogram* models. In *ST-RevSemK* framework, the dependencies among the meteorological and terrestrial *LULC* parameters are modeled through experimental spatio-temporal *semivariograms*. In this application, as the primary prediction parameter *Z* is the terrestrial *LULC* distribution, the *semantic kriging* modeling of this semantic knowledge is carried out. Therefore, the *semantic similarity* and the *spatial importance* metric of *SemK* quantifies this knowledge from the ontology (refer Chap. 3). The spatial and temporal *semivariogram* models then analyze the changes in the prediction parameter in space and time domain with respect to different combinations of the secondary parameters. Hence, for each of the three auxiliary parameters, the spatio-semantic and temporal-semantic *semivariogram* models (for spatial region Kolkata, WB, India) are depicted in Figs. 5.2, 5.4, 5.6, 5.3, 5.5, and 5.7 respectively. These temporal and spatial *semivariogram* models considers temporal lag and distance respectively as an independent variable along the X axis, the change in meteorological parameter as another independent variable along Y axis and the corresponding change in *LULC* as the dependent variable along Z axis.

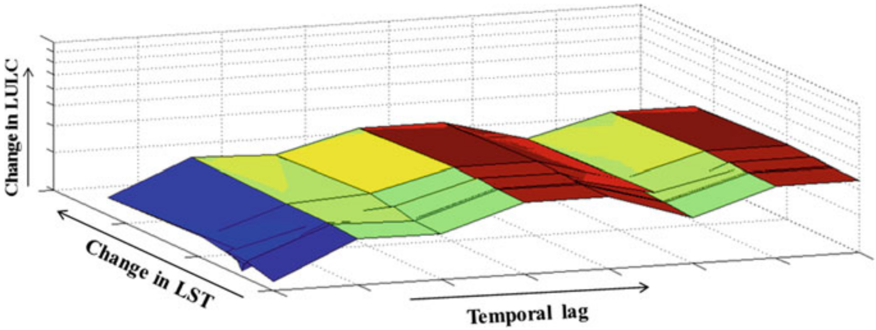


Fig. 5.3 Temporal-semantic *semivariogram* (Parameter: *LST*) [3]

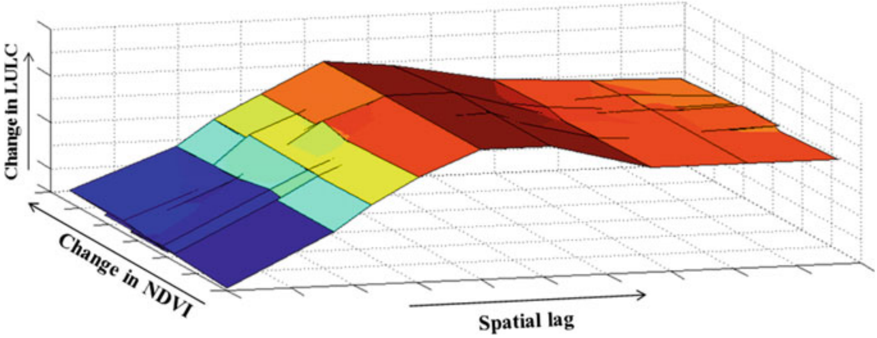


Fig. 5.4 Spatio-semantic *semivariogram* (Parameter: *NDVI*) [3]

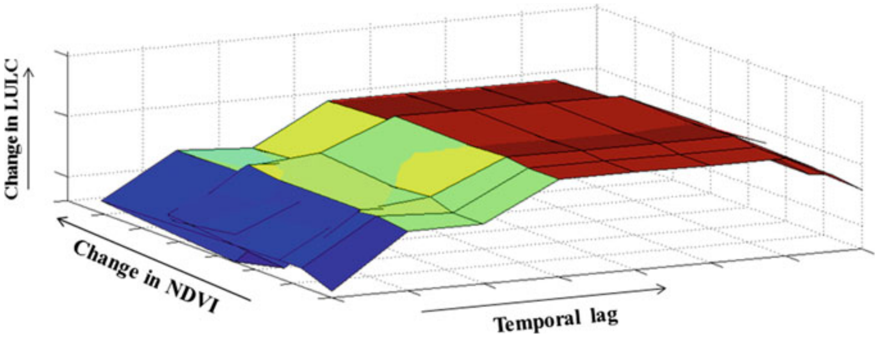


Fig. 5.5 Temporal-semantic *semivariogram* (Parameter: *NDVI*) [3]

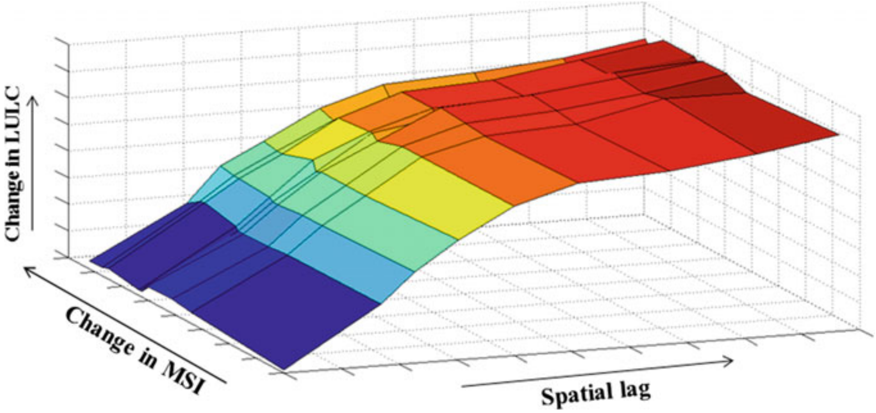


Fig. 5.6 Spatio-semantic *semivariogram* (Parameter: *MSI*) [3]

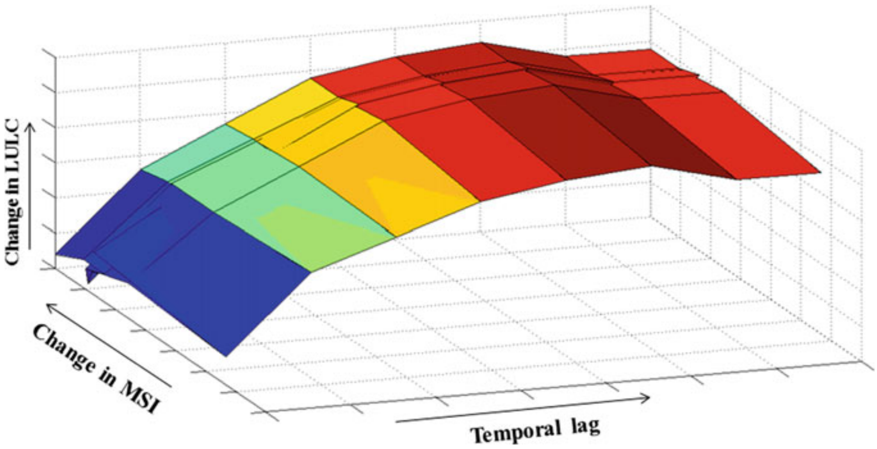


Fig. 5.7 Temporal-semantic *semivariogram* (Parameter: *MSI*) [3]

However, different combinations or all the secondary meteorological parameters together can be considered to evaluate the change in *LULC* with respect to the change in these secondary parameters. For this analysis, the *semivariogram* models considering two or more *semivariogram*s or *cross-semivariogram*s [6, 23] should be modeled further. Anisotropy ratios can be introduced to model the trade-off among the secondary parameters in spatio-temporal domain. Figures 5.8 and 5.9 show the change in semantic *LULC* information with respect to three secondary parameters in spatial and temporal domain respectively.

After modeling these *semivariogram*s, the separable spatio-temporal *SemK* is carried out further. The traditional spatio-temporal *semivariance matrix* ($[\mathbf{C}^{ST}]_{ij_{[N\times N]}}$)

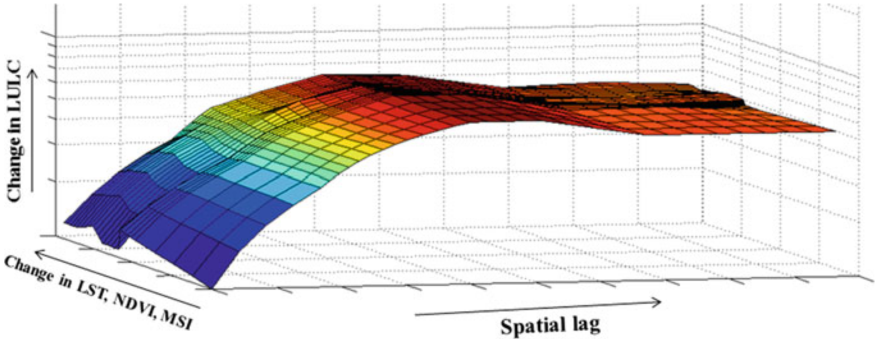


Fig. 5.8 Spatio-semantic *cross-semivariogram* (Parameter: *LST, NDVI, MSI*) [3]

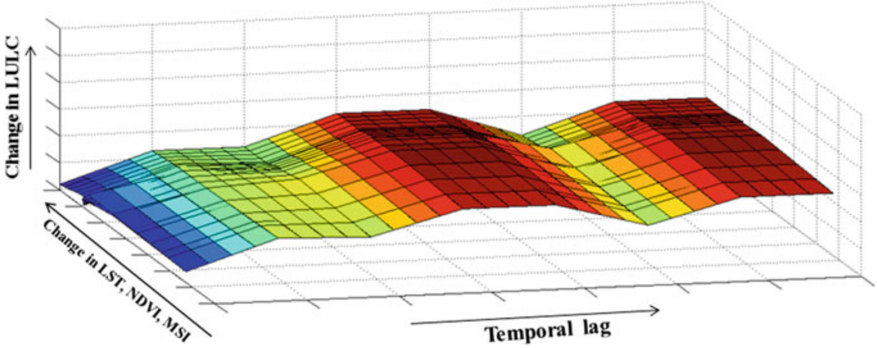


Fig. 5.9 Temporal-semantic *cross-semivariogram* (Parameter: *LST, NDVI, MSI*) [3]

(considering the interpolating points) and the *distance matrix* ($[\mathbf{D}^{ST}]_{0i[N\times 1]}$) (considering the interpolating and the prediction points) are undated with the semantic measures, which results in $[\mathbf{C}]^{ST-RevSemK}$ and $[\mathbf{D}]^{ST-RevSemK}$ respectively for the prediction of *LULC*. The individual element of these two matrices, $C_{ij}^{ST-RevSemK}$ in $[\mathbf{C}]^{ST-RevSemKK}$ and $D_{0i}^{ST-RevSemK}$ in $[\mathbf{D}]^{ST-RevSemK}$ are modified as: $C_{ij}^{ST-RevSemK} = \Delta f_{ij}$ and $D_{0i}^{ST-RevSemK} = \Delta f_{0i}$, where Δf_{pq} denotes the amount of change in *LULC*, calculated from the *semivariograms*. Hence, The weight vector produced by *ST-RevSemK* framework is given as follows:

$$\mathbf{W}^{ST-RevSemK} = \begin{bmatrix} [\mathbf{C}^{ST}]_{ij} \\ -\cdot-\cdot- \\ \left([\mathbf{SI}^{ST}]_{ij} \circ [\mathbf{SS}]_{ij}\right) \end{bmatrix}^{-1} \left[\begin{bmatrix} [\mathbf{D}^{ST}]_{0i} \\ -\cdot-\cdot- \\ \left([\mathbf{SI}^{ST}]_{0i} \circ [\mathbf{SS}]_{0i}\right) \end{bmatrix} - \lambda^{ST-RevSemK} \mathbf{1} \right] \tag{5.10}$$

where $\lambda^{ST-RevSemK}$ is the spatio-temporal *Lagrange multiplier* of *ST-RevSemK* for the prediction of *LULC*.

5.4 Empirical Proof for ST-RevSemK

For the empirical validation of the *ST-RevSemK* approach, an experiment is performed with meteorological data in two *RoI*s Kolkata, WB, India and Dallas, TX, USA. For the case study, the same zones, as considered in Figs. 3.3 and 3.4 in Chap. 3, are chosen here. The *NDVI*, *LST* and *MSI* time-series data are derived from the raw satellite imagery. The supervised classification is performed with the satellite imagery to get the *LULC* information of the *RoI*s. Six *LULC* classes are considered to perform this classification, which are given as follows: *settlements*, *river/deep waterbodies*, *waterlogged areas*, *moist land*, *agriculture* and *unclassified*. Different signature sets have been used for different time instances. In this analysis, past eleven years data of meteorological and terrestrial parameters, for the duration 2005–2015, are used. Here, past ten years data (duration 2005–2014) are considered to plot the *semivariogram*s, to forecast the *LULC* distribution in 2015. For the accuracy analysis of *ST-RevSemK*, the data (*LULC*) of the year 2015 is assumed to be missing.

The following constraints can be stated as the experimental specifications of this study. A one kilometer (1 km) radius is considered against each prediction point for the selection of interpolating points. In this study, twenty interpolating points has been randomly selected from different past and present time instances within this radius for each prediction point. The experimental semantic *semivariogram*s are modeled by taking approximately 500 sample points within lag distance $h = 5$ km and temporal lag $t = 10$ years with 1 year interval. The result is specified for five selected zones from Kolkata, WB, India and Dallas, TX, USA. Four types of external drifts have been reported, such as: three individual secondary parameters *LST*, *NDVI* and *MSI* and all of them altogether. Here, the error in prediction/forecasting cannot be measured by the error metrics as specified in earlier chapters. The meteorological parameters are continuous variable. However, the prediction of *LULC* class of a pixel is either correct or incorrect. Hence, for each of the zones, binary error metric is evaluated such that *%Error* = (*total number of incorrectly predicted pixels/total number of pixels*) of that zone. The %Error of each of the zones of both the spatial regions are depicted in Fig. 5.10 and Fig. 5.11 respectively. It is observed from the results that *ST-RevSemK* approach performs better when the external drift of multiple correlated meteorological parameters are considered, compared to the drift with single parameter.

5.4.1 Discussions on Empirical Proof

The mapping imagery of forecasted *LULC* distribution of ten spatial zones of Kolkata, WB, India and Dallas, TX, USA are presented in Tables 5.3 and 5.4. In these tables, the actual *LULC* imagery are shown for different zones and the corresponding predicted imagery are also reported. The predicted imagery are constructed by *ST-RevSemK* approach, using four type of drifts such as the bivariate drift of *LST*, *NDVI*, *MSI* and

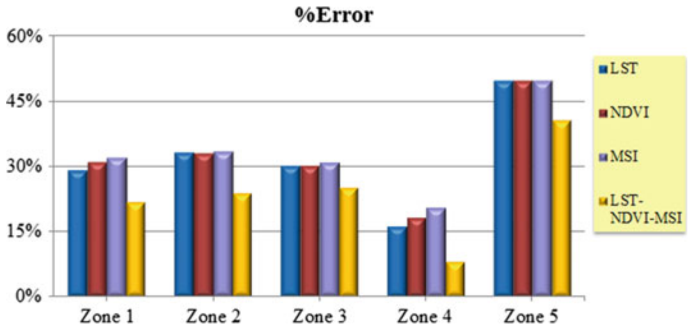


Fig. 5.10 Comparison study with error graph for *ST-RevSemK* (Region: Kolkata, WB, India)

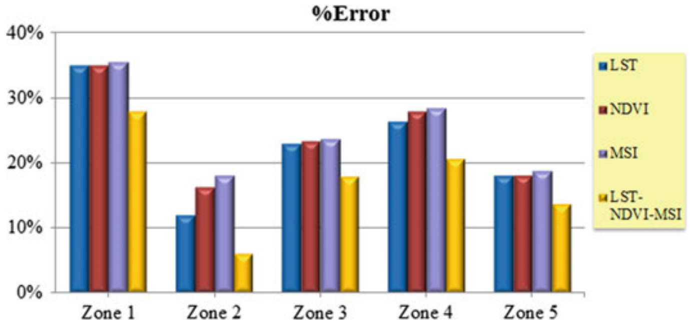


Fig. 5.11 Comparison study with error graph for *ST-RevSemK* (Region: Dallas, TX, USA)

the multivariate drift considering all the parameters altogether. The same symbology, as mentioned in Chap. 1, has been considered for all the imagery. For better pictorial representations, each of the imagery is presented with *stretched* symbology. The error surfaces are also reported in binary scale, where the black pixel represents wrongly predicted pixels and white pixels are the ones which are correctly predicted.

The Figs. 5.10, 5.11 and Tables 5.3, 5.4 advocate that the *ST-RevSemK* considering more meteorological parameters as secondary information, generates better mapping images and reports lesser amount of error compared to single parameter drift (≈6–13% higher for Kolkata, WB, India and ≈4–12% higher for Dallas, TX, USA). This actually matches with our *F-test* results for both the regions, where evaluated *F statistic* declare that the drift of all the secondary parameters together can predict the future *LULC* distribution most accurately. Among the single parameters drifts, it is also observed that the performance of individual parameter is almost similar.

Table 5.3 Comparison study for *ST-RevSemK* (Region: Kolkata, WB, India)

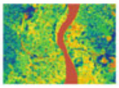
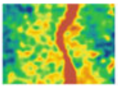
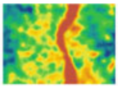
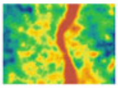
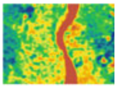
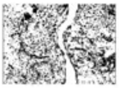
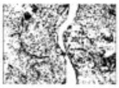
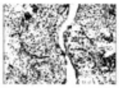
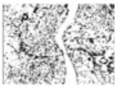
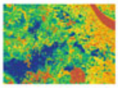
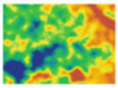
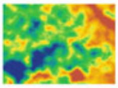
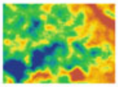
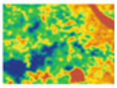
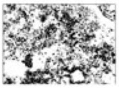
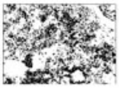
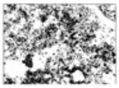
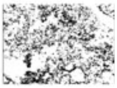
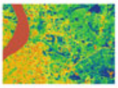
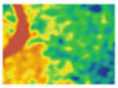
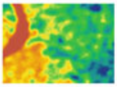
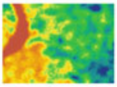
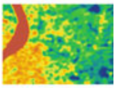
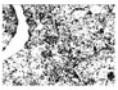
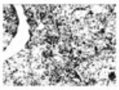
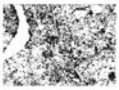
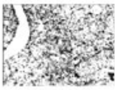
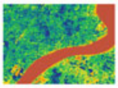
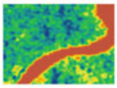
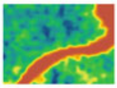
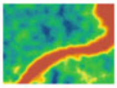
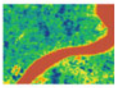
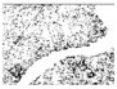
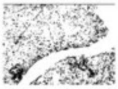
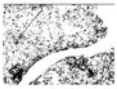
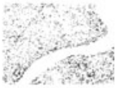
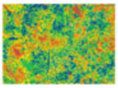
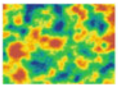
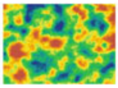
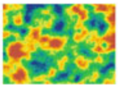
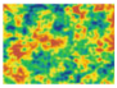
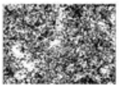
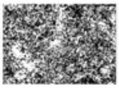
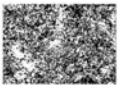
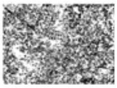
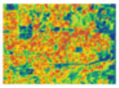
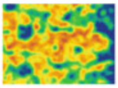
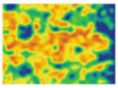
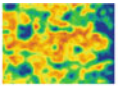
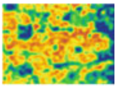
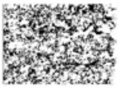
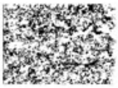
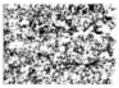
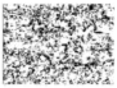
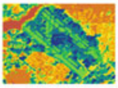
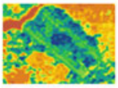
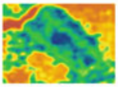
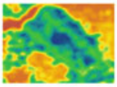
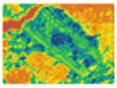
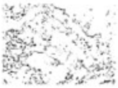
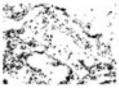
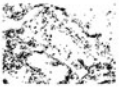
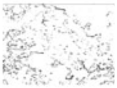
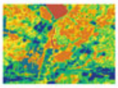
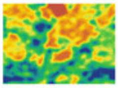
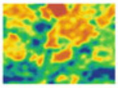
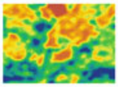
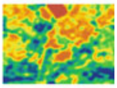
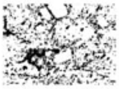
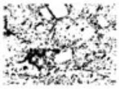
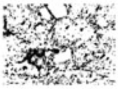
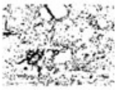
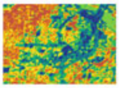
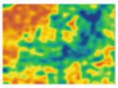
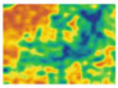
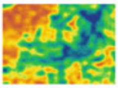
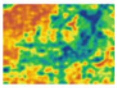
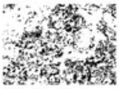
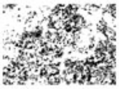
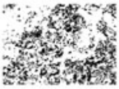
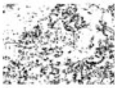
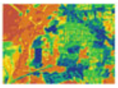
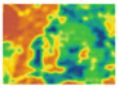
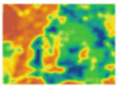
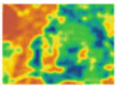
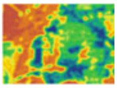
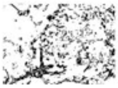
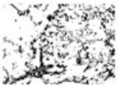
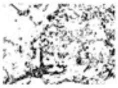
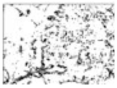
Zone	Actual image	Predicted image			
		using LST	using NDVI	using MSI	using LST, NDVI, MSI
Zone 1	BB: [(88°21′56.75″E 22°53′3.338″N); (88°26′11.684″E 22°56′20.576″N)]				
					
	Error surfaces (In binary scale) ■ Error □ No error				
		%Error=29%	%Error=31%	%Error=32%	%Error=22%
Zone 2	BB: [(88°16′37.345″E 22°43′16.086″N); (88°20′51.518″E 22°46′33.581″N)]				
					
	Error surfaces (In binary scale) ■ Error □ No error				
		%Error=33%	%Error=33%	%Error=33%	%Error=24%
Zone 3	BB: [(88°21′3.362″E 22°34′43.896″N); (88°25′17.437″E 22°38′1.121″N)]				
					
	Error surfaces (In binary scale) ■ Error □ No error				
		%Error=30%	%Error=30%	%Error=31%	%Error=25%
Zone 4	BB: [(88°10′24.797″E 22°29′45.122″N); (88°14′38.046″E 22°33′3.055″N)]				
					
	Error surfaces (In binary scale) ■ Error □ No error				
		%Error=16%	%Error=18%	%Error=21%	%Error=8%
Zone 5	BB: [(88°24′21.572″E 22°24′41.309″N); (88°28′35.399″E 22°27′58.463″N)]				
					
	Error surfaces (In binary scale) ■ Error □ No error				
		%Error=50%	%Error=50%	%Error=50%	%Error=41%

Table 5.4 Comparison study for *ST-RevSemK* (Region: Dallas, TX, USA)

Zone	Actual image	Predicted image			
		using LST	using NDVI	using MSI	using LST, NDVI, MSI
Zone 1	BB: [(96°48′52.626″W 32°53′19.252″N); (96°46′4.031″W 32°55′15.003″N)]				
					
	Error surfaces (In binary scale) ■ □ Error No error				
		%Error=35%	%Error=35%	%Error=36%	%Error=28%
Zone 2	BB: [(96°52′26.827″W 32°49′41.121″N); (96°49′38.61″W 32°51′36.489″N)]				
					
	Error surfaces (In binary scale) ■ □ Error No error				
		%Error=12%	%Error=16%	%Error=18%	%Error=6%
Zone 3	BB: [(96°44′48.861″W 32°47′22.359″N); (96°42′0.958″W 32°49′17.557″N)]				
					
	Error surfaces (In binary scale) ■ □ Error No error				
		%Error=23%	%Error=23%	%Error=24%	%Error=18%
Zone 4	BB: [(96°50′50.616″W 32°44′1.516″N); (96°48′2.552″W 32°45′56.856″N)]				
					
	Error surfaces (In binary scale) ■ □ Error No error				
		%Error=26%	%Error=28%	%Error=28%	%Error=21%
Zone 5	BB: [(96°44′27.666″W 32°42′20.309″N); (96°41′39.631″W 32°44′15.74″N)]				
					
	Error surfaces (In binary scale) ■ □ Error No error				
		%Error=18%	%Error=18%	%Error=19%	%Error=14%

5.5 Further Discussions

Urban and city planning have attracted significant research interest in the area of climatological and geospatial analysis for last few decades. It primarily focuses on the prediction and forecasting of distribution pattern of *LULC* in the terrain. Hence, estimating the future distribution of *LULC* with high degree of accuracy is a major research challenge. This work attempts to forecast the *LULC* distribution pattern in future considering different correlated meteorological parameters. The proposed framework (*ST-RevSemK*) learns and models the historic pattern of three meteorological parameters, *LST*, *NDVI*, and *MSI*. By evaluating the semantic dependencies among the *LULC* and different correlated parameters, the future trend of *LULC* is predicted. The *ST-RevSemK* follows the notion of spatio-temporal *semantic kriging* and is a multivariate extension of it. The empirical experimentation with meteorological data proves the efficacy of the proposed *ST-RevSemK* method and advocates the fact that the modeling of more auxiliary meteorological parameter enhance the estimation accuracy of future *LULC* pattern.

A major challenge for a multivariate analysis is to efficiently select the correlated parameters from the pool of available ones, which may enhance the accuracy of prediction. The evaluation of causal dependencies among the spatio-temporal time-series data is important for any spatial analysis. Similarly, in *ST-RevSemK*, the forecasting of the terrestrial parameter *LULC* is modeled by inferring the causal dependencies between *LULC* with the influential meteorological parameters. This work has considered a preprocessing framework of *CTF* to check and identify the secondary parameters that influences the primary parameter significantly and rank them according to their influence quotient. To evaluate the influence of the auxiliary parameters in *ST-RevSemK* approach, a hierarchical *GC* test is performed on the secondary parameters individually or in a group. It is observed from the results that the group considering all the secondary parameters together is the most influential one for the forecasting of *LULC*, which is further proved by the empirical analysis as well. The contributions of *ST-RevSemK* framework is stated as follows:

- utilizing the *SemK* interpolation method and its variants for an application, i.e., the spatio-temporal prediction/forecasting of *LULC* distribution of the terrain for urban landscape modeling.
- extending the separable spatio-temporal *SemK* approach in multivariate scenario for *LULC* prediction.
- finding the causal linkages among the groups of meteorological and terrestrial parameters to choose the most influential one to achieve enhanced accuracy.
- experimentation with meteorological and terrestrial data proves that proper identification and utilization of meteorological data can enhance the prediction accuracy of future *LULC* pattern.
- this application has a direct implication urban and city planning, prediction of urban heat island, etc.

Though the separable spatio-temporal *SemK* approach is considered for extending it to *ST-RevSemK* approach, however, the non-separable one also could have been used for the same. In this chapter, the basic concept of *SemK* and spatio-temporal *SemK* have been considered for an application, i.e., the change pattern analysis of *LULC* distribution. Being this framework a spatio-temporal multivariate interpolation technique, any other forecasting application also could have been tested. For the considered application, the meteorological and terrestrial factors should be influential to that application and together must influence the prediction parameter. Though the basic structure of the spatial and spatio-temporal base methods will remain same, some other modifications in the *ST-RevSemK* approach are obvious for other applications.

References

1. Attanasio A, Pasini A, Triacca U (2013) Granger causality analyses for climatic attribution. Atmos Clim Sci 3:515
2. Bhattacharjee S, Ghosh SK (2015) Exploring spatial dependency of meteorological attributes for multivariate analysis: a granger causality test approach. In: 2015 eighth international conference on advances in pattern recognition (ICAPR). IEEE, pp 1–6
3. Bhattacharjee S, Ghosh SK (2015b) Spatio-temporal change modeling of LULC: a semantic kriging approach. ISPRS Ann Photogramm Remote Sens Spat Inf Sci 1:177–184
4. Bhattacharjee S, Ghosh SK (2015) Time-series augmentation of semantic kriging for the prediction of meteorological parameters. In: 2015 IEEE international geoscience and remote sensing symposium (IGARSS 2015), pp 4562–4565
5. Bhattacharjee S, Mitra P, Ghosh SK (2014) Spatial interpolation to predict missing attributes in GIS using semantic kriging. IEEE Trans Geosci Remote Sens 52(8):4771–4780
6. Cressie N, Wikle CK (1998) The variance-based cross-variogram: you can add apples and oranges. Math Geol 30(7):789–799
7. Deardorff J (1978) Efficient prediction of ground surface temperature and moisture, with inclusion of a layer of vegetation. J Geophys Res Ocean (1978–2012) 83(C4):1889–1903
8. Dutta B, Ray A, Pal S, Patranabis DC (2009) A connectionist model for rainfall prediction. Neural Parallel Sci Comp 17(1):47–58
9. Findell KL, Eltahir EA (1997) An analysis of the soil moisture-rainfall feedback, based on direct observations from illinois. Water Resour Res 33(4):725–735
10. Gamerman D, Moreira AR (2004) Multivariate spatial regression models. J Multivar Anal 91(2):262–281
11. Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econ J Econ Soc, 424–438
12. Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics, pp 423–447
13. Hengl T, Heuvelink GB, Tadić MP, Pebesma EJ (2012) Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. Theor Appl Climatol 107(1–2):265–277
14. Humme A, Lindenbergh R, Sueur C (2006) Revealing celtic fields from LIDAR data using kriging based filtering. In: Proceedings of the ISPRS commission V symposium
15. Kodra E, Chatterjee S, Ganguly AR (2011) Exploring Granger causality between global average observed time series of carbon dioxide and temperature. Theor Appl Climatol 104(3–4):325–335
16. Lozano AC, Li H, Niculescu-Mizil A, Liu Y, Perlich C, Hosking J, Abe N (2009) Spatial-temporal causal modeling for climate change attribution. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 587–596

17. Rogan JC, Keselman H (1977) Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: an investigation via a coefficient of variation. Am Educ Res J 14(4):493–498
18. Salvucci GD, Saleem JA, Kaufmann R (2002) Investigating soil moisture feedbacks on precipitation with tests of Granger causality. Adv Water Resour 25(8):1305–1312
19. Sfetsos A, Vlachogiannis D (2010) A new approach to discovering the causal relationship between meteorological patterns and PM_{10} exceedances. Atmos Res 98(2):500–511
20. Shukla J (1998) Predictability in the midst of chaos: a scientific basis for climate forecasting. Science 282(5389):728–731
21. Smirnov DA, Mokhov II (2009) From granger causality to long-term causality: application to climatic data. Phys Rev E 80
22. Triacca U (2005) Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature? Theor Appl Climatol 81(3–4):133–135
23. Vauclin M, Vieira S, Vachaud G, Nielsen D (1983) The use of cokriging with limited field soil observations. Soil Sci Soc Am J 47(2):175–184
24. Yang D, Gu C, Dong Z, Jirutitijaroen P, Chen N, Walsh WM (2013) Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging. Renew Energy 60:235–245

Chapter 6
Summary and Future Research

Abstract This monograph focuses on the prediction and forecasting of the meteorological parameters that are related to the earth surface. These parameters are mainly derived from the raster satellite imagery, and generally contain missing and erroneous pixels, line gaps, and cloud covers. These issues are considered as the major hindrances to generate complete raster surface for these parameters. In this situation, the spatial interpolation methods are reported to be the most efficient choice in many literature. This monograph attempts to incorporate the LULC-based contextual knowledge of the terrain for the interpolation process of the meteorological parameters.

This monograph focuses on the prediction and forecasting of the meteorological parameters that are related to the earth surface. These parameters are mainly derived from the raster satellite imagery, and generally contains missing and erroneous pixels, line gaps, cloud covers. These issues are considered as the major hindrances to generate complete raster surface for these parameters. In this situation, the spatial interpolation methods are reported to be the most efficient choice in many literature. The *kriging*-based geostatistical interpolation methods can handle the spatial properties of the terrain most efficiently. Spatial autocorrelation is one such property, which can be defined as the dependency that exists among the sampled locations with respect to a parameter. The *kriging*-based interpolation methods are the most effective choice to model complete autocorrelation model for the prediction of spatial parameters.

It is also observed that, these geostatistical interpolation methods model spatial autocorrelation in terms of *Euclidean* distances between sampled locations, by following Tobler's law of spatial proximity. However, for the meteorological parameters, that are nearby to the earth surface (e.g., *LST*, *NDVI*, *MSI*, etc.), the domain knowledge of the terrain (e.g., land-use/land-cover *LULC* distribution) plays a crucial role for land–atmospheric interaction modeling, hence for the parameters as well. As existing interpolation methods fail to integrate this knowledge into the

S. Bhattacharjee et al., *Semantic Kriging for Spatio-temporal Prediction*, Studies in Computational Intelligence 839,
https://doi.org/10.1007/978-981-13-8664-0_6

prediction process, it leads the methods to be unrealistic for real-life applications. This monograph attempts to address this issue for the interpolation process of the meteorological parameters. It assumes this contextual knowledge of the terrain to be modeled as the “semantic" property of the sampled locations. Hence, the main objectives of this study are to quantify the contextual *LULC* knowledge of the earth surface with some proposed metrics, and then combining the traditional *kriging* (as it is the most popular, pragmatic, efficient, and widely used technique) based interpolation process with this semantic knowledge. Hence, addressing these objectives would extend the present state of the art of spatial interpolation process (which capture numeric input and yield numeric output) to higher dimension (which can capture numeric as well as contextual input and yield numeric output). Theoretical, empirical efficacy analyses are also the major contributions of this monograph, which establishes the usefulness of the proposed scheme for the meteorological parameters' prediction.

A new spatial interpolation method, namely *semantic kriging* (*SemK*) is proposed, which extends *ordinary kriging* with the contextual *LULC* knowledge of the terrain for the prediction of *land surface temperature*. Further, some variants of the basic *SemK* method are also proposed, which can be regarded as the enhancement of *SemK* for some advanced geospatial applications such as spatio-temporal prediction, forecasting, multivariate analysis, etc. The following section presents the detailed contributions of *SemK* and its variants.

6.1 Summary

The primary goal of this monograph is to propose a more pragmatic interpolation method for the meteorological parameters. Assuming the influencing auxiliary information or covariates to be useful to enhance the prediction accuracy, it has been investigated that the *LULC* distribution of the sampled locations are significant for different parameters. For incorporating this knowledge into the prediction process, the *semantic kriging* and its variants are proposed for different categories of prediction. The primary contributions of each of the variants are specified in the following subsections.

Spatial Semantic Kriging

It is the base framework for spatial interpolation. That is, to carry out the prediction at a certain location, the sampled interpolating points are considered from the same time instance. An ontology-based semantic hierarchy organizes the semantic *LULC* knowledge for further quantification. The two proposed metrics processes this hierarchy and measures the semantic distance between sampled locations. It is further blended with traditional interpolation (*ordinary kriging*) process to achieve enhanced accuracy. Hence, the broad contributions of spatial *SemK* can be listed as follows:

- presenting a novel generic framework to quantify contextual semantic knowledge with ontology hierarchy.

- providing semantic metrics by analysis of the semantic hierarchy and extending the traditional process with these metrics.
- proposing 3D spatio-semantic *semivariogram* model (to model spatial autocorrelation) by extending traditional 2D *semivariogram* with semantic knowledge.
- presenting a *Euclidean* vector analysis based approach to theoretically compare the proposed and existing interpolation techniques.

Fuzzy Bayesian Semantic Kriging

This approach is a probabilistic variant of basic *SemK* process as *SemK*'s correlation analysis process between each pair of *LULC* classes can be improved further by considering the mutual effect of other nearby *LULC* classes. For this purpose, a fuzzy *Bayesian network* based approach is adopted which gives a graphical model to analyze the causal dependency among the *LULC* classes. By processing this directed acyclic graph, the effect of other classes on a particular *LULC* can be inferred. Then, the actual correlation analysis is carried out with the influenced values (by other *LULC*s) of the sampled locations. The overall contributions of this extension can be stated as follows:

- dependency analysis between each pair of *LULC* classes and proposing a directed acyclic graph for this dependency.
- applying fuzzy *Bayesian network* based approach on the *LULC* classes to evaluate a-posterior correlation between each pair.
- incorporating this probabilistic correlation analysis into *SemK* process to enhance the prediction accuracy of *SemK*.
- presenting empirical evaluations for the comparison with other methods and *SemK* to check whether this added computation is actually beneficial.

Spatio-Temporal Reverse Semantic Kriging

It is a multivariate extension of separable spatio-temporal *SemK* approach to apply the proposed *SemK* (or its variants) for a forecasting application. For this purpose, the forecasting of *LULC* distribution of a terrain is considered in this study, which has direct impact on urbanization, city planning, and other socioeconomic activities. The meteorological data of multiple parameters have been considered as the auxiliary information for this multivariate model. A causality testing component is proposed to investigate whether a parameter is causal to the *LULC* (primary parameter) distribution of the terrain, else pruning it for further analysis. The empirical analysis is carried out with different combinations of parameters' drift. Therefore, the major contributions of this variant of the proposed work are given as follows:

- extending the notion of univariate separable spatio-temporal *SemK* method in multivariate approach for the forecasting of *LULC* distribution of the terrain.
- proposing a causality testing module to identify the meteorological parameters that are actually causal to the *LULC* distribution of the terrain from a group of parameters.
- modeling 3D spatio-semantic and temporal-semantic *semivariogram* models for different combinations of parameters drift.

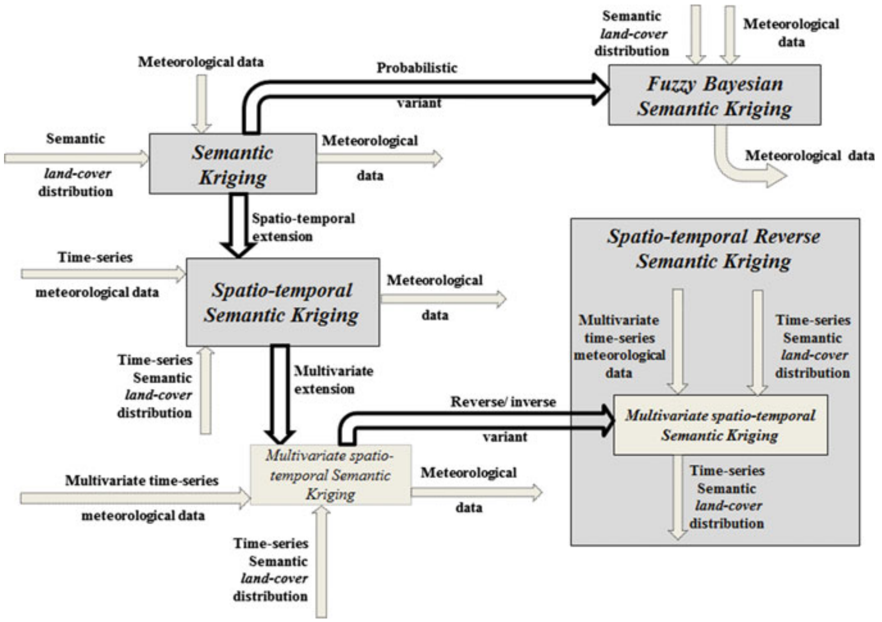


Fig. 6.1 *Semantic kriging* and its variants [1]

- identifying the best combination of the parameters (from the causality testing module and also from the empirical analysis) that are most causal to the *LULC* and forecast *LULC* distribution most accurately.

Therefore, the interrelationships among the *semantic kriging* method and its variants are depicted through Fig. 6.1 [1]. The input and output specifications of each of the components are also presented in the figure. The hallow arrow (⟹) represents the type of extension from the base component to the extended component.

6.2 Future Research

Some of the research challenges which may be taken up as the future extensions of this work are as follows:

- verification of *SemK* and its variants with other *LULC* datasets (such as NLCD: National Land Cover Data from USGS), different *LULC* ontologies, other sample techniques than uniformly random sampling (such as area weighted sampling, stratified sampling, etc.), to check its effect on the prediction accuracy.
- parallelization of *SemK* and making it scalable to be applied for large volumes of raster datasets.

- the theoretical performance evaluations have been carried out for the base method, i.e., spatial *semantic kriging*. It has explored the cases such as the conditions in which *SemK* would perform better than other existing methods, impact of the granularity of the ontology hierarchy for the prediction accuracy, ability of *SemK* to accurately capture and incorporate the semantic knowledge into the interpolation process, etc. However, the same theoretical performance evaluations can be more extensively carried out for all its variants also. Therefore, this theoretical analysis for *FB-SemK* and *ST-RevSemK* can be considered as the future work of this monograph.
- an extensive empirical analysis can be carried out to choose the fuzzy membership function that is considered for the *FB-SemK* method. For the present approach, we have relied upon domain experts' knowledge to fuzzify the meteorological parameters. However, depending on the empirical dataset, a data preprocessing can be carried out to check which is the best suitable fuzzy membership function that reports highest accuracy for *FB-SemK*- based interpolation.
- the non-separable spatio-temporal *SemK* approach can also be extended further in multivariate model to model *ST-RevSemK*.
- the proposed *SemK* framework its other variants can be deployed to predict and forecast some hazardous climatic events such as drought, urban heat island, etc. by incorporating other semantic knowledge in its land–atmospheric interaction modeling.

Reference

1. Bhattacharjee S (2016) Semantic kriging: a semantically enhanced approach for spatial interpolation. PhD thesis, Indian Institute of Technology (IIT) Kharagpur, India