UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
The School of Mathematical Sciences

Developing Statistical Methods for Spatial and Spatio-Temporal Prediction with Applications to Air Quality Data

by

LIFENG YANG

BSc, MSc

A thesis for the degree of
Doctor of Philosophy

July 2021

University of Southampton

Abstract

Faculty of Social Sciences
The School of Mathematical Sciences

Doctor of Philosophy

Developing Statistical Methods for Spatial and Spatio-Temporal Prediction with Applications to Air Quality Data

by LIFENG YANG

Prediction at an unobserved location for spatial and spatial time-series data, also known as Kriging, with complex structures is flourishing in a wide range of disciplines lately. It acts as a powerful tool capable of revealing meaningful insights by studying, seemingly isolated, spatial information on the subjects of interest. Despite the vast demand, methods for exploring spatial data collected at irregularly spaced sampling locations remain limited to mostly parametric linear techniques owning to the multi-lateral nature of space. We aim to provide semiparametric nonlinear alternatives to these applications.

The current linear spatial prediction methods for spatial data, conventionally are based on an assumption that the underlying spatial data-generating process can be decomposed into two components: a deterministic linear trend function[1] and a Gaussian stochastic process[2]. In practice, such an assumption may not be reasonable as the linear-structured spatial trend function and the Gaussian stochastic process may not be true. We hence develop new ideas in this thesis. Firstly, a nonparametric-trend universal Kriging (NTUK) method is proposed by replacing the deterministic linear component[1] with a nonparametric local linear fitting regression function, as such the solution space of the trend function is vastly enlarged. Secondly, we adopt a semiparametric model structure, i.e., the model averaging marginal regression approximation for Kriging. Through a nonparametric estimation of spatial probability density functions, an affine combination of one-dimensional conditional marginal regression functions is used for approximation in Kriging. By suggesting a K-radius averaging function to the Kriging, the stochastic process[2] part which is not assumed Gaussian is also predicted. A complete semiparametric spatial nonlinear prediction procedure is thus developed.

In spatial time-series setting, we further extend our developed methods above to the prediction of the future data at an unobserved location. We integrate the above semiparametric spatial nonlinear prediction procedure with a semiparametric spatio-temporal nonlinear regression model, which allows the spatio-temporal random field to be non-stationary over space (but stationary along time; for time series, say, through differencing) while the sampling spatial grids can be irregular. Hence the proposed model uses a two-phase framework performing firstly a spatio-temporal forecasting for a future time at the observed locations, followed by our spatial nonlinear prediction procedure stated above.

Empirical applications to air quality data are demonstrated. The performances of the proposed models are evaluated against those obtained from linear methods with significant improvement.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:..... Date:.....

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisors, Professor Zudi Lu and Professor Hou-Duo Qi, who have been supporting me throughout my PhD studies with their patience, knowledge and kindness. This thesis would not have been possible without their support and guidance. Furthermore Professor Lu's professionalism and determination, with Professor Qi's proverb of 'do one thing better than everyone else does', will go a long way with me.

Secondly, I would like to thank all the mentors who helped me during my time in the School of Mathematical Sciences. Particular thanks go to Professor Wei Liu in the Statistics team, Professor Huifu Xu, Dr Navid Izady, Professor Christine Currie and Dr Patrick Beullens in the Operational Research team who have coached me in the past years. Also I like to send my appreciation to the Faculty Graduate School Office, especially to Ms. Kulvir Bouri the Postgraduate Research Administrator for the School of Mathematical Sciences, for their extraordinary assistance. Further, I would like to thank fellow PhD students, postdoctoral researchers whom I met and worked with across the School of Mathematical Sciences, Southampton Statistical Sciences Research Institute, Southampton Business School and the School of Economic, Social and Political Sciences, for their support and some good exchange of ideas over the years.

Moreover, my appreciation goes out to my parents for their love and unreserved support. Lastly, I would like to pay special thanks to Ziqing Ho for her accompany in the past four years, I wish her very success on her PhD journey.

Definitions and Abbreviations

Y	A univariate response variable
S	The studied spatial space
s	A random location in S
$u(s)$	The spatial trend function
$X(s)$	The stochastic spatial process
d	The spatial lag between two locations
$\|d\|$	The spatial distance between two locations
u	The temporal lag between two time points
S'	The set of sampled locations (in Chapter 4)

Chapter 1

Introduction

The centre of this research is spatial prediction, more specifically, we develop semiparametric nonlinear procedures for spatial prediction with applications to non-Gaussian data sets from irregular sampling grids. There are two main parts in this thesis, Chapters 1 to 4 focus on purely spatial prediction at a single point of time while in Chapter 5 we include time dimension into consideration, that is the spatio-temporal prediction we will be dealing with.

1.1 Background

The study of spatial analysis has been increasing rapidly in the recent decades when spatial information is recorded and discussed more often than ever. Lately the question of 'how much at where' is asked more commonly than merely the 'how much' question stated by Schabenberger et al. (2005). Cressie and Wikle (2011) echoed that in order to disclose true reasons behind a problem, science should also bring the element of spatial location into the equation and address the 'where it occurs' at the same time. Briefly, spatial analysis combines variables that denote the spatial location at 'where' the particular response was observed together with information about the attributes of interest, then discovers the best possible relationships among them. The applications of spatial analysis appear in a broad spectrum of fields for example in environment, climate, socio-economic, business and health sciences. Results of these kinds contribute greatly to both large organisations, such as governments, global institutions and multinational corporations, and small groups like local communities, regional health, rescue networks, etc.

1.1.1 Spatial prediction

Spatial prediction, the main focus in this thesis, is one of the most important elements in spatial analysis. It is a process of modelling underlying spatial relationships, subsequently using the nearby sampled information (the input) to predict the value of the

target variable at a new location of interest. The methods of spatial prediction vary substantially from their origins (Schabenberger et al., 2005). Some of these methods were developed outside the mainstream statistics, such as in geology, geography and other subject-related areas. Some are rooted in traditional statistical areas, for instance the linear models and response surface theory. Others were derived from time series approaches or stochastic processes theory around which this thesis will be primarily centred. Depending on the nature of the studied problems, applications of spatial prediction commonly reflect the characterisations of their fields of study. Typical examples of these include regional flooding forecast, weather-related prediction and nature resource exploration, which have attracted vast interests from the theoretical and applied areas of this topic.

In general, spatial prediction studies the spatial dependence among the entities of interest, how can they be mathematically modelled based upon the knowledge (conditions) anticipated from the studied problems. Commonly under the assumption of spatial continuity, it is helpful to believe that nearby spaces share greater similarities in the observations than those faraway have (Cressie and Wikle, 2011), this statistical characterisetic of dependent data is oftern referred to as Tobler's first law of geography (see Miller (2004)). We consider this as a 'local effect', distance weighting methods are often used in such spatial prediction problems. Another widely established assumption is the stationary attributes, regarded as 'global effect' in comparison with the former. Assuming the underlying spatial trend is less significant or can be modelled separately, some predefined stationary conditions will offer practical solutions to these prediction tasks, thus autocorrelation, covariance, variogram and a collection of other statistical methods were introduced under this belief. These two assumptions will be studied intensively in this thesis.

1.1.2 The Kriging methods

Among methods developed in the above beliefs, Kringing has been extensively focused since this family of methods were firstly introduced as the best linear unbiased prediction (BLUP) by Matheron (1963). The name of Kriging was coined after a mining engineer D. G. Krige who carried out an empirical work evaluating the contents of mineral resources at the Witwatersrand reef complex in South Africa (Krige, 1951). Different from the deterministic interpolation methods, such as the inverse distance weighted (IDW) method and spline interpolation, Kriging methods take a stochastic processes viewpoint to the prediction problems.

The statistical idea of Kriging methods is, in brief, to interpolate across the space according to a spatial lag relationship that contains both systematic and random components (Cressie and Wikle, 2011). More specifically, it treats the studied space as a random field defined with certain statistical assumptions while the sampled values together with the optimal estimation at any locations of interest are seen as a single

realisation of an underlying spatial process together with a deterministic trend, which can often be modelled in a linear regression form under assumptions. Due to its statistical properties, Kriging methods are capable of generating a spatial prediction surface/mapping by repeating this process in the studied space. Meanwhile the variance of the estimation methods can also be quantified. Because of these advantages, Kriging methods are appealing for many practitioners. Software packages for calculating basic Kriging methods, i.e. linear Krigings, are widely available in R and other software platforms for academic and commercial purposes.

Despite the convenience, these tools often do not produce accurate solutions to a problem. There are two common mistakes noticed in real cases. On the one hand, the assumptions for a specific model are not always verified prior to the application, on the other hand, the limited collection of linear Kriging models is merely a (often oversimplified) mimic of the true underlying processes. Thus, mis-specifications are commonly observed in reality when a high level of accuracy is required.

1.1.3 Summary

Following the initial introduction, the first main task of this research is to review the current spatial Kriging methods, referring to Chapters 1-2. We compare their model structures, assumptions and limitations for the reasons above. In Chapters 3-4, we will discuss the second main task by proposing two new semiparametric nonlinear Kriging methods, sequentially reveal their theoretical properties before being applied to empirical data for comparison. In Chapter 5 we will expand our focus into a spatio-temporal setting to develop a spatial prediction procedure for a future time.

In Chapter 1, we start with a brief literature review of spatial prediction and one of its principal families of methods, Kriging. The concept of Kriging and its three linear forms are presented in Section 1.2, followed by a short introduction of some developments in nonlinear Kirging methods. In Section 1.4 we highlight the contributions to be made in this thesis and outline the structure of this thesis in the following chapters.

1.2 The Kriging method and its common linear forms

Kriging denotes a body of techniques predicting the values of a response in an identified space of interest. It was originally coined by Matheron (1963) for optimal spatial linear prediction under the minimum mean squared error criterion. Its methodology is embedded in the framework of stochastic mean squared prediction, closely related to the earlier works by Norbert Wiener and Andrey Kolmogorov on their studies in time series, see a review in Stein et al. (2006). Since then, linear Kriging and its extensions such as generalised linear Kriging have been extensively developed in literature,

see Anselin (2004), Gelfand et al. (2010) and Cressie and Wikle (2011) for some latest reviews.

The basic aim of Kriging is to predict the value of the underlying (spatial) random field $Y = Y(s)$ at an arbitrary location of interest, $s = s_0$, from the sampled observations $\{Y(s_1), Y(s_2), ..., Y(s_N)\}$, N is the sample size. For the reason of simplicity, we record $Y(s_0)$ as Y_0, where s belongs to a space S.

Cressie (1993) and Anselin (2004) stated that in spatial series analysis, it is customary to decompose the random variable $Y(s)$ into two components: i) a deterministic components $\mu(s)$, the so-called spatial trend function, presenting the large-scale variation which is often in relation to its location s, and ii) a stochastic process component $X(s)$, with $E[X(s)] = 0$, models the smooth small-scale fluctuations or the irregular part of the variation. Under a linearity assumption, the spatial trend $\mu(s)$ is often modelled as a weighted sum of known variables $f_l(s)$ for $l = 0, 1, ..., L$, which leads to the well-known linear Kriging model,

$$Y(s) = \mu(s) + X(s) = \sum_{l=0}^{L} \beta_l f_l(s) + X(s). \tag{1.1}$$

It can be further written in a matrix form as $\boldsymbol{F}^T\boldsymbol{\beta} + X(s)$, where $\boldsymbol{F} = (f_0(s), f_1(s), ..., f_L(s))^T$ denotes a $(L+1)$-vector of the explanatory variables, $\boldsymbol{\beta}$ is a $(L+1)$-vector of the unknown weights of these variables. In this thesis, we use the superscript T to represent the transpose of a vector, or more generally of a matrix.

Built upon such structure, the statistical concept of linear Kriging is centred on the assumptions of distance weighting among s, the locations. Distance-related (or lag) statistics are purposely used to reveal the spatial relations of responses in the space S, and thereafter estimate the response values at a new location. Despite the commonality, various Kriging methods are derived from Model (1.1) by imposing different assumptions. In the following sections, three widely used linear Kriging methods (simply-, ordinary- and universal Kriging) will be briefly discussed. Further details of these methods will be presented in Chapter 2, along with an empirical application of these methods to the air quality data in England.

1.2.1 Three common linear Kriging methods

Simple Kriging, as its name suggests, is the simplest form of linear Kriging. It assumes a constant spatial trend which applies to the entire space S. Furthermore, the value of the spatial trend μ and the covariance function of the underlying spatial process are assumed known that may be obtained from some prior knowledge of this problem. Hence, intuitively we would like to employ this knowledge into Model (1.1) to predict $Y(s)$ at an arbitrary location s_0, see Kerry and Oliver (2007).

In reality, it is rare that the spatial trend μ and the covariance function $C(d)$ of the underlying process $Y(s)$ are known, where $d \subset \mathbb{R}^2$ refers to the spatial lag between two locations, and $|d|$ is the L2-norm of d representing the corresponding spatial distance. Ordinary Kriging method is therefore proposed to overcome this difficulty. Schabenberger et al. (2005) described it as the mostly used Kirging method in practice.

The previous simple- and ordinary Kriging models share the same assumption that a constant spatial trend u exists in the space S. However, this condition is often violated in real applications, for instance when the studied geographic space is large. Universal Kriging, on the contrary, provides a practical solution under the assumption of the existence of a varying trend $\mu(s)$. Effectively, under linearity assumptions the $Y(s)$ in this method is decomposed into a deterministic linear function as its spatial trend, and a random zero-mean component $X(s)$ (see Cressie (1993)). Based on the definition, the universal Kriging is recognised as the most general method among these three linear Krigings. Actually it is easy to notice that the simple- and ordinary Kriging are two special cases of the universal Kriging.

1.2.2 Characteristics of linear Kriging methods

In the section above, we briefly explain three linear Kriging methods under the form of Model (1.1). The reason of constructing this particular linear model originated from the concept of best linear unbiased prediction (BLUP), which is also what the Kriging methods are primarily known for in spatial statistics (Stein, 1999).

Firstly, we define a random field $Y = Y(s)$ with $s \in S \subset \mathbb{R}^2$, where $Y(s)$ is a random variable for each $s \in S$. We observe this random variable at N sample points $s_1, ..., s_N$. Let $\boldsymbol{Y} := (Y(s_1), ..., Y(s_N)) \in \mathbb{R}^N$ denote the random vector providing the random function $Y(s)$ evaluated at the sample locations, i.e., the random variables $Y(s_i)$, $i = 1, ..., N$. The basic aim of Kriging is to predict the value of the random variable at a new location of interest $s_0 \in S$, based on the available sampled observations $Y(s_1), Y(s_2), ...,$ $Y(s_N)$.

The linear predictor (Kriging) of $Y(s_0)$, denoted as $Y^*(s_0)$, is defined as a linear combination of a constant $\lambda_0 \in \mathbb{R}$, a weight vector $\boldsymbol{\omega} := (\omega_1, ..., \omega_N)^T \in \mathbb{R}^N$, and the random variable $Y(s)$ measured at all sampled locations:

$$Y^*(s_0) := \lambda_0 + \boldsymbol{\omega}^T \boldsymbol{Y} = \lambda_0 + \sum_{i=1}^{N} \omega_i Y(s_i). \tag{1.2}$$

A linear predictor $Y^*(s_o)$ is defined as the best linear unbiased predictor (BLUP) if a) it is unbiased, i.e., $E[Y^*(s_0) - Y(s_0)] = 0$, and b) it has the minimal prediction variance among all linear unbiased predictors. The prediction variance is as follow,

$$Var(Y^*(s_0) - Y(s_0)) = \underbrace{E[(Y^*(s_0) - Y(s_0))^2]}_{\text{mse}(Y^*(s_0))} - \underbrace{(E[Y^*(s_0) - Y(s_0)])^2}_{\text{bias}=0}. \quad (1.3)$$

Therefore, minimising the prediction variance $Var(Y^*(s_0) - Y(s_0))$ of an unbiased predictor is equivalent to minimising the $mse(Y^*(s_0))$ of the predictor $Y^*(s_0)$. We use this property intensively for the three linear Kirging methods above, under the Model (1.1), to achieve the best linearly estimated value of the response at the new location $s_0 \in S$.

Despite their popularity, significant risks in using these linear models can not be overlooked. Misspecification of explanatory variables is common in real applications when their underlying relations with the response are not linear correlations. Nonlinear approaches may therefore be more appropriate under these circumstances. In the next section, we will discuss some developments in nonlinear spatial methods. Even though their broad origins, we will mainly focus on approaches from the viewpoint of spatial processes.

1.3 Nonlinear spatial Kriging and its developments

The development of above linear Krigings requires no distribution assumptions other than those in relation to the first two moments of the random field, stated by Schabenberger et al. (2005). As a result, these methods will always produce the best linear unbiased predictor regardless of the true underlying distribution of the data. Because of this advantage, when the observed data does not fit a Gaussian profile, alternative solutions often start with transforming the date into a Gaussian distribution before pursuing other methods. Log-normal and Trans-Gaussian Krigings are two commonly used techniques in this approach, suggested by Cressie (1993).

In this research, however, we will draw attention to direct nonlinear approaches to non-Gaussian problems. Comparing with the development of linear Kriging, seeking nonlinear alternatives had a late start until the last decades in the 20th century, Yakowitz and Szidarovszky (1985) and Moyeed and Papritz (2002) were two examples as such. The former compared the prediction and error estimations between a kernel nonparametric regression and parametric Kriging methods, and the latter performed an empirical comparison among a collection of linear and nonlinear Kriging methods, e.g., ordinary Kriging, indicator Kriging and disjunctive kriging. Both papers later stated that neither comparisons between methods in nonlinear and its rival linear families produced a conclusive result over performance. Our research is therefore aiming to explore this comparison further with an empirical study on geostatistical data. Among the few nonlinear methods in literature, we will begin with a short introduction of one strand of methods using the class of conditional expectation methods

exampled by disjunctive Kriging (Matheron, 1963, 1976). In the second half of this section, our review will be extended to some latest development on nonlinear approach, i.e., nonparametric estimation of probability density function for irregular spatial data, the in-fill and domain-expanding asymptotic approach in specific by Lu et al. (2007) and Lu and Tjøstheim (2014).

1.3.1 Disjunctive nonlinear spatial Kriging

Disjunctive Kriging method was introduced by Matheron (1976), who described it as an intermediate method of being more powerful than the simple linear combinations, but less complex than the conditional expectation type of approaches. It was imposed to make the most of the available information in estimating the indicator variables (Schabenberger et al., 2005).

The concepts of disjunctive coding and simple function approximation are adopted in this Kriging method. Let $\{R_k\}$ be a partition of $\Re$, i.e., $R_i \cap R_j = \emptyset$ for $i \neq j$ and $\cup_k R_k = \Re$. The indicator variables are defined as:

$$I_k(s) = \begin{cases} 1 & \text{if } Y(s) \in R_k, \\ 0 & \text{otherwise.} \end{cases}$$

If the number of intervals k is sufficiently large, the function $g(Y(s))$ can be approximated by a linear combination of these indicator functions, that is

$$g(Y(s_0)) = g_1 I_1(s_0) + g_2 I_2(s_0) + ... + g_k I_k(s_0) + \cdots. \tag{1.4}$$

With the structure as such, $I_k(s_0)$ can be estimated using indicator Kriging of the data in the k^{th} set $\{I_k(s_i), i = 1, ..., n\}$ (Watson, 1977). Furthermore, to make use of all available information, it is imposed to make another prediction of $I_k(s_0)$ using all the indicator variables sets, $\{I_1(s_i), i = 1, ..., n\}, \cdots, \{I_k(s_i), i = 1, ..., n\}, \cdots$. The new predictor of the indicator variable $\hat{I}_k(s_0)$ can be shown as a linear combination of all available indicator data,

$$\hat{I}_k(s_0) = \sum_i \sum_k \lambda_{ik} I_k(s_i), \tag{1.5}$$

where λ_{ik} is a series of constants associated with i and k. Thus, a predictor of $g(Y(s_0))$ is

$$\hat{g}(Y(s_0)) = \sum_{i=1}^{n} \sum_{k=1} g_k \lambda_{ik} I_k(s_i) := \sum_{i=1}^{n} g_i(Y(s_i)). \tag{1.6}$$

The general model of the disjunctive Kriging predictor is obtained (Schabenberger et al., 2005).

In contrast to the Kriging methods, which obtains the best linear approximation of Y_0 by a linear combination of observed values from only estimating the covariance matrix of (Y_0, Y_i), Matheron (1976) narrated that disjunctive Kriging, however, requires the knowledge of the two-dimensional laws of the pairs (Y_0, Y_i) and (Y_0, Y_j) for $i, j \in N$.

In essence, to determine the functions $g_i(Y(s_i))$, one needs to use all the indicators' information in Eq (1.6), which may include the estimation and modelling of covariances and cross-covariances of all the indicator variables, as Schabenberger et al. (2005) highlighted. For these reasons, in practice, disjunctive Kriging relies on models that expand a given function in terms of other uncorrelated functions. To make the calculation feasible, this method necessitates knowledge in hermite (orthogonal) polynomials and assumptions of Gaussian distributed sample data set. Despite modified models developed to cope with non-Gaussian data, estimating the bivariate distribution with marginals on different locations does not come easy, see Schabenberger et al. (2005), Matheron (1984) and Jean-Paul and Pierre (1999).

In summary, even with its distribution-free property, in practice, disjunctive Kriging is yet a kind of parametric method with linear Kriging applied to a continuous Gaussian data set. Lack of software support may also be a shortcoming in applying this method. Hence the implementation of disjunctive Kriging in real applications are usually limited. In this study, therefore, we opt not to explore this method further, instead we treat it as a motivation to our research in density-based nonlinear Kriging methods.

1.3.2 Nonparametric estimation of probability density function for spatial data

Another nonlinear approach of predicting methods emerges recently, aiming to apply nonparametric approaches directly to spatial prediction problems. By doing so it will systematically avoid the drawbacks of misspecification from linear Kriging methods. Originated from time series, nonparametric estimation methods are well established and extensively used in this one-dimensional setting, see Fan and Yao (2003), Terasvirta et al. (2010). However when extending these methods to multi-dimensional applications such as in spatial series, their limitations become significant.

There are several reasons for this. Lu and Tjøstheim (2014) indicated that the most important one has to be the fact that the sampling points are often irregularly positioned in space. In time series, observed data are sampled or can be aggregated at regular time intervals. Yet in spatial series, this may not be the case where the monitoring sites are rarely located from an ideal regular grid due to constraints in nature.

In time series, density estimations, e.g., marginal or joint density estimation, are the main focused statistics in nonparametric modelling. Under certain stationary assumptions, for instance, one can estimates the joint density function $P_1(x_1, x_2)$ of consecutive observations (X_t, X_{t-1}) in a time series $\{X_t\}$. A kernel estimate can be shown as follows:

$$\hat{p}_1(x_1, x_2) = \frac{1}{n} \sum_{t=1}^{n} K_h(X_t - x_1) K_h(X_{t-1} - x_2), \tag{1.7}$$

where $K_h(x) = h^{-1}K(x/h)$ with K and h being a defined kernel function and its bandwidth respectively (Lu and Tjøstheim, 2014). To make the prediction feasible, ideally infinite pairs of (X_t, X_{t-1}) are required in time series. In the case of spatial series, where the monitoring sites are irregularly located, the estimation of joint density function between any spatial lag becomes very difficult. As such, currently the nonparametric spatial methods are centred with regular grid applications (Gao (2007); Hallin et al. (2001); Lu et al. (2007)). Some attempts of spatial prediction from irregular locations were made for other problems, for example Hall and Patil (1994) suggested nonparametric estimators of the autocovariance of a stationary random field and Matsuda and Yajima (2009) proposed a frequency domain approach for irregularly spaced data by extending the original definition of a periodogram for time series to to an irregularly spaced data set. Yet both methods still categorise a linear spatial relationship between spatial locations.

Lu and Tjøstheim (2014), however, proposed a nonparametric estimation of probability density functions for irregularly observed spatial data based on a new asymptotic framework, the so-called domain-expanding infill (DEI) asymptotics. It combines the properties of extending the domain of the measurement locations to infinity and simultaneously allowing the intensity of the observation locations in a fixed domain to increase indefinitely. As such, this method will enable the user to find the conditions under which the density estimates are consistent and asymptotically normal. The error limits and confidence intervals can be identified too. In literature, it may be the first step in this direction.

In this research, we will mainly follow this path and use the properties from this framework in the following chapters. Specifically, we will use some of its assumptions on spatial processes and kernel functions to show the asymptotic properties of a new Kriging method in Chapter 3. In Chapter 4 an nonlinear semiparametric regression model will be developed on the basis of this framework.

1.4 Outline of this thesis

In this chapter, so far we briefly reviewed some of the main methods in spatial interpolation, explicitly the linear Kriging methods and some further developments in nonlinear spatial prediction. We started with the introduction of three widely used linear models and their common characteristics. Despite the simplicity, the disadvantages of linear models are too significant. In reality, this is rarely the case that the spatial trend fits a linear combination of the selected explanatory variables. When the nonlinearity becomes obvious, the wrongly-fitted linear spatial trend may produce misleading results to the problem, thus one would always be encouraged to check the sampled data before applying such methods.

In general, when the observations does not follow a Gaussian profile, we should search for nonlinear or other nonparametric methods to model the spatial process. However, it is noticed that the development in this area is yet satisfactory. It is indeed when the number of unknown parameters is close or equal to the size of observations, especially when a large number of explanatory variables is involved, the curse of dimensionality may lead the problem to be practically unsolvable.

Another challenge lies in the spatial irregularity among sampling locations. The current nonlinear spatial prediction methods are most likely rooted from one-dimension problems such as applications in time series where the irregularity does not apply. Whereas in multiple-dimension problems, for instance spatial predictions, the sample locations are not restricted to a fixed grid/lattice. As a result, the same distance between two pairs of sample sites is hard to find even if there are a large quantity of samples, the situation can get worse when further spatial directions are added into consideration.

In this research, we will examine these restrictions and develop suitable semi- or nonparametric spatial prediction methods where appropriate to overcome these difficulties. In Section 1.4.1 we will highlight the three objectives (novelties) studied in Chapters 3 - 5 of this thesis.

1.4.1 The three objectives and main contributions

After an empirical application of linear Kriging methods to air quality data in Chapter 2, the first objective of our research is to propose an nonlinear regression method for the linearly modelled spatial trend function in the current Kriging methods. By doing so, we expect to significantly expand the solution space of the spatial trend to accommodate non-Gaussian data sets that are widely available in practice.

The second objective lies in the stochastic residual component $X(s)$. In linear Kriging methods, it is yet modelled as a linear combination of all sampled values. Our intention is to utilise a semiparametric one-dimensional nonlinear approximation model by incorporating nonparametric probability density function estimation techniques. This

method is proposed to ease the pressure from the curse of dimensionality and capable of estimating the spatial density functions from irregularly spaced sampling locations.

The third objective of this research is to integrate our semiparametric spatial nonlinear prediction method into spatio-temporal setting. Instead of an all-in-one spatio-temporal model that is often associated with simplified assumptions, we would like to propose a two-phase procedure performing separated spatio-temporal forecasting for a future time and spatial prediction at the future time, respectively.

In response to these objectives, we summarise the main contributions made in this thesis below.

For the first object, a nonparametric-trend universal Kriging (NTUK) is therefore proposed in Chapter 3, to replace the parametric linear trend function by a nonparamatric local linear fitting function estimated at each spatial location. We show that the predictor from the samples converges in probability to its equivalent from the population. Under this approach, the solution space for the deterministic spatial trend is vastly enlarged. In the fitting function, Kernel smoother is used to highlight the local information.

In Chapter 4, we adopt a model averaging marginal regression approximation method originated for time series, which employs an affine combination of one-dimensional conditional regression functions for approximation in Kriging. Under the domain-expanding infill (DEI) asymptotics framework, nonparametric estimation of spatial probability density functions at irregular locations are used to estimate these regression functions mentioned above. A k-radius averaging function is later introduced to the Kriging. By now, as the second contribution, the non-Gaussian stochastic process is predicted. Combing the first two contributions, a complete semiparametric spatial nonlinear Kriging is developed.

The third contribution is introduced in Chapter 5, the goal is to expand the use of the complete semiparametric spatial nonlinear Kriging stated above into spatial time-series prediction, i.e., to predict future observations at unobserved spatial locations. We integrate our complete spatial Kriging with a semiparametric spatio-temporal autoregressive partially nonlinear regression (STAR-PLR) model, which allows the spatio-temporal random field to be non-stationary over space (but stationary along time; for time series, say, through differencing) while the sampling spatial grids can be irregular. Hence, we propose a two-phase framework performing a spatio-temporal forecasting for a future time, then, as the second phase, a spatial nonlinear prediction procedure at the future time. Under such arrangement, complex assumptions can be made at each phase to cope with the diverse nature in real cases.

Empirical applications to air quality data are demonstrated. The proposed models outperform the linear methods with significant improvement.

1.4.2 Structure of this thesis

This thesis is divided into six chapters. In Chapter 1, we give a brief literature review on spatial prediction and introduce one of its main families of methods, Kriging. The concept of Kriging, its three linear forms and further developments in nonlinear Kirging are discussed, which lead to the contribution list of this study mentioned above. To begin this journey, Chapter 2 shows an empirical application of the linear spatial Kriging to the air quality data set in England, which becomes the baseline of comparison in our research.

It is then followed by Chapter 3, in which a new nonparametric-trend universal Kriging is introduced. We reveal the asymptotic properties of this method and its prediction outcome is then compared with the baseline result. In Chapter 4 we develop a semiparametric procedure of model averaging marginal Kriging for the de-trended process $X(s)$, by now a full semiparametric spatial interpolation method is proposed. In Chapter 5 our focus will be expanded to the realm of spatial-temporal prediction at a future time.

Finally, in Chapter 6, we will summarise the contributions that have been highlighted in the previous chapters, and outline the outlooks of our research where further developments can be made.

Chapter 2

Empirical Application of Linear Krigings to Air Quality Data

The purpose of this chapter is twofold: present the case study that will be used throughout this research, i.e., the measurement of air quality in England; Sections 2.2 and 2.3 introduce the details of the three linear Kriging methods, and apply them directly to the air quality data to be the baseline for our comparison.

2.1 Air quality problem

In spatial analysis, the task of estimating response values at unobserved locations plays an important role in many scientific disciplines. In many situations, it is impossible to measure the interested entity at any arbitrary location due to either practical reasons or physical constraints. One common way of solving this problem is to utilise known observations from the surrounding areas to predict the unknown by applying suitable statistical methods.

Among numerous applications, in this thesis we select the spatial prediction of air quality data as our case study for both its conventional and practical reasons. The problem of poor air quality is a global threat to a sustainable development for future. The UN (2016) [1] highlighted that air pollution is having serious adverse impacts on the quality of life, in particular on human health, environment and economy. There are some 6.5 million people dying annually from air pollution and 92 percent of the world's population living in places where air pollution level exceeds the recommended limit.

The UK's Department for Environment, Food & Rural Affairs (Defra) echoed that a cleaner, healthier environment benefits local people and the country's economy [2]. Clean

[1] https://www.theguardian.com/environment/2016/may/12/air-pollution-rising-at-an-alarming-rate-in-worlds-cities

[2] https://uk-air.defra.gov.uk/air-pollution/

air is vital for people's health and the environment, whilst polluted air can cause both short-term and long-term negative effects on human health. Currently, the Met Office (2017) [3] in the UK operates a 10-level (four bands) daily air quality index (DAQI) system to characterise the level of air pollution at a local level. The Index 1 stands for the least polluted air quality while the Index 10 indicates that the air quality reaches the fourth band the 'Very High' polluted air as shown in Figure 2.1. Health advice and recommended actions are published in corresponding to each air quality band. Therefore, to ensure correct countermeasures are in place, accurate prediction of air quality is critical to local inhabitants.

Bandings for the Daily Air Quality Index

Band	Index	Ozone	Nitrogen Dioxide	Sulphur Dioxide	$PM_{2.5}$ Particles	PM_{10} Particles
		Running 8 hourly mean	Hourly mean	15 minute mean	24 hour mean	24 hour mean
		µg m^{-3}	µg m^{-3}	µg m^{-3}	µg m^{-3}	µg m^{-3}
Low	1	0-33	0-67	0-88	0-11	0-16
	2	34-66	68-134	89-177	12-23	17-33
	3	67-100	135-200	178-266	24-35	34-50
Moderate	4	101-120	201-267	267-354	36-41	51-58
	5	121-140	268-334	355-443	42-47	59-66
	6	141-160	335-400	444-532	48-53	67-75
High	7	161-187	401-467	533-710	54-58	76-83
	8	188-213	468-534	711-887	59-64	84-91
	9	214-240	535-600	888-1064	65-70	92-100
Very High	10	241 or more	601 or more	1065 or more	71 or more	101 or more

FIGURE 2.1: The daily air quality index (DAQI) with the measured pollutants.

The UK's regional daily air quality index (DAQI) is forecasted through a collection of comprehensive systems developed by the NCAS, GMR and the Defra. Without going through the detailed methodology, we understand that Ozone (ug/m^3) , Nitrogen Dioxide (ug/m^3), Sulphur Dioxide (ug/m^3), PM 2.5 particles (ug/m^3) and PM 10 particles (ug/m^3) are the five chemical pollutants locally measured for computing the DAQI. Figure 2.1 also shows the relations between the DAQI level and the concentration for each pollutant published by Met Office (2007) [4].

[3] https://uk-air.defra.gov.uk/air-pollution/daqi
[4] http://www.metoffice.gov.uk/guide/weather/air-quality#Air-quality-index

Currently, the Defra records air quality data from multiple automated monitoring networks in the UK, some of them take local readings every hour. In addition to these networks, there are also over one hundred non-automated monitoring stations in operation, sampling air quality data on a daily, weekly and monthly basis[5]. Despite the large number, these monitoring stations are mainly located in cities, major towns and by the main highways as shown in Figure 2.3. A large percentage of local areas is actually far away from their nearest station. For this reason, advanced spatial prediction methods are critical for producing accurate air quality forecast in these areas.

2.2 Three linear Kriging methods

Before applying the linear Kriging to empirical data in Section 2.3, it is important to expand our knowledge of the three linear Kriging methods briefly mentioned in Section 1.2. For each method, in the following sections, we will show its detailed assumptions, unbiasness conditions and the Kirging predictor together with its prediction variance.

2.2.1 Simple Kriging

Simple Kriging (SK), as its name suggests, is the simplest form of linear Kriging having a constant known mean μ by assumption for an underlying random process $Y(s)$, where the μ is valid for the entire space S. This spatial trend μ may be obtained from the existing knowledge of this problem, intuitively we would like to integrate this knowledge into Model (1.1) to predict the process $Y(s)$ at a new location s_0 (Kerry and Oliver, 2007).

To begin the introduction of simple Kriging, Cressie and Wikle (2011) listed its two assumptions as follows:

1. The $\mu \in \mathbb{R}$, mean of $Y = Y(s)$ for $s \in S$, is a known constant, i.e., $E\left[Y(s)\right] = \mu$, $\forall\, s \in S$,

2. $Y = Y(s)$ is assumed to be secondary-order stationary with a known covariance function $C\,(d) := Cov\,(Y(s), Y(s+d)) = E[Y(s)Y(s+d)] - \mu^2$, $\forall\, s, s+d \in S$.

The simple Kriging predictor $Y_{\omega}^{SK*}(s_0)$ of $Y(s)$ at the prediction point s_0 is defined as the sum of the spatial mean μ and the weighted differences of the random function $Y(s)$ evaluated at each sample point s_i and the mean μ, i.e.,

$$Y_{\omega}^{SK*}(s_0) := \mu + \sum_{i=1}^{N} \omega_i (Y(s_i) - \mu) = \mu + \boldsymbol{\omega}^T (\boldsymbol{Y} - \mu \mathbb{1}), \tag{2.1}$$

[5]https://uk-air.defra.gov.uk/air-pollution/

where $\omega_i \in \mathbb{R}$ being the weight of the corresponding residual $Y(s_i) - \mu$ and $\boldsymbol{\omega} := (\omega_1, \omega_2, ..., \omega_N)^T \in \mathbb{R}^N$, the vector containing all the weights from the N observations. We denote $\mathbb{1}$ as the identity vector of order N, i.e., $\mathbb{1} := (1, ..., 1)^T$.

We calculate the mean prediction error for simple Kriging and get

$$E[Y_{\boldsymbol{\omega}}^{SK*}(s_0) - Y(s_0)] = \mu + \sum_{i=1}^{N} \omega_i E[Y(s_i) - \mu] - E[Y(s_0)] = \mu - \mu = 0.$$

It is noticed that the predictor is unbiased so the imposing of constraints is not required in the simple Kriging model, see Wackernagel (2003). It is understood that the construction of the predictor of simple Kriging itself spontaneously guarantees this unbiasedness (Lichtenstern, 2013).

Furthermore, the variance of the prediction error can be measured using its mean squared prediction error, i.e., $E[(Y_{\boldsymbol{\omega}}^{SK*}(s_0) - Y(s_0))^2]$. The calculation of the prediction variance $\sigma_E^2(s_0)$ for simple Kriging follows Wackernagel (2003):

$$\begin{aligned} \sigma_E^2(s_0) :&= \mathrm{Var}(Y_{\boldsymbol{\omega}}^{SK*}(s_0) - Y(s_0)) = E[(Y_{\boldsymbol{\omega}}^{SK*}(s_0) - Y(s_0))^2] \\ &= C(\mathbf{0}) + \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i \omega_j C(s_i - s_j) - 2\sum_{i=1}^{N} \omega_i C(s_i - s_0) \\ &= C(\mathbf{0}) + \boldsymbol{\omega}^T \boldsymbol{\Sigma} \boldsymbol{\omega} - 2\boldsymbol{\omega}^T \boldsymbol{c}_0 \geq 0, \end{aligned} \tag{2.2}$$

where $C(\mathbf{0}) = Cov(Y(s_0), Y(s_0))$, $\boldsymbol{\Sigma}$ is a $N \times N$ symmetric covariance matrix for any two locations $s_i, s_j \in S$, and $\boldsymbol{c}_0$ is a N-vector with $Cov(Y(s_i), Y(s_0))$ as its i^{th} element.

By taking the derivative of the prediction variance with respect to $\boldsymbol{\omega}$, the condition for a minimal prediction variance σ_E^2 is $\boldsymbol{\Sigma}\boldsymbol{\omega}_{SK} = \boldsymbol{c}_0$, where $\boldsymbol{\omega}_{SK} := (\omega_1^{SK}, ..., \omega_N^{SK})^T \in \mathbb{R}^N$ denotes the vector providing the simple Kriging weights (Wackernagel, 2003). The Hessian matrix satisfies the condition of a positive second-order derivative of σ_E^2 with respect to $\boldsymbol{\omega}$. With this result, the prediction variance σ_{SK}^2 and the estimated simple Kriging predictor of $Y_{\omega}^{SK*}(s)$ at the location s_0 shown by Cressie (1993) are as follows:

$$\begin{aligned} \sigma_{SK}^2(s_0) &= C(\mathbf{0}) - \boldsymbol{c}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{c}_0 = C(\mathbf{0}) - \sum_{i=1}^{N} \omega_i^{SK} C(s_i - s_0), \\ y_{\boldsymbol{\omega}_{SK}}^{SK*}(s_0) &= \mu + \sum_{i=1}^{N} \omega_i^{SK}(Y(s_i) - \mu) = \mu + \boldsymbol{c}_0^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mu\mathbb{1}), \end{aligned} \tag{2.3}$$

where $\boldsymbol{y} = (y_1, ..., y_N)^T$, i.e., the sampled observations.

2.2.2 Ordinary Kriging

In reality, it is rare that the mean μ and the covariance function $C(d)$ of the underlying spatial process $Y(s)$ can be assumed as known variables. The method of ordinary Kriging (OK) is therefore developed to get across this difficulty. Schabenberger et al. (2005) described it as the most commonly used Kirging method in practice.

Cressie (1993) and Wackernagel (2003) proposed weaker assumptions for ordinary Kriging than that for simple Kriging, they are

1. The global constant mean $\mu \in \mathbb{R}$ of the random process $Y(s)$ is unknown,
2. The observations come from an intrinsically stationary random process $Y(s)$ with known semivariogram function $\gamma(d)$, i.e.,

$$\gamma(d) = \frac{1}{2} Var(Y(s+d) - Y(s)) = \frac{1}{2} E[(Y(s+d)) - Y(s))^2].$$

Wackernagel (2003) defined the predictor of ordinary Kriging $Y_\omega^{OK*}(s_0)$ of the value $Y(s)$ at the location s_0 as the linear combination of $Y(s)$ evaluated from all sample locations $s_i, i = 1, ..., N$,

$$Y_\omega^{OK*}(s_0) := \sum_{i=1}^{N} \omega_i Y(s_i) = \boldsymbol{\omega}^T \boldsymbol{Y}, \tag{2.4}$$

where $\boldsymbol{Y} = (Y_1, ..., Y_N)^T$ and $\boldsymbol{\omega} := (\omega_1, \omega_2, ..., \omega_N)^T$ provide the unknown weights $\omega_i \in \mathbb{R}$ describing the influence of each variable $Y(s_i)$ for the calculation of $Y_\omega^{OK*}(s_0)$.

It is easy to show that the sum of all weighting factors ω_i equals to 1, or $\boldsymbol{\omega}^T \mathbb{1} = 1$, which is the condition for an unbiased ordinary Kriging predictor (Wackernagel, 2003).

Under the unbiasedness condition for the ordinary Kriging, the variance of the prediction error includes the semivariogram matrix, see Cressie (1993),

$$\begin{aligned} \sigma_E^2(s_0) : &= \mathrm{Var}(Y_\omega^{OK*}(s_0) - Y(s_0)) = 2\sum_{i=1}^{N} \omega_i \gamma(s_i - s_0) - \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i \omega_j \gamma(s_i - s_j) \\ &= 2\boldsymbol{\omega}^T \boldsymbol{\gamma}_0 - \boldsymbol{\omega}^T \Gamma \boldsymbol{\omega} = \boldsymbol{\omega}^T (2\boldsymbol{\gamma}_0 - \Gamma \boldsymbol{\omega}) \geq 0, \end{aligned} \tag{2.5}$$

where the symmetric variogram matrix $\Gamma_{i,j} := \gamma(s_i - s_j)$, $i, j = 1, ..., N$, $\boldsymbol{\gamma}_0$ is a $N \times 1$ vector whose i^{th} element is the semivariogram $\gamma(s_i - s_0)$ of $Y(s)$ between the observed location s_i and the new location s_0 (Webster and Oliver, 2007).

Similarly, the minimal prediction variance for ordinary Kriging under the unbiasedness condition can be obtained by taking the first and second derivatives of Eq (2.5) with respect to $\boldsymbol{\omega}$, that is

$$\underset{\omega}{\text{Minimise}} \quad \boldsymbol{\omega}^T(2\boldsymbol{\gamma}_0 - \Gamma\boldsymbol{\omega}), \text{ subject to } \quad \boldsymbol{\omega}^T\mathbb{1} = 1. \tag{2.6}$$

Lagrange multiplier $\lambda \in \mathbb{R}$ is introduced to satisfy the unbiasedness condition (Cressie, 1993), we define a function ψ to solve this problem such that

$$\begin{aligned} \psi : \mathbb{R}^N \times \mathbb{R} &\to \mathbb{R}, \\ (\boldsymbol{\omega}, \lambda) \mapsto \psi(\boldsymbol{\omega}, \lambda) &:= 2\boldsymbol{\omega}^T\boldsymbol{\gamma}_0 - \boldsymbol{\omega}^T\Gamma\boldsymbol{\omega} - 2\lambda(\boldsymbol{\omega}^T\mathbb{1} - 1). \end{aligned} \tag{2.7}$$

By setting the derivatives of ψ with respect to the weight vector $\boldsymbol{\omega}$ and λ as zero, we can solve Eqs (2.6) and (2.7), and get the condition for the minimal prediction variance as specified below

$$\begin{aligned} \boldsymbol{\omega}_{OK}(s_0) &= \Gamma^{-1}\Big[\boldsymbol{\gamma}_0 - \mathbb{1}\big(\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\big)\Big], \\ \lambda_{OK}(s_0) &= \frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}. \end{aligned}$$

Then we have both the unbiased minimal prediction variance $\sigma^2_{OK}(s_0)$ and the ordinary Kriging predictor of the $Y^{OK*}_{\boldsymbol{\omega}}(s)$ at the location s_0 as follows

$$\begin{aligned} \sigma^2_{OK}(s_0) &= \lambda_{OK} + \sum_{i=1}^{N} \omega_i^{OK}\gamma(s_i - s_0), \\ y^{OK*}_{\boldsymbol{\omega}_{OK}}(s_0) = \sum_{i=1}^{N} \omega_i^{OK} y(s_i) &= \boldsymbol{\omega}_{OK}^T\boldsymbol{y} = \Big[\boldsymbol{\gamma}_0 - \mathbb{1}\big(\frac{\mathbb{1}^T\Gamma^{-1}\boldsymbol{\gamma}_0 - 1}{\mathbb{1}^T\Gamma^{-1}\mathbb{1}}\big)\Big]^T\Gamma^{-1}\boldsymbol{y}. \end{aligned} \tag{2.8}$$

For ordinary Kriging, the mean μ of the process $Y(s)$ is assumed as an unknown constant. However in practice, this condition may yet be true. In the next section, we will discuss the universal Kriging with more relaxed assumptions.

2.2.3 Universal Kriging

The previous simple- and ordinary Kriging methods share a common assumption that a constant mean u exists in the space S. In reality, this condition is often violated, for instance when the geographic space is large. Universal Kriging (UK), on the contrary,

provides an effective solution with the assumption of a non-constant mean. Within the linearity setting of this method, the random field can be decomposed into a linear combination of some deterministic functions (often referred as a non-stationary trend or systematic component), and a random component as shown in model (1.1), see Cressie and Wikle (2011).

The assumptions for universal Kriging are listed as follows (c.f. Cressie (1993), Wackernagel (2003))

1. Assume that $Y(s)$ can be decomposed into a deterministic trend function $\mu(s)$ and a real-valued residual random process $X(s)$, such that

$$Y(s) = \mu(s) + X(s). \tag{2.9}$$

2. The stochastic term $X(s)$ is supposed to be intrinsically stationary with a zero mean, and the known semivariogram function $\gamma_X(d)$ is called residual semivariogram function of $Y(s)$, $\forall s, s+d \in S$.

$$E[Y(s)] = E[\mu(s)] + E[X(s)] = \mu(s),$$
$$\gamma_X(d) = \frac{1}{2} Var[X(s+d) - X(d)] = \frac{1}{2} E[(X(s+d) - X(d))^2].$$

3. Let $f_0, f_1, ..., f_L$ be deterministic functions, such as of the geographical coordinates $s \in S$, with L being the number of known and selectable basic functions $f_l : S \to \mathbb{R}$, $l = 0, 1, ..., L$. It is assumed that $\mu(s)$ is a linear combination of these functions evaluated at s, $\mu(s) = \sum_{l=0}^{L} a_l f_l(s)$ with unknown coefficients $a_l \in \mathbb{R}\backslash 0$ for all $l = 0, ..., L$, with $f_0(s) = 1, \forall s$ by convention. We define F as a $N \times (L+1)$ matrix with its $(i, l+1)$ element equals to $f_l(s_i)$ for $i = 1, ..., N$ and $l = 0, 1, ..., L$.

The universal Kriging predictor $Y_{\omega}^{UK*}(s_0)$ of $Y(s_0)$ at the locations of interest s_0 is defined as follows,

$$Y_{\omega}^{UK*}(s_0) := \sum_{i=1}^{N} \omega_i Y(s_i) = \boldsymbol{\omega}^T \boldsymbol{Y}, \tag{2.10}$$

where the individual weights $\omega_i \in \mathbb{R}$, $i = 1, ..., N$ corresponding to each observation of the random function $Y(s_i)$ at the sample point s_i and $\boldsymbol{\omega} := (\omega_1, ..., \omega_N)^T$, $\boldsymbol{Y} = (Y_1, ..., Y_N)^T$.

The unbiasedness condition for the universal Kriging stated by Kitanidis (1997) is

$$\sum_{i=1}^{N} \omega_i f_l(s_i) = f_l(s_0) \quad \text{for } l = 0, ..., L \Leftrightarrow F^T \boldsymbol{\omega} = \mathbf{f}_0, \tag{2.11}$$

where $\mathbf{f}_0 := (1, f_1(s_0), ..., f_L(s_0))^T \in \mathbb{R}^{L+1}$. Matheron (1971) named it as the universality condition.

Cressie and Wikle (2011) showed the variance of the prediction error of the universal Kriging that contains the residual semivariogram function $\gamma_X(d)$,

$$\begin{aligned} \sigma_E^2(s_0) &= Var(Y_\omega^{UK*}(s_0) - Y(s_0)) = E[(Y_\omega^{UK*}(s_0) - Y(s_0))^2] \\ &= 2\sum_{i=1}^{N} \omega_i \gamma_X(s_i - s_0) - \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i \omega_j \gamma_X(s_i - s_j) = 2\boldsymbol{\omega}^T \boldsymbol{\gamma}_{X,0} - \boldsymbol{\omega}^T \Gamma_X \boldsymbol{\omega} \geq 0, \end{aligned} \tag{2.12}$$

with symmetric residual semivariogram matrix $\Gamma_X \in \mathbb{R}^{N \times N}$, $(\Gamma_X)_{i,j} := \gamma_X(s_i - s_j), i, j = 1, ..., N$ and $\boldsymbol{\gamma}_{X,0} := (\gamma_X(s_1 - s_0), ..., \gamma_X(s_N - s_0))^T \in \mathbb{R}^N$.

Similar to ordinary Kriging, Lagrange parameter vector, $\boldsymbol{\lambda} := (\lambda_0, \lambda_1, ..., \lambda_L)^T \in \mathbb{R}^{L+1}$ providing the $L + 1$ Lagrange multipliers for each single condition in Eq (2.11), is used to solve the minimal prediction variance for this universal method:

$$\underset{\omega}{\text{Minimise}} \quad 2\boldsymbol{\omega}^T \boldsymbol{\gamma}_{X,0} - \boldsymbol{\omega}^T \Gamma_X \boldsymbol{\omega} \qquad \text{subject to } \boldsymbol{\omega}^T F = \mathbf{f}_o^T.$$

Cressie (1993) provided the solution of this optimisation problem for universal Kriging, which are

$$\begin{aligned} \boldsymbol{\omega}^{UK}(s_0) &= \Gamma_X^{-1}[\boldsymbol{\gamma}_{X,0} - F(F^T \Gamma_X^{-1} F)^{-1}(F^T \Gamma_X^{-1} \boldsymbol{\gamma}_{X,0} - \mathbf{f}_0)], \\ \boldsymbol{\lambda}^{UK}(s_0) &= (F^T \Gamma_X^{-1} F)^{-1}(F^T \Gamma_X^{-1} \boldsymbol{\gamma}_{X,0} - \mathbf{f}_0). \end{aligned} \tag{2.13}$$

Finally, by applying the above conditions, the universal Kriging variance σ_{UK}^2 and the predictor of $Y_\omega^{UK*}(s)$ at the location s_0 are

$$\begin{aligned} \sigma_{UK}^2(s_0) &= \sum_{i=1}^{N} \omega_i^{UK} \gamma_X(s_i - s_0) + \sum_{l=0}^{L} \lambda_l f_l(s_0), \\ y_{\boldsymbol{\omega}_{UK}}^{UK*}(s_0) = \sum_{i=1}^{N} \omega_i^{UK} Y(s_i) &= \left[\boldsymbol{\gamma}_{X,0} - F(F^T \Gamma_X^{-1} F)^{-1}(F^T \Gamma_X^{-1} \boldsymbol{\gamma}_{X,0} - \mathbf{f}_0)\right]^T \Gamma_X^{-1} \boldsymbol{y}. \end{aligned} \tag{2.14}$$

We conclude this section with Figure 2.2, a linear Kriging decision flowchart from the R package *gstat* user's manual [6]. In summary, universal Kriging is the most general model among the three linear Krigings introduced, and we are now ready to apply these methods to the air quality data.

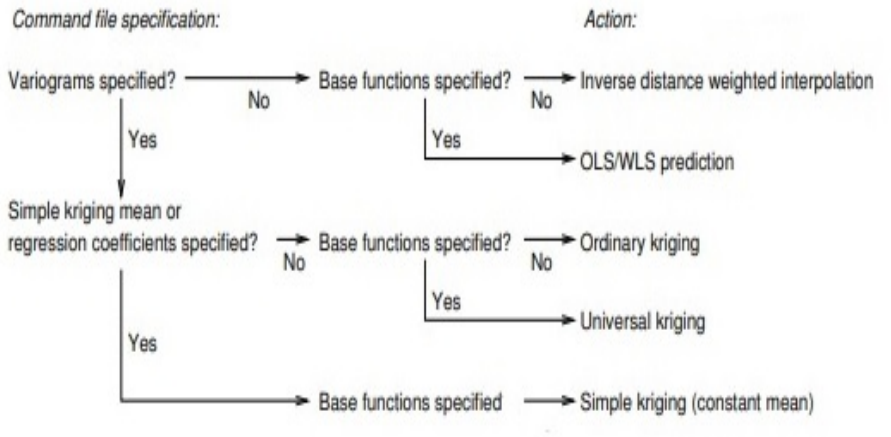

FIGURE 2.2: The decision tree for default programme action

2.3 Empirical application to air quality data

In this section, we will apply the three linear Kirging methods directly to the air quality data in England. By using the data collected from the monitoring stations (data archives from the Defra), we aim to compare the prediction statistics and then map the results to show the visual differences among these three Kriging methods.

2.3.1 The air quality data set

It is noticed that the current monitoring stations are unevenly distributed in the UK. England has the highest number of stations among the three nations in the main British isle [7]. Also by considering the very different geographical conditions between England and the other two nations, Scotland and Wales, we choose England as the studied space S in this research.

Among the five measured pollutants, we take Nitrogen Dioxide (NO_2) to be the response $Y(s)$ for this study. The date 18/04/2017 was exemplarily picked as a typical workday when it is generally believed of having a higher level of air pollution than that a weekend has. We use the air quality data from the validated 105 monitoring stations on that day for this research [8] [9]. Figure 2.3 shows the locations of the total 105 monitoring stations in England.

[6] http://www.gstat.org/gstat.pdf

[7] https://uk-air.defra.gov.uk/networks/

[8] All Air Quality raw data used in this report are downloaded from Defra UK, licenced under the Open Government Licence (OGL). © Crown 2017 copyright Defra via uk-air.defra.gov.uk.

[9] OGL: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/

FIGURE 2.3: Locations of the included air quality monitoring sites in England.

The R package *gstat* is used for performing the linear Krigings in this chapter. Briefly, we divide the Kriging task into two steps:

1. Fitting the theoretical semivariogram from the observed data,
2. Empirical application of simple-, ordinary- and universal Krigings.

In the next section, we will demonstrate the results at both steps.

2.3.2 Fitting the theoretical semivariogram

Before applying Kriging methods, we need to estimate the required properties of the underlying spatial process by fitting a theoretical semivariogram model from the observed data. To complete this step, we will first introduce three basic parameters for a varigram: the nugget, sill, and range defined in accordance with Matheron (1963) and Cressie (1993).

- Nugget: If the empirical semivariogram does not start at the origin, i.e. $\gamma(d) \to c_0 > 0$ as $|d| \to 0$, then the height of the jump c_0 is called the nugget, or nugget effect, representing the value which could be caused by measurement error or some microscale variation. Note that $\gamma(d) = \gamma(-d)$ by definition in this research.
- Sill: The value $\gamma(\infty) := \lim_{|d|\to\infty} \gamma(d)$ is called the sill.

- Range: The distance at which the semivariogram $\gamma(d)$ exceeds the sill value for the first time.

An illustrative example for these parameters is shown in Figure 2.4.

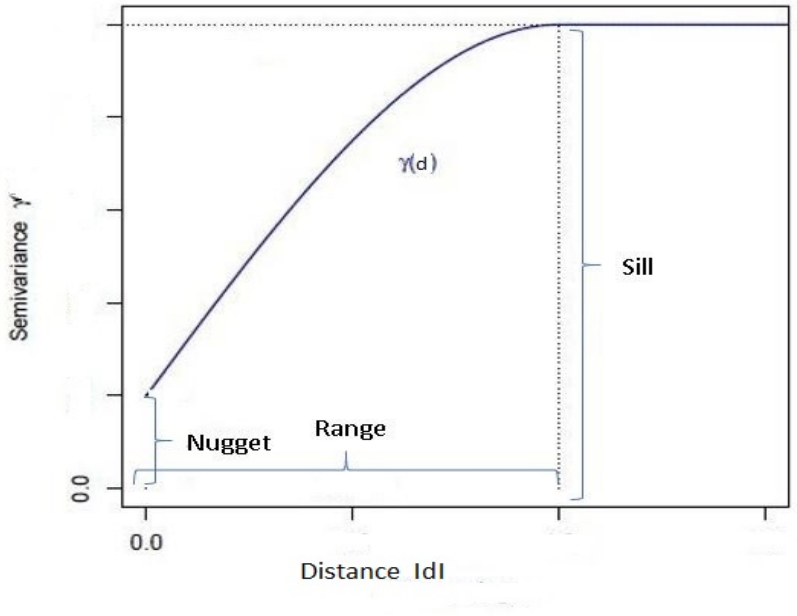

FIGURE 2.4: Semivariogram parameters: the nugget, sill and range.

Note that in practice, the range is often defined as the distance at which the semivariogram reaches about 95% of its sill value, called the effective range (Wackernagel, 2003).

To show the fitting steps for a theoretical semivariogram from the samples, a detailed procedure can be found in Section 4.2.3. It uses the least squares fitting method to compare the matching errors from a collection of candidate models to the experimental semivariogram generated from the sample data. Figure 2.5 shows a list of valid variogram model families often in use, see Appendix B or alternatively one can refers to the *Gstat* (R package) manual. It is worthwhile to mention that as Figure 2.5 reveals that the selection influences the prediction values, particularly when the shape of the curve near the origin differs significantly, such as the steeper the curve near the origin, the more influence the closest neighbours will have on the prediction. In fact, the setup of the Cutoff Distance for the experimental semivariogram in Section 4.2.3 and the weighted least squares options are two typical techniques offering practitioners flexible and more accurate prediction results.

Among the common model families tested, we fit the experimental semivariogram function using the log-transformed sample data (a typical data transformation in this kind of applications) to an exponential fitting model, i.e., it has the smallest sum of squared errors than the other tested models have, see Figure 2.6. The alternatives

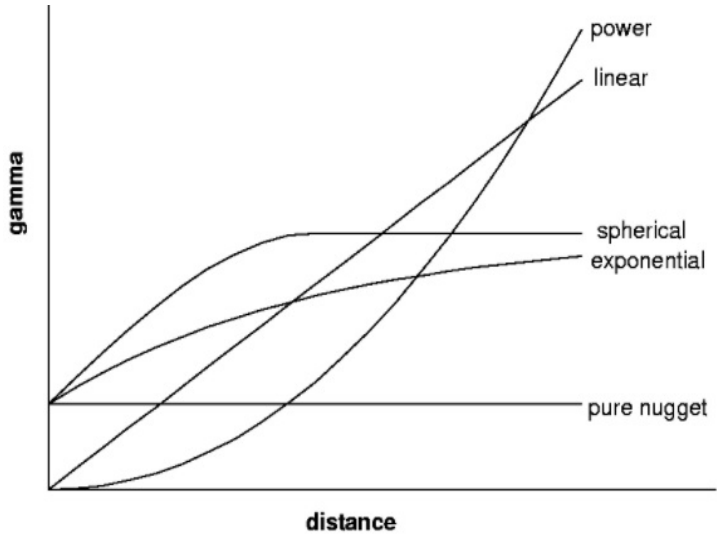

FIGURE 2.5: A list of commonly used semivariogram fitting models.

include nugget-effect model, bounded linear model, spherical model and Gaussian model families. We show the covariance and semivariogram functions of the exponential model below.

Let $\gamma_{a,b}(d)$ denotes the semivariogram function, $C_{a,b}(d)$ refers to the corresponding covariance function with lag d and $a, b > 0$ are the parameters of each model, where a represents the range parameter and b is the sill value, we have the exponential model as follows:

$$\gamma_{a,b}^{\exp}(d) := b\left(1 - \exp(-\frac{|d|}{a})\right),$$

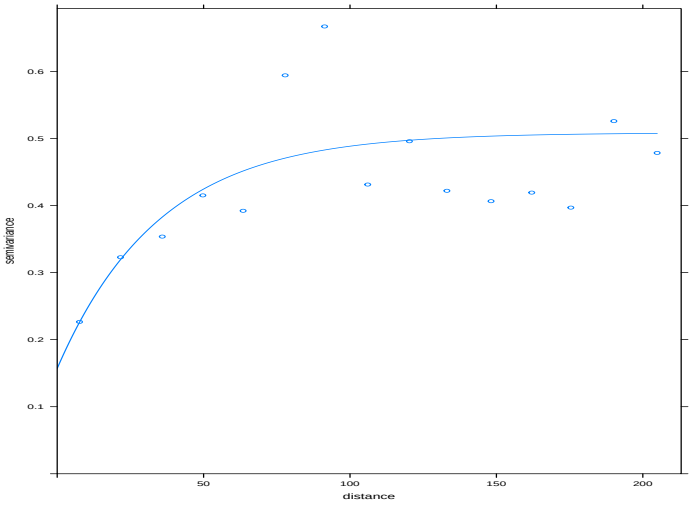

FIGURE 2.6: The fitted theoertical exponential semivariogram for the observed data.

$$C^{\exp}_{a,b}(d) := b \exp(-\frac{|d|}{a}).$$

2.3.3 Empirical applications to air quality data

By assuming the fitted semivariogram function as the theoretical underlying process, in turn we apply the three linear Krigings to the air quality data. Figures 2.7, 2.8 and 2.9 show the maps of the predicted values and prediction variances at each spatial location by simple-, ordinary- and universal Kriging. All three prediction results confirm the air quality problem around major cities in England such as London and Manchester; whereas rural areas including national parks enjoy good air quality in general. The changes of the colour contour may be used to identify the boundaries where abrupt shifts on air quality are likely to happen. Among the three prediction maps, all Kriging methods show similar performance with no clear dominant colour, albeit the colour patterns are visually different. It is also noticed that the predictions in the centred area and where the most monitoring sites are have finer changes in colour than those from the surrounding edges, which is understandable. The results may be different when comparing the prediction variances, yet no strong visual evidence suggests which the best model is.

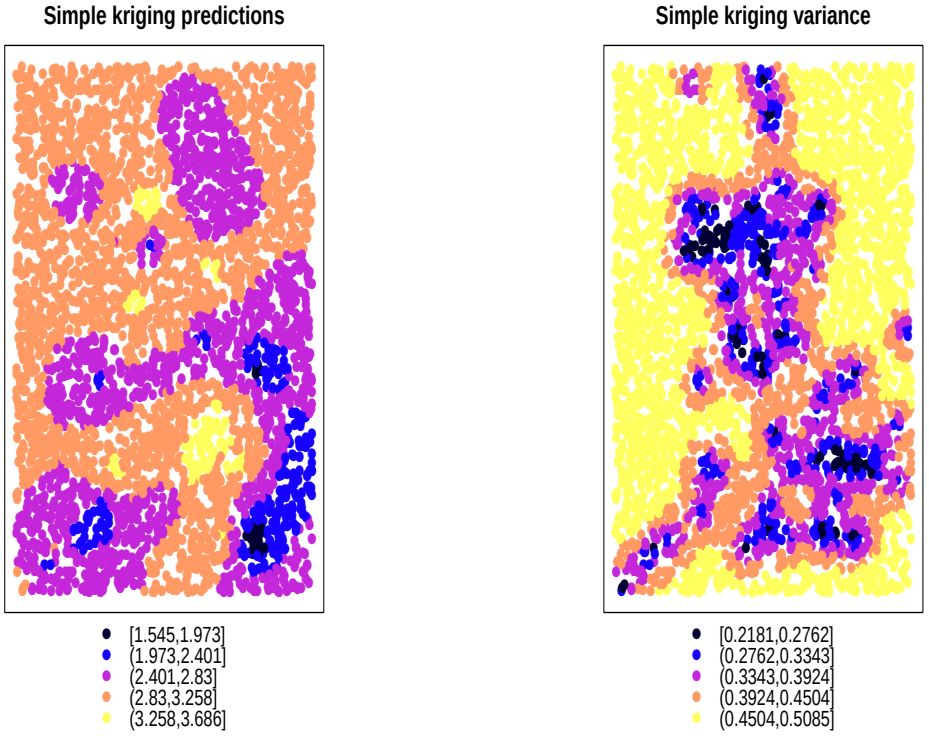

FIGURE 2.7: Map of the predicted values and variances using simple Kriging.

In addition to the visual check, we compare the prediction performance from the three linear Krigings using the cross validation criterion. Leave-one-out Cross Validation (LOOCV), a specific form of cross validation technique, where the number of folds

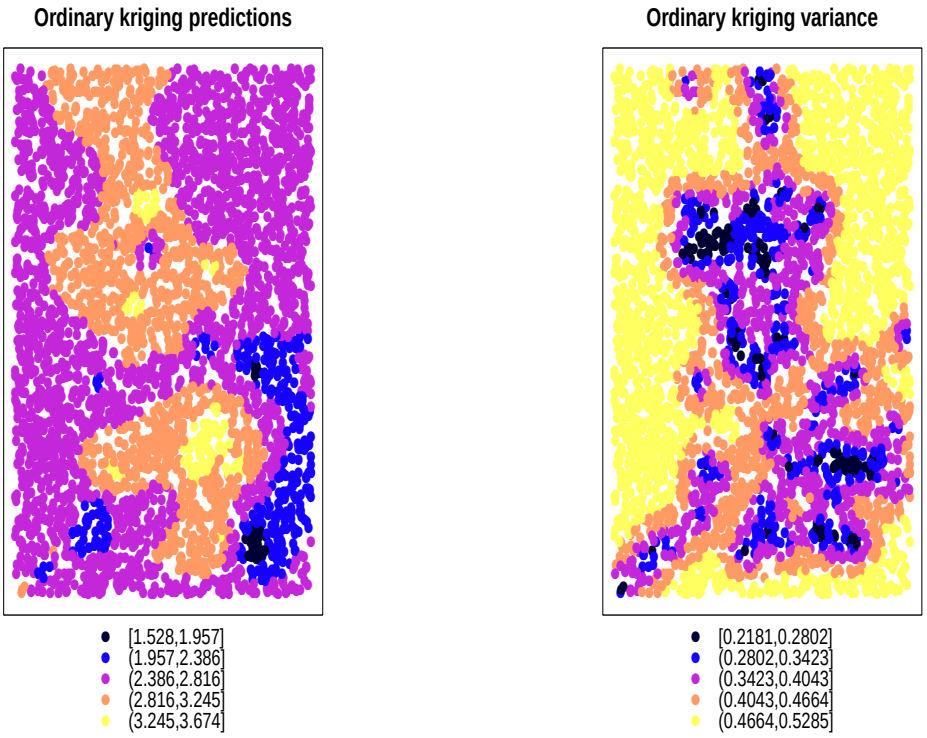

FIGURE 2.8: Map of the predicted values and variances using ordinary Kriging.

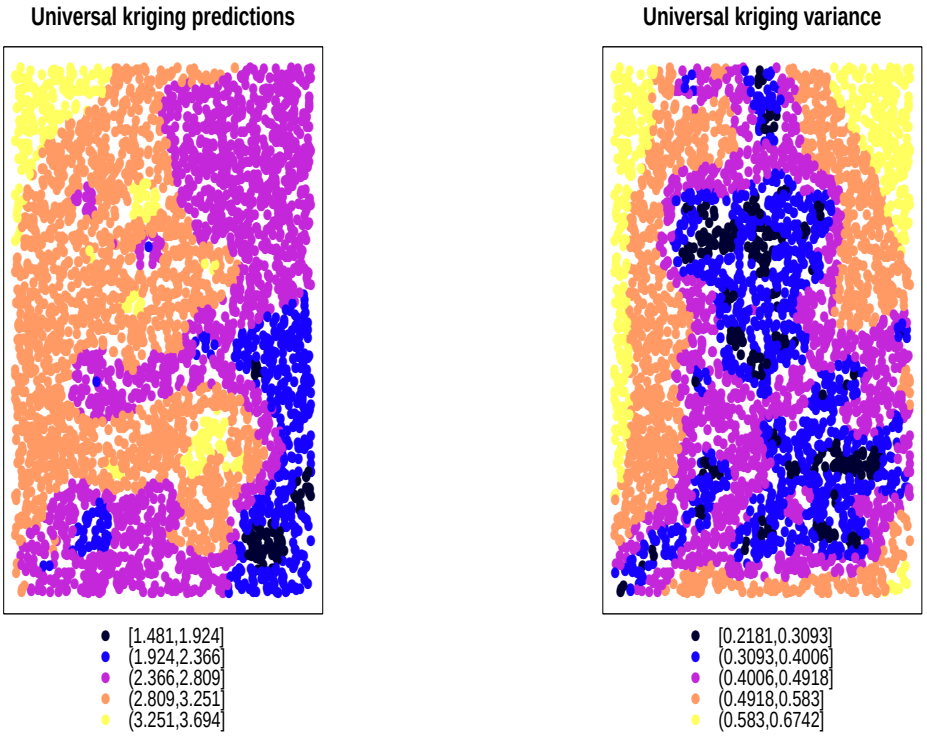

FIGURE 2.9: Map of the predicted values and variances using universal Kriging.

equals to the number of observations in the data set, estimates how accurate a predictive model performs on real data (Seymour, 1993) . In each iteration, one sampled data $Y(s_j), j \in [1, ..., N]$ is selected as the validation dataset while all the other samples: $Y(s_1), ..., Y(s_{j-1}), Y(s_{j+1})$, ..., $Y(s_N)$, or simply $Y(s_{-j})$, form the training dataset for establishing the estimation model, consequently to predict the $\hat{Y}(s_j)$. Each sampled location s_j is used for one iteration only.

Mean squared prediction error (MSPE) as shown in Eq (2.15) calculates the mean squared difference between the observed value $Y_{obs}(s)$ and its estimated value $\hat{Y}(s)$ obtained from the LOOCV. The modelling method with the smallest MSPE, as believed, has the highest prediction accuracy.

$$\text{MSPE} = \frac{1}{N}\sum_{j=1}^{N}(\hat{Y}(s_j) - Y_{obs}(s_j))^2, \text{ where } i = 1, ..., N. \tag{2.15}$$

Table 2.1 shows the comparison of the mean squared prediction error (MSPE) obtained from the three Kriging methods.

TABLE 2.1: Comparison of the mean squared errors from the three linear Kriging methods.

Kriging methods	Mean Squared Prediction Error
Direct Simple Kriging	231.9582
Direct Ordinary Kriging	232.7272
Direct Universal Kriging	233.0495

Based on the results, simple Kriging performs the best among them. However, there is no significant performance variation among these three methods on this air quality data set, which is in line with the visual comparison shown above. This indifference thus motivates us to develop alternative Kriging methods.

Chapter 3

Nonparametric-Trend Universal Kriging Method

After the initial introduction of spatial prediction and the current linear Kriging methods, as our first contribution, in this chapter we will propose a nonparametric Kriging method with an adaptive nonlinear function as the spatial trend component $\mu(s)$ in model (1.1) [1].

3.1 Background

In Chapter 2, linear Kriging methods are directly applied to the air quality data for predicting the value of the response $Y(s)$ at a new location of interest $s_0 \in S$. When these processes are repeated at a large number of locations, consequently we obtain visual plots (maps) of the predicted results. Despite clear benefits of these methods, the limitations of linear Kriging methods are likewise significant, which are essentially resulted from the linearity assumption as well as their linear prediction models (2.1), (2.4) and (2.10), where the predictors are defined as linear regressions of a set of variables, explicitly the observed values at the sampled locations.

Matheron (1976) described that since the space of a linear combination is much smaller than the space from an arbitrary measurable function, the approximation using linear Kriging methods is restricted, which may likely lead to a significant misspecification on the original data. Hence, by choosing a solution space larger than the linear combination space but small enough to allow the computation feasible, we hope to develop a nonlinear method for the air quality case study. In fact, the air quality data set from Chapter 2 indeed suggests a non-Gaussian profile as shown in Figure 3.2 later, which may explains the undesired performance in Chapter 2.

[1] A talk on this method was given by the author at the 12th International Conference on Computational and Financial Econometrics (CFE 2018), at the University of Pisa, Italy, on 14 December 2018.

To overcome the misspecification of applying linear Kriging methods directly to non-Gaussian data set, in this chapter, a modified predicting method, the nonparametric-trend universal Kriging (NTUK) model, is thereby proposed by employing a nonparametric adaptive regression function for estimating the spatial trend $\mu(s)$ and leaving the de-trended stochastic term $X(s)$ in model (1.1) for Kriging. In the following sections, we will firstly introduce this method, followed by an introduction of its asymptotic properties as for the estimator from the population aspect in Section 3.3. Finally, in Section 3.4 its empirical performance is compared with those from the methods in Chapter 2.

3.2 Methodology

We re-examine the most general linear Kriging method introduced so far, the universal Kriging. As shown in Eq (2.9), with the linearity assumption in the air quality case study, the spatial trend $\mu(s)$ is modelled as a linear regression of the two spatial coordinates of a location, then kriges the stochastic residual terms $X(s)$ using the variogram function estimated from the residual data. A clear drawback of this method is to restrict the spatial trend to a linear solution space. In this chapter, we propose to replace it with an nonparametric regression function to model the spatial trend $\mu(s)$, by which the deterministic trend does not take a predetermined form but to construct it according to the derived information from the samples. As a result, intuitively, this modified prediction model is more suitable for a general non-Gaussian data set, where we believe the novelty of this method lies, and most importantly we would like to show asymptotically the estimator converges in probability to the true response in the population.

Within the family of nonparametric methods, the local linear regression (LLR) is chosen for this method due to its superior accuracy over the other common method, i.e., the local constant regression (LCR), when the sample size is moderately large (Fan and Gijbels, 1996).

To obtain the estimated value of $\mu(s)$ using the LLR, we apply Taylor expansion of Eq (2.9) in the neighbourhood of a new location $s_0 := (u_0, v_0)$, where the distance of s_i to s_0, $\|s_i - s_0\|$, is small. It is obtained that,

$$Y(s_i) = \mu(s_i) + X(s_i) \approx \mu(s_0) + \mu_u'(s_0)(u_i - u_0) + \mu_v'(s_0)(v_i - v_0) + X(s_i), \qquad (3.1)$$

where $s_i = (u_i, v_i)$, u_i, v_i represent the spatial coordinates of the observation location $s_i, i = 1, ..., N$. It is commonly assumed that in this type of spatial applications, the first and second order derivatives of the underlying function $\mu(s)$ exist for the explanatory variables u_i and v_i (Hallin et al., 2004). Hence, the response at the concerned location

s_0 can be estimated from the sampled information using least squares method for C as follows:

$$C := \sum_{i=1}^{N} \Big(Y(s_i) - a_0 - a_1(u_i - u_0) - a_2(v_i - v_0) \Big)^2 K(\frac{s_i - s_0}{h_N}), \\ \underset{(a_0,a_1,a_2)}{\text{Minimise}} \; C, \tag{3.2}$$

where a_0, a_1, a_2 represent $\mu(s_0)$, $\mu_u^{'}(s_0)$ and $\mu_v^{'}(s_0)$, respectively, $K(\cdot)$ refers to a selected symmetric bivariate kernel probability density function with bounded support, and h_N, or h for short, is a bandwidth that tends to 0 as $N \to \infty$. The solution of a_0 for Eq (3.2), denoted as $\hat{\mu}(s_0)$, provides the estimated value of $\mu(s_0)$ at the concerned location s_0.

By taking each s_i, $i = 1, ..., N$ in turns as the s_0, the above nonparametric function produces the estimated spatial trends $\hat{\mu}(s_i)$. By subtracting this trend $\hat{\mu}(s_i)$ from the original observations $Y(s_i)$, we get the value of the de-trended stochastic term $\hat{X}(s_i)$, as shown below,

$$\hat{X}(s_i) = Y(s_i) - \hat{\mu}(s_i), \tag{3.3}$$

which can be treated as the estimate of the de-trended (residual) component $X(s_i)$ at s_i, $i = 1, ..., N$.

With the same assumptions that the underlying process of $\hat{X}(s)$ is intrinsically stationary with a zero mean and its variogram function $\gamma_{\hat{X}}(d)$ is known, as introduced in Chapter 2, the same ordinary (linear) Kriging, or simplified as OK method, can therefore be applied to the de-trended $\hat{X}(s_i)$ in order to get the $\hat{X}^{OK}(s_0)$, an ordinary Kriging estimator of $\hat{X}(s_0)$ at the location s_0. Subsequently, by adding the $\hat{X}^{OK}(s_0)$ to the $\hat{\mu}(s_0)$ from the nonparametric regression in Eq (3.2) , we get the estimated value of $\hat{Y}^{UK}(s_0)$, which is called the nonparametric-trend universal Kriging (NTUK) of $Y(s_0)$, the proposed new method of predicting the air quality value at a new location s_0 in this chapter. We name it as an universal Kriging because it has a varying trend similar to that of the linear universal Kriging. Overall, it follows the form of

$$\hat{Y}^{UK}(s_0) = \hat{\mu}(s_0) + \hat{X}^{OK}(s_0). \tag{3.4}$$

In summary, by employing the modified nonparametric spatial trend function in the current universal Kirging, this proposed NTUK method aims to enlarge the solution

space of the spatial trend from its original linear form. We believe it is a more general and convenient Kriging method specifically for non- Gaussian data sets.

3.3 Asymptotic theory

As the most important part in this chapter, we will conduct some theoretical investigations in this section concerning this proposed nonparametric-trend universal Kriging method, and provide related asymptotic proofs where needed.

Eqs (3.2), (3.3), and (3.4) from the last section show the process of how this NTUK new method using the sampled data to predict the response $Y(s_0)$ at a concerned location s_0. In brief, the NTUK process can be demonstrated as follows:

$$\hat{Y}^{UK}(s_0) = \hat{\mu}(s_0) + \hat{X}^{OK}(s_0), \text{ where } \hat{X}(s_1), ..., \hat{X}(s_N) \overset{OK}{\Longrightarrow} \hat{X}^{OK}(s_0). \tag{3.5}$$

Note that, $\hat{X}^{OK}(s_0)$ is predicted based on the nonparametrically de-trended stationary process $\hat{X}(s_i), i = 1, ..., N$. In other words, from the sample point of view, by adding the ordinary Kriging predictor with a nonparametric trend, we can estimate the response value at the new location s_0.

Following on, from the population point of view, we use Eq (3.6) to show the process of how a theoretical universal Kriging alone predict the response $\tilde{Y}^{UK}(s_0)$ by supposing the value of the true spatial trend $\mu(s_0)$ is known, as

$$\tilde{Y}^{UK}(s_0) = \mu(s_0) + \tilde{X}^{OK}(s_0), \text{ where } X(s_1), ..., X(s_N) \Longrightarrow \tilde{X}^{OK}(s_0), \tag{3.6}$$

where $X(s_i), i = 1, ..., N$ stand for the true values of the residual process at the N observation sites, $\mu(s_0)$ and $\tilde{Y}^{UK}(s_0)$ represent the true value of the spatial trend and the predicted value of the response $Y(s_0)$ at a concerned location s_0, respectively.

In this section, we will show that the pair of $\hat{\mu}(s_0)$ and $\hat{X}^{OK}(s_0)$ as shown in Eq (3.5) converge (in probability) to $\mu(s_0)$ and $\tilde{X}^{OK}(s_0)$ in Eq (3.6) respectively, when the number of the observation sites N tends to infinity. Consequently under the same conditions, the NTUK estimator $\hat{Y}(s_0)$ in Eq (3.5) estimated from the sampled observations converges to the $\tilde{Y}(s_0)$ in Eq (3.6) represented theoretically from the true population. Hence, we propose the following theorem,

Theorem 3.1: Assume that the above conditions hold, for $s_0 \in S$, we have

$$\hat{Y}^{UK}(s_0) - \tilde{Y}^{UK}(s_0) \xrightarrow{\text{P}} 0, \tag{3.7}$$

as the sample size $N \to \infty$.

Before formally prove this theorem, we need to summarise its theoretical backgrounds as preparation: (a) the definition of the Domain-Expanding Infill (DEI) asymptotics framework introduced by Lu and Tjøstheim (2014) and (b) the main conditions and assumptions of this theorem.

(a) The framework of the domain-expanding infill (DEI) asymptotics is defined as,

$$\begin{aligned} \delta_N &= \max_{1\leq j\leq N} \delta_{j,N} \to 0, \\ \text{with } \delta_{j,N} &= \min\{||s_i - s_j|| : 1 \leq i \leq N, i \neq j\}, \end{aligned} \tag{3.8}$$

that is, all the distance between neighboring observations tends to 0 as $N \to \infty$, and

$$\begin{aligned} \Delta_N &= \min_{1\leq j\leq N} \Delta_{j,N} \to \infty, \\ \text{with } \Delta_{j,N} &= \max\{||s_i - s_j|| : 1 \leq i \leq N, i \neq j\}, \end{aligned} \tag{3.9}$$

that is, the domain at each location expands to infinity as $N \to \infty$, where $\|\cdot\|$ is the Euclidean norm. Conveniently, we name the δ_N and Δ_N as the infilling distance and expanding distance of the spatial sites, respectively. The DEI asymptotics framework is a reconciliation of the traditional domain-fixed infill (DFI) asymptotics and the domain-expanding (DE) asymptotics which has the benefits from both frameworks while in many applications, it may be natural as a result of the data structure.

(b) For the sake of simplicity, the main conditions assumed for this theorem are summarised on the random field $\{X(s) : s \in \mathbb{R}^2\}$ and the kernel $K(\cdot)$ is used in estimation. We divide these assumptions into four categories: (A) spatial process, (B) sampling sites, (C) kernel function, and (D) the bandwidths.

For any collection of site $S \subset \mathbb{R}^2$, $\mathcal{B}(S)$, the Borel σ-field generated by $\{Y(s)|s \in S\}$, and for each couple $S^{'}$,$S^{''}$, let $d(S^{'}, S^{''}) := \min\{||s^{'} - s^{''}|| \mid s^{'} \in S^{'}, s^{''} \in S^{''}\}$ be the distance between $S^{'}$ and $S^{''}$, where $||s|| := (u^2 + v^2)^{1/2}$, for $s = (u, v) \in \mathbb{R}^2$. Finally, the cardinality of S is denoted by $\text{Card}(S)$ (Lu and Tjøstheim, 2014).

Assumption (A) (spatial processes):

(i) $X(s), s \in \mathbb{R}^2$ is a strictly stationary spatial process satisfying the α-mixing property that there exists a function φ such that $\varphi(t) \downarrow 0$ as $t \to \infty$, and a function $\psi : \mathcal{N}^2 \to \mathbb{R}^+$ that is symmetric and increasing in each of its two arguments such that

$$\begin{aligned}&\alpha(\mathcal{B}(S^{'}),\mathcal{B}(S^{''}))\\&:=\sup\{|P(AB)-P(A)P(B)|,A\in\mathcal{B}(S^{'}),B\in\mathcal{B}(S^{''})\}\\&\leq\psi(\mathrm{Card}(S^{'}),\mathrm{Card}(S^{''}))\varphi(d(S^{'},S^{''})),\end{aligned}\tag{3.10}$$

for any $S^{'},S^{''}\subset\mathbb{R}^2$. Moreover, the function φ is such that

$$\lim_{t\to\infty}t^{\gamma}\sum_{j=t}^{\infty}j^2\{\varphi(j)\}^{\kappa/(2+\kappa)}=0,\tag{3.11}$$

for some constant $\gamma>\max\{1,2\kappa/(2+\kappa)\}$ and some $\kappa>0$.

(ii) Denote by $f(x,y;s_0)$ the joint density function of $X(s)$ and $X(s+s_0)$, where $s_0\neq(0,0)$. $f(x,y;s)$ is uniformly a continuous function of (x,y) with respect to $s\in\mathbb{R}^2$, and it has second-order partial derivatives with respect to x,y and s, which are continuous.

(iii) The marginal and joint probability density function for X_i and (X_i,X_j), $f(x)$ and $f_{i,j}(x,y)$ satisfy $|f_{i,j}(x,y)-f(x)f(y)|\leq C$ uniformly for $i\neq j$ and $(x,y)\in\mathbb{R}^2$, where C is a generic positive constant.

Assumption (B) (sampling sites):

The observation sites are located at $\{s_i,\ \ i=1,...,N\}\subset\mathbb{R}^2$, for which Eq (3.8) and (3.9) hold with $\min_{1\leq j\leq N}\delta_{j,N}/\delta_N\geqslant c_1>0$ and $\max_{1\leq j\leq N}\Delta_{j,N}/\Delta_N\leq C_1<\infty$ for all N, and there exists a continuous sampling intensity function (density function) f_S defined on $\mathbb{R}^2$ such that:

(i) for any measurable set $A\subset\mathbb{R}^2$, $N^{-1}\sum_{i=1}^{N}I(s_i\in A)\to\int_A f_S(s)ds$ as $N\to\infty$.

(ii) $f_S(s)$ is bounded and it has second derivatives which are continuous on $\mathbb{R}^2$.

Assumption (C) (kernel functions):

(i) the kernel function $K(\cdot)$ satisfies $\int K(u)du=1$, $\int uK(u)du=0$, and $\mu_{K,2}=\int u^2K(u)du<\infty$, $v_K=\int K^2(u)du<\infty$.

(ii) the kernel function $L(\cdot)$ has a bounded support such that $\int_{\mathcal{R}^2}L(s)ds=1$, $\int_{\mathcal{R}^2}sL(s)ds=0$, and $b\mu_{K,2}=\int_{\mathcal{R}^2}ss^TL(s)ds<\infty$, $v_L=\int_{\mathcal{R}^2}L^2(s)ds<\infty$.

Assumption (D) (bandwidths):

(i) As $N\to\infty$, (1) $h\to0$, (2) $N\delta_N^2h^4\to\infty$, and (3) $(N\delta_N^2)^{(\gamma+2)}h^8\to\infty$.

Assumption (A) concerns the conditions on the spatial data-generating process which are standard in the context of the problem under study. Assumption (B) offers the conditions on the spatial sites where observations are irregularly located under the DEI asymptotics framework. Assumption (C) specifics the conditions for the two nonparametric kernels used. Assumption (D) lists the conditions on the bandwidth which will be used later.

In the rest of this section, we will show the theoretical proofs of this theorem revealing the asymptotic properties of this newly proposed NTUK method.

Note that in Eq (3.5), substracting the nonlinear spatial trend $\hat{\mu}(s_i), i = 1, ..., N$, from the observations $Y(s_i)$, we get the de-trended residual terms, $\hat{X}(s_i)$. By applying ordinary Kriging on these residuals with the mild assumption that the de-trended spatial process is intrinsically stationary, as introduced in Chapter 2, we get

$$\hat{X}^{OK}(s_0) = [\gamma_{\hat{X},0}]^T \Gamma_{\hat{X}}^{-1} \hat{X}, \tag{3.12}$$

where the element of the symmetric de-trended residual semivariogram matrix is $(\Gamma_{\hat{X}})_{i,j}$ $:= \gamma_{\hat{X}}(s_i - s_j) = \frac{1}{2}E(\hat{X}(s_i) - \hat{X}(s_j))^2, i, j = 1, ..., N$ and $\gamma_{\hat{X},0} := (\gamma_{\hat{X}}(s_1 - s_0), ..., \gamma_{\hat{X}}(s_N - s_0))^T$. Here $\gamma_{\hat{X},0}$ is the semivariogram function of the de-trended residual process measured between the concerned location s_0 and the sampled site $s_i, i = 1, ..., N$.

The benefit of an intrinsically stationary process has been described earlier, that is the semivariogram can be represented as a function which is in relation to the distance between each pair of locations in the space S (Cressie, 1993).

For Eq (3.6), we can write a similar equation to calculate the estimated residuals $\tilde{X}^{OK}(s_0)$ at any location s_0 from the unknown true residuals at the sampled locations, that is

$$\tilde{X}^{OK}(s_0) = [\gamma_{X,0}]^T \Gamma_X^{-1} X, \tag{3.13}$$

where the element of the true residual semivariogram matrix is $(\Gamma_X)_{i,j} := \gamma_X(s_i - s_j) = \frac{1}{2}E(X(s_i) - X(s_j))^2, i, j = 1, ..., N$ and $\gamma_{X,0} := (\gamma_X(s_1 - s_0), ..., \gamma_X(s_N - s_0))^T$. Here $\gamma_{X,0}$ is the true semivariogram function of the residual process $X(s)$ measured between the concerned location s_0 and the sampled site $s_i, i = 1, ..., N$.

To reveal the asymptotic relations between $\hat{X}^{OK}(s_0)$ and $\tilde{X}^{OK}(s_0)$, and consequently between $\hat{Y}^{UK}(s_0)$ and $\tilde{Y}^{UK}(s_0)$ in Eqs (3.5) , (3.6) and (3.7), we would like to study the relations between:

(i) $\hat{\mu}(s_0)$ and $\mu(s_0)$ at any location $s_0 \in S$,

(ii) $\gamma_{\hat{X}}(s_i - s_j)$ and $\gamma_X(s_i - s_j)$, the semivariogram functions of the detrended residual process $\hat{X}(s_i)$ and the true residual process $X(s_i)$, respectively.

We start with the investigation on (i). To solve Eq (3.2) for all s_i, $i = 1, ..., N$, it can be written in matrix form,

$$\underset{\boldsymbol{\beta}}{\text{Minimise}} \ (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})^T \boldsymbol{\omega} (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}), \tag{3.14}$$

where $\boldsymbol{Y} := (Y(s_1), ..., Y(s_N))^T$, $\boldsymbol{Z} \in \mathbb{R}^{(N\times 3)}$, $(\boldsymbol{Z})_i = (1, u_i - u_0, v_i - v_0)$, $\boldsymbol{\omega}$ is a diagonal matrix, whose i^{th} diagonal element is $K(\frac{s_i - s_0}{h_N})$, or simply $K_{i,0}$, and $\boldsymbol{\beta} = (a_0, a_1, a_2)^T$. Lastly, we denote $\hat{\boldsymbol{\beta}} = (\hat{a}_0, \hat{a}_1, \hat{a}_2)^T$ the optimal solution of Eq (3.2); in this application, we are interested in its first element $\hat{a}_0$, the estimator of $\mu(s_0)$.

Fan and Gijbels (1996) provided the solution of $\hat{\boldsymbol{\beta}}$ for Eq (3.2) and (3.14) by weighted least squares theory, furthermore Hallin et al. (2004) expanded the solution of Eq (3.2) to a n-dimensional case, that is,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T \boldsymbol{\omega} \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{\omega} \boldsymbol{Y}. \tag{3.15}$$

Thus, from Eq (3.15) and Eq (3.2), we obtain

$$\begin{aligned} \hat{\mu}(s_0) - \mu(s_0) &= (1,0,0)[(\boldsymbol{Z}^T \boldsymbol{\omega} \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{\omega} (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})] \\ &= (1,0,0)[\underbrace{(\frac{1}{Nh_N^2} \boldsymbol{Z}^T \boldsymbol{\omega} \boldsymbol{Z})^{-1}}_{\boldsymbol{D}^*} \underbrace{\frac{1}{Nh_N^2} \boldsymbol{Z}^T \boldsymbol{\omega} (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta})}_{\boldsymbol{D}}], \end{aligned} \tag{3.16}$$

where $\mathbf{Z}, \boldsymbol{\omega}$,$\mathbf{Y}$ and $\boldsymbol{\beta}$ are defined as above, and $\mathbf{D}^*$ is a deterministic 3×3 symmetric matrix related with the known locations of the monitoring sites.

To solve Eq (3.16), we explore the asymptotic solution of $\mathbf{D}$. It is easy to show that $\mathbf{D}$, the second term in Eq (3.16) is defined as

$$\begin{aligned} \mathbf{D} &:= \frac{1}{Nh_N^2} \mathbf{Z}^T \boldsymbol{\omega} (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}) \\ &= \begin{bmatrix} \frac{1}{Nh_N^2} \sum_{i=1}^N K_{i,0} (\mu(s_i) + X(s_i) - a_0 - a_1(u_i - u_0) - a_2(v_i - v_0)) \\ \frac{1}{Nh_N^2} \sum_{i=1}^N K_{i,0} (u_i - u_0)(\mu(s_i) + X(s_i) - a_0 - a_1(u_i - u_0) - a_2(v_i - v_0)) \\ \frac{1}{Nh_N^2} \sum_{i=1}^N K_{i,0} (v_i - v_0)(\mu(s_i) + X(s_i) - a_0 - a_1(u_i - u_0) - a_2(v_i - v_0)) \end{bmatrix} \end{aligned} \tag{3.17}$$

Imposing the Assumption A (ii) that the spatial regression function $\mu(s)$ is twice differentiable, we expand Eq (3.1) as below, with $|\gamma| < 1$,

$$\begin{aligned}\mu(s_i) =&\mu(s_0) + h\mu_u^{'}(s_0)(\frac{u_i - u_0}{h}) + h\mu_v^{'}(s_0)(\frac{v_i - v_0}{h}) \\ &+ \frac{h^2}{2}\left[\mu_{uu}^{''}\left(s_0 + h\gamma(\frac{s_i - s_0}{h})\right)(\frac{u_i - u_0}{h})^2 + 2\mu_{uv}^{''}\left(s_0 + h\gamma(\frac{s_i - s_0}{h})\right)\right. \\ &\left.(\frac{u_i - u_0}{h})(\frac{v_i - v_0}{h}) + \mu_{vv}^{''}\left(s_0 + h\gamma(\frac{s_i - s_0}{h})\right)(\frac{v_i - v_0}{h})^2\right].\end{aligned} \tag{3.18}$$

Note that $a_0 = \mu(s_0)$, $a_1 = \mu_u^{'}(s_0)$ and $a_2 = \mu_v^{'}(s_0)$ as in Eq (3.2). Plugging Eq (3.18) into Eq (3.17) , the first element of $\mathbf{D}$, denoted as D_{11}, can be shown as

$$\begin{aligned}\mathbf{D}_{11} = \frac{h^2}{2}\Bigg[&\underbrace{\frac{1}{Nh^2}\sum_{i=1}^{N} K(\frac{s_i - s_0}{h})(\frac{u_i - u_0}{h})^{'}\mu_{uu}^{''}(s_0)(1 + o(1))}_{D11.a} \\ &+ \underbrace{\frac{1}{Nh^2}\sum_{i=1}^{N} K(\frac{s_i - s_0}{h})(\frac{u_i - u_0}{h})(\frac{v_i - v_0}{h})\mu_{uv}^{''}(s_0)(1 + o(1))}_{D11.b} \\ &+ \underbrace{\frac{1}{Nh^2}\sum_{i=1}^{N} K(\frac{s_i - s_0}{h})(\frac{v_i - v_0}{h})^2\mu_{vv}^{''}(s_0)(1 + o(1))}_{D11.c}\Bigg] \\ &+ \underbrace{\frac{1}{Nh^2}\sum_{i=1}^{N} K(\frac{s_i - s_0}{h})X(s_i)}_{D11.d}.\end{aligned} \tag{3.19}$$

Starting with the residual term D11.d, we take its squared form and get

$$\begin{aligned}E(\frac{1}{Nh^2}\sum_{i=1}^{N} K(\frac{s_i - s_0}{h})X(s_i))^2 = &\underbrace{\frac{1}{N^2h^4}(\sum_{i=1}^{N} K^2(\frac{s_i - s_0}{h})E(X^2(s_i)))}_{\mathbf{E}} \\ &+ \underbrace{\frac{1}{N^2h^4}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} K(\frac{s_i - s_0}{h})K(\frac{s_j - s_0}{h})E\left(X(s_i)X(s_j)\right)}_{\mathbf{F}}.\end{aligned} \tag{3.20}$$

We adapt the Assumption B (i) of the intensity function as in Eq (3.21) to **E**, the first RHS term in Eq (3.20), where $f_S(s)$ is the intensity function of the sample at location $s = (u, v)$. As $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N} I(S_i \in A) \to \int_A f_S(s)ds. \tag{3.21}$$

Therefore, noting $E(X^2(s_i)) = \sigma_X^2$, we get,

$$\begin{aligned}
\mathbf{E} &= \frac{1}{N^2h^4}\sum_{i=1}^{N} K^2(\frac{s_i - s_0}{h})\sigma_X^2 = \frac{\sigma_X^2}{Nh^4}\int K^2(\frac{s - s_0}{h})f_S(s)ds(1+o(1)) \\
&= \frac{\sigma_X^2}{Nh^4}\iint K^2(\frac{u-u_0}{h}, \frac{v-v_0}{h})f_S(u,v)dudv(1+o(1)) \\
&= \frac{\sigma_X^2}{Nh^2}\iint K^2(u',v')f_S(hu'+u_0, hv'+v_0)du'dv'(1+o(1)) \\
&= \frac{\sigma_X^2}{Nh^2}\iint K^2(u,v)f_S(u_0)f_S(v_0)dudv(1+o(1)) \\
&= \frac{\sigma_X^2 f_S(u_0)f_S(v_0)}{Nh^2}\iint K^2(u,v)dudv(1+o(1)) = O(\frac{1}{Nh^2}),
\end{aligned} \tag{3.22}$$

with the condition $N \to \infty$ and $\int K^2(s)ds$ is bounded as in Assumption D (i) and C (i), $\mathbf{E} \to 0$ as $N \to \infty$.

Unlike the straightforward proof of **E**, the properties of α-Mixing in Assumption A (i) is required to prove the part **F** in Eq (3.20), two separate spatial locations s_i and s_j are involved.

We split **F** into two parts i) when the distance between the two locations s_i and s_j is less than or equals to a set distance, denoted as P_N , and ii) the distance between s_i and s_j is greater than P_N, that is,

$$\begin{aligned}
\mathbf{F} = &\frac{1}{N^2h_N^4}\sum_{i,j:0<d(s_i,s_j)\le P_N} K(\frac{s_i - s_0}{h})K(\frac{s_j - s_0}{h})E(X(s_i)X(s_j)) \\
&+ \frac{1}{N^2h_N^4}\sum_{i,j:d(s_i,s_j)>P_N} K(\frac{s_i - s_0}{h})K(\frac{s_j - s_0}{h})E(X(s_i)X(s_j)),
\end{aligned} \tag{3.23}$$

where the procedure of selecting a suitable P_N will be specified later in this section, following the method suggested by Lu and Tjøstheim (2014).

The proof of the asymptomatic property of Eq (3.23) is quite involved. The preparation starts from the Cauchy-Schwarz inequality in Eq (3.24), similarly the Hölder's inequality in Eq (3.26) with Lemma 3.1 (Eq 3.25), which is often applied in similar cases (Ibragimov et al. (1971), Deo (1973)). We also requires the assumptions from the $\alpha-$Mixing properties in Assumption A (i) and the bandwidth conditions in Assumption D (i),

$$|E(X_iX_j)| \leq (EX_i^2)^{\frac{1}{2}}(EX_j^2)^{\frac{1}{2}} = \sigma_X^2. \tag{3.24}$$

Lemma 3.1. Let $\mathcal{L}_r(\mathcal{F})$ denotes the class of $\mathcal{F}$-measurable random variables ξ satisfying $\|\xi\|_r := (E|\xi|^r)^{\frac{1}{r}} < \infty$. Let $\mathbf{U} \in \mathcal{L}_r(\mathcal{B}(S))$ and $\mathbf{V} \in \mathcal{L}_r(\mathcal{B}(S^{'})$, where $\mathcal{B}(S)$ and $\mathcal{B}(S^{'})$ denote the σ-fields generated by $\{Y(s) : s \in S\}$ and $\{Y(s) : s \in S^{'}\}$, respectively. Then, for any $1 \leq r, s, t < \infty$ such that $\gamma^{-1} + s^{-1} + t^{-1} = 1$,

$$|E(UV) - E(U)E(V)| \leq C \parallel U \parallel_\gamma \parallel V \parallel_s [\alpha(\mathcal{S}, \mathcal{S}^{'})]^{1/t}, \tag{3.25}$$

where $\alpha(S, S^{'}) = sup\{|P(AB) - P(A)P(B)| : A \in \mathcal{B}(S), B \in \mathcal{B}(S^{'})\}$.

The Hölder's inequality shows as,

$$|Cov(X, Y)| \leq ||X||_p||Y||_q \alpha^{\frac{1}{\gamma}}(X, Y), with \ \frac{1}{p} + \frac{1}{q} + \frac{1}{\gamma} = 1. \tag{3.26}$$

Taking $p = q = 2 + \kappa$ in Lemma 3.1, Eq (3.26) can be rewritten as,

$$|E(X_iX_j)| \leq ||X_i||_{2+\kappa}||X_j||_{2+\kappa} \ \alpha^{\frac{\kappa}{2+\kappa}}(d(S_i, S_j)). \tag{3.27}$$

Now we show the proof for **F** in Eq (3.23).

Since $s_i \neq s_j$, **F** follows that

$$\begin{aligned}
\mathbf{F} &= (N^2h_N^4)^{-1} \sum_{i,j:0<d(s_i,s_j)\leq P_N} K(\frac{s_i - s_0}{h})K(\frac{s_j - s_0}{h})E(X(s_i)X(s_j)) \\
&+ (N^2h_N^4)^{-1} \sum_{i,j:d(s_i,s_j)>P_N} K(\frac{s_i - s_0}{h})K(\frac{s_j - s_0}{h})E(X(s_i)X(s_j)) \\
&\leq C(N^2h_N^4)^{-1} \sum_{i,j:0<d(s_i,s_j)\leq P_N} K(\frac{s_i - s_0}{h})K(\frac{s_j - s_0}{h})\sigma_X^2 \\
&+ C(N^2h_N^4)^{-1} \sum_{i,j:d(s_i,s_j)>P_N} K(\frac{s_i - s_0}{h})K(\frac{s_j - s_0}{h})\sigma_X^2,
\end{aligned} \tag{3.28}$$

where $\sum_{i,j:0<d(s_i,s_j)\leq P_N}$ refers to the summation over $\{(i, j) : 1 \leq i, j \leq N, 0 \leq d(s_i, s_j) \leq P_N\}$, the cardinality of which is controlled by $C(P_N/\delta_N)^2$, and C is a generic finite positive constant that may differ at different places. Thus, the above Eq (3.28) continues

$$
\begin{aligned}
\mathbf{F} &\leq (N^2h_N^4)^{-1}CN(P_N/\delta_N)^2 \int\int K(\frac{s-s_o}{h})K(\frac{s^{'}-s_o}{h})f_S(s)f_S(s^{'})dsds^{'} \\
&+ c(N^2h_N^4)^{-1}\sum_{t=P_N}^{\infty}\alpha(t)^{\kappa/(2+\kappa)}\sum_{i=1}^{N}\sum_{j:d(s_i,s_j)\leq t}K(\frac{s_i-s_0}{h})K(\frac{s_j-s_0}{h}) \\
&\leq (N^2h_N^4)^{-1}CN(P_N/\delta_N)^2O(h^4) \\
&+ c(N^2h_N^4)^{-1}\sum_{t=P_N}^{\infty}\alpha(t)^{\frac{\kappa}{2+\kappa}}N(\frac{t}{\delta_N})^2\frac{1}{N(\frac{t}{\delta_N})^2}\sum_{i=1}^{N}\sum_{j:t\leq d(s_i,s_j)\leq(t+1)}K(\frac{s_i-s_0}{h})K(\frac{s_j-s_0}{h}) \\
&\leq O(1)\{(P_N/\delta_N)^2N^{-1}\} + O(1)\{\delta_N^2Nh^4\}^{-1}\sum_{t=P_N}^{\infty}t^2\alpha(t)^{\kappa/(2+\kappa)}
\end{aligned}
\tag{3.29}
$$

Let P_N be the integer part of $(\delta_N^2Nh^4)^{-1/\gamma}$ for the $\gamma > 0$ specified in Assumption A (i), by which the second part of Eq (3.29) tends to zero as $N \to \infty$. Now note that

$$
(P_N/\delta_N)^2N^{-1} = \{\delta_N^2Nh^4\}^{-\frac{2}{\gamma}}(\delta_N^2N)^{-1} = ((N\delta_N^2)^{(\gamma+2)}h^8)^{-1/\gamma}, \tag{3.30}
$$

which tends to zero, following Assumption D (i) (3). Combining another two assumptions in Assumption A (i) and D (i) (2), Eq (3.29) tends to zero when N goes to infinity. Therefore, from Eq (3.30) and Eq (3.22), D11.d in Eq (3.20) tends to zero as $N \to \infty$.

To prove the remaining terms in Eq (3.19), i.e., D11.*a*, D11.*b* and D11.*c*, we can firstly simplify them as follows

$$
\begin{aligned}
D11.a =& \mu_{uu}^{''}(s_0)\iint u^2K(u,v)dudv(1+o(1)), \\
D11.b =& \mu_{uv}^{''}(s_0)\iint uvK(u,v)dudv(1+o(1)) \\
D11.c =& \mu_{vv}^{''}(s_0)\iint v^2K(u,v)dudv(1+o(1)).
\end{aligned}
\tag{3.31}
$$

Similar to the proof of D11.*d*, the three elements in Eq (3.31) can be proven converging to 0 as the sample size N tends to infinity. So as the remaining two elements D21 and D31 in Eq (3.17). We have shown the relations between the $\hat{\mu}(s_0)$ and $\mu(s_0)$ in (i), that is

$$
E(\hat{\mu}(s_0) - \mu(s_0))^2 \to 0, \text{ as } N \to \infty, \tag{3.32}
$$

which deduces that

$$\hat{\mu}(s_0) \xrightarrow{\text{P}} \mu(s_0), \text{ as } N \to \infty. \tag{3.33}$$

We now focus on (ii). $\gamma_{\hat{X}}(s_i - s_j)$ and $\gamma_X(s_i - s_j)$ are the two semivariogram functions from the de-trended residual process $\hat{X}(S_i)$ and the true residual process $X(S_i)$ from the population, respectively.

Under the assumption of intrinsically stationary, by definition, the semivariogram function of the de-trended residual process can be shown as

$$2\,\gamma_{\hat{X}}(s_i - s_j) = E(\hat{X}(s_i) - \hat{X}(s_j))^2. \tag{3.34}$$

Similarly, we write the semivarigram function of $\gamma_X(s_i - s_j)$ as

$$2\,\gamma_X(s_i - s_j) = E(X(s_i) - X(s_j))^2. \tag{3.35}$$

By the decomposition rule introduced in Chapter 2, we have

$$Y(s_i) = \hat{\mu}(s_i) + \hat{X}(s_i) = \mu(s_i) + X(s_i). \tag{3.36}$$

Plugging Eq (3.36) into Eq (3.34), we obtain

$$\begin{aligned} 2\,\gamma_{\hat{X}}(s_i - s_j) =& E((\underbrace{\mu(s_i) - \mu(s_j) + \hat{\mu}(s_j) - \hat{\mu}(s_i)}_{A})^2 + (\underbrace{X(s_i) - X(s_j)}_{B})^2 \\ &+ 2(\underbrace{\mu(s_i) - \mu(s_j) + \hat{\mu}(s_j) - \hat{\mu}(s_i)}_{A})(\underbrace{X(s_i) - X(s_j)}_{B}) \\ =& E(\underbrace{\mu(s_i) - \mu(s_j) + \hat{\mu}(s_j) - \hat{\mu}(s_i)}_{A})^2 + E(\underbrace{X(s_i) - X(s_j)}_{B})^2 \\ &+ 2E(\underbrace{\mu(s_i) - \mu(s_j) + \hat{\mu}(s_j) - \hat{\mu}(s_i)}_{A})(\underbrace{X(s_i) - X(s_j)}_{B}). \end{aligned} \tag{3.37}$$

From Eqs (3.35) and (3.37) , we obtain the difference between the two semivariograms $\gamma_{\hat{X}}(s_i - s_j)$ and $\gamma_X(s_i - s_j)$ as:

$$2(\gamma_{\hat{X}}(s_i - s_j) - \gamma_X(s_i - s_j)) = E(\underbrace{\mu(s_i) - \mu(s_j) + \hat{\mu}(s_j) - \hat{\mu}(s_i)}_{A})^2 + 2E(\underbrace{\mu(s_i) - \mu(s_j) + \hat{\mu}(s_j) - \hat{\mu}(s_i)}_{A})(\underbrace{X(s_i) - X(s_j)}_{B}). \tag{3.38}$$

Given the result from Eq (3.33), we can conclude that the difference between two variograms in Eq (3.38) tends to zero when the number of sampled locations N goes to infinity, which is,

$$\gamma_{\hat{X}}(s_i - s_j) - \gamma_X(s_i - s_j) \xrightarrow{\text{p}} 0, \quad \text{as } N \to \infty, \tag{3.39}$$

hence there is the completion of the proof for (ii).

Applying the result in Eq (3.39) into Eqs (3.12) and (3.13), which share the same structure and elements from the two identified spatial processes. Based on their definitions, it is understood that

- $\gamma_{\hat{X},0} = (\gamma_{\hat{X}}(s_0 - s_i), ..., \gamma_{\hat{X}}(s_0 - s_N))^T$ and $\gamma_{X,0} = (\gamma_X(s_0 - s_i), ..., \gamma_X(s_0 - s_N))^T$. With Eq (3.39), we have $\gamma_{\hat{X},0} \xrightarrow{p} \gamma_{X,0}$, as $N \to \infty$.
- Similarly, $\Gamma_{\hat{X}} = (\gamma_{\hat{X}}(s_i - s_j)) \xrightarrow{p} \Gamma_X = (\gamma_X(s_i - s_j))$, as $N \to \infty$.
- $E(\hat{X}_i - X_i)^2 = E[(Y_i) - \hat{\mu}(s_i) - (Y_i - \mu(s_i))]^2 = E(\hat{\mu}(s_i) - \mu(s_i)) \overset{(3.33)}{\to} 0$, as $N \to \infty$. Thus $\hat{\boldsymbol{X}} = (\hat{X}_1, ..., \hat{X}_N)^T \xrightarrow{p} \boldsymbol{X} = (X_1, ..., X_N)^T$, as $N \to \infty$..

Therefore, it is concluded that,

$$\hat{X}^{OK}(s_0) - \tilde{X}^{OK}(s_0) = \gamma_{\hat{X},0}\Gamma_{\hat{X}}^{-1}\hat{X} - \gamma_{X,0}\Gamma_X^{-1}X \xrightarrow{\text{p}} 0, \text{ as } N \to \infty. \tag{3.40}$$

Plugging the asymptotic results of Eqs (3.33) and (3.40) into Eqs (3.5) and (3.6), we prove the Theorem 3.1.

3.4 Application of NTUK method to air quality data

In this section, we will apply the proposed NTUK method to the same benchmark air quality data used in Chapter 2.

3.4.1 Examination of the air quality data

It is well known that, in reality, observed data most likely does not strictly follow a Gaussian profile. For this reason, it is essential to check the input data before assigning it to a linear model. In this section, the normality of the de-trened air quality data is firstly tested, we will then apply the nonparametric-trend universal Kriging (NTUK) method to the de-trended data.

We use the nonparametric regression function available in the R package *sm* to estimate a nonlinear spatial trend $\hat{\mu}(s)$ by applying the local linear fitting method (Bowman and Azzalini, 1997), the result of the estimated trend in the region of England is shown in Figure 3.1. It appears that the spatial trend tends to be relatively stable in the centre area which is understandable, whilst the variations around the edges of the studied space are noticeably large.

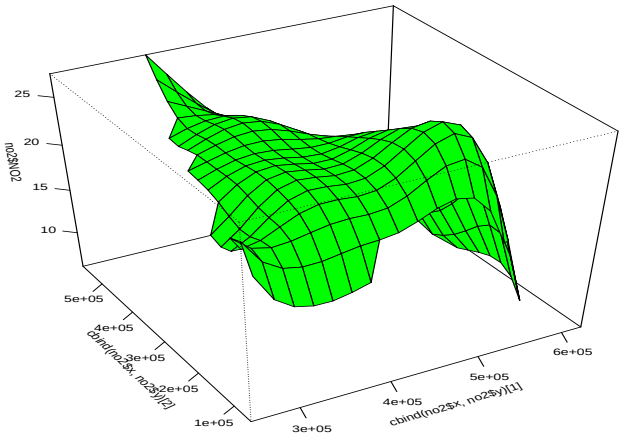

FIGURE 3.1: Nonparametric estimation of spatial trend $\hat{\mu}(S)$.

A kernal density estimation of the de-trended data, subtracting the estimated spatial trend from the observations, is shown in Figure 3.2 in which a dotted Gaussian density curve is added with the equal mean and variance from the de-trended data. It is shown that the de-trended residual part $\hat{X}(s)$ does not fit the matching Gaussian distribution.

3.4.2 NTUK to air quality data and comparison

The second step of the NTUK method is to apply linear ordinary Kriging to the de-trended residual data, despite its proven non-linearity property. We compare its prediction performance with the three linear Krigings using the cross validation criterion, more specifically the leave-one-out Cross Validation (LOOCV) technique as introduced in Chapter 2.

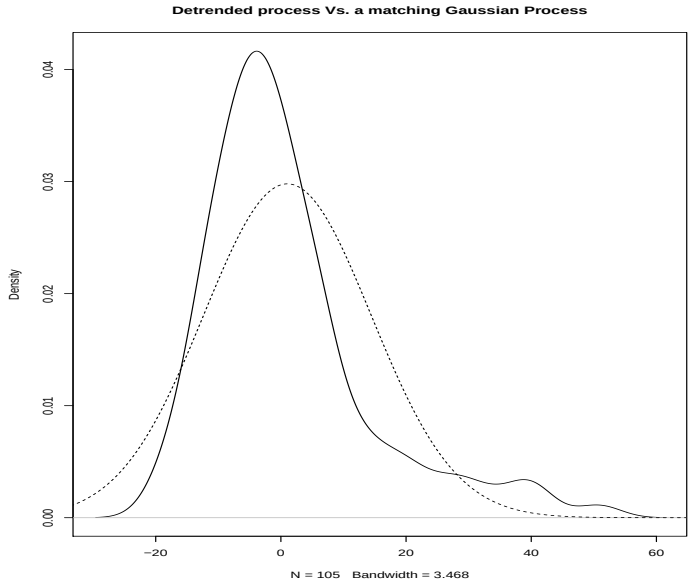

FIGURE 3.2: De-trended air quality profile Vs. a matching Gaussian curve.

Mean squared prediction error (MSPE), in Eq (3.41), calculates the mean squared differences between the observed values $Y_{obs}(S_i)$ and their estimators $\hat{Y}(S_i)$ from the LOOCV criterion,

$$\text{MSPE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}(S_i) - Y_{obs}(S_i))^2. \tag{3.41}$$

Table 3.1 shows a comparison of the mean squared prediction error results from the four Kriging methods. We get the MSPE value from the NTUK method as 213.6436, which is significant less than the same measurements from three linear Krigings in Chapter 2. We conclude that the newly proposed NTUK method outperforms the group of linear Kriging methods introduced so far on the air quality data set.

Despite the nonlinearity in the de-trended residual data, linear ordinary Kriging is still used for the NTUK method in this chapter for the reason that there is no suitable nonlinear method available up to this point. To overcome this constraint, we will show a novel family of semiparametric model averaging marginal Kriging methods in the

TABLE 3.1: A comparison of mean squared errors from the NTUK and three linear Kriging methods.

Kriging Methods	Mean Squared Prediction Error
Nonparametric-trend universal Kriging	213.6436
Simple Kriging (linear)	231.9582
Ordinary Kriging (linear)	232.7272
Universal Kriging (linear)	233.0495

next chapter, aiming to introduce nonlinear models for the de-trended residual process $X(s)$.

Chapter 4

Semiparametric- Model Averaging Marginal Kriging

In Chapter 3, the linear spatial trend is replaced by a nonparametrically fitted function $\hat{\mu}(s)$, see Eq (4.1). However, linear Kriging methods are still employed in the NTUK procedure for Kriging the stochastic process $X(s)$. In this chapter, our contribution is to compare and propose a new semiparametric Kriging model for the stochastic process $X(s)$ by taking some potential nonlinear features of $X(s)$ into account, following the initial works from Lu and Tjøstheim (2014) and Li et al. (2015).

$$Y(s) = \mu(s) + X(s). \tag{4.1}$$

4.1 Background

We assume that there are N spatial sampling locations, denoted as $S^{'} := \{s_1, ..., s_N\}$, in the studied space S. The sampled data of the stochastic process $\hat{X}(s)$ is obtained by subtracting the predicted spatial trend data $\hat{\mu}(s)$ from the observed values of $Y(s)$ as Eq (4.1). The benefit of such a decomposition is that stationarity properties and other assumptions may thus be supposed to make prediction of this stochastic spatial process feasible from the sampled observations. One could otherwise make no progress with only one set of sampled data available (Bivand et al., 2008). In this section, we will review some possible nonparametric prediction methods for this task.

From the model selection perspective, we re-examine this research work. Between two common categories of regression models, the parametric and nonparametric models, the linear Kriging methods introduced in Chapters 2 and 3 belong to the former within which the response takes the form of a weighted sum of the observations of all individual predictors. The benefits of such models are discussed intensively in Chapter 2. However, when the described underlying process does not follow a Gaussian profile,

i.e., the linearity assumption is violated, the disadvantages of this category of models become significant such as higher prediction variances when extreme values exist, and substantial misfittings around the edges of the space.

Alternative parametric methods were developed for more complex cases, often when the number of regressors is large, for instance in time and spatial series cases, a large number of lags can enter into the model. One common approach by the generalised linear models (GLMs) is to test a set of candidate models, then rank their prediction results by some comparison criteria, e.g., the Akaike Information Criterion (AIC) (Akaike, 1998) and Bayesian Information Criterion (BIC) (Schwarz et al., 1978).

Recently, introducing penalisation devices into a model becomes another focus in research for dimensionality reduction. Following some selection criteria, this approach forcefully lessens the solution space by assigning zero to the weights of many noise variables. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (Lasso) by imposing the L_1 penalty. Further, L_p penalty and other penalty forms on likelihood estimation were expanded, see Fu (1998) and Fan and Li (2001). For more insights on model averaging and covariates selection, Claeskens et al. (2008) and Fan and Lv (2010) are two recent references of this kind.

Despite the development in parametric methods, however in temporal and spatial modeling, the numbers of regressors and time or spatial lags may take infinite values and restrictive assumptions are hard to make. Moreover, evidently in many research fields involving live or dynamic systems, such as Geostatistics or Econometrics where the data usually exist as the outcome of a stochastic process, as White (2014) claimed, the response and regressors may exhibit nonnormality or heteroskedasticity and often serial correlation of unknown form. When the classical assumption of linear model is violated, however some nonparametric approaches have been proved to be more effective for owning less restrictive forms and assumptions. In this chapter, we will focus on this category of models.

Let $\{X(s)\}$ be a stationary spatial series process, we denote $X_s := (X(s_1), X(s_2), ..., X(s_N))$ as a N-dimensional random row vector representing observations at N sampled locations, and X_0 represents the value of the response at a new random location s_0 , i.e. $X_0 = X(s_0)$, $s_0 \in S$. Applying nonparametric methods, the regression function of $E(X_0|X_s = x_s)$ with $x_s = (x(s_1), x(s_2), ..., x(s_N))$, as Li et al. (2015) stated, can be well estimated when the dimension N is small , i.e., $N \leq 3$. When a higher dimension is involved, the result however turns to be far less reliable due to the impact of curse of dimensionality. More detailed discussion on this phenomenon can be found in Chapter 5. Among the recent literature on this topic (see Fan and Yao (2003); Li and Racine (2007)), we focus on one specific framework introduced by Li et al. (2015), the model averaging marginal regression (MAMAR) procedure. It was introduced as a semiparamatric multivariate regression function for time series and constructed by

an affine combination of one-dimensional marginal regression functions, often called the smooth functions. The asymptotic normality of this model under the condition of having fixed number of regressors was established with a convergence rate of $\sqrt{n}$. In this chapter, we aim to adopt this model structure into a spatial series setting for spatial interpolation (Kriging).

4.2 Semiparametric- full model averaging marginal Kriging

We consider a scenario in which only one observation is available at the location s_1, and we intend to make use of its information alone for predicting the response at a random new location s_0, for $s_1, s_0 \in S$. Denoting a Borel measurable function g_1 of $X(s_1)$ on $\mathbb{R}$, we seek the minimiser $\hat{g}_1(X(s_1))$ that minimises the mean squared prediction error satisfying

$$\underset{g_1}{\text{Minimise}} \quad E[X(s_0) - g_1(X(s_1))]^2. \tag{4.2}$$

Intuitively, the optimal solution of minimising this mean squared prediction error (MSPE), denoted as $Z(s_1)$, can be shown using the conditional expectation on $X(s_1) = x(s_1)$,

$$Z(s_1) := g_1^0(X(s_1)) = E[X(s_0)|X(s_1) = x(s_1)]. \tag{4.3}$$

Now escalating the above idea to a more general N observations case, i.e., $x(s_1), x(s_2)$,..., $x(s_N)$ are observed at N spatial locations $s_1, s_2, ..., s_N \in S$, we can update Eqs (4.2) and (4.3) in the form of N-observations and its solution are as follows:

$$\underset{G}{\text{Minimise}} \quad E[X(s_0) - G(X(s_1), X(s_2), ..., X(s_N))]^2, \tag{4.4}$$

$$G^0(X) = E[X(s_0)|X(s_1) = x(s_1), X(s_2) = x(s_2), ..., X(s_N) = x(s_N)], \tag{4.5}$$

where $G^0(X)$ refers to the Borel function on $\mathbb{R}^N$ with all the N known observations as the condition that minimises the MSPE in Eq (4.4).

In the interest of finding a solution for this multivariate conditional expectation problem in Eqs (4.4) and (4.5), from countless possibilities, we implement the model averaging marginal regression (MAMAR) method proposed by Li et al. (2015) under the approach of additive modelling. The general idea of this method is to approximate this conditional expectation result by a sum of some low-dimensional conditional expectations, i.e., $E(X_0 \,|X_{(k)} = x_{(k)})$ with $x_{(k)}$ represents a subset of $(x_1, x_2, ..., x_N)^T$ with the size of k, which can be well estimated by nonparametric methods when the dimension

k is small (say less than 4), we can thus approximate the conditional expectation in Eq (4.5) by

$$G^0(X) \approx \omega_0 + \sum_{k=1}^{K} \omega_k^0 \, E(X_0 | X_{(k)} = x_{(k)}), \tag{4.6}$$

where ω_0 is the constant trend (or mean), and the summand is an affine combination of conditional regression functions with weight $\omega_k \in \mathbb{R}$ to each $X_{(k)}$, $1 \leq k \leq K \leq N$. Here each $X_{(k)}$ can be a subset of the random vector X. Theoretically these $X_{(k)}$ have the choice of taking a combination of any elements in X, however in reality one would not set them to be high-dimensional terms, in order to avoid the effect of curse of dimensionality.

Li et al. (2015) described this method as an approximation of the true conditional regression function. Explicitly, within the family of models in Eq (4.6), there is the true solution to be approximated. Moreover as its name suggests, structurally this approach can be treated as a model averaging method. By identifying $\omega^0 := (\omega_0^0, \omega_1^0, \omega_2^0, ..., \omega_K^0)^T$, the optimal solution of Eq (4.6) can therefore be approximated.

Among all possible models by this framework, a special solution was proposed with $K = N$, and $X_{(k)} = X_k$, the exact k-th component in X, thereby no overlappings among the covariates. Let $m(x) := E(X_0 | X = x)$ denotes the true conditional expectation, we transfer Eq (4.6) into the form of

$$m^0(x) = \omega_0^0 + \sum_{k=1}^{K=N} \omega_k^0 E(X_0 | X_k = x_k), \tag{4.7}$$

where $m^0(x)$ is an approximation of $m(x) := E(X_0 | X = x)$ and $1 \leq k \leq K = N$.

In this N one-dimensional conditional marginal regression setting, Li et al. (2015) claimed that a unique solution ω^0 is generally true; subsequently with these expectation components estimated by some nonparametric techniques, we can approximate the true response value at the new location of interest s_0. In the next section, we will follow this context introducing a semiparametric approximation procedure for spatial Kriging.

4.2.1 Approximation

Following the initial introduction, we will implement this model averaging marginal regression (MAMAR) method to our spatial series data, i.e., the air quality residual process $X(s)$. The aim is to predict $X(s_0)$ from the de-trended observations of $X(s_1)$, $X(s_2)$,..., $X(s_N)$ sampled at the N known locations. Applying the Eqs (4.4) and (4.7), we solve the problem of minimising the mean squared prediction error of $X(s_0)$,

$$\underset{\omega_k,\ 1\leq k\leq N}{\text{Minimise}} \quad E\Big[X(s_0)-\omega_0-\sum_{k=1}^{N}\omega_k E(X(s_0)|X(s_k)=x(s_k))\Big]^2, \tag{4.8}$$

where $\omega_0^0 = 0$ after the detrended step. In fact, the nonparametric detrending process in Chapter 3 can be viewed as the answer for estimating ω_0^0. The elements in the vector ω_k^0, $k = 1, 2, ..., N$ are the approximated weights of the N one-dimensional expectation components. To solve Eq (4.8), ideally, we wish to estimate all the coefficients in $\{\omega_k^0\}$, $k = 1, ..., N$.

Similar to the settings for linear Krigings in Chapter 2, two assumptions for the process of $X(s)$ and a newly identified process $Z(s)$ are made for the new Kriging method:

1. The detrended residual process $X(s)$ is supposed to be a strictly stationary process, for $s \in S, s, d_1, d_2, ..., d_n \subset \mathbb{R}^2$, the joint distribution of $\{X(s), X(s+d_1), X(s+d_2), ..., X(s+d_n)\}$ depends only on the spatial lag d, but not on its initial location s (see Davidson (1994)). Furthermore, after removing the spatial trend $\mu(s) = E(Y(s))$, the global mean of $X(s)$ is constant and centred, i.e., $E[X(s)] = 0$,

2. The data of $Z(s) := E(X(s_0)|X(s))$, as defined in Eq (4.3), comes from a second-order stationary process with a finite covariance function $C(d) := Cov(Z(s), Z(s+d)), \forall s\ , s+d \in S$,

where d denotes the spatial lag between any pair of spatial locations in space S. In this thesis, the Euclidean distance (L^2-norm of d) is used as the measurement of the spatial distance between any pair of spatial locations.

The solution of this multiple linear regression problem, as Wackernagel (2003) and Li et al. (2015) stated, can be shown as

$$\begin{pmatrix}\omega_1^0\\ \omega_2^0\\ \vdots\\ \omega_N^0\end{pmatrix} = \begin{pmatrix} cov(Z(s_1),Z(s_1)) & \cdots & cov(Z(s_1),Z(s_N))\\ cov(Z(s_2),Z(s_1)) & \cdots & cov(Z(s_2),Z(s_N))\\ \vdots & \vdots & \vdots\\ cov(Z(s_N),Z(s_1)) & \cdots & cov(Z(s_N),Z(s_N))\end{pmatrix}^{-1} \begin{pmatrix} cov(Z(s_1),X(s_0))\\ cov(Z(s_2),X(s_0))\\ \vdots\\ cov(Z(s_N),X(s_0))\end{pmatrix}, \tag{4.9}$$

where the process $Z(s_k) := E(X(s_0)|X_k = x(s_k))$, $k = 1, 2, ..., N$, is supposed to be second-order stationary. In addition to the positive semi-definiteness of the covariance matrix in Eq (4.9), we assume it is also a non-degenerate covariance matrix, i.e., it is strictly positive definite. This is commonly recognised in this kind of applications to ensure a feasible inverse operation.

We can re-form the $Cov(Z(s_k), X(s_0))$ vector as $Cov(Z(s_k), X(s_0) - E(X(s_0)|X(s_k)) + E(X(s_0)|X(s_k)))$. Based on the orthogonality condition, this covariance between the two processes, $Cov(Z(s_k), X(s_0))$ can be further derived into $cov(Z(s_k), Z(s_k))$, which is indeed the variance of $Z(s_k)$, the diagonal elements of the covariance matrix. Now we have the updated equation for the weights vector ω^0 as

$$\begin{pmatrix} \omega_1^0 \\ \omega_2^0 \\ \vdots \\ \omega_N^0 \end{pmatrix} = \begin{pmatrix} cov(Z(s_1), Z(s_1)) & \cdots & cov(Z(s_1), Z(s_N)) \\ cov(Z(s_2), Z(s_1)) & \cdots & cov(Z(s_2), Z(s_N)) \\ \vdots & \vdots & \vdots \\ cov(Z(s_N), Z(s_1)) & \cdots & cov(Z(s_N), Z(s_N)) \end{pmatrix}^{-1} \begin{pmatrix} Var(Z(s_1)) \\ Var(Z(s_2)) \\ \vdots \\ Var(Z(s_N)) \end{pmatrix}. \quad (4.10)$$

With the predicted weights, the model averaging marginal Kriging prediction of $X(s_0)$ finally takes the form of,

$$\tilde{X}(s_0) = \sum_{k=1}^{K=N} \omega_k^0 E[X(s_0)|X_k = x(s_k)] = \sum_{k=1}^{K=N} \omega_k^0 Z(s_k). \quad (4.11)$$

We need to uncover the process $Z(s) := E(X(s_0)|\ X(s))$ and its covariance matrix to solve Eq (4.11). Since only one realisation of the detrended residual process $X(s)$ is available, the result may not be directly estimated due to the extreme curse of dimensionality, a N-variables Versus N-observations scenario. Adopting the additive approximation concept from Lu et al. (2007) and Mammen et al. (1999), in this section we will establish a three-stage iterative procedure for this task, which are (1) the Marginal regression function estimation, (2) Spatial prediction of covariance matrix, and (3) Nonparametric kernel bandwidth selection. We name this procedure the Semiparametric-full Model Average Marginal Kriging (SFMAMK). Here, the full refers to $K = N$; we will later introduce a model with $K < N$, i.e., the K-radius neighbouring average based marginal (KNAMA) Kriging, in Section 4.3.

4.2.2 Marginal regression function estimation

We start this procedure by estimating the conditional marginal function of $Z(s_k) := E(X(s_0)|X(s_k))$, $k = 1, 2, ..., N$. The name marginal refers to the N one-dimensional conditional functions in Eq (4.8).

Based on the definition, assuming a continuous conditional density function $f(\cdot)$ exists in space S, the studied individual regression mean function can be shown as

$$E\left[X(s_0)|\ X_k = x_k\right] = \int u f_{0|k}(u|x_k)du, \quad k = 1, 2, ..., N, \quad (4.12)$$

where $X_k := X(s_k)$, $x_k := x(s_k)$ and $f_{0|k}(u|x_k)$ is the conditional probability density function of X_0 given $X_k = x_k$. Our aim is to use the de-trended residual values to estimate this conditional density function at s_0.

We approximate the integral in Eq (4.12) by a discrete form,

$$\hat{E}[X(s_0)|X_k = x_k] \approx \sum_{\ell=2}^{N} u_\ell \hat{f}_{0|k}(u_\ell|x_k)(u_\ell - u_{\ell-1}), \quad k = 1, 2, ..., N, \tag{4.13}$$

where the u_ℓs are the order statistics of the detrended values, i.e., $x(s_\ell)s$ are sorted in an ascending order, which are $u_1 := min_{1\leq k\leq N}\{\hat{x}_i\}$, $u_N := max_{1\leq k\leq N}\{\hat{x}_i\}$ and $u_{\ell-1} < u_\ell$, for $\ell = 2, ..., N$.

To estimate the conditional density functions in Eq (4.13), we define

$$\hat{f}_{0|k}(u|x_k) = \frac{\hat{f}_{0,k}(u, x_k)}{\hat{f}(x_k)}, \tag{4.14}$$

where $\hat{f}(x_k)$ is the estimator of the marginal density function $f(x)$ of X_k that can be easily constructed as

$$\hat{f}(x) = N^{-1} \sum_{i=1}^{N} K_h(X_i - x), \tag{4.15}$$

where $K_h(x) = h^{-1}K(x/h)$ with a kernel function $K(\cdot)$ on $\mathbb{R}$ and a bandwidth $h = h_N \to 0$, as $N \to \infty$.

The numerator is the estimator of a joint density function $f(x, y; s_0)$ of $X(s)$ and $X(s + s_0)$ that characterises the nonlinear, non-Gaussian spatial dependence, with $s_0 \neq (0, 0)$. Intuitively, it can be shown as

$$\hat{f}(x, y, s_0) = \frac{1}{n_0} \sum_{j,\ell \in S_0} K_h(X_j - x)K_h(X_\ell - y), \tag{4.16}$$

where $S_0 := \{(j, \ell) : s_j - s_\ell = s_0, j, \ell = 1, ..., N\}$ and $n_0 =^\sharp S_0$, the cardinality of the set S_0. However, as Lu and Tjøstheim (2014) pointed out in practice, this cardinality can be very small for most spatial distances and may equals to 0. An alternative solution has to be introduced. In this study, we adopt a modified definition stated by Lu and Tjøstheim (2014), that is

$$\hat{f}_{i,k}(x, y) = \hat{f}(x, y; s_i - s_k) = \frac{\sum_{j,\ell=1}^{N} L_b(s_j - s_\ell - (s_i - s_k))K_h(X_j - x)K_h(X_\ell - y)}{\sum_{j,\ell=1}^{N} L_b(s_j - s_\ell - (s_i - s_k))}, \tag{4.17}$$

where $L_b(s) = b^{-2}L(s/b)$ with $L(\cdot)$ is a kernel function on $\mathbb{R}^2$, and a bandwidth $b = b_N \to 0$, as $N \to \infty$.

The idea of introducing this spatial smoothing in Eq (4.17) is associated with the estimation of nonlinear dependence of the spatial process, when the monitoring locations are irregularly positioned which is different from the case of regularly gridded data. As a result, information from any pairs of observations with similar distances are retained and furthermore highlighted. Yet the estimation itself becomes more theoretical- and compute-intensive with the complexity of selecting multiple bandwidths. In Sections 4.2.4 and 4.4, we will propose approaches for bandwidth selection.

Until now, we are able to estimate all the elements on the RHS in Eq (4.13) to obtain the N conditional marginal functions $\hat{E}[X(s_0)|X_k = x_k]$, $k = 1, 2, ..., N$. In Section 4.2.3, we will estimate the ω^0 vector in Eqs (4.10) and (4.11).

4.2.3 Prediction of the spatial covariance matrix

In this section, we will show the derivation of the covariance matrix from the N estimated conditional functions, $\hat{Z}(x_k)$. It is clear that with only one set of estimations, one can not directly compute the required covariance matrix. Suitable stationary assumptions are thereby required to construct the spatial dependence in S.

The introduction of using spatial Kriging methods for this task is made intensively in Chapters 1-2. Imposing the assumption of second-order stationary for the process $Z(s)$, we suppose that it has an finite covariance function whose value is only related to the spatial lag d between any pair of geographic locations, i.e., $C(d) := Cov(Z(s), Z(s+d))$, $\forall s\ , s + d \in S$.

The objective of this stage is to obtain the $N \times N$ covariance matrix in Eq (4.10). We need to estimate $Cov(Z(x_i), Z(x_j))$, its (i, j) element, for $i, j = 1, ..., N$. By imposing suitable stationary assumptions, we are able to estimate a covariance function $\hat{C}(d)$ with respect to the lag d between spatial locations. Hence, the covariance matrix and the variance vector in Eq (4.10) can be discovered. In Chapter 2, we have explained the use of spatial variogram and covarance functions for Kriging. It is worthwhile to mention an important proposition describing the equivalence relation between the variogram and covariance functions (see Wackernagel (2003)).

Proposition: (Equivalence of variogram and covariance functions)

1. If $Z(s)$ is second-order stationary, i.e., there exists a covariance function $C(d)$ of $Z(x)$, then a varigram function γ_d can be deduced from $C(d)$ according to the formula below,

$$\gamma(d) = C(0) - C(d). \tag{4.18}$$

2. If $Z(x)$ is intrinsically stationary with a bounded variogram γ_d, i.e., $\gamma(\infty) := lim_{|d|\to\infty}\gamma(d) < \infty$, which denotes the lowest upper bound of an increasing variogram function, then a covariance function $c(d)$ can be specified as

$$C(d) = \gamma(\infty) - \gamma(d). \tag{4.19}$$

3. For a second-order stationary process $Z(x)$, both two properties stated above hold, and the variogram and the covariance are said to be equivalent.

Cressie (1993) stated that if $Z(s)$ is second-order stationary and if $C(d) \to 0$ as $|d| \to \infty$, then $\gamma(d) \to C(0)$ as $|d| \to \infty$ due to the stationary criterion. The value $C(0)$, or the sill, is equal to the variance of $Z(s)$. With this proposition, we should be able to covert one from the other between these two spatial functions.

As a beneficial supplement to Chapter 2, we expand the brief introductions made in Section 2.3.2 on fitting the theoretical Variogram based on the observations. There are three main steps for gaining the fitted variogram model from the sampled data.

The first step, described as a 'useful diagnostic tool' by Cressie (1993), is to draw the Variogram Cloud. Wackernagel (2003) named it as a measurement of the dissimilarity $\gamma^*_{i,j}$ of readings between two sampled locations. It is defined as,

$$\gamma^*_{i,j} := \frac{(\hat{z}(s_i) - \hat{z}(s_j))^2}{2}, \tag{4.20}$$

where $\hat{z}(s_i)$ and $\hat{z}(s_j)$ are estimated values at s_i and s_j, respectively, $i, j = 1, ..., N$.

Adding the stationarity assumptions, the calculation is updated depending on their spatial lag d, that is,

$$\gamma^*(d) := \frac{(\hat{z}(s_i) - \hat{z}(s_i + d))^2}{2}. \tag{4.21}$$

The numerator of Eq (4.21) ensures $\gamma^*(d)$ to be symmetric with respect to d. The presentation of the variogram cloud is to plot the dissimilarity against the spatial lags. There are a total of C_2^N entries shown on this scatter cloud providing an initial spatial information of this underlying process $Z(s)$. Figure (4.1) shows an illustrative example of this cloud using our estimated $\hat{z}(s)$ data.

Now we move to the second step to draw the Experimental Variogram. It is noticed that the dissimilarities in the variogram cloud are unevenly distributed, i.e., many lags d are not covered by the limited number of pairs of samples, also a small number of lags have more than one calculated dissimilarities from different pairs of locations. To overcome this challenge, Wackernagel (2003) suggested to divide all spatial distances

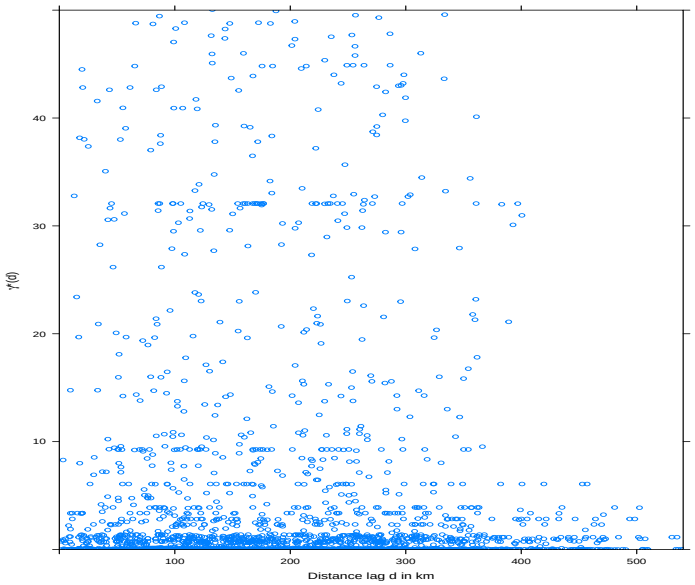

FIGURE 4.1: An illustration of the variogram cloud drawn from the estimated $\hat{z}(s_0)$, where $s_0 = s_2$

into G classes D_g, $g = 1, ..., G$, also known as lag intervals or bins, so $\cup_{g=1}^{G} D_g$ covers all the distances up to the cutoff distance, which Cressie (1993) recommended to take one-third of the maximum sampled distance, i.e., $\frac{1}{3} max_{i,j=1,...,N} |s_i - s_j|$ from the samples. The reason is that any dissimilarities after the cutoff distance may contribute little information for modelling the process (Bivand et al., 2008).

For each distance class D_g the corresponding average dissimilarity $\tilde{\gamma}^*(d)$ can therefore be determined. The $N(D_g)$ denotes the cardinality of the g-th distance class. As such, we have

$$\tilde{\gamma}^*(d) := \frac{1}{2N(D_g)} \sum_{N(D_g)} (z(s_i) - z(s_j))^2, \tag{4.22}$$

where $N(D_g) = \{(s_i, s_j) : |s_i - s_j| \in D_g \text{ for } i, j = 1, ..., N\}$ represents the set of all pairs of locations having the distances inside the class D_g.

One practical note though, the experimental variogram varies significantly to how the distance class is selected, so does to the choice of the cutoff value. The latter is the maximum distance chosen in the experimental variogram, which it is normally set as a far smaller value than the maximum distance measured on the samples. [1]

Before moving to the next step, we also like to discuss another pair of important concepts presented in Cressie (1993): the isotropic and anisotropic. In the isotropic case, the distance class D_g depends only on the Euclidean distance between two locations, or

[1] The practitioners are strongly advised to check the default settings for these parameters when using a spatial package in R.

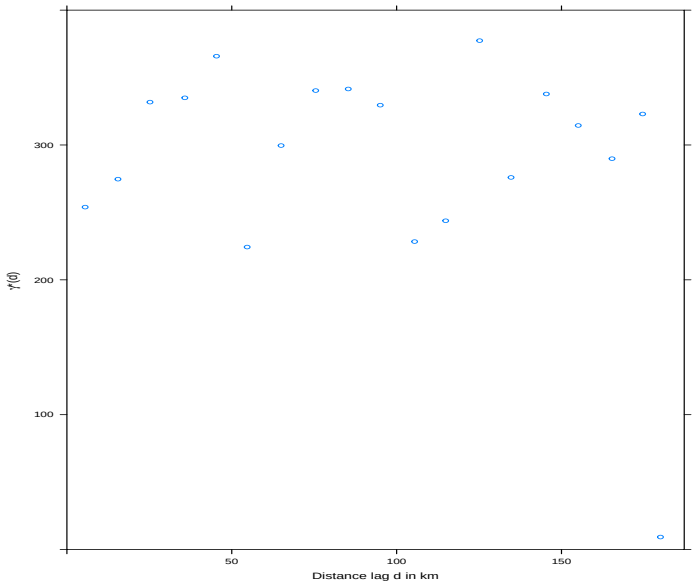

FIGURE 4.2: An illustration of the experimental variogram drawn from the estimated $\hat{z}(s_0)$, where $s_0 = s_2$

explicitly the studied property that invariant in relation to a particular direction; whilst in the anisotropic case, it has a second variable taken into account for identifying the distance class, which is the direction of the lag d. Webster and Oliver (2007) described that in Geostatistics when the underlying process performs differently between the horizontal and vertical directions, anisotropic model is more suitable than the isotropic. In our air quality case, we confine our attention to the former for the reason of simplicity without introducing extra factors such as the short-term wind direction.

The last step in this stage is to fit the captured experimental variogram from a few valid parametric covariance/variogram models, i.e. to replace the experimental variogram by a theoretical model. Among the countless possible fitting models, there are some models recommended often chosen in applied geostatistics. A list of the common candidates can be found in Wackernagel (2003) and Webster and Oliver (2007).

In general, there are some good signs from the empirical variogram indicating which theoretical model families to choose from. When estimating the parameters in a likely model, the ordinary or weighted least square fitting methods are widely applied. We notice that in the recommended models there are only a limited number of parameters left to be estimated; among them, the commonly used models are the sill, range and nugget, which were discussed in Chapter 2. In addition to the exponential model shown in Section 2.3.2, we add another commonly used fitting model, i.e., the spherical model.

Let $\gamma_{a,b}(d)$ denote the variogram function, $C_{a,b}(d)$ the corresponding covariance function with lag d and $a, b > 0$ the parameters of each model, where a, b represent the range parameter and the sill value, we have

$$\gamma_{a,b}^{\text{sph}}(d) := \begin{cases} b(\frac{3}{2}\frac{|d|}{a} - \frac{1}{2}(\frac{|d|}{a})^3), & \text{if } |d| \leq a, \\ b & \text{, otherwise.} \end{cases}$$

$$C_{a,b}^{\text{sph}}(d) := \begin{cases} b(1 - \frac{3}{2}\frac{|d|}{a} - \frac{1}{2}(\frac{|d|}{a})^3), & \text{if } |d| \leq a, \\ 0 & \text{, otherwise.} \end{cases}$$

With the estimated range a , sill b, and often combined with the nugget, the covariance matrix in Eq (4.10) can be easily computed, so does the variance vector. Subsequently the weight vector ω^0 is uncovered for estimating the N marginal regression functions.

4.2.4 Nonparamatric Bandwidth Selection

Combining the previous two stages, we now have a nearly complete iterative procedure for the proposed full Model Averaging Marginal Kriging (SFMAMK) method. The 'nearly' refers to the two free parameters yet to be decided: the spatial kernel bandwidths b and h in estimating the marginal and joint probability density functions in Eqs (4.15) and (4.17), respectively.

The purpose of introducing these spatial smoothing kernels, $K_h(x)$ and $L_b(s)$, is to construct the nonlinear spatial dependence of the $Z(s)$ process, in particular when the sampled locations are irregularly positioned. Intuitively, in this nonlinear Kriging method, we intend to emphasise the information extracted from the observed data, which is closely related to Z_0. The information includes two sources: one is the information of closeness in site with spatial dependence, i.e., the data Z_i whose site s_i is close to s_0; the other is the information of the closeness in value of spatial variable, that is the data Z_i whose value is close to Z_0 even though this is unobserved. These two kernels established will assist us in achieving these purposes.

In practice, there are various types of kernel functions to choose from, e.g., the Gaussian, uniform, triangular, Epanechnikov, biweight, etc. However, it is worth mentioning that the selection of a specific Kernel type has no significant impact on the prediction results when the sample size is large. The Epanechnikov function is the one selected in this chapter.

One feasible solution for the selection of the optimal bandwidths is cross-validation method, in particular, the Leave-one-out-cross-validation (LOOCV) that we have used before. Cross-validation, also called rotation estimation, is an effective model selection technique for choosing the best candidate out of the possible models by comparing some predetermined testing statistics. As its name suggested, each repetition of this LOOCV method is executed by selecting $N-1$ sampled observations, denoted as $\{X_{-i}\} := (X_1, ..., X_{i-1}, X_{i+1}, ..., X_K)$, $i = 1, ...N$, to be the training set, leaving the only

one sample X_i as the validation set. This iterative process runs in total N repetitions for $\forall i$, before the preset criteria are satisfied.

Given that in the first stage, the estimation of the marginal and joint probability density functions are strongly influenced by the choice of the bandwidths. The outcome, the estimated marginal mean expectation $\hat{Z}(s_i)$, effectively becomes a function of the bandwidth variables h and b, i.e., $\hat{Z}(s)$ depends on $h, b > 0$. In the second stage, the characteristic of this functional relation is passed on, and remained in $\hat{X}(s)$ in Eq (4.11). Applying LOOCV, one can compute the Mean Squared Prediction Error (MSPE) between the $\hat{X}(s)$ and $X(s)$. At last, we have the MSPE, as defined in Eq (2.15), is the function of (h, b), say $f(h, b)$, $h, b > 0$. The selection of suitable bandwidths now effectively becomes an optimisation problem of finding the minimiser of MSPE.

Unlike in a continuous case, the MSPE objective function of our full Model Averaging Marginal Kriging does not have a closed-form expression, which means that the objective function can not be summarised in an analytic solution owing mainly to the estimation steps of the covariance matrix in the second stage, i.e.,

$$\underset{h,\, b>0}{\text{Minimise}} \quad f(h, b). \tag{4.23}$$

We select the Nelder Mead method, a widely used nonlinear programming technique (NLP), for this multivariate optimisation problem. Nelder and Mead (1965) introduced this method for a minimisation of a function of n variables, which depends on the comparison of function values at the $(n + 1)$ vertices of a general simplex, followed by the replacement of the vertex with the highest value by another point. The simplex adapts itself to the local landscape, and contracts on to the final minimum. Figure 4.3 shows an illustration of how this heuristic algorithm works for a two-dimentional problem (Ozaki et al., 2017). Overall, this method has some nice features such as capable of working on gradient-free, non-differentiable functions, which are required in this procedure.

It is recognised that the initial choice of the bandwidths has a noticeable impact for this type of compute-intensive tasks in general. The kernel density function estimation in our SFMAMK procedure has no exception. Reducing the solution plane from existing knowledge of the problem and some experimental tests will help to reduce the overall running time of the computation.

One approach for seeking the initial point is to start with the data itself. Dehnad (1987) proposed a normal distribution approximation of h^*, or Silverman's rule of thumb for a near Gaussian function,

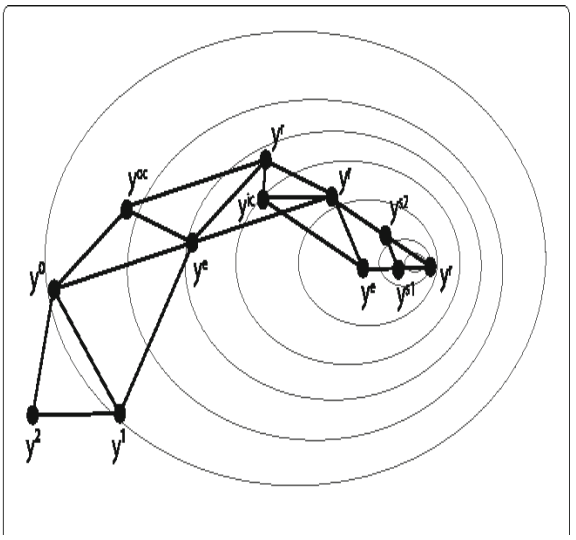

FIGURE 4.3: An illustration of how the Nelder Mead NLP method works in a two-dimensional local area.

$$h^* := (\frac{4\hat{\sigma}^5}{3N})^{-\frac{1}{5}}, \tag{4.24}$$

where the standard deviation $\hat{\sigma}$ and the sample size N are both from the observed data set. Despite the easy computation, this kind of estimator has to be applied with caution. Further adjustments on the bandwidths may be required for identifying more suitable initial starting point.

4.2.5 Trial run of the SFMAMK method to air quality data

So far, a complete procedure is developed for approximating the residual process of $\hat{X}(s)$. Specifically, we start estimating the full set of the one-dimentional marginal regression functions $\hat{Z}(s_i)$, for $i = 1, ..., N$, through the nonparametric estimation of probability density functions for irregularly observed spatial data proposed by Lu and Tjøstheim (2014). Then with the assumption of second-order stationarity, we fit the estimated $\hat{Z}(s)$ with a parametric covariance function to calculate the ω^0 vector in Eq (4.11) before completing this procedure. In this section, we will test this method to our spatial air quality data.

Beginning with the estimation of marginal mean functions, to avoid the extreme response values at which the density functions can be poorly estimated, in practice we adapt the suggestions from Lu and Tjøstheim (2014) to focus on the estimators of the density function at the points, $(x, y) \in [a, A] \times [a, A]$, where a and A are chosen as 1% and 99% sample quantiles of x_i's, respectively. To further simplify the computation, we replace the μ_l in Eq (4.13) with a series of $M = 50$ response values, denoted as $\{x_m\}$, $m = 1, ..., M$, evenly distributed within the identified range.

Figure 4.4 shows some initial results of the estimations of the marginal mean functions. The horizontal and vertical axes of the graphs represent the values of x_m and the marginal mean estimators, respectively. The graphs use doted lines to display the estimators at two new locations, i.e., s_1 on the left and s_2 on the right. The estimators, shown on the both sides, vary for each x_i, $i = 3, ..., 105$, taken as the conditions. Multiple bandwidths are tested with comparable results similar to the selected set of $(h, b) = (0.07, 8)$. The outcomes are approximately the same at other new locations.

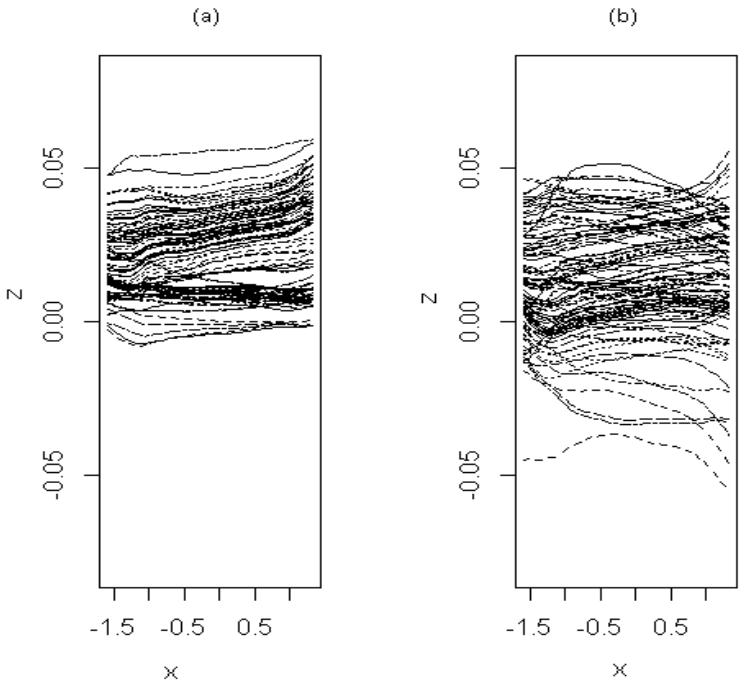

FIGURE 4.4: Estimates of $E(X(s_0)|X(s_k) = x)$, $k = 3, 4, ..., N = 105$, with $(h, b) = (0.07, 8)$: (a): $s_0 = s_1$, and (b): $s_0 = s_2$.

Less favourably, the result reveals a significant impact on the estimators when dissimilar locations are used as the condition, furthermore this dissimilarity is observed when s_0 varies irregularly in the sampled space. Both findings make the estimation of a unified covariance function in the second step, even if the functions are from the same model family, impossible at different new locations. The stationarity condition assuring a consistent covariance function may be violated in this case, i.e., extra factors might contribute to define the underlying covariance function other than merely the spatial distance. Further investigation in our air quality case is thereby required to seek for an alternative solution.

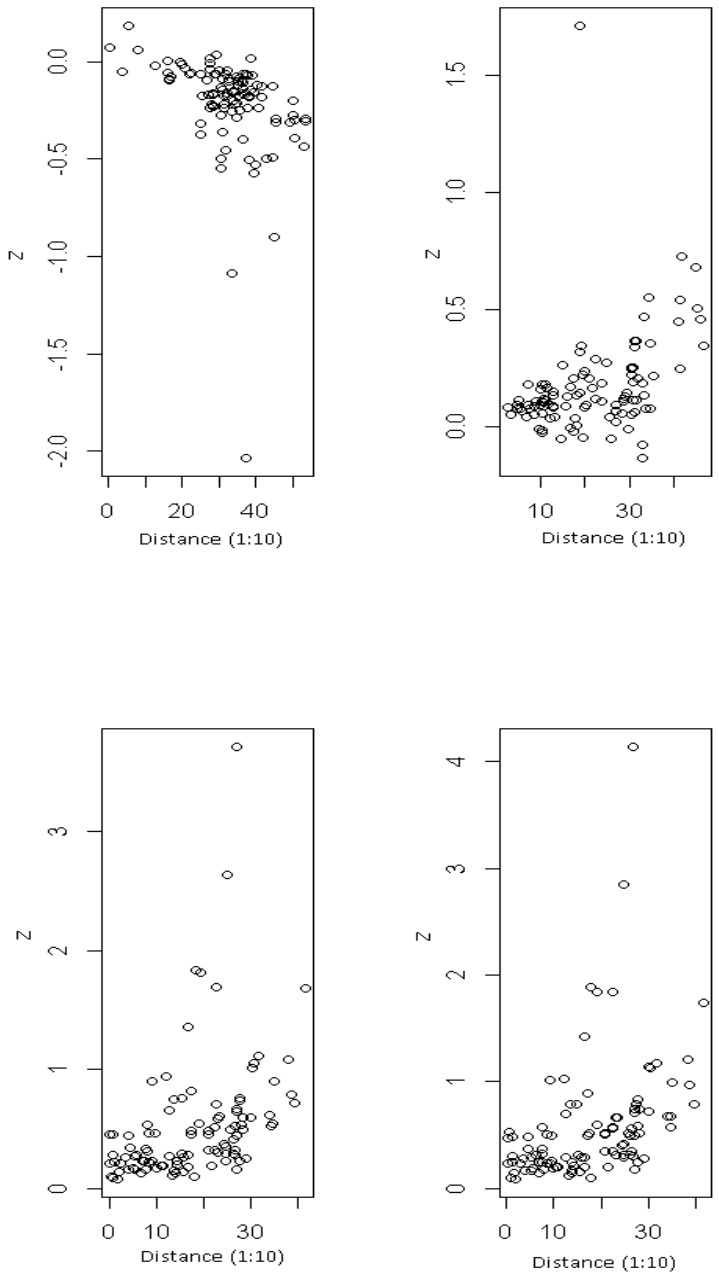

FIGURE 4.5: Estimates of $Z(s_k) = E[X(s)|X(s_k)]$ for $s = s_j$, $j = 97, 98, 101, 102$, in the four panels respectively, as a function of the distance between s_j and s_k, with $k \neq j$, based on their corresponding training set $\{s\}_{-j}$. Here $(h, b) = (0.07, 8)$ is the selected bandwidths for the estimations.

4.3 K-radius neighbouring average based marginal Kriging

We look further into our air quality case, an interesting sign of bifurcation, linked with the spatial distances between each pair of locations which gained our attention. Figure 4.5 illustrates a shared phenomenon that the estimates of the marginal mean functions

are more consistent when the conditioned locations are chosen from a close distance; conversely the estimates obtained from those far-distanced locations, say beyond a near 200 km mark, show substantial variations in the results. It is reasonably believed that the spatial locations of the monitoring sites may contribute to this challenge.

We recall the comments made by Bivand et al. (2008) and Cressie (1993) on the cutoff distance for the experimental variogram, that any dissimilarities after the cutoff distance may contribute little information for modelling the process, the 200 Km is indeed near the recommended cutoff distance in our air quality case, which takes one-third of the maximum sample distance.

The left side of Figure 4.6 shows the true locations of the air monitoring sites in England, whereas a detailed map of UK's metropolitan areas, posted by ESPON (2007), is on the right. It is clear that the monitoring sites are spread sparsely in rural areas, whereas a high density of sites are located in the main metropolitan districts. Moreover, from a nationwide point of view, the monitoring sites are also more centrally stationed in the centre and South East of England. It is a genuine belief that, because of many inherent relations, e.g., economic ties and geographical reasons, there are greater similarities shared in a same region than those from cross regions. Three major subnational divisions, London, West Midland and North West, are highlighted in Figure 4.6. As a reference factor, the direct distance between London and Manchester is about 260 Km.

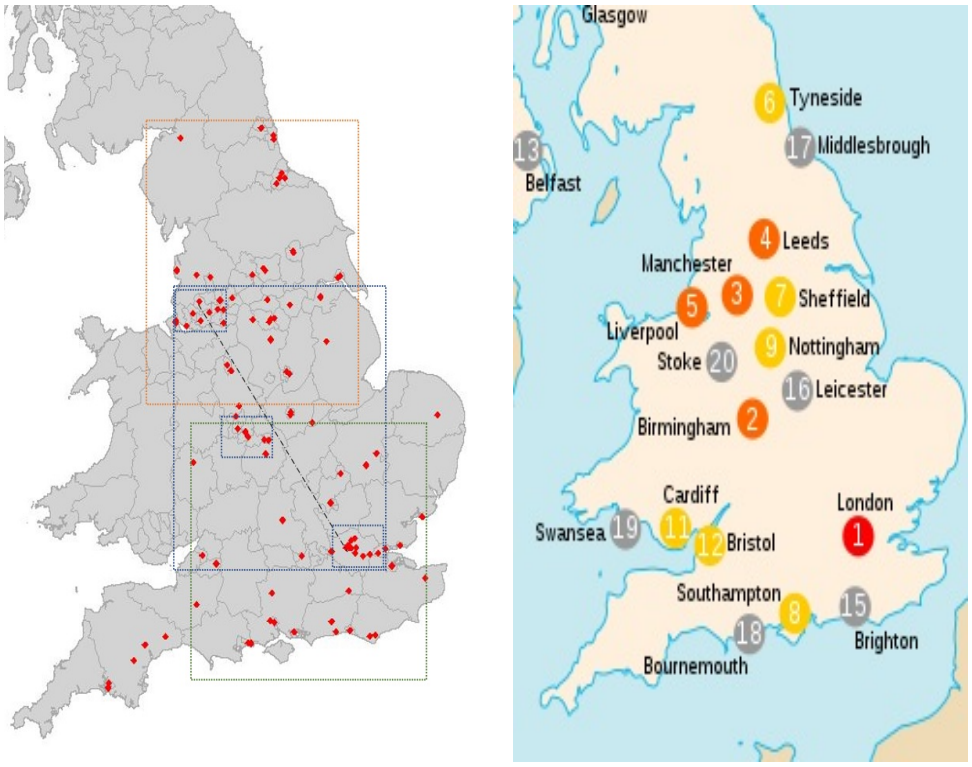

FIGURE 4.6: The left graph shows the locations of the air monitoring sites in England. The main metropolitan areas in England are indicated in the right graph.

We identify this bifurcate effect as a clustering problem where we intend to find our alternative solution. Clustering, often called clustering analysis, is a process of grouping similar objects together as a cluster from the other distinct samples. By the established criteria, the outcome of this process is that some new, often useful, insights of the subject are obtained from the clusters. Clustering analysis was originated during the 1930s and 40s from social sciences, such as in the field of anthropology by Driver and Kroeber (1932) and psychology by Cattell (1943). It has since then developed and became one of the major aproach of exploratory data analysis and statistical techniques for data analysis, in particular, in big data and machine learning era in the 21st century.

The task of implementing a distance-related clustering in this marginal Kriging procedure is to divide the full set of obtained marginal mean functions into two groups, and keep the group estimated from the nearby conditioned locations for the next averaging step. There are two popular methods, K-nearest neighbouring (KNN) and K-radius neighbouring (KRN), that we opt for the latter. After all the KRN approach satisfies the need from Figure 4.5. While by the KNN method, there is no fixed range of distance guaranteed at each location, undesired estimates from the distances far over 200 km may therefore be unwillingly included.

Figure 4.7 demonstrates the amendment made to the previous full- model averaging marginal Kriging procedure in Section 4.2. After the first step of the marginal regression function estimation, we replace the spatial prediction of covariance matrix by a K-radius neighbouring average function, in which a radius of value K will be identified so that those estimated $\hat{z}(s)$ conditioned from distances greater than K are discarded. Hence, the majority of estimates from cross major regions are excluded in the second step of the proposed procedure. Statistically, it is aware that there are 2385 pairs of monitoring sites having distances less than 200 km, counting as 52.3% of the total 4560 pairs of locations in our spatial air quality data set.

To illustrate this new procedure, Eq (4.11) can be reformulated in an averaging form of

$$\tilde{X}(s_0) = \frac{1}{n}\sum_{i=1}^{n} E[X(s_0)|X_i = x(s_i)] = \frac{1}{n}\sum_{i=1}^{n} Z(s_i), \tag{4.25}$$

where $\{s_i\}$ is a subset of $S^{'}$ satisfying $||s_0 - s_i|| < K$, and n is the cardinality of the set $\{s_i\}$. It is worth mentioning that both $\{s_i\}$ and n are location sensitive, which vary at each new location s_0.

In conclusion, combining Stage 1, the marginal regression function estimation introduced in Section 4.2.2 with Stage 2, the K-radius neighbouring average (Eq (4.25)), we now complete this new two-stage K-radius neighbouring average based marginal Kriging method, or KNAMK in short.

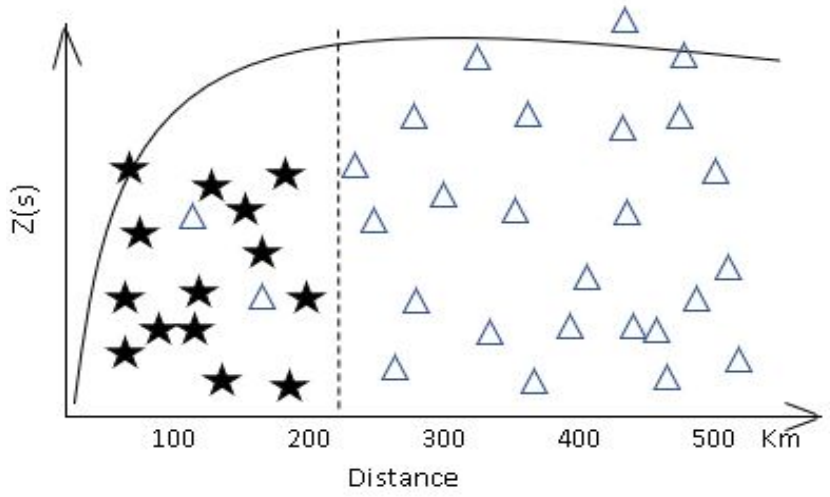

FIGURE 4.7: Illustration of the K-radius neighbouring average function, where the distance parameter K is to be estimated. The star symbol signifies that both s_0 and s_k are in a same major region, and the triangle indicates that s_0 and s_k are in different major regions in England.

4.4 Application of KNAMK method to air quality data

Following the theoretical introduction of the K-radius neighbouring average based marginal Kriging and its background knowledge explained, we will apply this method to our spatial de-trended data $\hat{x}(s)$.

4.4.1 Examination of the de-trended data

In Section 3.4.1, we compared the de-trended air quality data with a matching Gaussian profile using a nonlinear kernel density estimation by the *sm* function in R. Before embarking on testing our new Kriging method, we check our de-trended observation data $\hat{x}(s)$, as shown in Figure 4.8. Ideally, the data would be expected to fit the matching Gaussian process, yet deviations are detected, especially around the tails as well as an over-fitting near the mean. We will take this into account in the following sections.

4.4.2 Selection of the bandwidth and the K-radius

To perform the newly proposed two-stage KNAMK method detailed in Sections 4.2 and 4.3, there are critical parameters yet to be decided, i.e., the selection of the bandwidths h and b, and the K-radius. An optimal set of bandwidths contribute to a balanced estimation of probability density functions, while a suitable K-radius ensures the consistency of the selected marginal estimations in the second step. The MSPE under the cross-validation principle is used to evaluate the whole process.

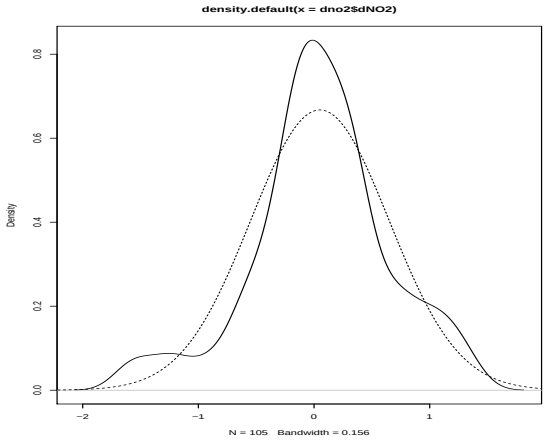

FIGURE 4.8: A comparison between the detrended air quality residual data with a matching Gaussian profile.

Among the three parameters to be estimated, the previous knowledge on K-radius can be adopted to set its initial value. For the bandwidths, the normal distribution approximation h^* in Eq (4.24) is used to initiate this task, that is we set $h_0 = h^*$, considering the data is relatively closed to a Gaussian profile (see Figure 4.8). Meanwhile we define the $b_0 = h^*$, which means that both kernels are treated equally at the beginning of this process. Instead of dedicating the entire computational tasks to an nonlinear optimisation solver, we adopt an idea from the design of experiments (see Fisher (1936)). Applying the concept of the orthogonal factorial design, in the initial test phase, we complete the KNAMK process using five sets of bandwidths, i.e., (h_0, b_0), $(2h_0, 2b_0)$, $(2h_0, 0.5b_0)$, $(0.5h_0, 2b_0)$ and $(0.5h_0, 0.5b_0)$. The comparison of the five calculated MSPEs shows the likely direction of where the MSPE descends, so we re-centre the (h_0, b_0) along this descending direction and repeat this testing routine. By a small number of iterations (two or three rounds), we can significantly reduce this two-dimensional solution plane of the h-b bandwidth set. Using this kind of 'long jump' test, the potential 'local traps' effect may be reduced which is very beneficial for tasks of this kind.

In the next step, however, we adapt the suggestion from Lu and Tjøstheim (2014) that we can not simultaneously select optimal b and h to minimise the MSPE, as a result from the asymptotic assumptions made for estimation of the probability density functions. After the new initial starting point is defined, we alter the values of b and h one at a time to calculate and further compare the MSPEs from each set. The vectorization function in R greatly assists this intensive computational work. Furthermore, we repeat the bandwidths selection process with multiple K-radius values until the estimated MSPE converges.

4.4.3 Numerical result

In this section we report the result of the K-radius neighbouring average based marginal Kriging method applied to our benchmark data set, the UK air quality data. The purpose is to illustrate that this semiparametric estimation method works reasonably well in a spatial additive regression case. Furthermore, we use the MSPE criterion to compare its result with those from the previous methods, where

$$\text{MSPE} = \frac{1}{N}\sum_{k=1}^{N}(\tilde{Y}(s_k) - Y_{obs}(s_k))^2, \text{ for } k = 1, ..., N. \tag{4.26}$$

Table 4.1 shows a comparison of the mean squared prediction error using the three Kriging methods as titled. By taking the bandwidths $(h, b) = (0.07, 8)$ and K-radius = 205 Km, the MSPE of KNAMK method is 189.4570. This is noticeably smaller than the linear Kriging, while 233.0495 is the MSPE of the direct universal Kirging method. It is also smaller than the result from the Nonparametic-trend Universal Kriging (NTUK) method proposed in Chapter 3.

TABLE 4.1: A comparison of Mean Squared Prediction Errors from three Kriging methods

Methods	Mean Squared Prediction Error
K-radius neighbouring average marginal Kriging	189.4570
Nonparametric-Trend universal Kriging (Chapter 3)	213.6436
Direct universal Kriging (Chapter 2)	233.0495

It is understood that due to the nature of the free parameters (the two bandwidths and the K-radius), achieving the global minimum for this specific data set is not the main purpose of this section. Indeed, the current comparison adequately concludes that the proposed KNAMK method meets our requirement and performs the best among all Kriging methods introduced so far, based on our benchmark air quality data set.

Up to now, our attention has been focused solely on spatial prediction. By proposing two nonparametric spatial interpolation models, the NTUK and KNAMK, we attempt to overcome the confined Gaussian assumptions of the current linear Kriging methods. The aim is to provide alternative Kriging methods suitable for more general non-Gaussian spatial data sets. In Chapter 5, we will expand our attention into the realm of nonlinear spatio-temporal modelling and prediction.

Chapter 5

Semiparametric Spatio-Temporal Nonlinear Prediction

In this chapter, we will add time dimension into the current study of purely spatial prediction at a single point of time. Integrating the nonlinear spatial prediction method in Chapter 4, the contribution of Chapter 5 is to develop a staged semiparametric spatio-temporal procedure performing future prediction (forecast) for nonlinear data at irregularly spaced sampling locations.

5.1 Introduction

So far in this research, we confined our focus on the spatial series prediction studying likely underlying relationships between spatial-indexed information among locations at a single point of time. In this chapter, we will expand our research to the field of spatio-temporal prediction. The objective of this attempt is to predict the future values of the response at any locations using spatial-indexed data sampled from the current and past times, i.e. adding time dimension into the spatial prediction. With the additional knowledge from nearby points in time, it is believed that some time dependence structure can be established, and subsequently applied in the prediction model for the future. Yet proposing a sensible model structure remains a challenging task of being adequate and still achievable for computation.

In Section 5.2, we will begin with a literature review on the history, development and principal methodologies of time Series and further the spatio-temporal series, with main focus on nonlinear approaches. From the review, a new two-staged SPKM procedure of nonlinear semiparametric spatio-temporal prediction will be proposed in Section 5.3. Finally in Section 5.4, we will apply this nonlinear procedure to the air quality data and compare the result with those from some current methods in literature.

5.2 Background knowledge

Recognising a substantial leap from the spatial series study, we believe that it is essential to expand our review on two new fields, namely the Time Series and Spatial Time Series. Then, these three modelling families, which are distinguished by their fields of applications, share immense similarities, comparable concepts and methods.

This section is divided into two main parts: the review of time series and the spatial time series, respectively. The aim of this section is to elaborate the knowledge in the past chapters and build the foundation for this new subject, which then leads to the introduction of the class of spatio-temporal autoregressive partially (non)linear regression (STAR-PLR) models in Section 5.3.

Prior to the review, we would like to briefly compare a few important concepts that are critical to this chapter, all of which will be mentioned repeatedly in the following sections.

- Linear and Non-linear model: In statistical modelling, when the finite sample data shows a linear correlation between the response and covariates, or they can be transformed (regularised) into a near-linear relationship, the well-studied linear modelling methods would usually be appropriate due to their superiorities in simple to implement and easy for interpretation. Despite the advantages, linearity features are not commonly observed in the real world, misspecification may occur when significant nonlinearity is presented in the data. Nonlinear modelling methods are therefore developed to meet the situations when non-Gaussian features exist.

- Parametric, Semiparametric and Nonparametric model: In statistical modelling, after a model family is selected, when the form of the probability law in a family is specified except for some finite dimensional defining parameters, such a model is referred to as a parametric model. In contrast, if infinite parameters are included or the form of the probability law is not all specified, such a model is called nonparametric model (Fan and Yao, 2003). Furthermore, a model with elements of both two forms is called semiparametric model. In general, parametric model is a model with a global focus, in which the defining parameters are well-defined in the studied space, whilst semi- and nonparametric models can be more flexible in structure and locally focused, resulted from that the defining parameters can vary as the time and location changes.

- Stationarity and Nonstationarity: Statistics is a science built upon assumptions that help to explore the unknown from the known. In the study of stochastic process, stationarity is a powerful assumption establishing statistical properties of the underlying data-generating process. Explicitly, it imposes in which pattern the studied process changes. Weak stationarity and strict stationarity are

two commonly used stationarities in our research. The former defines the first two moments of the underlying process which is sufficient in a linear modelling case while the latter are required for nonlinear models bringing the process additional properties. Modelling nonstationarity in nonlinear data-generating process is possible however it requires additional assumptions to support.

- Regular and Irregular sampling: Similar to the difference between discrete and continuous data, in a sense that the regular sampling refers to a method that collects data at equal time intervals (i.e., for Time Series) or from a lattice (i.e., a regular grid in Spatial Series). In many applications, as restricted by nature, the regular sampling may not be possible and one has to face irregularly sampled data. Interpolation is one of the methods often used to transforms irregularly sampled time series data, albeit it is much harder to implement in a spatial series case when advanced computational capacity is required.

5.2.1 Time series

Time is an indefinite continued progress of existence where it is a key component that forms the Universe around us. Recording time information together with their events has been familiar with us since ancient times; further in the Middle Ages scientists gradually postulated one of time's primary properties of consistency, i.e., on the Earth, the proceeding rate of time is constant. Despite the effort, however a scientific way of studying a series of data points indexed in time order, now called Time Series Analysis, did not begin until the 20^{th} century following the study in stochastic process. Time Series or Time Series Analysis in this research refers to a family of formalised methods processing time-indexed data, which is denoted by $\{X_t\}$, $(t = 1, ..., T)$. In specific, we treat these series as random realisations of their underlying stochastic processes for study. Overall, it comprises methods of describing the changes of the object from a temporal prospective, and often providing forecasting capability through proposed models for instance regression.

5.2.1.1 Linear time Series

Some of the pioneering work of this subject, i.e., the autoregressive (AR) modelling, was initiated by scholars such as G. U. Yule and J. Walker in the 1920s and 1930s, see for instance Slutzky (1937) and Yule (1921). It was then the works of Box and Jenkins (1970), in which the autoregressive moving average (ARMA) model was presented, confirms the formal introduction of linear time series models. It contains the methodology and complete modelling procedure for individual series. The Box-Jenkins method, named after its contributors, is regarded as the most commonly used time series method. It remains as the foundation of many its derivatives ever since. The standard linear ARMA model consists of three basic building blocks: white noise process, autoregressive(AR)

model and moving average (MA) model. Any subset of a complete ARMA model can be treated as an individual stochastic process with own characteristics.

The basic building block for an ARMA model, indeed for most linear time series models, is white noise (WN) process. It is defined by its first two moments. A sequence $\{\epsilon_t\}$, $t \in \mathbb{R}$ is a WN process if

$$E(\epsilon_t) = 0, \;\; E(\epsilon_t^2) = \sigma^2 \;\text{ and }\; E(\epsilon_t \epsilon_\tau) = 0, \;\text{ for }\; \forall \;\; t \neq \tau. \tag{5.1}$$

Adding stronger conditions, a specific independent white noise process, when $\epsilon_t \sim N(0, \sigma^2)$, is called Gaussian white noise process, denoted as $WN(0, \sigma^2)$, see Hamilton (1994).

Autoregressive(AR) model, the second building block of an ARMA model, is often regarded as the most popular time series model in practice for its easy implementation. As a random process, a p-order Autoregressive Process is defined as

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t, \tag{5.2}$$

where $\{X_t\}$ is a time series with a length of a positive integer p , and $\{\epsilon_t\} \sim WN(0, \sigma^2)$ is the unobservable component in this model. The main idea of this structure is to define a time varying relationship specifying how the current quantity of the monitored variable relies on its p immediate past values. As a consequence, it draws a recurrence relation within the response series. Since p past values are involved, this model is also called a AR(p) process, or the p-th order autoregressive process (Box and Jenkins, 1970).

For general case, Hamilton (1994) included a constant term (mean) in model Eq (5.2). Furthermore, to assign stationarity properties to the model, certain constraints on the weight parameters b_i, $i = 1, ..., p$, are required.

The final building block of an AMRA model is Moving Average (MA) process. A q-order MA process, often abbreviated as MA(q), is defined as

$$X_t = \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}, \tag{5.3}$$

where the $\{\epsilon_t\}$ is a WN process of length $q \in \mathbb{N}^+$, where q is a positive integer. Treating $(a_1, ...a_q)$ as linear weight factors, the MA process represents the moving average of the selected white noise sequence. The purpose of a moving average model is to record the impact from the past random noises to the observed value at the time t.

Combining the three random processes in one general linear form, it forms an autoregressive moving average (ARMA) time series model. It takes the form of

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}. \tag{5.4}$$

It is easy to notice that the ARMA(p, q) model captures both the univariate time series information from the past and the historical randomness registered by the MA model. Actually, in the light of the Wold's decomposition theorem (Wold, 1954), any covariance-stationary process can be revealed as a sum of two time series: one deterministic and another stochastic series. These two series are mutually uncorrelated processes. The ARMA linear model is indeed under such an arrangement and proven to be an effective linear regression method for a stationary time series.

However in a real time series, the time-invariant properties are often violated. Generally when there are level shifts or cyclic features in the population, current ARMA model is insufficient to capture these changes (Fan and Yao, 2003). Some common level shifts include seasonal patterns and other long-range trend functions. Thus, an modified autoregressive integrated moving average (ARIMA) model was proposed.

For a nonstationary time series $\{Y_t\}$, if its d-order difference process is a stationary ARMA(p,q) process, we name it an ARIMA process with order p, d and q, denoted as $\{Y_t\} \sim \text{ARIMA}(p, d, q)$, see Box and Jenkins (1970). The new parameter $d \geq 1$ is introduced as a backshift operator. The difference transformation, in effect, converts a nonstationary time series into a stationary time series, i.e., an ARMA process, by detaching underlying time-variant dynamics from the original data. Therefore, ARIMA model is a powerful generalisation of the ARMA model taking some long-range varying features into account.

5.2.1.2 Nonlinear time series

ARMA, or ARIMA alike, and other Gaussian processes are capable of performing well in linear time series analysis. However, in situations when substantial nonlinear features exhibit from the sampled data, nonlinear time series models may be more appropriate for the purpose of avoiding significant misspecification. Introducing the diverse nature of the nonlinearity, Fan and Yao (2003) showed a list of nonlinear features, such as nonnormality, asymmetric cycles, bimodality, nonlinear relationship between lagged variables, variation of prediction performance over the state-space, time irreversibility, sensitivity to initial conditions. Noticeably, many of these features may be more prominent in their own developed areas, as a result, the nonlinear time series modelling approaches vary greatly by the fields of their applications.

In fact, many parametric linear methods have their nonlinear counterparts. The widely-used ARCH and GARCH models, closed derivatives of the ARMA/ARIMA, are good

examples of catching the nonlinear volatility in studying financial time series. In autoregressive model, the random component, or the innovation, is defined by a second-order stationary white noise process. Yet in economic and finance data, large instability is most likely to occur around the peak values, and this feature is called conditional heteroscedasticity. To model this inconsistent variance (volatility in time series), autogressive conditional heteroscedastic (ARCH) was introduced by Engle (1982), which is defined as

$$X_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = a_0 + b_1 X_{t-1}^2 + \cdots + b_q X_{t-q}^2 \ , \tag{5.5}$$

where $a_0 \geq 0$, $b_j \geq 0$, $\{\epsilon_t\} \sim \text{IID}\ (0,1)$ and σ_t represents the volatility (the conditional standard deviation). In most linear models, a white noise is adequate to represent weak stationarity, whilst in a nonlinear case, stronger assumptions on stationarity are required, the white noise process is thus purposely replaced by a centred independent and identically distributed random process.

To enhance the ability of describing the persistence nature of volatility, Bollerslev (1986) proposed a useful extension to the ARCH model, namely the generalised autogressive conditional heteroscedastic (GARCH) model. It takes the form of

$$X_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = a_0 + a_1 \sigma_{t-1}^2 + \cdots + a_p \sigma_{t-p}^2 + X_{t-q}^2 + b_1 X_{t-1}^2 + \cdots + b_q X_{t-q}^2 \ , \tag{5.6}$$

where $a_i \geq 0$, $i = 0, 1, ..., p$ and $b_j \geq 0$, $j = 1, ..., q$. By adding the volatility data from the immediate p past lags, the persistence effect can be better addressed in this new model.

So far we have listed several well-studied time series models, which have fully developed parametric structure. However to model a nonlinear time series data, there are indeed infinite possible forms to choose from; restricting this task to families of parametric models may not always produce convincing results. To overcome this difficulty, nonparametric modelling methods have been largely developed since the 1980s, bringing powerful alternatives to the classical parametric modelling approach. The main notion of the nonparmetric is, instead of estimating the finite number of fixed parameters, to let the data speak for itself with limited assumptions on model structures.

Fan and Yao (2003) showed a generic form of nonparametric time series model: the nonparametric autoregressive conditonal heteroscedastic model (NARCH). It is defined as

$$X_t = f(X_{t-1}, ..., X_{t-p}) + \sigma(X_{t-1}, ..., X_{t-q})\, \epsilon_t \ , \tag{5.7}$$

where $f(\cdot)$ and $\sigma(\cdot)$ are unknown functions, and $\{\epsilon_t\}$ $\sim$IID$(0,1)$. Only some qualitative assumptions are made in this model such as that the unknown functions f and σ are smooth, and it allows the presence of heteroscedasticity. Despite its plain form, this model leans more towards a theoretical construct and is less practical when the number of covariates is large, i.e., $p, q \geq 3$. The reason behind this is often described as the curse of dimensionality, that was coined by Bellman (1961) when he studied problems in dynamic programming. The common phenomena of this term is that in a model when the dimensionality rises, the size of its solution space will vastly increase so that the sampled data become sparse. In the field of nonlinear time series, Fan and Gijbels (1996) explained that it is an essential requirement of having a certain number of local observations in order to implement nonlinear smoothing techniques. When performing a multivariate surface smoothing, the multi-dimensional surface expands exponentially that one has to enlarge the neighbourhood for the required data sets, which is in principle contradict this method itself.

In addition to the contradiction, when p, q increase, there is also a computational concern in dealing with models having a large number of covariates. Taking the local linear smoothing as an example, to define a high dimensional kernel function $\boldsymbol{K}$ and the bandwidth matrix associated, one may firstly encounter difficulties in deciding whether the smoothing settings for all covariates are same or not (see Section 4.2.3). This may rises a practical request for a higher computing power.

Among the countless possible solutions between saturated nonparametric models (see Eq (5.7)) and parametric models, a widely adopted effective dimensionality reduction technique is the additive modelling (AM) method (see Hastie and Tibshirani (1990)). It transforms this high-dimensional challenge into a problem of solving a group of low-dimensional or univariate functions that each of them can be estimated by well established nonparametric regression methods. Function- coefficent autoregressive (FAR) model and additive autoregressive (AAR) model are such two examples often seen in practice.

The function-coefficient autoregressive (FAR) model introduced by Chen and Tsay (1993) admits the form of

$$X_t = f_1(X_{t-d})X_{t-1} + \cdots + f_p(X_{t-d})X_{t-p} + \sigma(X_{t-d})\epsilon_t, \tag{5.8}$$

where $d > 0$, $\{\epsilon_t\}$ $\sim$IID$(0,1)$ and $f_1(\cdot)$,$\cdots$, $f_p(\cdot)$ are unknown coefficient functions. In general, the FAR model can be treated as an extension of the thres-hold model proposed by Tong (1990) viewing the nonlinear dynamics as a piecewise linear approximation via partitioning a state-space into several subspaces. It allows the coefficient functions to change gradually, therefore it is commonly used in ecological studies.

The additive autoregressive (AAR) model is a generalised extension of the linear AR(p) model. It is defined as

$$X_t = f_1(X_1) + \cdots + f_p(X_{t-p}) + \epsilon_t, \tag{5.9}$$

where $\epsilon_t \sim \text{IID}(0,1)$. The p unknown functions are set to be one-dimensional in this model, the curse of dimensionality effect is thereby significantly reduced.

For additive modelling, after a low-dimensional model was chosen, the second step of solving a nonlinear time series problem is to estimate the functions of the coefficients by smoothing. Local linearisation, spline fitting and orthogonal series methods are the three main approaches for a nonparametric modelling, see Fan and Yao (2003). We will focus on the first method only while the spline (the piecewise linear approximation) and the orthogonal series methods such as Fourier series and wavelets for spectral analysis will not be covered in this research. To explain the local linearisation method, we take the FAR model, Eq (5.8), as an example.

In the FAR model, limited qualitative properties are assumed for $f_1, \cdots, f_p$, i.e., these functions are smooth. Each of these can therefore be locally approximated by a constant or more generally a linear function, see Fan and Gijbels (1996); Fan and Yao (2003). It follows as such that for a given response value x_0, the coefficient functions f_1,..., f_p can be approximated locally by a linear approximation, that is

$$f_j(x) \approx a_j + b_j(x - x_0), \tag{5.10}$$

for $x \in x_0 \pm h$, where the local parameters a_j, b_j correspond to $f_j(x_0)$ and the local slope of f_j at x_0, respectively. It is easy to learn that the local constant fitting is a special case of the local linear approximation where the corresponding b_j equals to zero. The h is called a bandwidth identifying the size of the neighbourhood within which the linear approximation holds. This leads to the approximation model:

$$X_t \approx \{a_1 + b_1(X_{t-d} - x_0)\}X_{t-1} - \cdots - \{a_p + b_p(X_{t-d} - x_0)\}X_{t-p} + \sigma(x_0)\epsilon_t, \tag{5.11}$$

for $X_{t-d} \in x_0 \pm h$. By adding a weighting scheme $K((X_{t-d} - x_0)/h)$, such as a nonnegative unimodal kennel function in Eq (5.11), we can complete the model using the least squares method by minimising the locally weighted squares as

$$\sum_{t=p+1}^{T} \Big\{ X_t - [a_1 + b_1(X_{t-d} - x_0)]X_{t-1} - \cdots \\ - [a_p + b_p(X_{t-d} - x_0)]X_{t-p} \Big\}^2 K(\frac{X_{t-d} - x_0}{h}), \tag{5.12}$$

where the x_0-dependent minimiser represents the weight for each term based on the distance between x_{t-d} and x_0. If K has a support on $[-1, 1]$, the weighted regression uses the local data points in the neighbourhood $X_{t-d} \in x_0 \pm h$. Alternatively, K may not require a bounded support as long as its distribution has thin tails away from the centre. The range of x_0, denoted as [a,b], is common to be defined as a grid containing 100 to 400 intervals on the resolution required, as suggested by Fan and Yao (2003).

We show how the local linearisation method is performed using the Kernel density estimation. In general, smoothing acts as a powerful building block for nonparametric estimation in a multidimensional setting. In its early developing stage, Bartlett (1946) recorded and later showed that smoothed periodograms improve the spectral density estimation in time series, then smoothing techniques were widely developed in areas such as time domain time series, state domain time series and beyond. A recent example in literature was to use smoothing for marginal and joint probability density function estimations in irregularly observed spatial series data, see Lu and Tjøstheim (2014).

Kernel density estimation, an improved form out of the classical histogram method, aims to estimate the population distribution from a sampled data set. In theory, there are countless forms of Kernel functions to choose from, Figure 5.1 shows some common kernel functions in use. In fact, as long as the functions are symmetric and unimodal, the resulting kernel density estimator performs nearly the same when the bandwidth h is optimally chosen (Fan and Yao, 2003). The Gaussian and Epanechnikov kernel are two popular kernel functions chosen in this research.

The challenging task in Kernel density estimation is the selection of bandwidth, similar as the selection of bin size to a histogram. When a large bandwidth h is chosen, from missing out potential details, the outcome may include oversmoothed estimation and a large bias. On the contrary, a small bandwidth would lead to excessive local information kept in the estimator and a high variance, consequently produces a wiggly distribution curve. Trial and error thus is inevitable in seeking the optimal bandwidth, denoted as h_{opt}. Thankfully, recommended kernel selectors are available for candidate distributions. Taking the normal reference bandwidth selector as an example, see Silverman (1986); Bickel and Doksum (2015), it works for data sets ideally the Gaussian-distributed data, the h_{opt} can be approximated using statistics from the samples, e.g, the size T and its variance σ shown in the last chapter. However, the recommended

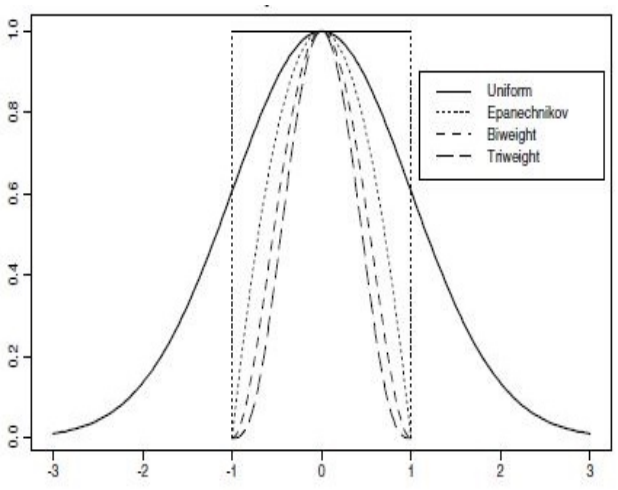

FIGURE 5.1: Five commonly used kernel functions normalised to have the same maximum height 1, the Gaussian kernel shows a wider support to the four named functions, see Fan and Yao (2003).

bandwidth merely acts as an initial choice of h, further fine tuning of the bandwidth is required by applying techniques such as the Cross-Validation criterion we used in this study.

Considering the inherent nature of modelling nonlinear time series, Fan and Yao (2003) noted that by performing preliminary nonparmateric methods, useful insights may be obtained prior to further parametric fitting, hence it is normal to perform a nonlinear time series analysis using a multiple-step procedure. Furthermore, semiparameteric models, which parametric and nonparmetric conpoments are both exhibited have been proven to be alternative options to ease the pressure from a full saturated nonparamteric one.

So far in this section, we introduced some basic knowledges of time series analysis, in particular the nonlinear time series, and highlighted how the time series was developed, its general ideas, common building blocks and techniques. For further reviews on nonlinear time series, readers are referred to Tong (1990), Fan and Yao (2003), Gao (2007), Douc et al. (2014), Tsay and Chen (2018) and references therein.

5.2.2 Spatial Time Series

We had separate reviews on spatial series in the first part of this dissertation, mainly in Chapters 1 and 2 , thereby additional reviews on such a subject will be integrated with the review of spatial time series in this section. Spatial time series analysis is a nature evolution from time series and spatial series, where spatial-indexed information from multiple time points are taken into account as one entity, however structurally much complex. The necessity of such study arises from a wide range of scientific fields such as earth and space science, economics, econometrics, public health policy, energy, natural resources, etc, where many data sets can be characterised by both temporal and spatial

properties. John Snow's London Cholera prediction was indeed a spatial time series case that can be traced back to the 1850s, see Ward (2008).

Despite the high availability of spatio-temporal data around us, it is only until the recent decades, in parallel with the emergence of Big Data, statistical approaches for spatial time series are flourished with the helps from modern computing techniques. Traditionally, as Cressie and Wikle (2011) suggested, deterministic approaches were popular in tackling such spatio-temporal problems. Taking physics science as an example, phenomena evolved in space and time often follows the known laws of physics. However when uncertainty and stochastic behaviour are involved, statistical modelling approach is proven to be a more suitable solution. Likewise, similar developments are witnessed in many other fields, as a result, we see diverse mathematical methods proposed in solving non-deterministic problems in general. In brief, it shares similar phenomenon as how the spatial series analysis was developed, as described in Section 1.1.1. In our research, the focus is on spatial time series with a univariate response, and the response data is treated as one realisation of the underlying spatio-temporal random process (or random field in a high-dimensional case).

Wikle et al. (2019), one of the latest reviews on this subject, outlined three main objectives expected from a spatial time series study: the prediction in space and time (filtering and smoothing), inference on parameters, and forecasting in time. It categorised many statistical approaches into two groups: the descriptive modelling and dynamic modelling. By recognising the difference between these two methods, one may be guided to choose suitable modelling options for their problems.

The descriptive modelling refers to the approaches aiming to characterise the spatio-temporal process in terms of its first two moments, i.e., the mean and covariance functions. A typical example of this approach is the Kriging methodology originated from the spatial analysis to reach the optimal linear prediction. It is commonly useful when the underlying mechanisms in the studied processes are less known to the modellers. Furthermore, this type of modelling is capable of: (1) capturing the large-scale trend of the spatio-temporal process, and (2) fitting less defined error terms, i.e., they can be statistically dependent in space and time. Yet its shortcomings are obvious too, since there are only the first two moments involved, this descriptive modelling method is limited to linear processes or less complex spatial time series.

The dynamic modelling for spatio-temporal analysis is based on the prior knowledge of how the system (a spatial process) changes over time. Statistically, we depict this category as a conditional approach by assuming the past is known, to model (often by approximation) the information evolving from the past to the present, and furthermore to the future, with helps from appropriate stationary assumptions on time and space. In practice, flexible classes of dynamic models are often used, since not all the underlying statistical or physical knowledge are known beforehand. One of the major challenges

for this group of methods is to reduce the pressures from high dimensionality, especially when a high level of dependencies exists. Hence, one widely-used structure for dimensionality reduction consists of two parts: a deterministic part of known covariates with their coefficients to be estimated, and another random part of known basic functions. Typical basic functions include polynomials, splines, wavelets and trigonometric functions, which are in associated with their applications.

The above two approaches cover the three main objectives of a spatial time series analysis. One is able to make a choice based on the available knowledge and/or assumptions made to the subject. We will briefly demonstrate two widely used methods: the deterministic inverse distance weighting (IDW) and the stochastic multivariate regression model.

In Chapter 1, we mentioned the IDW method for spatial prediction. In a spatial time series setting as shown in this chapter, the weighting coefficient will be adjusted according their spatio-temporal lag. Suppose a spatial time data set $\{Y(s_{11};t_1), Y(s_{21};t_1),$ $..., Y(s_{n_1 1};t_1), ..., Y(s_{1T};t_T)\ Y(s_{2T};\ t_T)\ ,..., Y(s_{n_T};t_T)\}$, where at each time t_j, there are n_j samples, for $j = 1, ...T$. The IDW predictor $\hat{Y}$ at a new location s_0 and time t_0 is defined as (see, Wikle et al. (2019))

$$\hat{Y}(s_0;t_0) = \sum_{j=1}^{T}\sum_{i=1}^{n_j} \omega_{ij}(s_0;t_0) Y(s_{ij};t_j), \tag{5.13}$$

where

$$\begin{aligned} \omega_{ij}(s_0;t_0) &\equiv \frac{\tilde{\omega}_{ij}(s_0;t_0)}{\sum_{k=1}^{T}\sum_{l=1}^{n_k}\tilde{\omega}_{lk}(s_0;t_0)}\ , \\ \tilde{\omega}_{ij}(s_0;t_0) &= \frac{1}{d((s_{ij};t_j),(s_0;t_0))^{\alpha}}\ , \end{aligned} \tag{5.14}$$

$d((s_{ij};t_j),(s_0;t_0))$ is the spatio-temporal lag between the sampled coordinates (s_{ij},t_j) and the new location (s_0,t_0), and α is a positive number controlling the level of smoothing. From Eq (5.14) , it is easy to read that the closer the sample point to the new coordinate, the larger the weight $\omega_{ij}(s_0;t_0)$ becomes. Further, the α can be decided by cross validation criterion on the mean squared prediction error (MSPE). This IDW method is particularly useful when the so-called Tobler's law is strongly observed.

The stochastic multivariate regression is a more common yet advanced method than the IDW. The model consists of a random term and a trend function, including covariates that represent the underlying spatio-temporal dependence. Taking one of the simplest regression models as an example, we define the data set $Y(s;t)$ observed at discrete times $\{t_j\}$, $j = 1, ..., T$ from spatial locations $\{s_i\}$, $i = 1, ..., n$ as

$$Y(s_i;t_j) = \beta_0 + \beta_1 X_1(s_i;t_j) + ... + \beta_p X_p(s_i;t_j) + \epsilon(s_i;t_j), \tag{5.15}$$

where β_k, $k = 1, ..., p$ is the coefficient to each selected covariate $X_k(s_i;t_j)$, β_0 is the intercept, and the random error $\epsilon(s_i;t_j)$ follows an i.i.d. $N(0,\sigma^2)$ for all its $(s_i;t_j)$ locations. In this model, the p covariates could be spatial and temporal dependent at the same time, and the response takes an additive form of all dependent variables. The advantage of such model is clear that standard regression techniques can be easily adapted to compute the coefficients and variances. Good candidates for the covariates include the spatio-temporal coordinates and their low-order interactions. The ordinary least squares (OLS) is capable of estimating of the coefficients that minimising the residual sum of squares (RSS).

The assumption of independent random error was made in Eq (5.15), however in a real data set, one may notice that the residual will most likely show a dependent relationship indexed in time and/or space. Wikle et al. (2019) commented that the OLS parameter estimates and predictions are still unbiased even if one has ignored the dependence in the error term, but the result tends to give inaccurate prediction errors.

In addition to the simple form of regression as in Eq (5.15), a generalised linear model (GLM) or more broadly, a generalised additive model (GAM) is popular in recent literature. There are two parts in a standard GAM model, which are a systematic component and a random component. Eq (5.16) shows the link function of the systematic component of the GAM model expressing the transformation of the mean response in relation to the covariates $X_1, ..., X_p$, that is as

$$g(Y(s;t)) = \beta_0 + f_1(X_1(s;t)) + f_2(X_2(s;t)) + ... + f_p(X_p(s;t)), \tag{5.16}$$

where the functions $\{f_k(\cdot)\}$ can take a parametric, semiparametric or nonparametric form with specified smoothing techniques. The random component of the GAM assumes that observations conditioned on their respective means and sometimes scaling parameters are independent and they come from the exponential family of distributions such as the normal (Gaussian), Possion, Binomial and Gamma distributions, see Wikle et al. (2019).

Moreover, it is gradually common to see that a large number of covariates are included in an initial spatio-temporal model. Among the many variable reduction techniques, the regularisation method in which a penalty term is added to the RSS from OLS is popular in practice, two commonly used regulation methods are the lasso (L_1-norm penalty) and ridge regression (L_2-norm penalty). Some of the further discussions on this topic are widely available in literature, e.g., Al-Sulami et al. (2019) and Zhu et al.

(2010). The same variable reduction topic was mentioned too in Section 4.1, in the spatial series setting.

So far, we did a brief review of the spatial time series analysis and introduced two groups of regular approaches, mainly about their possible model structures. However dealing with a real spatio-temporal problem, there are still many critical factors to consider. Taking Geographic Information System (GIS) as an example, which our air quality analysis belongs to, the first question to ask may be about its data category. Cressie (1993) showed three types of spatial data: geostatistical data, lattice data and spatial point patterns, it further emphasised that the spatial model is very much related to their data set, albeit some methods are shared between different data types. In this study, we focus on only the geostatistical data since the air quality data are collected at irregular monitoring sites, and we are interested in predicting the response value at a new spatial location s varying continuously over the studied geographical space, descried by a subset of $\mathbb{R}^2$.

Unlike the well-developed nonlinear time series analysis, the nonlinear approach in spatial series and spatial time series has been a challenging task. Gao et al. (2006) highlighted that an obvious reason for this is the curse of dimensionality, which we have discussed earlier in Section 5.2.1.2. In time series, the neighbourhood of a time point t_i is still in one-dimensional, whilst for spatial series, on a spatial grid evaluating the condional mean given, its closest neighbours requires a four-dimensional nonparametric regression, explicitly for spatila data $\{Y_{ij}\}$, the neighbourhood of Y_{ij} is constructed by $Y_{i-1,j}$, $Y_{i,j-1}$, $Y_{i+1,j}$ and $Y_{i,j+1}$. As such, in order to perform a nonparametric local fitting over space, it is clear that a vastly large sample size is essential. To ease such a difficulty on dimensionality, two widely used models were observed under this circumstance: the additive model and semiparametric model. In addition to this obstacle, the assumptions on regularity, linearity and stationarity (both in time and over space) of the underlying data-generating process are critical too, by which the asymptotic theory of a candidate model is constructed.

In the past two decades, limited theoretical attempts on nonlinear spatial data were made so far. Lu et al. (2009) and Lu and Tjøstheim (2014) were such two examples, where the former was extended from Fan et al. (2003) by proposing an adaptive varying-coefficient spatio-temporal model for data observed irregularly over space and regularly in time; and the latter offered a nonparametric estimation of probability density functions over irregularly positioned spatial data for both the marginal and joint density, which we have adopted in Chapter 3. Further readings in this and related subjects can be found in Cressie and Wikle (2011), Gao et al. (2006), Rao (2008), Sun et al. (2014), Wikle et al. (2019), to list a few.

On the application side, Al-Sulami et al. (2017) and Al-Sulami et al. (2019) were two recent papers sharing how a family of semiparametric nonlinear regression model (STAR-PLR) performs on irregularly located spatial time series data with applications in econometrics. The proposed regression model describes the nonlinear relationship between the response and covariates, which is location-based and both temporal-lag and spatial-neighbouring effects are considered. It is worthwhile to mention that despite the inherited extreme curse of dimensionality in a spatial time series, a computationally feasible two-step estimation procedure was developed using the ideas from Gao et al. (2006) and Lu et al. (2009). Because of the closed similarities to our case, we will adapt and integrate this family of regression models to our air quality forecast.

Despite the recent advancement in nonlinear spatial time series analysis, universal solutions of this approach is still lacking. Likewise there are few shelf-ready software packages available for the practitioners because of the distinct challenges in both the theoretical and computational aspects. The general approaches to these problems are yet overwhelmingly dominated by parametric techniques [1], e.g, the family of Kriging methods. Pebesma and Heuvelink (2016) commented that interpolation of spatial random fields is a common task in geostatistics and the corresponding approaches such as the inverse distance weighted predictions (IDW) and the Kriging procedures have routinely been applied for many years. As such, the Kriging in spatial series was the main benchmark method in the first part of this dissertation. It is a method of interpolation by which the values of the response are modelled by a prior covariance structure under the Gaussian assumption. Spatio-temporal Kriging is naturally an extended version of the spatial Kriging by adding the time dimension. Mathematically, it is a study of a Gaussian spatial time random field Y defined over a two-dimensional spatial domain $\mathcal{S}$ and a temporal domain $\mathcal{T}$, hence the observed $\mathcal{S} \times \mathcal{T}$ locations are denoted as $(s_1, t_1), ..., (s_n, t_m) \subset \mathbb{R}^2 \times \mathbb{R}$, the outcome of Kriging is then the values of response at any new locations in $\mathcal{S} \times \mathcal{T}$.

The process of implementing Spatio-temporal Kriging is similar to that for spatial Kriging introduced in Chapters 2, 3, and 4. It begins by fitting a spatio-temporal varigram cloud, the fitted mean squared error will determine the parameters in the theoretical covariance model before the final Spatio-temporal Kriging. However, instead of measuring the unilateral spatial distance, we will measure a combined distance in time and over space simultaneously, i.e., the spatial temporal distance between any pair of points $(s_{i1}, t_{j1}), (s_{i2}, t_{j2})$ is registered as $||s_{i1} - s_{i2}|| = |d|$ and $|t_{j1} - t_{j2}| = |u|$. With the assumptions of the random field Y to be stationary and spatially isotropic for example, a theoretical variogram can be constructed assuming the underlying covariance function C_{st} only responds to the space lag d and the time lag u. To simplify this complex spatio-temporal variogarm model, some basic classes of sptio-temporal covariance

[1] https://cran.r-project.org/web/views/SpatioTemporal.html (the SpatioTemporal CRAN task review)

functions are nominated such as separable, product-sum, metric, sum-metric and simple sum-metric model. Eq (5.17) shows that the covariance function from the separable covariance model can be represented as the product of a spatial and temporal terms, i.e.,

$$C_{\text{sep}}(d, u) = C_s(d)C_t(u). \tag{5.17}$$

More spatio-temporal covariance models can be found in Section B.2, Appendix B. Further reading on sptio-temporal Kriging method together with examples using *gstat* package in R, can be found in Pebesma et al. (2012),Pebesma and Heuvelink (2016), Pebesma and Gräler (2021). In addition, Frazier (2017) could be a good reference explaining how various Coordinate Reference Systems (CRS) are processed in R.

With the review so far on time-, spatial- and spatial time series, we will propose a new two-phase semiparametric nonlinear spatial-temporal forecast procedure in the next section.

5.3 The SPKM procedure for nonlinear spatio-temporal prediction

The objective of this chapter is to extend the K-radius neighbouring average based marginal spatial Kriging (KNAMK) model, as introduced in Chapter 4, to the realm of spatio-temporal series modelling. In this section, we will propose a computationally feasible method of two-phase procedure performing semiparametric nonlinear spatial time series prediction, then apply and evaluate this procedure to air quality data in Section 5.4.

Figure 5.2 exhibits a complete diagram of this research on a spatio-temporal space. The horizontal axis represents the time, for simplicity, we denote the present time as t_0, where the times on its left are from the past, and t_{0+1} is the next consecutive time in future. Distinguishing from s being a random location on the vertical axis, we use $s_1, ..., s_N$ to present the N monitoring sites irregularly located in the space S. In the 2-D spatial setting of this research, we denote these N known locations as $s_i = (u_i, v_i) \in S \subset \mathbb{R}^2$, $i = 1, ..., N$. It is worth mentioning that the orders of s_i do not indicate their locations in S rather reflect their existence. In Figure 5.2, there are two types of symbols used to categorise the spatio-temporal response $Y(s, t)$, the solid dots and the circles. A solid dot signifies that the value of the response is recorded from true measurement, hence they can only be on the left side of the diagram up to the present time t_0. A circle, in contrast, is used to show that the exact response at that location is predicted, it can thus be observed on all parts of the diagram.

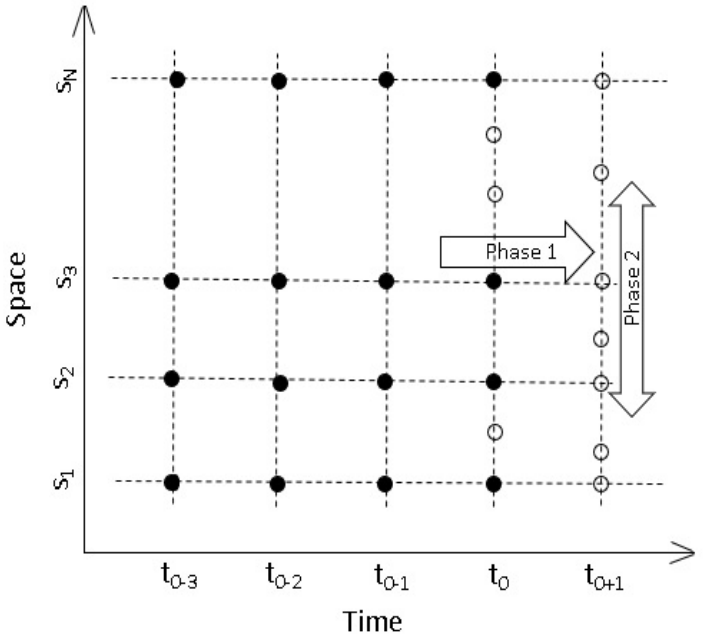

FIGURE 5.2: A two-phase procedure for spatio-temporal prediction in a future time.

In Chapters 2 to 4, the focus has been solely on spatial prediction at a single point of time t_{sp} that satisfies $t_{sp} \leq t_0$. At $t = t_{sp}$, we sample the values of the response $Y(s, t_{sp})$ from N known locations, and sequentially use these samples to predict the response at a new location of interest $s \in S$, i.e., along the vertical line of $t = t_{sp}$ in Figure 5.2. By implementing the KNAMK method developed in Chapter 4, nonlinear semiparametic spatial prediction can be performed at any time on or before t_0. However for the future time $t_0 + 1$, the same prediction task for $Y(s, t_{0+1})$ at any location s cannot be directly performed on the basis that existing knowledge of the future responses are absent. Intuitively, a forecasting step would assist this task by predicting future responses at the known locations from the past information, we call this Phase 1. Thus by combining the Phase 1 with the developed KNAMK method applied at the future time t_{0+1} of being Phase 2, the objective can now be fulfilled. We briefly outline this two-phase process as follows,

- Phase 1 (spatio-temporal forecast): Using the data collected from the sampled locations, a suitable spatio-temporal forecast method will be employed to predict the future response values at these known locations.
- Phase 2 (spatial prediction): The results from Phase 1 are used in the unilateral spatial prediction at any random locations at the future time $t = t_{0+1}$. The KNAMK method is assigned for this phase.

The purpose of this section is therefore to identify a suited spatio-temporal future prediction method for the Phase 1, specifically to the air quality case, which is required to model the data observed irregularly over space and regularly in time. Since the nonlinearity in our data was revealed earlier, most current methods such as the traditional

space-time Autogressive-moving-Average model (STARMA) assume linearity on the underlying data-generating process (e.g., Cressie and Wikle, 2011, pp. 449, and Wikle et al., 2019), we may need alternative approaches in dealing with this nonlinearity problem, e.g., by semiparametric and additive models.

Gao (2007) and Lu et al. (2009) are two theoretical works as such in the literature proposing nonlinear low-dimentional semiparametric regression models, and through density estimation techniques the curse of dimensionality is effectively circumvented. In Gao (2007), the focus was on spatial regression for lattice data. It emphasised that in a nonlinear spatial case following the spirit of conditional models, see Besag (1974), one must live with the approximative aspect. Explicitly, semiparametric and additive models can be seen as approximations to the required conditional mean in dealing with nonlinear data. In the companion paper by Lu et al. (2009), the attention was extended to spatio-temporal models for data sampled irregularly over space and regularly in time. An adaptive varying-coefficient spatio-temporal model was proposed, which was the first attempt to address spatio-temporal nonlinear dependence structures for possible nonlinearity and nonstationarity in the targeted data-generating process since its original linear form was developed by Fan et al. (2003). The construction of this semiparametric model allows one-dimensional smoothing in estimating the coefficients for its variables, which include possible exogenous variables. When implementing these families of models, both papers recommended staged approaches for estimation. The main reasons for this are to ease the pressure from the dimensionality in a spatio-temporal setting, and meanwhile to improve the total accuracy on estimators by applying additional spatial smoothing techniques at each stage.

Al-Sulami et al. (2017) and Al-Sulami et al. (2019) took this approach and offered a class of location-dependent spatio-temporal autoregressive partially (non)liner regression (STAR-PLR) models with applications including an econometric case study in the US housing market. To show the model in Al-Sulami et al. (2017) specifically, let $Y_t(s)$ and $X_t(s)$ denote two spatio-temporal processes at discrete time point t=1,...,T and continuous location s in a two-dimensional spatial domain $S \subset \mathbb{R}^2$, respectively. The relationship between Y and X is of interest, denoting Y the response and X the covariate vector of dimension d, albeit d may be limited to be a small value in applications. It is assumed that both processes are observed at regular time intervals from the N spatial sampling locations $s_j = (u_j, v_j) \in S$ for $j = 1, ..., N$ on an irregular spatial grid. Hence, the data comprise $\{(Y_t(s_j), X_t(s_j)) : t = 1, ..., T \text{ and } j = 1, ..., N\}$. The STAR-PLR model now has the form of

$$Y_t(s_j) = g(X_t(s_j), s_j) + \sum_{i=1}^{p} \lambda_i(s_j) Y_{t-i}^{sl}(s_j) + \sum_{l=1}^{q} \alpha_l(s_j) Y_{t-l}(s_j) + \epsilon_t(s_j), \tag{5.18}$$

where $g(X_t(s_j), s_j)$ is a nonparametric function varying by location and describing the relationship between the response Y and the exogenous covariates X. A spatial lagged response variable, $Y_t^{sl}(s_j) = \sum_{k=1}^{N} w_{jk} Y_t(s_k)$, is defined, where w_{jk} is a spatial weight for $1 \leq j,k \leq N$ such that $w_{jj} = 0$ and the spatial weight matrix $W = (w_{jk})_{j,k=1}^{N}$ is assumed to be priori, the idea of which is well known in applied econometrics, see Anselin (1988), to reflect the true underlying spatial interaction. Two temporally lagged response variables, $Y_{t-i}^{sl}(s_j)$ and $Y_{t-l}(s_j)$ are included in the model to account for the temporal effects. The former involves neighbouring locations to s_j and the latter is s_j itself with temporal lags of i up to the pth lag and l up to the qth lag, respectively. Both $Y_{t-i}^{sl}(s_j)$ and $Y_{t-l}(s_j)$ are in linear relation to $Y_t(s_j)$ with spatially-varying autoregressive coefficients $\lambda_i(s_j)$ and $\alpha_l(s_j)$. The innovation term $\epsilon_t(s_j)$ is assigned to be distribution free and also independently and identically distributed (iid) over time with a zero mean and spatially varying variance $\omega^2(s_j)$. The processes $Y_t(s_j)$, $Y_{t-i}^{sl}(s_j)$ and $X_t(s_j)$ are assumed to be stationary over time and independent of the innovation process $\epsilon_t(s_j)$ for any t and s_j.

As the key nonparametric part of this model, the function $g(x_t(s_j), s_j)$ is left undefined, it provides higher flexibility than a spatio-temporal linear regression. Further, $g(x_t(s_j), s_j)$, the coefficients $\lambda_i(s_j)$, $\alpha_l(s_j)$ and the variance of innovation process all vary by location, hence despite of being stationary in time, the STAR-PLR model family captures nonstationary over space.

Next, the unknown function g and the autoregressive coefficients $\lambda_i(s_j)$, $\alpha_l(s_j)$ are to be estimated. We rewrite Eq (5.18) as

$$Y_t(s_j) = g(X_t(s_j), s_j) + Z_t(s_j)^T \beta(s_j) + \epsilon_t(s_j), \tag{5.19}$$

where $Z_t(s_j) = (Y_{t-1}^{sl}, ..., Y_{t-p}^{sl}, Y_{t-1}(s_j), ..., Y_{t-q}(s_j))^T$ and $\beta(s_j) = (\lambda_1(s_j), ..., \lambda_p(s_j), \alpha_1(s_j), ..., \alpha_q(s_j))^T$ denote the vector of spatio-temporally lagged variables and the cosponsoring vector of autoregressive coefficients, respectively, and $t = r+1, ..., T$ for $r = max\{p, q\}$.

Taking expectation conditional on the covariate in Eq (5.19) leads to

$$g(X_t(s_j), s_j) = E([Y_t(s_j)|X_t(s_j)] - E[Z_t(s_j)|X_t(s_j)]^T \beta(s_j). \tag{5.20}$$

which can be estimated by

$$\underbrace{\hat{g}(X_t(s_j), s_j)}_{g_0} = \underbrace{\hat{E}[Y_t(s_j)|X_t(s_j)]}_{g_1} - \underbrace{\hat{E}[Z_t(s_j)|X_t(s_j)]^T}_{g_2} \hat{\beta}(s_j). \tag{5.21}$$

The task now is equivalent to the estimation of g_0 and $\hat{\beta}(s_j)$ at each location $s = s_j$, providing the two conditional mean g_1 and g_2 can be well approximated by nonlinear methods.

Lu et al. (2009) proposed a two-step procedure for this task, which is computationally feasible to deal with even when the quantity of spatial time series data is relatively large, that is

- Step 1 (Time-series based estimation): For each s_j, we conduct time-series based estimation.
  - (i) $E([Y_t(s_j)|X_t(s_j)]$ and $E[Z_t(s_j)|X_t(s_j)]$ are estimated by a local linear regression method,
  - (ii) The estimators g_1 and g_2 from (i) are then used to estimate the unknown vector of autoregressive coefficients, $\beta(s_j)$, by least square method.
- Step 2 (Spatial smoothing): The estimation results from the above step are further improved by pooling information from neighbouring locations.

The step-by-step process of the estimation can be found in Al-Sulami et al. (2017). The final time series based estimators for Step 1 can be reached as

$$\begin{aligned} \hat{\beta}(s) &= \left\{ \sum_{t=r+1}^{T} \hat{Z}_t(s)\hat{Z}_t(s)^T \right\}^T \left\{ \sum_{t=r+1}^{T} \hat{Z}_t(s)\hat{Y}_t(s) \right\}, \\ \hat{g_0}(x,s) &= \hat{g}_1(x,s) - \hat{g}_2(x,s)^T\hat{\beta}(s). \end{aligned} \tag{5.22}$$

In Step 2, as the two estimators $\hat{g}_0$ and $\hat{\beta}(s_j)$ are obtained for all N known locations, these information can now be pooled together for spatial smoothing at a new location of interest $s_0 \in S$ (Lu et al., 2009). The spatial smoothing estimators for $g(x, s_0)$ and $\beta(s_0)$, denoted as $\tilde{g}(x, s_0)$ and $\tilde{\beta}(s_0)$ can be shown as

$$\tilde{g}(x,s_0) = \sum_{j=1}^{N} \hat{g}(x,s_j)\tilde{K}^*_{h,j}(s_0) \quad \text{and} \quad \tilde{\beta}(s_0) = \sum_{j=1}^{N} \hat{\beta}(s_j)\tilde{K}^*_{h,j}(s_0), \tag{5.23}$$

where $\tilde{K}^*_{h,j}(s_0)$ represents a weight function on $\mathbb{R}^2$, associated with $h = h_N > 0$, a spatial kernel bandwidth in relation to the sample size N.

The asymptotic properties of this model were examined with the results showing that the asymptotic variances of both the spatial smoothing estimators $\tilde{g}(x, s_0)$ and $\tilde{\beta}(s_0)$ are of a smaller order than those of the time series based estimators $\hat{g}(x, s_j)$ and $\hat{\beta}(s_j)$ from Step 1. The same effect to the mean squared error (MSE) between those of $\tilde{\beta}(s_0)$

and $\hat{\beta}(s_j)$ is observed under some preconditions on the smoothing bandwidth, see Al-Sulami et al. (2017).

Finally combining this two-step, i.e., the STAR-PLR model and the KNAMK spatial prediction method as Phases 1 and 2 , our newly proposed nonlinear spatio-temporal prediction procedure, STAR-PLR-KNAMA is now completed, or the SPKM for short.

The summary of SPKM is as follows:

- Phase 1 (STAR-PLR): Nonlinear spatio-temporal forecast

  - Step 1 (Time-series based estimation): For each s_j, we conduct time-series based estimation.

    (i) $E([Y_t(s_j)|X_t(s_j)]$ and $E[Z_t(s_j)|X_t(s_j)]$ are estimated by a local linear regression method,

    (ii) The estimators g_1 and g_2 from (i) are then used to estimate the unknown vector of autoregressive coefficients, $\beta(s_j)$, by least square method.

  - Step 2 (Spatial smoothing): The estimation results from the above step are further improved by pooling information from neighbouring locations.

- Phase 2 (KNAMK spatial prediction): The results from Phase 1 are then used for a unilateral spatial prediction for any random locations at the future time $t = t_{0+1}$.

We believe that the SPKM procedure is an early attempt in combining nonlinear adaptive semiparametric spatio-temporal regression with nonlinear marginal Kriging for a spatial time series forecast. In the next section, we will apply and evaluate this SPKM procedure to our air quality data.

5.4 Application of SPKM procedure to Air Quality Data

Tackling climate change and managing air quality have became ever more critical in the 21st century, especially in some parts of the world, the situation deteriorates fast and there are great negative impacts on their local inhabitants and future economy prospects. Encouragingly, the United Nations and global major economies are leading the way of shifting policies and regulations to guide a sustainable change on global climate [2] [3] [4] .

[2] UN climate change website: https://www.un.org/en/climatechange/

[3] China policy and action plan on tackling climate change, 2018 annual report: www.mee.gov.cn/ywgz/ydqhbh/qhbhlf/201811/P020181129539211385741.pdf

[4] UK Gov: Ways to tackle climate change: https://www.gov.uk/government/news/uk-to-go-further-and-faster-to-tackle-climate-change

Climate change includes global warming driven by both human emissions of greenhouse gases and the resulting large-scale shifts in weather patterns. By definition, a greenhouse gas is a gas that absorbs and emits radiant energy within the thermal infrared range causing the greenhouse effect [5]. The primary greenhouse gases in Earth's atmosphere include water vapor, carbon dioxide, methane, nitrogen oxide and ozone. Among them, a direct hazardous gas family to human recorded by the Defra UK, the Department for Environment Food and Rural Affairs, is nitrogen oxide including mainly nitrogen dioxide (NO_2), nitric oxide (NO) and other binary compounds of oxygen and nitrogen. To accurately measure and estimate local nitrogen oxide level, considering the scale of this problem is inevitably one of the key winning factors for this ever challenging task.

Prior to Chapter 5, the focus of this research has been on spatial prediction at a single point of time, i.e., on one spatial plane. Since time is continuous, the measurements of air quality are too continuous along the temporal horizon, which means that spatial time-series data are often readily available for making accurate predictions including even future forecast. We use prediction for estimating outcome of unseen data at the current and past time, while forecast is explicitly for making predictions for the future.

To perform spatio-temporal forecast using the new SPKM procedure to our air quality data, we access the data archive of 82 days in the early 2017, i.e., 27/01/2017-18/04/2017, from the Defra. The daily air quality data from 96 known monitoring sites in England are attained, which is slightly less than the total 105 stations in the spatial prediction case due to missing data.

In Phase 1, the spatio-temporal forecast, NO_2 values at the known locations are of interest, henceforth the response variable $Y_t(s_j)$ at the t-th time and j-th location represents the daily mean value of nitrogen dioxide NO_2 for $t = 1, ..., 82$ and $j = 1, ..., 96$, and the $s_j = (u_j, v_j)^T$ consists of the latitude and longitude of the j-th monitoring site.

The exogenous variable of interest is the daily mean value of nitric oxide NO defined as $X_t(s_j) = x_{t-1}(s_j)$ for $t = 1, ..., 82$ and $j = 1, ..., 96$, the $s_j = (u_j, v_j)^T$ of $X_t(s_j)$ is the same as that of $Y_t(s_j)$. Figure 5.3 shows that strong nonlinearity exists among these observations. Note the $X_t(S_j)$ in Al-Sulami et al. (2017) varies only by time, whereas in our case, it does by both time and location simultaneously. The nonlinear relations between the values of NO_2 , $Y_t(s_j)$, and NO, $X_t(s_j)$, for $t = 1, ..., 82$ and $j = 1, ..., 96$, can now be assessed by specifying a STAR-PLR model.

In Model (5.18), when specifying the spatial weights $W_{j,k}$ for the spatially lagged variable $Y_t^{sl}(s_j) = \sum_{k=1}^{N} w_{jk} Y_t(s_k)$, a common practice in econometrics is to use the inverse distance between the locations, that is $W_{j,k} = 1/d_{j,k}$, where $d_{j,k}$ is the Euclidean distance

[5] https://en.wikipedia.org/wiki/Climate_change

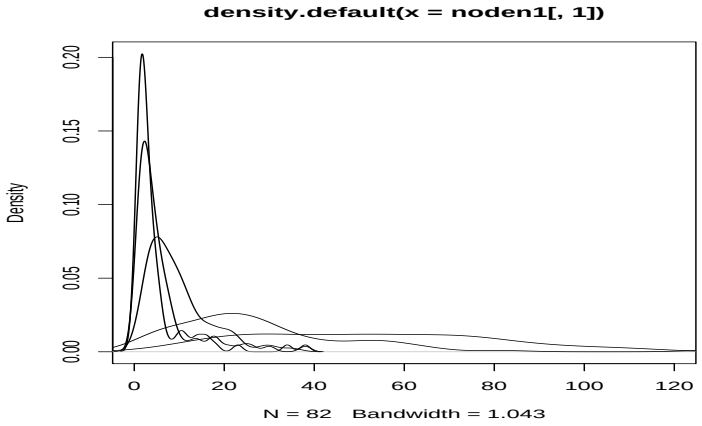

FIGURE 5.3: The $X_t(s)$ density plots for the first 5 locations, i.e., $s = s_1, ...s_5$.

between two monitoring sites s_j and s_k for $j \neq k$, otherwise, $W_{j,j} = 0$ (c.f., Wilhelmsson, 2002). It is clear that the deterministic spatial weight matrix $W = (w_{j,k})_{j,k=1}^{N}$ is symmetric with zeros on the diagonal, and furthermore is row-standardised so that $\sum_{k=1}^{N} w_{j,k} = 1, \forall j$.

We choose $t = 82$, i.e., 18/04/2017, as the future date in this case study, which acts as the time $t = t_{0+1}$ in Figure 5.2. Thus the Step 1 in Phase 1 of our proposed SPKM procedure, effectively becomes the prediction of $Y_{t=82}(s_j)$ from $Y_{t_i}(s_j)$ and $X_{t_i}(s_j)$, where $i = 1, ..., 81$ and $j = 1, ..., 96$. The prediction performance is evaluated by the mean squared prediction error (MSPE $_{\text{Phase1}}$) between $\tilde{Y}_{t=82}(s_j)$ and the true values of $Y_{t=82}(s_j)$, for $j = 1, ..., 96$.

To determine the orders of temporally lagged variables p and q in this phase, we take a direct approach from the LOOCV criterion to minimise the MSPE $_{\text{Phase1}}$. Table 5.1 shows that the minimiser, $\hat{p} = \hat{q} = 3$, has the smallest MSPE $_{\text{Phase1}}$ of 138.7398* for $p, q = 1, ..., 6$. So far, we have completed the Phase 1, the spatio-temporal forecast at the 96 observed locations.

TABLE 5.1: Selection of the orders of temporally lagged variables p and q.

	q=1	q=2	q=3	q=4	q=5	q=6
p=1	145.6658	149.9671	148.0758	178.4326	182.0250	181.1680
p=2	149.6226	148.6551	148.7689	178.6921	181.7329	180.2127
p=3	159.8895	158.8167	138.7398*	165.5006	167.5490	168.5271
p=4	178.7669	177.0459	155.2526	151.4803	153.5513	151.9619
p=5	171.3528	169.3229	148.0613	147.1323	142.7587	144.4958
p=6	170.7553	171.5596	149.5391	148.4697	145.0092	148.0188

In Phase 2, the task of performing a unilateral nonlinear spatial prediction at the time $t = 82$ will be conducted by the KNAMK procedure developed in Chapter 4, that is to predict the values of $Y_{t=82}(s_0)$, $s_0 \in S$ based on the 96 forecasted values of $\tilde{Y}_{t=82}(s_i)$, $i =$

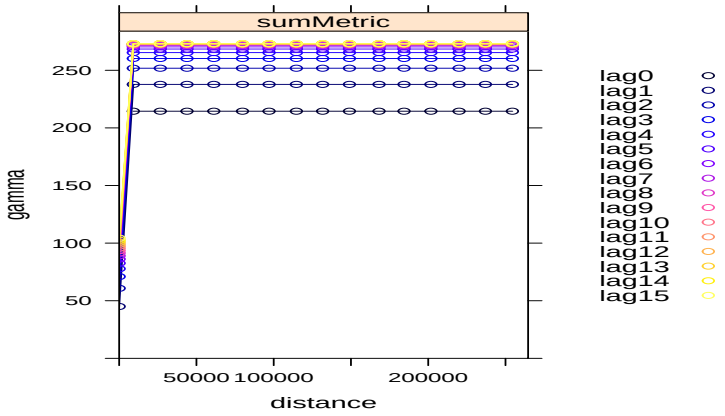

FIGURE 5.4: A variogram plot of the sum-metric model

1, ..., 96 from Phase 1. Following Sections 4.2.1 and 4.4, we achieved a MSPE $_{\text{Phase 1+2}}$ of 202.3867 by the LOOCV criterion with the K-Radius $=$ 205 Km and the two bandwidth h, b are chosen as 0.42 and 2.4, respectively.

The selected benchmark method for comparison is the Spatio-Temporal Kriging, which has a long development history in the field of Geostatistical modelling and interpolation. Cressie (1993), Cressie and Wikle (2011) and Wikle et al. (2019) are a series of books introducing this family of methods. Like spatial Kriging, spatio-temporal Kriging employs parametric spatial time covariance/variogram function to describe possible underlying spatio-temporal dependence structures. The advantages of this concept are its simplicity and easy to implement, but this family of methods poses strong Gaussian assumptions on the process which may leads to misspecification.

Among Geostatistics R packages, *gstat* gains its popularity for being continuously developed, and together supported by the *R-sig-geo* online forum [6], a mailing list for discussing the development and use of R functions and packages for handling and analysis of spatial and particularly geographical data, attracting a large number of users from the academia and industry. In *Gstat* package, the authors reuse the spacetime classes from the *Spacetime* R package for the estimation of spatio-temporal covariance/variogram models for spatio-temporal interpolation, see Pebesma et al. (2012) and Pebesma and Heuvelink (2016).

We follow the three steps of a spatio-temporal Kriging: (1) the selection of the covariance model, (2)model parameter estimation and (3) the Kriging. Among the five recommended variogram models, the sum-metric model has the minimum mean square error (MSE). Its covariance function is a combination of spatial, temporal and a metric model including an anisotropy parameter k as follows

[6] stat.ethz.ch/mailman/listinfo/r-sig-geo

TABLE 5.2: The MSPE results and the corresponding the number of temporal lags included in our spatio-temporal model

No. of temporal Lags having	3	6	10	20	30
MSPE	241.5	235.4	232.0	221.8	221.6
No. of temporal Lags having	40	50	60	70	80
MSPE	217.5	216.0*	217.7	218.0	217.8

$$C_m(d,u) = C_s(d) + C_t(u) + C_{\text{joint}}\left(\sqrt{d^2 + (k*u)^2}\right), \tag{5.24}$$

where d and u are the spatial and temporal lags mentioned earlier. Figure 5.4 serves as a demonstration showing the impact on the variograms from different temporal lags.

The last undecided parameter is the number of temporal lags to be included in this spatio-temporal Kriging. Ten sets of spatial lags were tested with the results showing their MSPE values in Table 5.2. The model with 50 temporal lags has the smallest MSPE. For this reason, it is selected for the comparison in Table 5.3.

Finally, we show the MSPE values obtained from the proposed SPKM procedure and the other three Kriging methods applied to the air quality spatial time series data for a comparison. The results in Table 5.3 suggest the SPKM procedure has the smallest MSPE value among the four methods. We conclude that our newly proposed STAR-PLR + KNAMK (SPKM) procedure outperforms the other methods. The mentioned naive Kriging in the third method takes the sample mean as the predicted outcome.

TABLE 5.3: A comparison of MSPEs from the spatio-temporal forecast methods in this chapter, these methods are demonstrated on the air quality data.

Methods	Mean Squared Prediction Error
STAR-PLR + KNAMK [SPKM]	202.3867
STAR-PLR + Linear Kriging	256.1796
STAR-PLR + Naive Kriging	267.9072
Spatio-temporal Kriging (R:Gstat)	215.9958*

Chapter 6

Conclusion and The Outlooks of this Research

In recent decades, scale has became one of the key attributes found in many scientific fields and subjects, e.g., climate change, earth sciences, astronomy, renewable energy, telecommunications, logistics and supply chain, big data, to list a few. It is where recently many new research opportunities are uncovered. With the scale often described by measured spatial information, isolated data/events could be indeed meaningful in revealing insights that would not otherwise be possible. Since the 1980s, the study of spatial statistics and more lately spatio-temporal statistics are flourished, Wikle et al. (2019) stated that there has been an exponential increase in the number of papers dealing with spatio-temporal data analysis, not only in statistics, but also in many other branches of science.

Despite the vast demand, tools and techniques for spatial and spatio-temporal series remain limited. Unlike time series analysis where nonlinear methods have been well developed for non-Gaussian data, in spatial series, the applications can be constrained by strong assumptions on stationarity and their sampling methods owning to the multilateral of space, i.e., the curse of dimensionality effect led by spatial interactions from multiple directions. The situation turns worse for spatio-temporal series, as a result, development on nonparametric analysis for such data sets is still at its early stage (Al-Sulami et al., 2017).

Our study is hereby an attempt seeking semiparametric solutions under this circumstance for analysis of nonlinear spatial and spatio-time series data collected from irregular spaced sampling grids. Furthermore, we apply our proposed new methods/procedures to the air quality data in England and compare their results with those obtained from the current conventional methods.

We divide this chapter into two sections. The first section summarises the work of this thesis and reports the main research finding and contributions. We then outline the outlooks of our research in Section 6.2 discussing the potential improvements that may be considered in future work.

6.1 Summary of the contributions

Echoing Section 1.4.1, we will by turns review the progress we have made so far. All three main contributions are evaluated with the air quality data that they outperform the popular linear Kriging methods.

6.1.1 First contribution

The first contribution is the nonparametric-trend universal Kriging (NTUK) method proposed in Chapter 3. In linear Kriging methods, by imposing linearity assumptions, the spatial trend $\mu(s)$ is modelled as a linear combinations of explanatory variables. To overcome the misspecification when dealing with nonlinear data, we purposely replace it by a nonlinear spatial trend estimated by a nonparametric local linear regression fitting. As opposed to a global linear function, we demonstrate by real data that this nonparametric estimation of spatial trend at each location offers great flexibility to the model, which allows the local information speaks for itself.

6.1.2 Second contribution

In Chapter 4, aiming to develop a nonlinear method for predicting the random residual process, we adopt a semiparametric model structure called the model averaging marginal regression (MAMAR), which was proposed by Li et al. (2015) originally for forecasting of time series. Through a nonparametric estimation of spatial probability density functions, we estimate the concerned variable at a new location s_0 as an affine combination of one-dimensional conditional regression functions based on the data. A K-radius function will then be performed by averaging the estimators conditioned from those locations within a K radius to s_0. Finally the result will be used as an predicted value of the process $X(s_0)$ at this new location. We deem this semiparametric K-radius neighbouring average based marginal Kriging (KNAMK) procedure as our second contribution. By combining the two contributions, we have a complete semiparametric spatial nonlinear Kriging method (procedure), which we have shown it performs well in spatial prediction.

6.1.3 Third contribution

We further extend our focus from spatial prediction to the realm of spatio-temporal forecasting in Chapter 5, which brings up the third contribution. Integrating our developed semiparametric spatial nonliear Kriging method with a semiparametric spatio-temporal nonlinear regression model, which allows the spatio-temporal random field to be non-stationary over space (but stationary along time; for time series, say, through differencing) but the sampling spatial grids can be irregular. A two-phase semiparametric nonlinear prediction SPKM procedure is proposed to offer a spatio-temporal forecasting for a future time at a new unobserved location.

All three main contributions are evaluated by the air quality data set, and they perform better than the common linear methods in use today.

6.2 The outlook of this research

This research is a journey of making improvements for spatial prediction from the existing linear Kriging methods. From the three contributions made in this thesis, we look further for areas for future research also the possible opportunities for applying these methods.

6.2.1 Areas for future research

Among the three contributions, two direct improvements could be made as the future research: varying smoothing parameters (bandwidths) in the estimation of the spatial probability density functions in Chapter 4, and the penalised lag effects for the semiparametric spatio-temporal nonlinear regression (Phase 1 of SPKM procedure) in Chapter 5.

We start with the first possible improvement. In KNAMK procedure, a K-radius is defined to penalise, or shrink, the estimated long-distanced marginal regression functions to zero at a new location. Alternatively, we can achieve this by introducing varying smoothing parameters, i.e., the bandwidths in the estimation of spatial probability density functions. In Chapter 4, Epanechnikov kernel is used for both spatial bandwidths, $h \in \mathbb{R}$ and $b \in \mathbb{R}^2$. Hence, bandwidths with a distance-related tier(tag) system could be adopted for this purpose. Another development in literature about this topic is the method of adaptive bandwidth choice for spatial density function (Jiang et al., 2020), in which the bandwidth varies based on local data and therefore adaptively conforms with local features of the spatial data. A spatial cross-validation (SCV) choice was proposed to facilitate this method. However, it is noticed that this method was by far performed on spatial lattice. Significant work may be required to adopt it to the air quality case sampled from irregular grid.

The second possible area for future research is in the spatio-temporal autoregressive partially nonlinear regression (STAR-PLR) model, i.e., the Phase 1 Step 1 of the SPKM

procedure in Chapter 5. In Model (5.18), a spatial lagged response variable, $Y_t^{sl}(s_j) = \sum_{k=1}^{N} w_{jk} Y_t(s_k)$ is defined to model the spatial interactions among the sampling sites, within which the spatial weight matrix $W = (w_{jk})_{j,k=1}^{N}$ is assumed to be priori, the value of each w_{jk} is defined by the distance between s_j and s_k. However, as Al-Sulami et al. (2019) stated, this pre-specified neighbourhood structure can be subjective despite of being a common approach in applied econometrics. A more involved model can therefore be defined by extending model (5.18) to model (6.1) into the form of

$$Y_t(s_j) = g(X_t(s_j), s_j) + \sum_{i=1}^{p} \sum_{k=1}^{N} \lambda_{jk,i} Y_{t-i}(s_k) + \sum_{l=1}^{q} \alpha_l(s_j) Y_{t-l}(s_j) + \epsilon_t(s_j), \qquad (6.1)$$

where $\sum_{i=1}^{p} \lambda_i(s_j) Y_{t-i}^{sl}(s_j)$, the second term on the RHS of model (5.18), is modified by adding the parametric spatio-temporal lag interactions. A penalized procedure utilising adaptive Lasso was developed for the identification and estimation of such lag interactions (Al-Sulami et al., 2019). Another model of applying adaptive lasso for spatial lattice data can be found in Zhu et al. (2010).

6.2.2 Possible areas for applications

The final words of this thesis go to some thoughts on possible applications of this research. The study of air quality data is a conventional choice with specific purposes, which requires few additional efforts to introduce. However, we believe that there are a wide range of practical applications for our proposed methods. We would like to list two possible areas as the examples for future applications: one from the operational research point of view, and the other from the finance and econometrics aspect.

In operational research, after the descriptive analysis about 'what is it?', the ultimate goal is to conduct predictive and prescriptive analysis, that answers the questions such as 'what is likely to happen?' and 'what should we do?'. Thus, studies on resource allocation and energy distribution can be two good candidates for our methods.

On resource allocation, for instance when the NHS forecasts the changing number of patients at each local area during a pandemic, or answers how the NHS and its contractor arrange logistics for distributing each batch of vaccine. The proposed spatio-temporal models in this research can be adapted for such purposes. Another example could be for city planning or for relocation of a warehouse or a facility, where making decisions with spatial viewpoint would be extremely useful for local councils or business owners.

From the finance and econometrics aspect, green finance and energy pricing are two good examples both emphasising the long-term sustainability, which is inevitably associated with time and stationary conditions. Furthermore, we also notice that spatio-temporal analysis is used for risk control and profitability management as well.

In summary, spatial time series analysis is the future in this data-driven era, offering vast opportunities in the increasingly connected world today.

Appendix A

Air Quality Data Set

Empirical applications and the evaluation of our spatial and spatio-temporal prediction methods are executed by the statistical software R, using the air quality data from the UK-AIR database archived by the Department for Environment Food & Rural Affairs (Defra), the UK.

In Chapters 1 - 4, we take exemplarily the date 18/04/2017 as the single point of time for spatial prediction. We choose the monitoring stations from Defra's Hourly networks in England (see Table A.1 for an overview), where the first 35 out of the total 105 stations and their corresponding values are listed. The Easting/Northing and Latitude/Longitude are two standard geographic coordinates systems used to position a location on Earth, which are commonly used in scientific software. It is noticed that the monitoring stations are irregularly positioned in England, as shown in Figure 2.3. Furthermore, neither the observations nor the detrended residuals follow a Gaussian profile, as seen in Figure A.1 and Figure 3.2.

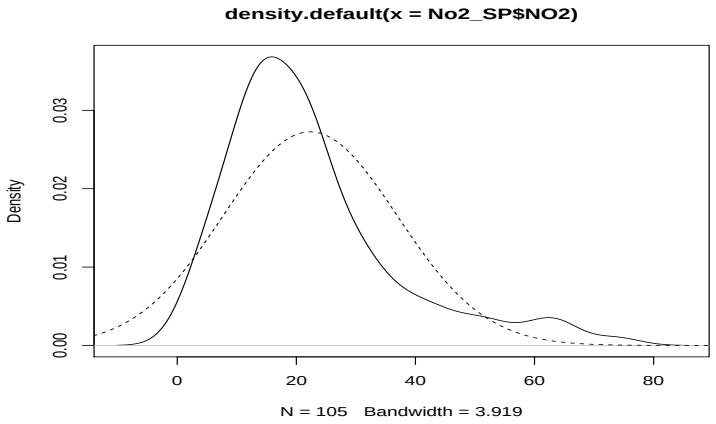

FIGURE A.1: A density function of the observed data on 18/04/2017 with a matching Gaussian profile

For the spatio-temporal forecast case in Chapter 5, we set 18/04/2017 as the future date, and use its previous 81 temporal lags (days) as the possible training set for Phase 1 of the SPKM procedure. The number of monitoring stations is reduced from 105 to 96 due to missing data.

TABLE A.1: The first 35 monitoring stations in the spatial prediction data set for Chapters 1 - 4.

Stations	Easting	Northing	Latitude	Longitude
Barnsley Gawber	432524	407478	53.56292	-1.510436
Bath Roadside	375455	165847	51.391127	-2.354155
Billingham	446928	523597	54.60537	-1.275039
Birkenhead Borough Road	331926	388453	53.388511	-3.025014
Birmingham A4540 Roadside	408586	286470	52.47609	-1.875024
Birmingham Acocks Green	411654	282146	52.437165	-1.829999
Blackburn Accrington Road	370242	428026	53.747751	-2.452724
Blackpool Marton	333768	434759	53.80489	-3.007175
Bournemouth	412322	93343	50.73957	-1.826744
Bradford Mayo Avenue	415931	430572	53.771245	-1.759774
Brighton Preston Park	530524	106225	50.840836	-0.147572
Bristol St Paul's	359492	173925	51.462839	-2.584482
Bristol Temple Way	359523	173383	51.457968	-2.583975
Bury Whitefield Roadside	380637	406974	53.559029	-2.293772
Cambridge Roadside	545279	258142	52.20237	0.124456
Camden Kerbside	526633	184390	51.54421	-0.175269
Cannock A5190 Roadside	401394	309957	52.687298	-1.980821
Canterbury	616187	157319	51.27399	1.098061
Carlisle Roadside	339469	555976	54.894834	-2.945307
Charlton Mackrell	352196	128768	51.05625	-2.68345
Chatham Roadside	577437	166993	51.374264	0.54797
Chesterfield Loundsley Green	436470	372039	53.244131	-1.454946
Chesterfield Roadside	436348	370658	53.231722	-1.456944
Chilbolton Observatory	439390	139078	51.149617	-1.438228
Christchurch Barrack Road	415559	92894	50.735454	-1.780888
Coventry Allesley	430011	279376	52.411563	-1.560228
Coventry Binley Road.	434785	278978	52.407708	-1.490082
Doncaster A630 Cleveland Street	457247	402812	53.518868	-1.138073
Eastbourne	560155	103150	50.805778	0.271611
Exeter Roadside	291929	92838	50.725083	-3.532465
Glazebury	368755	396030	53.46008	-2.472056
Haringey Roadside	533894	190707	51.5993	-0.068218
Honiton	315749	99874	50.792287	-3.196702
Horley	528206	142331	51.165865	-0.167734
Hull Freetown	509482	429322	53.74878	-0.341222

Appendix B

Parametric Variogram Models

B.1 Parametric models for spatial data

In Sections 2.3.2 and 4.2.3, the process of fitting theoretical variogram is intensively discussed. Webster and Oliver (2007) commented that bounded models are more commonly in use than the unbounded variation from experience. In this section, we present a short list of commonly used bounded, isotropic and valid variogram model families, which can be found in Wackernagel (2003) and Webster and Oliver (2007).

Let $\gamma_{a,b}(d)$ denotes the semivariogram function, $C_{a,b}(d)$ the corresponding covariance function with spatial lag d, and $a, b > 0$ the parameters of each model, represent the range and sill parameters, respectively.

(i) Nugget-effect model:

$$\gamma^{nug}_{a,b}(d) := \begin{cases} 0, & \text{if } |d| = 0, \\ b, & \text{otherwise.} \end{cases}$$

$$C^{nug}_{a,b}(d) := \begin{cases} b, & \text{if } |d| = 0, \\ 0, & \text{otherwise.} \end{cases}$$

(ii) Bounded linear model:

$$\gamma^{lin}_{a,b}(d) := \begin{cases} b(\frac{|d|}{a}), & \text{if } |d| \leq a, \\ b, & \text{otherwise.} \end{cases}$$

$$C^{lin}_{a,b}(d) := \begin{cases} b(1 - \frac{|d|}{a}), & \text{if } |d| \leq a, \\ 0, & \text{otherwise.} \end{cases}$$

(iii) Spherical model:

$$\gamma_{a,b}^{sph}(d) := \begin{cases} b(\frac{3}{2}\frac{|d|}{a} - \frac{1}{2}(\frac{|d|}{a})^3), & \text{if } |d| \leq a, \\ b & \text{, otherwise.} \end{cases}$$

$$C_{a,b}^{sph}(d) := \begin{cases} b(1 - \frac{3}{2}\frac{|d|}{a} - \frac{1}{2}(\frac{|d|}{a})^3), & \text{if } |d| \leq a, \\ 0 & \text{, otherwise.} \end{cases}$$

(iv) Exponential model:

$$\gamma_{a,b}^{exp}(d) := b\ (1 - \exp(-\frac{|d|}{a})).$$

$$C_{a,b}^{exp}(d) := b\ \exp(-\frac{|d|}{a})).$$

(v) Gaussian model:

$$\gamma_{a,b}^{gau}(d) := b\ (1 - \exp(-\frac{|d|^2}{a^2})).$$

$$C_{a,b}^{gau}(d) := b\ \exp(-\frac{|d|^2}{a^2})).$$

When analysing data from real applications, often the variogram appears to be more complex. Webster and Oliver (2007) suggested that it is common to combine some of the basic models to achieve a better fit. The most common combination of this kind is to add a nugget parameter c_0 into another model, e.g., $\gamma_{a,b,C_0}^{gau}(d) := c_0 + \gamma_{a,b}^{gau}(d)$ represents a modified Gaussian model with a nugget component, the sill now becomes the sum of b and c_0, where the value b is named partial sill (Cressie, 1988).

The *gstat* package in R is used to perform linear Kriging in this research by which the sum of squared errors under least squares fitting method is calculated for each model family. The results are then compared for identifying the best theoretical model.

B.2 Parametric models for spatio-temporal data

In Section 5, the focus expands to applications with spatial-time data sets. To perform spatio-temporal interpolation using *gstat* R package, we list four basic model classes: the separable, product-sum, metric and sum-metric spatio-temporal covariance functions from Pebesma and Heuvelink (2016):

a) The *separable* covariance model assumes that the spatio-temporal covariance function can be shown as the product of a spatial and temporal term:

$$C_{\text{sep}}(d,u) = C_s(d) + C_t(u),$$

where d and u represent the spatial lag and time lag respectively.

b) Extended from the above model, the *product-sum* covariance model takes the form as

$$C_{\text{ps}}(d,u) = kC_s(d)C_t(u) + C_s(d) + C_t(u),$$

with $k > 0$.

c) Assuming identical spatial and temporal covariance functions except for spatio-temporal anisotropy, the spatio-temporal *metric* covariance model employs a single covariance model C_{joint},

$$C_{\text{m}}(d,u) = C_{\text{joint}}(\sqrt{d^2 + (k*u)^2}),$$

where k is an anisotropy correction parameter.

d) Combining spatial, temporal and a metric model, the *sum-metric* covariance model is

$$C_{\text{sm}}(d,u) = C_s(d) + C_t(u) + C_{\text{joint}}\Big(\sqrt{d^2 + (k*u)^2}\Big).$$

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

D. Al-Sulami, Z. Jiang, Z. Lu, and J. Zhu. Estimation for semiparametric nonlinear regression of irregularly located spatial time-series data. *Econometrics and Statistics*, 2:22–35, 2017.

D. Al-Sulami, Z. Jiang, Z. Lu, and J. Zhu. On a semiparametric data-driven nonlinear model with penalized spatio-temporal lag interactions. *Journal of Time Series Analysis*, 40(3):327–342, 2019.

L. Anselin. Spatial econometrics: Methods and models. 1988.

L. Anselin. Exploring spatial data with geodatm: a workbook. *Urbana*, 51(61801):309, 2004.

M. S. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41, 1946.

R. E. Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 1961.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

P. J. Bickel and K. A. Doksum. *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press, 2015.

R. S. Bivand, E. J. Pebesma, V. Gomez-Rubio, and E. J. Pebesma. *Applied spatial data analysis with R*, volume 747248717. Springer, 2008.

T. Bollerslev. Glossary to arch (garch. In *in Volatility and Time Series Econometrics Essays in Honor of Robert Engle. MarkWatson, Tim Bollerslev and Je¤ rey*. Citeseer, 1986.

A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

G. E. Box and G. M. Jenkins. Time series analysis: Forecasting and control holden-day. *San Francisco*, page 498, 1970.

R. B. Cattell. The description of personality: Basic traits resolved into clusters. *The journal of abnormal and social psychology*, 38(4):476, 1943.

R. Chen and R. S. Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993.

G. Claeskens, N. L. Hjort, et al. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

N. Cressie. Spatial prediction and ordinary kriging. *Mathematical geology*, 20(4):405–421, 1988.

N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2011.

N. A. Cressie. Statistics for spatial data: Wiley series in probability and mathematical statistics. *Find this article online*, 1993.

J. Davidson. *Stochastic limit theory: An introduction for econometricians*. OUP Oxford, 1994.

K. Dehnad. Density estimation for statistics and data analysis, 1987.

C. M. Deo. A note on empirical processes of strong-mixing sequences. *The Annals of Probability*, pages 870–875, 1973.

R. Douc, E. Moulines, and D. Stoffer. *Nonlinear time series: Theory, methods and applications with R examples*. CRC press, 2014.

H. E. Driver and A. L. Kroeber. *Quantitative expression of cultural relationships, by HE Driver and AL Kroeber*. University of California Press, 1932.

R. Engle. Arch with estimates of variance of united kingdom inflation. *Econometrica*, 50 (4):987–1007, 1982.

ESPON. Espon project 1.4. 3 study on urban functions. final report, march 2007, 2007.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

J. Fan and Q. Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer, New York, 2003.

J. Fan, Q. Yao, and Z. Cai. Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 65(1):57–80, 2003.

R. A. Fisher. Design of experiments. *Br Med J*, 1(3923):554–554, 1936.

M. Frazier. Overview of coordinate reference systems (crs) in r national center for ecological analysis and synthesis, 2017.

W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

J. Gao. *Nonlinear time series: semiparametric and nonparametric methods*. CRC Press, 2007.

J. Gao, Z. Lu, D. Tjøstheim, et al. Estimation in semiparametric spatial regression. *The Annals of Statistics*, 34(3):1395–1435, 2006.

A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

P. Hall and P. Patil. Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, 99(3):399–424, 1994.

M. Hallin, Z. Lu, L. T. Tran, et al. Density estimation for spatial linear processes. *Bernoulli*, 7(4):657–668, 2001.

M. Hallin, Z. Lu, L. T. Tran, et al. Local linear spatial regression. *The Annals of Statistics*, 32(6):2469–2500, 2004.

J. D. Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, NJ, 1994.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

I. Ibragimov et al. Independent and stationary sequences of random variables. 1971.

C. Jean-Paul and D. Pierre. Geostatistics: modeling spatial uncertainty. *John Wiley & Sons Inc., New York*, page 695, 1999.

Z. Jiang, N. Ling, Z. Lu, D. Tjøstheim, and Q. Zhang. On bandwidth choice for spatial data density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):817–840, 2020.

R. Kerry and M. Oliver. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma*, 140(4):383–396, 2007.

P. K. Kitanidis. *Introduction to geostatistics: applications in hydrogeology*. Cambridge university press, 1997.

D. G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6): 119–139, 1951.

D. Li, O. Linton, and Z. Lu. A flexible semiparametric forecasting model for time series. *Journal of Econometrics*, 187(1):345–357, 2015.

Q. Li and J. S. Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

A. Lichtenstern. Kriging methods in spatial statistics. *Technische Universität München*, 2013.

Z. Lu and D. Tjøstheim. Nonparametric estimation of probability density functions for irregularly observed spatial data. *Journal of the American Statistical Association*, 109 (508):1546–1564, 2014.

Z. Lu, A. Lundervold, D. Tjøstheim, and Q. Yao. Exploring spatial nonlinearity using additive approximation. *Bernoulli*, pages 447–472, 2007.

Z. Lu, D. J. Steinskog, D. Tjøstheim, and Q. Yao. Adaptively varying-coefficient spatiotemporal models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 71(4):859–880, 2009.

E. Mammen, O. Linton, J. Nielsen, et al. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27 (5):1443–1490, 1999.

G. Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

G. Matheron. The theory of regionalized variables and its applications, vol. 5. *Paris: École National Supérieure des Mines*, 211, 1971.

G. Matheron. A simple substitute for conditional expectation: the disjunctive kriging. In *Advanced geostatistics in the mining industry*, pages 221–236. Springer, 1976.

G. Matheron. Isofactorial models and change of support. In *Geostatistics for natural resources characterization*, pages 449–467. Springer, 1984.

Y. Matsuda and Y. Yajima. Fourier analysis of irregularly spaced data on rd. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):191–217, 2009.

H. J. Miller. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284–289, 2004.

R. A. Moyeed and A. Papritz. An empirical comparison of kriging methods for nonlinear spatial point prediction. *Mathematical Geology*, 34(4):365–386, 2002.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

Y. Ozaki, M. Yano, and M. Onishi. Effective hyperparameter optimization using nelder-mead method in deep learning. *IPSJ Transactions on Computer Vision and Applications*, 9(1):20, 2017.

E. Pebesma and B. Gräler. Introduction to spatio-temporal variography. 2021.

E. Pebesma and G. Heuvelink. Spatio-temporal interpolation using gstat. *RFID Journal*, 8(1):204–218, 2016.

E. Pebesma et al. spacetime: Spatio-temporal data in r. *Journal of statistical software*, 51 (7):1–30, 2012.

S. S. Rao. Statistical analysis of a spatio-temporal model with location-dependent parameters and a test for spatial stationarity. *Journal of Time Series Analysis*, 29(4):673–694, 2008.

O. Schabenberger, C. A. Gotway, et al. Statistical methods for spatial data analysis. 2005.

G. Schwarz et al. Estimating the dimension of a model. *Annals of statistics*, 6(2):461–464, 1978.

G. Seymour. Predictive inference. *0412034719Chapman and Hall, New York*, 1993.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

E. Slutzky. The summation of random causes as the source of cyclic processes. *Econometrica: Journal of the Econometric Society*, pages 105–146, 1937.

A. Stein, F. D. van der Meer, and B. Gorte. *Spatial statistics for remote sensing*, volume 1. Springer Science & Business Media, 2006.

M. L. Stein. Interpolation of spatial data: some theory for kriging. Technical report, 1999.

Y. Sun, H. Yan, W. Zhang, Z. Lu, et al. A semiparametric spatial dynamic model. *Annals of Statistics*, 42(2):700–727, 2014.

T. Terasvirta, D. Tjostheim, C. W. Granger, et al. Modelling nonlinear economic time series. *OUP Catalogue*, 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

H. Tong. *Non-linear time series: a dynamical system approach*. Oxford University Press, 1990.

R. S. Tsay and R. Chen. *Nonlinear time series analysis*, volume 891. John Wiley & Sons, 2018.

H. Wackernagel. External drift. In *Multivariate Geostatistics*, pages 21–23. Springer, 2003.

M. Ward. Spatial epidemiology: Where have we come in 150 years? In *Geospatial Technologies and Homeland Security*, pages 257–282. Springer, 2008.

G. S. Watson. Advanced geostatistics in the mining industry., 1977.

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

H. White. *Asymptotic theory for econometricians*. Academic press, 2014.

C. K. Wikle, A. Zammit-Mangion, and N. Cressie. *Spatio-temporal Statistics with R*. CRC Press, 2019.

M. Wilhelmsson. Spatial models in real estate economics. *Housing, theory and society*, 19 (2):92–101, 2002.

H. O. A. Wold. *A Study in the Analysis of Stationary Time Series: With an Appendix*. Almqvist & Wiksell, 1954.

S. Yakowitz and F. Szidarovszky. A comparison of kriging with nonparametric regression methods. *Journal of Multivariate Analysis*, 16(1):21–53, 1985.

G. U. Yule. On the time-correlation problem, with especial reference to the variate-difference correlation method. *Journal of the Royal Statistical Society*, 84(4):497–537, 1921.

J. Zhu, H.-C. Huang, and P. E. Reyes. On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3): 389–402, 2010.