Non-linear Univariate and Multivariate Spatial Modelling and Optimal Design

Gnai Nishani Musafer

B.Sc (Hons)

Principal supervisor: Dr. Helen Thompson

Associate supervisors: Prof. Erhan Kozan, Prof. Rodney Wolff

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy (Research)

School of Mathematical Sciences

Faculty of Science and Engineering

Queensland University of Technology (QUT), Australia

(2016)

Keywords

Spatial, Copula, Pair-copula, Geostatistics, Artificial neural networks, Mining, Non-linear spatial dependency, Non-linear principal component analysis, Multivariate spatial statistics, Spatial design, Multivariate spatial design, Monte Carlo simulation, Sequential simulation, Non-linear transformation, Sequential design.

Abstract

The research in this thesis was motivated by challenges that arose in investigating optimal designs for additional drillings in the field of mining. However, these challenges can be generalised to any field that deals with spatial data.

In mine projects, more knowledge about the ore body, in addition to the knowledge obtained from initial drillings, is required for strategic mine planning. Hence additional drilling campaigns are carried out and optimal design concepts are applied in order to balance the benefit between drilling costs and additional information. Optimal design for additional drills for one variable based on conventional geostatistical models, such as kriged models, is a well understood problem. However, it has been identified that it is not only the grade but also other variables, such as concentration of deleterious elements and hardness, that play significant roles in the evaluation of the cost and revenue of mine projects. Moreover, these variables are unlikely to be totally independent and the dependence between these variables can be non-linear. In addition, in reality, the spatial dependence structure of an individual variable can also be non-linear.

This thesis aims to develop general methodology for the optimal design of additional sampling based on a geostatistical model that can preserve both multivariate non-linearity and spatial non-linearity present in spatial variables. This methodology can be applied in mining or any other field that deals with spatial data. We focus on copula-based geostatistical models since these models offer a solution to modelling non-linear spatial dependence in individual spatial variables. Specifically, the pair-copula model, among other simple copula-based models, has more flexibility to capture the non-linear dependence structure. The four contributions of this thesis to research, based on pair-copulas, are as follows.

Firstly, the existing pair-copula is improved by developing an algorithm to optimally determine the distance classes required in pair-copula modelling. Within the algorithm, a goodness-of-fit test used to compare two classical copulas is extended to compare two spatial copulas. The results of two case studies show an improvement in fit of the pair-copula model based on distance classes using the proposed algorithm compared to using distance classes of equal width, as implemented in the literature.

Secondly, new methodology for modelling non-linear multivariate spatial data is developed based on non-linear principal components analysis (NLPCA) and the pair-copula model. The results from two case studies illustrate that the proposed methodology preserves both multivariate non-linearity and spatial non-linearity present in the spatial variables.

Thirdly, a new sequential adaptive optimal design for univariate spatial data based on the pair-copula model, in order to reduce the uncertainty in spatial prediction, is proposed. The sequential design is a simulation-based design. The performance of the proposed methodology is evaluated by partially redesigning an existing spatial design. The results demonstrate, in the case study presented, that the proposed design methodology outperforms a traditional kriging based design.

Finally, methodology for the optimal design of additional sampling is proposed based on the non-linear multivariate model in order to simultaneously reduce the uncertainty of multiple variables in multivariate spatial prediction. Based on simulation results and results of the case studies using the proposed methodology, it can be conjectured that selecting optimal locations for new samples based on the correct model which honour the *in-situ* dependence of the spatial data will improve the precision of multivariate prediction in the spatial random field.

Ultimately, results from each contribution indicate that the pair-copula model, its extensions and sampling optimal designs based on these shows promising improvement over existing methods.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

QUT Verified Signature

Signature:

Date: 30/04/2016

Acknowledgement

First at most, I would like to express my deepest gratitude to my principal supervisor, Dr. Helen Thompson. It has been an honour and a priviledge to be one of her PhD students. The advices, guidance and the encouragements given by her as the supervisor, as a mentor and sometimes as a friend truly helped me to accomplish my research goals. Undoubtedly, working with Dr. Helen Thompson was the most exciting and rewarding experience I've ever had in my academic life.

Moreover, I sincerely thank my assistant supervisors, Prof.Erhan Kozan and Prof. Rodney Wolf for their help and valuable suggestions during my PhD work. My special thanks to Prof. Voh Anh for his kind recommendation and help given for obtaining this valuable opportunity.

I would like to acknowledge CRC ORE (Cooperative Research Centre optimising resource extraction) for the financial support given. Additionally, I also would like to acknowledge the High performance computing and research support unit (HPC) at QUT for resources and support that I have used during this research.

Further, I would like to extend my thanks to all colleagues and friends who adviced and encouraged me during this three years period.

Last but not the least I wish to thank my family members for their consistent love, support and encouragement which are motivated me to be strong during any hard times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

This research is mainly motivated by the challenges that arose in investigating optimal designs for additional drillings in the field of mining. However, these challenges can be generalised to any field that deals with spatial data.

In mine projects, more knowledge about the ore body, in addition to the knowledge obtained from initial drillings, is required to make accurate decisions during strategic and tactical mine planning. Hence, additional drilling campaigns are carried out and optimal design concepts are applied in order to balance the benefit between production cost and additional information. Optimal design for additional drillings based on a geoscientific variable, such as metal grade, is a well understood problem (Walton and Kauffman [1982], Scheck and Chou [1983], Koppe et al. [2011]). However, it has been identified that it is not only the grade but also the geometallurgical variables, such as concentration of deleterious elements and hardness, that play significant roles in the evaluation of the cost and revenue of mine projects (Dunham and Vann [2007]). Therefore, it is rquired model the dependency of geometallurgical variables and metal grade simultaneously where the variables may be correlated. Moreover, the dependence between these variables can be non-linear. In addition, the spatial dependence of individual variables can be non-linear (See the definition in Appendix A). Thus, it is required to develop an optimal design based on a geostatistical model that can

reflect the non-linear spatial dependence structure between these variables.

For instance, suppose that the actual relationship between variables is non-linear and spatial dependence within an individual variable is non-linear. If a geostatistician fits a model that ignores these non-linearities and uses that model to develop an optimum sampling design for additional drillings, and subsequently estimates the ore reserve using the additional drilling information, the final estimation of the ore reserve will be inaccurate. There is no improvement that can be gained with an optimum design without a valid model. It should be noted that the dependency of an optimal design on the assumed model is not unique to spatial data. These challenges are not specific to the field of mining and are general to any field that deals with spatial data.

In traditional geostatistics, even though optimal design targeting one variable is a well understood problem, optimal design for one variable with a non-linear dependence structure (See the definition in Appendix A) is rarely addressed. By considering non-linearity in the dependence structure, a more complete estimate of uncertainty in spatial field prediction can be gained. Here "complete" uncertainty estimation means a measurement that can capture not only the variation of the configuration of the spatial locations but also the variation of the measured values for those spatial locations. Hence, an optimal design for additional samples based on a model that can capture the non-linear dependence will result in more precise estimates. However, most univariate geostatistical models use the variogram to model spatial dependence (Kazianka and Pilz [2010b]). The variogram measures the dissimilarity, or increasing variance (decreasing correlation), of the variable of interest at different locations (King [2011]). Hence this can be considered as a measure of linear dependence over the distribution of the variable for a given spatial distance. Therefore this tool is inappropriate if non-linear dependence is present. Moreover, some other limitations of the variogram have also been discussed in the literature, such as sensitivity to extreme values and inability to provide more than a single measure of dependence (Li [2010]). Consequently, any model that employs the variogram in the estimation process

may not be able to provide accurate estimation for real world phenomena.

Optimal design for additional samples in the multivariate setting is also poorly addressed in the literature. Even though there are a few multivariate optimal designs proposed in the literature, most designs use multivariate geostatistical models that only model linearity between variables and also ignores the non-linear dependence in the individual variables.

The challenges and problems in the above discussion motivate this research to contribute new methodologies in both spatial modelling and sampling design. This research will develop a novel geostatistical multivariate modelling technique that captures both multivariate non-linearity and spatial non-linearity present in spatial variables (See the definition in Appendix A). We focus on copula-based geostatistical models since these models offer a solution to modelling non-linear spatial dependence in individual spatial variables. Specifically, the pair-copula model, among other simple copula-based models which were introduced by Bárdossy and Li [2008], has more flexibility to capture the non-linear dependence structure (Gräler and Pebesma [2011]). Optimal sampling design strategies are then developed for additional samples based on this modelling approach. The model, and subsequent optimal sampling design, will enable richer information, i.e., information over the entire distribution of any variables of interest with greater precision in estimates, to be obtained from the spatial process. The richer information from the spatial study will consequently enhance all elements of the spatial process.

The mining data provided by the funding organisation for this research did not contain the expected spatial complexities, hence data from environmental applications, such as soil contaminations and forest inventory attributes are additionally used, to demonstrate the proposed methods for complex spatial data.

1.2 Aims and Objectives of the Research

1.2.1 Aim

The ultimate aim of this research is to increase knowledge of the spatial process, in particular information on spatial characteristics, through enhanced statistical modelling of the spatial process and through improved collection of additional information. The novel multivariate geostatistical model will enable more accurate estimation of characteristics of spatial variables. The sampling design for additional samples, based on this model, will sample locations of the spatial domain to provide richer information on the spatial process than would otherwise have been collected trough sub-optimal sampling or based on an inferior model.
Specifically, the main aim is to develop general methodology for the optimal design of additional sampling based on a geostatistical model that can preserve both multivariate non-linearity and spatial non-linearity present in spatial variables, which can be applied in mining or any other field that deals with spatial data. However, the main aim can only be achieved by achieving the specific aims which are related to modelling and design as follows.

1. Aims related to modelling.

   (a) Univariate model: Improve the existing copula-based spatial model proposed by Gräler and Pebesma [2011] to estimate characteristics of a single spatial variable with non-linear spatial dependence.

   (b) Multivariate model: To extend the copula-based spatial model proposed by Gräler and Pebesma [2011] to the multivariate setting to estimate characteristics of two or more spatial variables whilst capturing their non-linear relationship by applying a suitable transformation.

2. Aims related to optimal design.

   (a) Univariate design: Develop an optimal sampling design for additional samples based on the pair-copula model proposed by Gräler and Pebesma

[2011] with the objective of reduction of prediction uncertainty.

(b) Multivariate design: Develop an optimal sampling design based on the proposed multivariate model that can capture both multivariate non-linearity and spatial non-linearity with the objective of reduction of prediction uncertainty for all the variables simultaneously.

1.2.2 Objectives

The objectives related to each sub-aim are as follows.

1. Objectives related to modelling.

   (a) Univariate model: Improve the pair-copula model by introducing a new algorithm to define lag distances. In order to develop the algorithm, the test proposed by Rémillard and Scaillet [2009], which is used to compare non-spatial two copulas, is extended to the spatial framework.

   (b) Multivariate model: Extend the pair-copula model introduced by Gräler and Pebesma [2011] to the non-linear multivariate setting by integrating non-linear principal components analysis to remove the non-linear relationship among the variables of interest.

2. Objectives related to optimal design.

   (a) Univariate design: Develop an optimal design for additional samples based on the pair-copula model by modifying the approach proposed by Li et al. [2011] in order to reduce prediction uncertainty.

   (b) Multivariate design: Extend the univariate design approach to the multivariate setting in order to reduce the prediction uncertainty in all variables simultaneously based on the model proposed in objective 1.(b).

1.3 Research Questions

The following research questions related to spatial modelling and design are addressed in this research.

1. Modelling: How can a multivariate geostatistical model be developed to capture both multivariate non-linearity and spatial non-linearity present in spatial variables based on a pair-copula model?

2. Optimal design: How can an optimum sampling design be developed for additional samples based on this multivariate geostatistical model that accounts for both multivariate non-linearity and spatial non-linearity present in spatial variables?

1.3.1 Research sub-questions

The research questions above can be broken up into the following sub-questions.

1. Sub-questions related to modelling.

   (a) Univariate model:

      i. Will a pair-copula model produce better prediction than a conventional univariate geostatistical model?

      ii. How can the pair-copula model be improved to produce more precise prediction?

      iii. How much more accurate is the improved pair-copula model than the existing pair-copula model?

   (b) Multivariate model:

      i. How can the univariate pair-copula model be extended to the multivariate setting to capture non-linearity between spatial variables?

      ii. Will the property of capturing the non-linear dependece in individual variables of the pair-copula model be retained when applied to the multivariate setting?

iii. Can any improvement in prediction be gained by using non-linear multivariate modelling based on pair-copulas when compared to traditional multivariate geostatistical modelling approaches?

2. Sub-questions related to optimal design.

   (a) Univariate design:

      i. Based on the pair-copula model, what statistical criteria should be used to develop optimal designs?

      ii. What are the resultant designs for the additional samples?

      iii. How do the resultant designs based on the pair-copula vary from designs based on the conventional design approach?

      iv. How much more precise are estimates from a design based on pair-copulas than estimates from a design based on conventional geostatistical models?

   (b) Multivariate design:

      i. How can the univariate design in objective 2.(a) be extended to the multivariate case using the multivariate model from objective 1.(b)?

      ii. What statistical criterion should be used in multivariate design?

      iii. Will the resultant multivariate design be optimal for all the variables of interest?

      iv. How does this design vary from the univariate design based on the pair-copula model?

      v. Is there any difference between the proposed sampling locations from the design based on the model that can capture both non-linearity between variables and within individual variables and the design based on models that ignore the non-linearity between the variables and within the variables?

      vi. How much more precise is the design based on the model that can capture non-linearity than the design based on the model that

ignores non-linearity?

1.4 Significance of the Research

Ultimately, this research presents four main contributions to the field of geostatistics.

There is no well-defined procedure in the literature to define the distance classes required in pair-copula modelling. The first part of this research develops an efficient algorithm to define the distance classes required in the pair-copula model. When developing this new algorithm, a goodness of fit test that is used to compare the equality between two non-spatial copulas (Rémillard and Scaillet [2009]) is extended to the spatial setting. In addition to the algorithm, this extension is another new contribution to geostatistics. By developing a pair-copula based on the distance classes defined by this algorithm, more precise predictions can be obtained than the estimates obtain by the existing pair-copula model with equal width distance classes.

The second contribution of this research is the development of a novel geostatistical approach to model non-linear multivariate spatial dependence using non-linear principal components analysis (NLPCA) and pair-copulas. This work extends the work of Barnett and Deutsch [2012] and Barnett et al. [2014] by considering non-linear spatial data and, consequently, non-linear multivariate decomposition of non-linear spatial data that retains non-linearity of the spatial data. This work also extends the work of Gräler and Pebesma [2011] and Gräler [2014] by introducing the pair-copula to the multivariate framework. By applying this proposed geostatistical approach to spatial data, any non-linear dependence between variables and non-linear spatial dependence structure in individual variables can be captured. As a result, by employing the proposed modelling approach, simultaneous simulation or simultaneous interpolation would be more precise than the results from the existing approaches.

A new sequential adaptive optimal design for univariate spatial data based on the pair-copula model in order to reduce the uncertainty in spatial prediction is the

third contribution. As far as the author is aware spatial optimal design based on the pair-copula is considered for the first time. If the proposed design methodology is used, optimal locations for the additional samples can be obtained based on the spatial configuration of the observations and their measured values. Finally, the precision of prediction is increased if information on additional samples that are obtained based on the proposed methodology is used.

The final contribution to research is the development of an optimal spatial multivariate design for additional samples. The model in the second contribution is used in order to simultaneously reduce the uncertainty estimation of multiple variables. By employing this proposed multivariate designs methodology, precision of the prediction of more than one variable is increased when compared to traditional design approaches.

1.5 Scope of Organisation of the Thesis

The scope of this thesis is to improve spatial modelling, spatial interpolation and spatial design using the pair-copula based geostatistical model with the objective of increasing accuracy in decision making for spatial processes. The remainder of this thesis is organised is as follows.

A literature review is contained in Chapter 2 where traditional univariate geostatistical models and multivariate geostatistical models are briefly summarised. The strengths and weakness of these models are reviewed and compared against copula based models, specifically, pair-copula based models. Moreover, the general approaches of spatial designs are also summarised in Chapter 2. In addition to this, each Chapter contains a literature review related to its topic.

A detailed description of the pair-copula based geostatistical model, including its strengths and weaknesses, is discussed in Chapter 3 and an application of the pair-copula model to mining data is also presented.

Chapter 4 discusses an improvement of the pair-copula model by introducing an efficient algorithm to determine the lag distances of the pair-copula model. Within this algorithm, the test proposed by Rémillard and Scaillet [2009] is ex-

tended to the spatial framework. At the end of this chapter, the algorithm is applied to two case studies.

The extension of the pair-copula model to the multivariate setting using transformation methods is presented in Chapter 5, which includes two case studies.

The following chapters of this thesis are devoted to the development of the methodology for optimal spatial design for additional samples based on the pair-copula model and its application.

Chapter 6 deals with univariate optimal design based on the pair-copula model. In this chapter, methodology is developed to reduce the prediction uncertainty based on both the configuration of spatial observations and its values. Application of this methodology is presented and validity of the methodology is evaluated by partially redesigning an existing spatial design of a soil based case study.

The extension of this optimal design methodology to the multivariate setting, with the objective of reduction in prediction uncertainty for all the variables simultaneously, is provided in Chapter 7. Application of this methodology is demonstrated for two environmental case studies.

The first section of Chapter 8 provides a comparison between univariate modelling and multivariate modelling. A comparison between univariate designs and the corresponding multivariate design is discussed in the second section of Chapter 8. A brief discussion of each contribution is contained in the third section of the Chapter 8. At the end of this chapter, limitations of the proposed methodologies and recommendations for future work are discussed.

Chapter 2

Literature Review

This chapter contains extracts of the following refereed conference paper.

- Musafer, G.N., Thompson, M.H., Kozan, E., and Wolff, R.C. (2013). Copula-based spatial modelling of geometalurgical variables. In Dominy, S., editor, *Proceedings of The Second AUSIMM International Geometallurgy Conference (Geomet 2013)*, pp. 239–246. The Australasian Institute of Mining and Metallurgy(AusIMM), Brisbane, Australia.

The literature review is mainly classified into two sections. The first section concerns existing geostatistical models where areas of improvement are identified in relation to capturing non-linear dependence between spatial variables and non-linear spatial dependence within individual spatial variables. The copula based geostatistical model is also reviewed, including how the copula based model can address the limitations of traditional geostatistical models. The second section examines general procedures for optimal spatial designs based on conventional models.
The definitions of terminology used in this chapter can be found in Appendix A.

2.1 Geostatistical Models

Many models used in the spatial framework are based on the concept that the spatial data are generated by a random field. The term "field" is used here to

denote the higher dimension of the parameter space. If the parameter space is one dimensional, the random field is simply a random process or stochastic process (see Appendix A for technical definitions).

The most important aspect when dealing with geological data is spatial dependence. Spatial dependence is "the propensity for nearby locations to influence each other and to possess similar attributes" (Goodchild [1992]). This means that realisations of a variable of interest at nearby locations are more highly related than observations that are at far away locations. Hence, classical models cannot be used in the spatial setting because they assume that realisations of the same variable of interest are independent. Therefore, models developed for the spatial setting, called geostatistical models, should be capable of dealing with this spatial dependence (Noppé [1994]).

The main scientific goal of a geostatistical model is estimation of a variable of interest at unsampled locations whilst modelling spatial variability by using the limited sample data. In order to obtain this estimation, it is necessary to evaluate the conditional distribution of unsampled location conditioned on the nearby locations. Since only one observation available at each location, it is infeasible to assess the distribution function unless stationary is assumed on random process. In traditional geostatistics, second order stationary (See the definition in Appendix A) is commonly assumed, by implying that the two-point covariance exists and dependes only the separation vector h of that two points. When comes to the practical accepts, instead of covariance function, variogram is used to model the spatial variability in traditional geostatistics.

The value of the theoretical variogram function for lag h can be written as

$$\gamma(h) = \frac{1}{2} Var(Z(x) - Z(x+h)),$$

where $Z(x)$ is the spatial random variable of interest at location x. However, if the spatial variable of interest is stationary, the theoretical variogram can be defined as

$$\gamma(h) = \frac{1}{2}E(Z(x) - Z(x+h))^2.$$

This theoretical variogram can be estimated using the empirical variogram. The value of the empirical variogram at lag h can be calculated as

$$\hat{\gamma(h)} = \frac{1}{2N}\sum_{N}(z(x) - z(x+h))^2,$$

where $z(x)$ is the observed value of the variable of interest at location x and N is the number of pairs of sample points separated by the separation distance h. The variogram measures the dissimilarity, or increasing variance (decreasing correlation), of the variable of interest at different locations (King 2011). Hence, this measurement can be considered as a measure of linear dependence over the distribution of the variable for a given spatial distance. However, in reality, in most cases the spatial dependence structure may vary over the distribution of the variable of interest (Journel and Alabert 1989). Therefore, this method is inappropriate if non-linear dependence is present. Other than this main pitfall, some other limitations have also been discussed in the literature, such as sensitivity to extreme values and inability to provide more than a single measure of dependence (Li [2010]). Consequently, any model that employs the variogram in the estimation process may not be able to provide accurate estimation for most real world phenomena.

The first geostatistical model was developed by Matheron over six decades ago based on the work of mining engineer Danie Krige (Matheron [1963]). The aim of developing this model, called a kriged model, was to provide the best linear spatial estimate for unsampled locations based on the sample data by minimising the prediction variance. Since then, many models have been developed in the literature. Diggle et al. [1998] introduced a new term "model based geostatistics" to spatial statistics field. In model based geostatistics consists with three main parts; formulation of a statistical model to data, estimation of parameter of the model using maximum-likelihood method and prediction of spatial field using fitted model. Under the Gaussian assumption, classical geostatistical approach

(kriging) and model-based geostatistical approach produce similar prediction methodology (Diggle et al. [2003]). However for the non-Gaussian data resultant prediction methodology more accurate for model based geostatistics approach than classical geostatistical approach.

These models can basically be divided into two types: linear models and non-linear models. The most commonly used models are based on Matheron's kriged model. Linear models, such as ordinary kriging, can be used when spatial dependence of the variable of interest is linear. This means that a linear model is suitable when the relationships of the observations of nearby locations are only influenced by the configurations of the locations. Non-linear models, such as indicator kriging, can be used if the relationship with observations at nearby locations is influenced by both configuration of observations and the value of the observations (non-linear spatial dependence) (Vann and Guibal [2001]). Moreover, non-linear models can be used if the objective is to estimate the distribution of a random variable at unsampled locations.

2.1.1 Linear geostatistical models

Ordinary kriging

Since the ordinary kriging (OK) model was developed, it has become popular in different spatial fields, such as mining and the petroleum industry to hydrology, meteorology, oceanography, environmental control, landscape ecology and agriculture. Simply, the OK estimator of the value of the variable $Z(x_0)$ of interest at unsampled location x_0 can be written as a linear combination of nearby samples as follows

$$\hat{Z}(x_0) = \sum_{i=1}^{n} w_i Z(x_i).$$

The weights w_i are obtained by minimising the error variance σ_R^2 under the constraint $\sum_{i=1}^{n} w_i = 1$ to ensure the unbiased property of the estimator. Moreover, the weights w_i are depended on x_0. This means same location will receive different weight for different estimation location.

The error variance σ_R^2 is

$$\begin{aligned}\sigma_R^2 &= Var[\hat{Z}(x_0) - Z(x_0)] \\ &= \sigma^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C_{ij} - 2\sum_{i=1}^{n} w_i C_{i0},\end{aligned}$$

where $\sigma^2 = Var[Z(x)]$, x is any sampled location, $C_{ij} = Cov[Z(x_i), Z(x_j)]$ and $C_{i0} = Cov[Z(x_i), Z(x_0)]$. Hence w_i can be calculated by solving the following system of equations (Isaaks and Srivastava [1989]): $\sum_{j=1}^{n} w_j C_{ij} + \mu = C_{i0}$ for all $i = 1, \ldots, n$ and $\sum_{i=1}^{n} w_i = 1$, where μ is the Lagrange multiplier. The aim of the Lagrange parameter is to obtain weights that produce minimum variance. Moreover, $Cov[Z(x_i), Z(x_j)]$ and $Cov[Z(x_i), Z(x_0)]$ are estimated using variogram modelling. Since the OK method employs the minimum variance concept, $\hat{Z}(x_0)$ is called the "best linear estimator" of the spatial variable of interest at unsampled locations x_0.

In mining, the estimation of a block is more applicable than point estimation. Blocks can be estimated using the ordinary kriging system by replacing the right hand side term C_{i0} (covariance between the i-th sample location and the unsampled location) by C_{iA} (covariance between the i-th sample location and the block). C_{iA} is equal to the average covariance between the i-th sample location and points within the block (Isaaks and Srivastava [1989]).

The most important assumption of ordinary kiriging is that the data generating process is second order stationary (see Appendix A). Estimation is optimised when the data generating process is Gaussian. Also, this method assumes that spatial dependence is linear (Vann and Guibal [2001]) by using the variogram to model the spatial variability. If an ordinary kriging system is used for a skewed distribution, then, as with any other linear interpolation system, the ordinary kriging estimates will be sensitive to extreme values.

Lognormal kriging

If geological variables of interest are positively skewed, in the early stages of the development of a geostatistical model, a log transformation is applied to the data

that renders them Gaussian. Then the ordinary kriging method is applied to produce estimates at the unsampled locations and back-transformation is carried out to obtain the estimates on the original scale. This method is simply named lognormal kriging. Roth [1998] discusses the bias of back-transformed estimation. In more detail, if lognormal kriging overestimates the standard error at unsampled locations, when it back transforms to the original scale, the standard error becomes larger due to exponentiation and finally leads to a seriously overestimated prediction (Roth [1998]).

Multigaussian kriging

In multigaussian kriging, the normal score transformation is used to transform the data to Gaussain (Saito and Goovaerts [2000]) and the OK system is applied to the transformed data. According to Saito and Goovaerts [2000], if strong stationarity (see Appendix A) of the random field can be guaranteed, this method is able to provide more accurate estimation than lognormal kriging. However, these two models assume linearity of the autocorrelation by employing the ordinary kriging system for the estimation process, and the original autocorrelation structure is overlooked during the back transformation.

Furthermore, the minimum variance concept of OK introduces conditional bias to the estimator (Seo [2013]). Due to this conditional bias, the kriging estimator overestimates lower values and underestimates higher values (McLennan and Deutsch [2004]). Even though some solutions have been suggested to reduce the conditional bias to some extent, the problem of conditional bias can't be removed completely from the estimator (McLennan and Deutsch [2004]).

Linear multivariate geostatistical models

When considering linear multivariate geostatistical models, the universal kriging (UK) model developed by Matheron [1963] can be used if a relationship between the variable of interest and the spatial coordinates is present (Goovaerts [1997]), that is, there is a trend in the random field. Kriging with external drift (KED) is very similar to UK, in that it allows use of a secondary variable in the estimation process, but here the secondary variable is what would be considered a traditional

variable, not the spatial coordinates. In KED, the secondary variable has to be available for all the sampled spatial locations and all the unsampled spatial locations of interest. In contrast, in co-kriging (CK) it is not essential to have all the information from the secondary variable either at the sampled locations or unsampled locations (Isaaks and Srivastava [1989]). The advantage of additional information from a secondary variable is that estimates of the primary variable will be more accurate and complete than compared to ordinary kriging estimation. A disadvantage of KED and CK is that they assume the relationship between the primary and secondary variable is linear. Hence, KED and CK are unable to capture non-linear dependence among the variables.

Generally, the following limitations are present in linear geostatistical modelling, as discussed in Vann and Guibal [2001]:

1. this method can only be used to estimate the expected value unless make an assumtion on distribution;

2. if the variable of interest has a skewed distribution, the estimates produced from linear methods are not appropriate due to the effect of extreme values;

3. the lack of appropriateness for situations in which arithmetic means are not suitable.

Moreover Vann and Guibal [2001] suggest that non-linear estimation is a suitable method to overcome the above mentioned limitations.

2.1.2 Non-linear geostatistical models

From a geostatistical point of view, most of the non-linear models are able to produce the distribution of the variable of interest at unsampled locations conditional on the observations of nearby locations. This can be simply defined as the conditional distribution of the variable of interest at unsampled locations. Moreover, for non-linear modelling, no assumption on the distribution is needed to obtain estimates of variable of interest at unsampled location or over the area of interest.

Most non-linear models transform the original variable to an indicator variable based on a cut-off value before starting the modelling process, as below:

$$I(x, Z_c) = \begin{cases} 0; & Z(x) \leq Z_c, \\ 1; & Z(x) > Z_c \end{cases}$$

where $Z(x)$ denotes the variable of interest at location x and Z_c denotes the chosen universal cut-off. Hence, the resulting distribution of the indicator variable is binary. As a result, extreme values cannot influence this model. This means that indicator kriging is useful for dealing with skewed distributions (Triantafilis et al. [2004]).Moreover, geostatistical modelling for discrete data such as Poisson process model can be found in Diggle et al. [1998].

Indicator kriging

Indicator kriging (IK) is simply ordinary kriging for indicator variables. However, the variogram is constructed using an indicator variable and, so, is called an indicator variogram. For example, in mining, the variable of interest is usually grade of a metal and the cut-off value is the level of metal grade that is used to determine the economic feasibility of the ore to mine. The resulting estimate should lie on the interval $[0, 1]$ and can be interpreted as the probability that the grade is above a specified cut-off or the proportion of the block above the specified grade cut-off (Vann and Guibal [2001], Triantafilis et al. [2004]). However this method tends to produce unacceptable estimates, such as probabilities outside $[0, 1]$.

Multiple indicator kriging

Multiple indicator kriging (MIK) is similar to IK but allows multiple cut-off grades and provides the facility to calculate the expected grade. This method is explained here using an example. For instance, assume that we need to produce estimates for the recoverable ore reserve of three dimensional (3D) blocks for three different cut-off grades $Z_{c,1}, Z_{c,2}, Z_{c,3}$ where $Z_{c,1} < Z_{c,2} < Z_{c,3}$. Hence, three indicator variables will be used to perform the MIK:

$$I_1(x, Z_{c,1}) = \begin{cases} 0; & Z(x) \leq Z_{c,1} \\ 1; & Z(x) > Z_{c,1} \end{cases},$$

$$I_2(x, Z_{c,2}) = \begin{cases} 0; & Z(x) \leq Z_{c,2} \\ 1; & Z(x) > Z_{c,2} \end{cases},$$

$$I_3(x, Z_{c,3}) = \begin{cases} 0; & Z(x) \leq Z_{c,3} \\ 1; & Z(x) > Z_{c,3} \end{cases}.$$

Indicator kriging is performed for each indicator variable. This method can be adopted to find the expected grade for the unsampled location by a weighted average of the empirical means over the intervals bounded by the cut-off values. MIK may lead to the estimation of more recoverable metal at higher cut-off grade compared to lower cut-off due to the inconsistency of indicator models from one cut-off to another as a result of the indicator variables being treated separately. This issue is called the order relation problem (Vann and Guibal [2001]). The other main issue with this method is that it assumes that the shape of the distribution of grade of 3D blocks to be estimated is identical to that of the samples. In reality, variation of the grade in a small volume (drill hole) is higher than that of a larger volume (3D blocks). Consequently, MIK ignores the change that may occur in the shape of distribution when there is a change in the size of the volume upon which estimates are calculated. In technical terms this is called ignoring the change support. Also, MIK is only capable of estimating one variable of interest. A multiple indicator co-kriging model (disjunctive kriging) can be used when considering multiple variables and their cross-relationship. However according to De-Vitry et al. [2007] the multiple indicator co-kriging system is not widely used in spatial applications such as mining due to the high computational requirements of modelling the variograms and cross-variograms.

Uniform conditioning

Uniform conditioning (UC) is a non-linear method that is a practical approach to estimate multiple variables of interest with block support even given that the

size of the block support is much smaller than the space between two sampled locations. In this method, the first step is to transform the sample data using a Normal score transformation if the sample data don't follow the Gaussian distribution. Secondly, estimation of the ore reserve of a panel that contains a number of small blocks is conducted. The estimate of the panel is produced by ordinary kriging using sampled locations. The ordinary kriging estimates are more reliable for larger volumes than for smaller volumes. Finally, conditioned on the estimated panel value, the probability that the grade is above a specified cut-off for a block within the panel can be estimated as in Wackernagel [2003]. However, this method also has the order relation issue. But UC is the only method that considers the change support for multivariate non-linear methods (De-Vitry et al. [2007]).

Although non-linear models with indicator variables are able to address the issue of estimating the distribution of the variable of interest, they are not able to quantify the in-situ non-linear dependence between variables and non-linear spatial dependence within the individual variables due to binary transformation. First, much more information is lost due to the binary transformation. Hence, statistical power of identifying the real relationship between the variables is reduced (Royston et al. [2006]). Secondly, the binary transformed model is then only able to determine whether there is relationship or not. Any information about the strength of the relationship or type of the relationship (linear or non-linear) cannot be obtained. This means that binary transformation distorts the relationship between the original variables. As a result, the effect of the true relationship between the variables cannot be included in the estimation procedure based on the non-linear models developed using indicator variables.

2.1.3 Conditional simulation

The goal of the kriged model is to estimate the value of the variable of interest at unsampled locations. As described earlier, kriged models estimate the unsampled locations by minimising the prediction error variance under a constraint

imposed to secure the unbiased property of the estimator. As a result, the variance of the kriging estimator $Var(\hat{Z}(x_0))$ is always lower than the actual variance $Var(Z(x_0))$. Therefore, the actual variability of the estimate cannot be quantified by using kriged models (Goovaerts [1997]).

Conditional simulation is a tool whose objective is to demonstrate the variability and the uncertainty of the estimation by generating several realisations of estimates based on the the limited observed data (Larocque et al. [2006], Sidler [2003]). Most commonly, this set of realisations of block estimates is called the "equally likely" images of 3D blocks. However, this equally likely concept is not correct if the underlying marginal distribution of the variable of interest is skewed. In the mining industry, this technique has been used to represent the uncertainty of an ore body and review the risk involved in various decisions (Khosrowshahi and Shaw [2001]). However, one realisation from the simulation is not adequate to use as an estimate of a block in terms of the minimum variance. The average of all the realisations may tend to provide a good estimator if the number of simulations is large enough. Since several realisations are obtained for a block, this can be considered as the distribution of the variable of interest for a block. However, accuracy of the local distribution is highly dependent on the number of simulations. Finding the optimal number of simulations is still to be investigated and further research may be required. Since this conditional simulation technique is based on conventional geostatistical models such as ordinary kriging and indicator kriging, all negative aspects related to these models are inherent in conditional simulation.

2.1.4 Transformation of multiple correlated variables into uncorrelated variables

As discussed in the subsections above, modelling of multiple variables (multi-variables) with spatial cross-relationships is complex and time consuming when compared to single variable modelling. Multi-variables can be transformed to spatially uncorrelated variables (factors) by using a suitable transformation method

(Rondon [2012]). Hence, univariate geostatistical modelling can be performed on each factor separately. Minimum/maximum autocorrelation factors transformation (MAF) (Rondon [2012]) and stepwise conditional transformation (SCT) (Leuangthong and Deutsch [2003]) are the most commonly used transformation methods. Recently, Barnett et al. [2014] introduced a new transformation method called Projection Pursuit Multivariate Transform (PPMT) to the spatial framework to remove the non-linearity between spatial variables.

Bandarian et al. [2010] describe how the minimum/maximum autocorrelations factor transformation can be used to remove the correlation between the variables for all lag distances using principal components analysis. Principal components analysis can be used to obtain uncorrelated factors from correlated variables (Wackernagel [2003]). However, Rondon and Tran [2008] proved that MAF cannot be used to produce uncorrelated factors if variables are non-linearly related. The SCT and PPMT methods transform the original variables to multivariate Gaussian variables with no cross-relationship at zero lag distance (Leuangthong and Deutsch [2003]). Hence cross-correlation of the transformed variables at lag $h > 0$ may be present. Therefore, the interpolation or simulation process should be carried out after verifying that there is the zero correlation between the variables for any lag distance.

The most important advantage of this methods is the ability to transform the non-linear multivariate distribution to the multivariate Gaussian distribution (Leuangthong and Deutsch [2003]). Even though the SCT and PPMT methods can be used to remove any kind of relationship between the variables at lag zero, there is no guarantee of removing spatial dependence at lags that are greater than zero. More details about these transformation methods and their strengths and weaknesses are reviewed in Chapter 5.

2.1.5 Copula based geostatistical models

Most of the pitfalls in the above mentioned traditional geostatistical models motivated Bárdossy and Li [2008] to develop a new non-linear geostatistical model

based on copulae. This section discusses the literature review regarding copula based models.

The development of the copula based spatial model was motivated by the restrictive assumption of linear spatial dependence when using the variogram and covariance function. Additionally, sensitivity of the variogram and covariance function to extreme measurements and their inability to change the dependence structure over the distribution of the variable of interest also influenced the development of copula based spatial models (Li [2010]).

Although copula based modelling is a new avenue for spatial statistics, it has been widely used in non-spatial applications in fields where it is essential to deal with non-linear dependence, such as in finance and actuarial sciences (Bárdossy [2006]). Since this method is comparatively new to geostatistics, relatively few papers have been published relating to this area.

Sklar (1959) introduced copula theory. A copula describes the dependence structure between random variables. A copula does not need any information about the marginal distribution of the random variables to describe the dependence structure. An introduction to copula theory can be found in Nelsen [2006] and Trivedi and Zimmer [2007]. An applied review of copulas can be found in Boardman and Vann [2011].

A copula can be defined as a multivariate distribution function of uniformly distributed random variables on the interval $[0, 1]$. Therefore it has the same properties as any distribution function. For multivariate distribution function $C(u_1, \ldots, u_n)$ to be a copula, it must satisfy the following conditions:

1. $U_1, \ldots, U_n \sim Uniform(0, 1)$;
2. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for every $i \leq n$ in $[0, 1]$;
3. $C(u_1, \ldots, u_n) = 0$ if $u_i = 0$ for any $i \leq n$;
4. C is an n-increasing function.

Sklar's Theorem

Sklar's theorem describes the relationship between the copula and the joint distribution function $F(z_1, \ldots, z_n)$ of an n-dimensional vector of random variables $(Z_1, \ldots, Z_n)$ as follows:

$$F(z_1, \ldots, z_n) = C(F_1(z_1), \ldots, F_n(z_n))$$

where $F_i(z_i)$ represents the i-th one-dimensional marginal distribution function of z_i.

This theorem states that, for a given joint distribution, a copula can be found to model the multivariate structure of a vector of random variables by using their marginal distributions. Moreover, the copula will be unique if all the marginal distributions are continuous.

Based on Sklar's theorem, the joint density function $f(z_1, \ldots, z_n)$ of the random variables $(Z_1, \ldots, Z_n)$ can be derived by applying partial derivation to the joint density function $F(z_1, \ldots, z_n)$. Hence, the joint density function can be fragmented into its univariate margins and dependence structure as follows:

$$\begin{aligned} f(z_1, \ldots, z_n) &= \frac{\partial F(z_1, \ldots, z_n)}{\partial z_1, \ldots, \partial z_n} \\ &= \frac{\partial C(F_1(z_1), \ldots, F_n(z_n))}{\partial z_1, \ldots, \partial z_n} \\ &= \frac{\partial C(u_1, \ldots, u_n)}{\partial u_1, \ldots, \partial u_n} \times \prod_{i=1}^{n} \frac{\partial F_1(z_i)}{\partial z_i} \\ &= c(u_1, \ldots, u_n) \times \prod_{i=1}^{n} f_i(z_i) \end{aligned}$$

where, $u_i = F_i(z_i)$ for $i = 1, \ldots, n$ and $c(u_1, \ldots, u_n)$ denotes the density function of the copula. In other words, the copula density can be expressed as the dependence structure of the random vector $Z_1, \ldots, Z_n$.

Bárdossy [2006] used the bivariate copula to describe spatial dependence for water analysis systems for given distances. Bárdossy [2006] describes the advantages of using a copula to quantify the spatial dependence when compared to using the covariance or variogram. The ability to quantify any kind of dependence is one

of the major advantages of using copulae. Moreover, monotonic transformation of the marginal distribution, such as the Box-Cox transformation and Normal score transformation, cannot influence the copula of a multivariate distribution. This means that the copula is not affected by the marginal distribution. This is the second major advantage when compared to traditional methods like the variogram, because variograms are highly dependent on the marginal distribution. Furthermore, copulae can be used to describe the dependence structure over any percentile of the variable of interest for a given spatial distance. Bárdossy and Li [2008] adopted this model in their estimation process at unsampled locations. From this point, we only consider copulae to describe the joint multivariate distribution of the variable of interest between the unsampled location and nearby spatial locations.

However, as with conventional geostatistics some assumptions are required to apply the copula based model. As with conventional geostatistics, copula based modelling assumes that the set of measured values of the variable of interest are realisations of a random function (Bárdossy and Li [2008]). However, when applying copula based models, a strong stationary random field (see Appendix A) is assumed over the domain of interest. This assumption is stronger than the conventional linear geostatistical assumption of a second order stationary random field over the domain of interest because the copula based model requires all the moments of the data generating process be unaffected by a change of spatial distance. However, copula based modelling has more advantages when compared to geostatistical modelling, even though it requires a more limiting assumption, such as the capability to obtain the full conditional distribution, ability to remove the influences of marginal distributions when modelling the dependence structure and ability to model the non-linear spatial dependence (Haslauer et al. [2010]).

Based on this strong stationary assumption, the marginal distributions of the variable of interest for each location in the domain are identical, i.e.,$F_i(z_i) = F(z_i)$. The empirical bivariate copula can be used to explore the spatial variability. As with the variogram, it is assumed that the bivariate spatial copula C_s at any two

locations only depends on the separation vector $\boldsymbol{h}$ (and is independent of the locations x) (Bárdossy [2006], Bárdossy and Li [2008]), that is

$$\begin{aligned} C_s(u,v) &= Pr(F(Z(x)) < u, F(Z(x+\boldsymbol{h})) < v)) \\ &= C_{\boldsymbol{h}}(F(Z(x)), F(Z(x+\boldsymbol{h}))) \end{aligned}$$

Moreover, not just any copula model can be used as a spatial copula (Bárdossy and Li [2008]). There are requirements that should be fulfilled by a copula to be a spatial copula (Bárdossy and Li [2008], Kazianka and Pilz [2010a]). Generally, as with conventional geostatistics, it is assumed that the spatial dependence between location x_1 and location x_2 is the same as the dependence between location x_2 and location x_1. Hence, this symmetrical property should be a feature of the spatial copula. Another requirement is that the dependence structure of the copula must be able to be parameterised in order to be described as a function of $\boldsymbol{h}$. Furthermore, the well-known spatial property of no dependence between far distant observations and high dependence between near observations can be represented in copula based models as follows:

1. $C(u_1, \ldots, u_n) = \prod_{i=1}^{n} u_i$ when $\|\boldsymbol{h}\| \to \infty$;

2. $C(u_1, \ldots, u_n) = min(u_1, \ldots, u_n)$ when $\|\boldsymbol{h}\| \to 0$.

The most readily available copulae in the literature are unable to be extended to higher dimensions and some copulas that do have that ability do not provide good parameterisation for the dependence structure to reflect the spatial configuration of the data points (Bárdossy and Li [2008]). Even though the most popular copulas, such as Gaussian and Student t copulas fulfil both requirements, these copulae cannot be used to model asymmetric dependence structure. As a result, Bárdossy [2006] introduced the non-central chi copula to model asymmetric dependence structures. However, this model is computationally very expensive when fitting non-central chi-squared copulas to large scale data sets. For example, if n is the number of observations, calculation of 2^n terms are needed in the

process of spatial interpolation to estimate the value of the variable of interest at an unsampled location.

If the copula employed is Gaussian or Student t then there are no difficulties in applying the maximum likelihood method directly in estimating the copula parameters. However, calculation of the copula density for higher dimensions may be difficult if the copula is a non-central chi-squared copula. As with the Gaussian and Student t copulae, the correlation matrix is required in obtaining the copula parameter estimates for the central chi-squared copula. This correlation matrix may be difficult to estimate for higher dimensions. As a solution to this, Kazianka and Pilz [2010b] propose finding the correlation matrix for higher dimensional copulae using the correlations from the bivariate copulae, assuming independence of different pairs of observations. That is, the entries of the correlation matrix for the higher dimensional copula are simply given by the correlation between pairs of observations with the same distance. However these estimated parameters are not efficient compared to estimates obtained by applying the maximum likelihood method directly to the higher dimensional copula.

More generally, for higher dimensional copulae, not necessarily restricted to central chi-squared copulae, goodness of fit of the higher dimensional copula to the data can be measured by comparing the observed data to data simulated from the fitted multivariate copula, where several simulated data sets are obtained. A test of the difference between the copula observed from the random field and that from simulated random fields can be carried out using the method proposed by Malevergne and Sornette [2003], and as demonstrated by Bárdossy and Li [2008]. However, using bivariate copulae to construct the higher dimensional multivariate copula, as described above, may not necessarily give the best fit to the joint multivariate distribution.

Finally, by fitting the joint multivariate copula to the unsampled location and the nearby locations, it is possible to derive the copula density of the unsampled location conditioned on the nearby locations as follows:

$$c(u_0 \mid u_1, \ldots, u_k) = \frac{c(u_0, u_1, \ldots, u_k)}{\int_0^1 c(v, u_1, \ldots, u_k)dv}$$

where $u_i = F(z_i)$, k is number of nearby locations and u_0 denotes the marginal distribution at the unsampled location.

Consequently, any estimator can be obtained from the estimated conditional density. As an example, the expected value or the median of the conditional distribution can be obtained using the following equations:

$$\begin{aligned} \text{Expected value } &= \int_0^1 F^{-1}(u)c(u|u_1, ., u_n)du, \\ \text{Median value } &= F^{-1}(u = C_n^{-1}(0.5|u_1, ., u_n)), \end{aligned}$$

where F^{-1} is the inverse marginal distribution function and C^{-1} is the inverse copula.

This copula based model has been used in a few different spatial applications, for example, to model hydrology properties (Bárdossy and Li [2008]), soil properties (Marchant et al. [2011]) and air pollutants (Kazianka and Pilz [2010b]). The authors demonstrated that more realistic estimation can be obtained using copula based geostatistical modelling when compared with ordinary and indicator kriging. Additionally, Kazianka and Pilz [2010a] developed copula based modelling for random fields with trend and for random fields of discrete random variables. In addition, these two authors attempt to fit the copula based model using a Bayesian framework as well (Kazianka and Pilz [2011]).

However, only a small number of copula families, such as Gaussian, Student t and the non-central chi-squared, have been used for modelling. From these families, only one copula family is used to capture the complex dependence. Moreover, the same copula family is assumed at each separating vector $\boldsymbol{h}$ and multivariate dependence, which is required in the interpolation process, is also modelled using the same family of higher dimensional copula. A new geostatistical model based on the pair-copula construction was introduced by Gräler and Pebesma [2011] and this pair-copula construction allows the use of different types of families when modelling spatial dependency for different separating vectors and for higher order

dependencies as well. As a result, multivariate dependence can be modelled by this sophisticated copula model, which has full flexibility to capture the complex dependence. A detailed explanation of pair-copula based geostatistical model is given in Chapter 3.

2.2 Optimal Design

2.2.1 Non-spatial optimal design

In classical statistics, the aim of optimum design is to obtain estimates of statistical model parameters in an unbiased way with minimum variance using a smaller number of experimental runs than non-optimal design. As a result, experimental cost can be reduced. Optimal experimental design is model dependent. This means that the optimum design developed based on one statistical model may not be optimal for another statistical model. Optimality of experimental design is usually evaluated based on the Fisher information (inverse of the variance-covariance matrix of the estimators) (Fedorov and Hackl [2012]).
Muller [2007] discusses the reasons why optimal design based on the classical framework, even for continuous variables, cannot be adopted by the spatial framework. This is due to the following two reasons:

- classical optimal designs are unable to capture the spatial correlation between the observations;
- it is difficult to obtain replications from the spatial experimental setting.

Therefore, the information matrix should be replaced by using different techniques, such as mean prediction error, which should be able to capture the spatial correlation.

2.2.2 Optimal spatial sampling design

Optimal spatial sampling design can be simply defined as optimal allocation of sampling points to spatial coordinates (Pilz and Spöck [2008]). The optimum

sampling design will vary according to the scientific goal, such as parameter estimation of the model and predictions using the geostatistical model (Diggle and Lophaven [2006], Diggle and Ribeiro [2007b]). If prediction of the random field is the aim, then optimality of the sampling design is evaluated based on the maximum or average mean square prediction error of the predicted locations (Diggle and Ribeiro [2007b]). Van Groenigen and Stein [1998] used Monte Carlo methods, such as simulated annealing, to optimise different objective functions, such as maximising the spatial spread of the sample locations rather than the existing objective function of minimisation of average prediction error. Most of the optimal spatial designs in the literature are based on two dimensional (2D) space.

The main aim of collecting the spatial samples over the study domain in the initial phase is to obtain good geostatistical coverage and projection of the variable of interest. Usually, thereafter, a systematic pattern is commonly used to collect the spatial sample for areas without access problems. However, the decision of the sampled locations for the next phase can be derived using the statistical information obtained from the first phase (Moon and Whateley [2006]). This means that information obtained from the initial phase can be used to develop an appropriate geostatistical model and additional samples can be used to improve the quality of the predictions, which reduces the uncertainty of prediction. This improvement is not limited to the prediction process. As an example, in mining all the process are related to each other. Consequently, reduction of error in the prediction will benefit all the interdisciplinary processes formed in mining, such as mine design, mine scheduling and financial evaluation (De Souza et al. [2004], Soltani and Hezarkhani [2013]). This thesis focuses on optimising designs for additional samples after the initial phase.

Moreover, generally, the common purpose of taking additional samples is to increase the precision of the random field predictions. Since kriged models are the most commonly used models for prediction in spatial applications, most of the developed optimal designs consider functions related to kriging variance as a sta-

tistical criterion for most environmental based applications. In the mining field also, Walton and Kauffman [1982] attempted to develop a design for additional drilling using the kriged model. The aim of their method was to improve the accuracy of the estimate of grade and tonnage of the ore reserve. According to their proposed method, kriging variances of all the blocks are calculated. Then the block with the highest kriging variance is selected as the next drill location. This procedure is repeated until acceptable global estimation of the variance is obtained. Later, Scheck and Chou [1983] introduced an iterative procedure based on fixed point theory, which is a mathematical optimisation method used to select the number and the location of drill holes. However, their method, which uses the maximum kriging variances, is unable to produce the optimum locations for additional drill holes.

Average kriging variance over interpolated grids is the most commonly used statistical criterion to obtain the optimal design based on the following two assumptions (Saikia and Sarkar [2006]:

1. the variogram model used to compute the kriging variance is the correct one;

2. the model of the variogram and estimated population mean of the variables of interest are not affected by the additional samples.

The selection of the model is likely dependent on the experience of the geostatistician. Therefore, the fitted model for the variogram may be inadequate if the geostatistician is not sufficiently experienced. Hence, reliability of the optimal design produced from this method is doubtful. On the other hand, criteria related to kriging variance are popular due to insensitivity to the variability of sampled values under the Gaussian assumption. Kriging variance is only sensitive to the spatial configuration and the fitted variogram model and ignores effects of variability of the sample values (Journel and Alabert [1989], Goovaerts [1997]). Hence, the actual value of additional samples is not required to evaluate their impact on the estimation process if kriging variance is used as the uncertainty

measurement (Deutsch [1993]). As long as a linear dependence structure assumption is valid, values of the samples are not required to estimate the uncertainty in a spatial framework. But, in reality, this assumption is rarely fulfilled. Thus, it is essential to consider the sample values for estimation of uncertainty when a complex dependence structure is present.

Moreover, in the literature, many authors discuss the pitfalls of using a kriging variance as an uncertainty measure. Because kriging variance ignores effects of variability of the values, Pilger et al. [2001] introduced optimal design for additional samples based on the uncertainty measurement produced from stochastically conditional simulation in mining. They stated that, through this procedure, uncertainty of resource estimation can be calculated by considering the configuration of sample points and their values. By applying stochastic simulation conditioned on sample data, possible realisations of estimates of the ore reserve for each grid are able to be obtained. One realisation can be considered as one possible image of the ore body. Thus, these realisations can be used to define uncertainty indices of the estimates of each grid, such as conditional variance, interquartile range (IQR) and coefficient of variation. Pilger et al. [2001] used the IQR in their research. The grid point with maximum IQR is selected as the new sample location. Then, one value of a realisation (randomly selected) of the particular grid point is assigned to the new sampled location. Then, again, conditional simulations are carried out using the new sample value. The average IQR of the nearby grid points to the grid point with the new sample location is calculated as the local evaluation criteria and that of all the grid points used as the global criteria. This process is repeated until the global reduction of IQR is stabilised. The same methodology was adopted by Koppe et al. [2011] to compare two infill spatial patterns. They assumed that the pattern locations for the initial sampling is regular. In the first pattern, additional samples are scattered over the region of interest, whilst in the second pattern they are located at the grid points with higher uncertainty related to the variable of interest. However, they selected the spatial pattern that produced the lowest uncertainty of net present

value as the best spatial locations for the additional samples. Moreover, they assumed a constant number of additional samples. This means that they only considered the optimisation of limited sample patterns. Also, this methodology, based on conditional simulation, evaluates optimality by using one possible value for the candidate location. A similar concept of finding the optimal pattern for additional locations in a different application, such as soil sampling and plants sampling, can be found in work presented by Van Groenigen et al. [1999] and Emery et al. [2008]. However, these papers evaluate the optimality of a spatial pattern based on a statistical criteria related to kriging variance.

There may be specific purposes for different spatial projects, for example, Koppe et al. [2011] developed an optimal design in the mining field to reduce the uncertainty of net present value for a new mine. Hassanipak and Sharafodin [2004] introduced another strategy to find the optimal design for additional samples with the aim of improving the reliability of resource classification and improving the estimates of grade and tonnage of the ore reserve. They introduced a function called GET as the criteria to select the locations for additional drill holes. GET is a function of three variables: the average estimation error of block grade (E), the average estimated block grade (G) and compounded thickness of ore blocks (the total thickness of the block that has been identified as ore) (T). This method is the first method that considers the 3D extension of the ore body. However, this method is only able to produce some suggested points for the additional drillings. These suggested points may not be the optimal design for additional drillings.

2.2.3 Optimal design for copula based geostatistical model

This section reviews the literature on optimal design based on spatial copulae. It appears that only two papers have been been published up to now covering this area. Li et al. [2011] were the first to develop sampling design based on a copula based geostatistical model. The aim of their research is to add observation locations to an existing water observation network. Their methodology allows one to capture the variability of sample values when making decision regarding

additional locations (Li [2010]). The locations that give the minimum expected penalty of making an incorrect decision among the other points are the new locations for observations. In their research, the penalty for an incorrect decision is decided by the researcher and the incorrect decisions are: using the water when water is not clean and not using the water when water is clean. Marchant et al. [2013] adopted the same procedure to add new locations for a soil based application with the objective of minimising the expected loss in misclassifying the soil contamination status. This method can be adopted to develop a strategy to decide the optimal design for additional samples with the objective of maximising the expected return based on the given cut-off value. This thesis aims to develop an optimal design for the additional samples with the objective of reducing the uncertainty estimation in prediction. This means our statistical criteria to optimise the additional samples should be related to the precision of prediction.

2.2.4 Limitations of previous work

It has been identified that conventional geostatistical modelling cannot capture non-linear spatial dependence by employing the variogram, which just produces two point-statistics. Moreover, the variogram is sensitive to the extreme values. Furthermore, conventional linear kriging only produces optimal results when the random field is Gaussian, which is not statisfied in most real world applications. Hence, conventional linear geostatistical models are unable to produce accurate prediction (interpolation and simulation) for real world case studies by modelling the spatial dependence incorrectly with use of the variogram. Even though non-linear kriged models, such as indicator kriging, are a solution for the non-Gaussian random field, due to the binary transformation, this method has loss in statistical power to detect the true relationship between the variables.

In conventional geostatistics, the uncertainty estimation used to quantify the precision of the prediction is kriging variance. As discussed in the literature, this only depends on the configuration of observations and the fitted variogram model. But in reality, the uncertainty estimation used for prediction is expected to behave

differently for the different quantiles observed for the additional samples.

Copula based geostatistical modelling is a good solution to overcome the problems of conventional kriged models. Copula based models relax the Gaussian assumption used in conventional geostatistical models and it has the ability to produce the full conditional distribution at unsampled locations. The most important feature of the copula based model is the ability to model the non-linear dependence structures. In other words, copula based models can produce uncertainty estimation for prediction based on both configuration of the observations and their measured values. The pair-copula model introduced by Gräler and Pebesma [2011] is a more flexible model than the simple copula based model.

Since the pair-copula model was only recently introduced to geostatistics, no improvements to the model fitting process have been considered to improve the pair-copula model, such as defining an efficient way to define the distance classes. Moreover, the pair-copula model has still not been used in multivariate geostatistics. Spatial optimal design approaches based on the pair-copula model have also not been considered in the litreature. This research intends to fill these research gaps.

Chapter 3

Application of the Pair-copula Model to Spatial Data

This chapter is based on the paper detailed below, which was presented at the 11th Engineering Mathematics and Applications Conference (2013), Brisbane, Australia, and is currently under review with the Journal of Applied Statistics. The core contributions of the paper are: a detailed description of the spatial pair-copula methodology and its first-time application in the mining field.

- Musafer, G.N., Thompson, M.H., Wolff, R.C., and Kozan, E. (n.d). Pair-copula modelling of grade in ore bodies. *Journal of Applied Statistics. Under review.*

Abstract

Conventional kriged models are the most commonly used for estimating grade, or other spatial variables. These models use the variogram or covariance function to model the spatial correlation required in the process of estimation. The variogram and covariance function produce one single average value to represent the spatial dependence of grade for a given distance. The underlying assumption behind this oversimplified measurement of dependence structure is linear spatial correlation of grade. In reality, the dependence structure of metal grade may be non-linear and complex. Hence, inaccurate estimation of the ore reserve may result if a

kriged model is used for estimating grade at unsampled locations when non-linear spatial correlation is present. Pair-copula based methods may offer a solution to modelling non-linear spatial dependence in a more flexible way when compared with simple copula based models. This solution will additionally benefit the ore reserve estimation and simulation processes where non-linear dependence may be present. In addition, since pair-copula based models are capable of producing the full distribution of an ore characteristic, such as grade, at unsampled locations, estimation of uncertainty is possible and this uncertainty estimation will be more complete than the uncertainty estimation obtained from a kriged model. The pair-copula model is applied to a real world mining application in this chapter for the first time. The performance of the pair-copula model is compared with a conventional linear geostatistical model.

3.1 Introduction

This chapter provides practitioners with instructions outlining the steps involved in fitting a pair-copula model to spatial data. In addition, for the first time, a geostatistical model based on pair-copulas is applied to real world mining data with the purpose of illustrating the advantages of pair-copula based spatial models over traditional kriged models in mining.

One of the most important aspects of modelling a geological variable, such as metal grade, is spatial correlation. Spatial correlation describes the relationship between realisations of a geological variable sampled at different locations (Getis [2007]). Any method modelling a geological variable should be capable of accurately estimating the true spatial correlation. The variogram (see definition in Diggle and Ribeiro [2007a]) and covariance function are the most common methods used to capture the spatial dependence structure of a geological variable (Gräler and Pebesma [2011], Kazianka and Pilz [2010a]). These methods are only capable of providing one simple average measurement of dependence and also assume linear dependence over the distribution of the variable of interest. However, in reality, in most cases the spatial dependence structure may vary over

the distribution of the variable of interest (Journel and Alabert [1989]). In other words, the spatial dependence structure of the variable of interest may be complex. Therefore, conventional geostatistical models, such as kriging, which uses the variogram to model spatial dependency, are unable to produce accurate estimators of distributional properties of the variable at unsampled locations when a complex dependence structure is present. Bárdossy and Li [2008] introduced a new geostatistical model based on copulas that uses bivariate copulas to model spatial dependence. The development of this copula based spatial model was motivated by the restrictive assumption of linear spatial dependence when using the variogram and covariance function. Additionally, sensitivity of the variogram and covariance function to extreme measurements and their inability to change the dependence structure over the distribution of the variable of interest also influenced the development of copula based spatial models (Li [2010]).

Moreover, unlike the kriged model, the copula based model has the ability to estimate the full conditional distribution of the variable of interest at unsampled locations. This means that it is possible to obtain all the possible realisations of estimates while preserving the observed data. This is very similar to the process of conditional simulation (Larocque et al. [2006]). However, the conditional simulation technique is based on conventional geostatistical models. Consequently, all negative aspects related to these models are inherent in conditional simulation. Therefore, realisations obtained from copula based models demonstrate the variability and uncertainty of the estimation more accurately than those from conditional simulation. Hence, by using copula based models, it is possible to represent the uncertainty of an ore body more accurately and thus obtain a more robust measure of the risk involved across the mining process than compared with conditional simulation.

Gramacy and Lee [2008] introduced treed Gaussian process models to spatial data framework. These models allow modelling the non-stationary, and heteroscedasticity relationship of dependent variable and independent variables. This is done by splitting the study domain in to regions in order to fit the Gaussian process

model to dependent variable and independent variables each split region. In this kind of modelling approach address the non-linearity between spatial variables. Hence, it is not appropriate to compare this model with copula based geostatistical model which is address the non-linearity in the spatial dependency.

However, the most readily available copulae in the literature are unable to be extended to higher dimensions. Additionally, some copulae that do have the ability to be extended to higher dimensions do not provide good parameterisation for the dependence structure to appropriately reflect the spatial configuration of the data points (Bárdossy and Li [2008]). Even though the most popular copulae, such as Gaussian and Student t copulae, fulfil both requirements, these copulae cannot be used to model asymmetric dependence structures. As a result, Bárdossy [2006] introduced the non-central chi copula to model asymmetric dependence structures. However, this model is computationally very expensive when fitting non-central chi-squared copulae to large scale data sets. For example, if n is the number of observations, 2^n calculations are needed in the process of spatial interpolation to estimate the value of the variable of interest at unsampled locations. Moreover, the same copula family is assumed for each separation vector $\boldsymbol{h}$. Also, multivariate dependence, which is required in the interpolation process, is also modelled using the same family of higher dimensional copula. Therefore, this method lacks the flexibility to capture more complex spatial dependence structures.

A new geostatistical model based on a pair-copula construction was introduced by Gräler and Pebesma [2011]. This pair-copula construction allows the use of different types of copula families when modelling the spatial dependency for different separating vectors and for higher order dependencies as well. As a result, multivariate dependence can be modelled more accurately by this sophisticated copula model, which has full flexibility to capture complex spatial dependence. Moreover, Gräler [2014] applied a pair-copula model to a skewed spatial random field.

Although copula based modelling is a new avenue for spatial statistics, it has been widely used in non-spatial applications in fields where it is essential to deal

with non-linear dependence, such as in finance and actuarial sciences (Bárdossy [2006]). Since this method is comparatively new for geostatistics, relatively few papers have been published relating to this area (Kazianka and Pilz [2010a]). As far as the author is aware, in the literature, simple copula models have been used in only a few spatial applications, for example, to model hydrology properties (Bárdossy and Li [2008]), soil properties (Marchant et al. [2011]), air pollutants (Kazianka and Pilz [2011]) and mining (Musafer et al. [2013]) and the pair-copula model has been used in only a few spatial (Gräler and Pebesma [2011], Gräler [2014]) and spatial-temporal applications (Erhardt et al. [2015a,b]). However, the pair-copula model has not yet been used in mining applications.

The main objectives of this chapter is to fit a pair-copula model to estimate the metal grade of an ore reserve obtained from a real mine site, and to estimate the distribution of metal grade at unsampled locations, conditional on the local neighbourhood of sampled locations. Moreover, the pair-copula model is compared with an ordinary kriging model to evaluate the performance of the pair-copula model.

3.2 Theory

This section contains detailed explanation of the basic classical statistical theories, utilised by Gräler and Pebesma [2011], that underpin the construction of geostatistical models based on pair-copulas.

3.2.1 Copula

Copula theory, which was introduced by the Sklar (1959), is the base theory for any copula based spatial modelling. A copula describes the dependence structure between random variables. A copula does not need any information about the marginal distribution of the random variables to describe the dependence structure. A copula can be defined as a multivariate distribution function of uniformly distributed random variables. Conversely, the copula can be constructed using the multivariate distribution function. An introduction to copula theory can be

found in Nelsen [2006] and Trivedi and Zimmer [2007]. For an applied review of copulas, the reader is referred to Boardman and Vann [2011].

3.2.2 Pair-copula

Although the process of modelling bivariate distributions using copulae is straightforward, modelling high dimensional distributions using copulae is a complicated task. Moreover, there are many bivariate copulas in the literature, most of which lack the flexibility for extension to higher dimensions except for a few well known copulas such as the Gaussian and Student t copulas.

The pair-copula model can be classified as a hierarchical model building concept. Aas et al. [2009] initially introduced this method to estimate the joint multivariate distribution of random variables using a set of bivariate copulas based on the work of Joe [1996], Bedford and Cooke [2002], and Kurowicka and Cooke [2006]. Aas et al. [2009] present a worked example for the construction of a multivariate distribution for four random variables. To provide a simple demonstration of Aas et al. [2009]'s method, a small example for three variables is given below.

Let the joint density function of X_1, X_2, X_3 be $f_{123}(x_1, x_2, x_3)$. This can be factorised as

$$f_{123}(x_1, x_2, x_3) = f_3(x_3) f_{2|3}(x_2|x_3) f_{1|23}(x_1|x_2, x_3). \tag{3.1}$$

From Sklar's theorem, any multivariate distribution function F with marginals $F_1(x_1), \ldots, F_n(x_n)$ can be written as

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)) \tag{3.2}$$

where C is an n dimensional copula. Hence the joint density function can be written as

$$f(x_1, \ldots, x_n) = c_{1,2,\ldots,n}(F_1(x_1), \ldots, F_n(x_n)) \times f_1(x_1) \times \ldots \times f_n(x_n) \tag{3.3}$$

where $c_{1,2,\ldots,n}$ is the copula density.

Using Eq.(3.3), the second term of Eq.(3.1) can be written as

$$\begin{aligned} f_{2|3}(x_2|x_3) &= \frac{f(x_2, x_3)}{f(x_3)} \\ &= \frac{c_{23}(F_2(x_2), F_3(x_3)) \times f_2(x_2) \times f_3(x_3)}{f_3(x_3)} \\ &= c_{23}(F_2(x_2), F_3(x_3)) \times f_2(x_2). \end{aligned} \tag{3.4}$$

Again, using Eq.(3.3), the third term of Eq.(3.1) can be written as

$$f_{1|23}(x_1|x_2, x_3) = c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \times c_{12}(F_1(x_1), F_2(x_2)) \times f_1(x_1). \tag{3.5}$$

Substituting Eqs.(3.4) and (3.5) into Eq.(3.1) gives

$$\begin{aligned} f_{123}(x_1, x_2, x_3) = f_1(x_1) \times f_2(x_2) \times f_3(x_3) \times c_{12}(F_1(x_1), F_2(x_2)) \\ \times c_{23}(F_2(x_2), F_3(x_3)) \times c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)). \end{aligned}$$

This equation states that the density of the three dimensional copula can be decomposed into a set of three bivariate copulas. The copulas $c_{12}(F_1(x_1), F_2(x_2))$ and $c_{23}(F_2(x_2), F_3(x_3))$ are unconditional bivariate copulas (unconditional pair-copulas) and $c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))$ is a conditional bivariate copula (conditional pair-copula). Here, three pair-copulas have been used for the decomposition. In general, to decompose an n-dimensional density function, $n(n-1)/2$ pair-copulas are required. Marginal conditional distributions are required when constructing the conditional pair-copula. Joe [1996] showed that

$$F(x \mid \boldsymbol{v}) = \frac{\partial C_{x,v_j|\boldsymbol{v_{-j}}}(F(x \mid \boldsymbol{v_{-j}}), F(v_j \mid \boldsymbol{v_{-j}}))}{\partial F(v_j \mid \boldsymbol{v_{-j}})} \tag{3.6}$$

where $\boldsymbol{v}$ is a d dimensional vector, v_j is one arbitrarily selected variable and $\boldsymbol{v_{-j}}$ denotes the vector $\boldsymbol{v}$ excluding v_j. If $\boldsymbol{v}$ is univariate such that $\boldsymbol{v} = v$, then

$$F(x \mid v) = \frac{\partial C_{x,v}(F(x), F(v))}{\partial F(v)}.$$

However, this pair-copula decomposition is not unique, for example, there are 240 different constructions for a five dimensional density. Each decomposition approximates the full copula density differently (Aas et al. [2009]). A graphical model, called a regular vine model, was developed by Kurowicka and Cooke [2006] to organise the large number of pair-copula constructions. Canonical vines and D-vines are special cases of regular vines. Canonical vines can be used if one can identify the key variable that governs the interaction of the data set. If dependence between variables needs to be treated in a specific order, D-vines can be used.

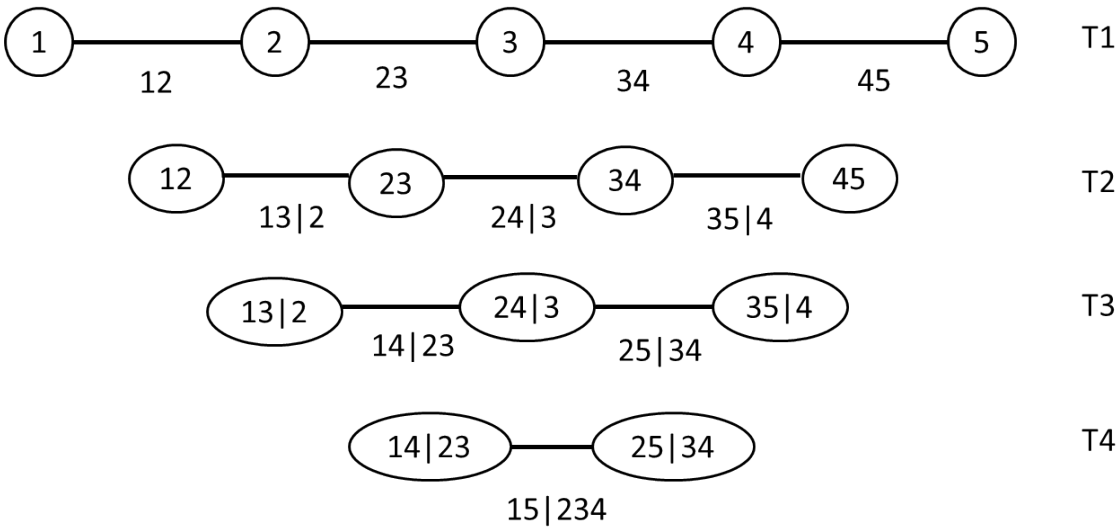


Figure 3.1: A D-vine (5 variables).

Figures 3.1 and 3.2, which are reproduced from Aas et al. [2009], represent the graphical model used to illustrate the D-vine and a canonical vine, respectively, for five variables. Each figure consists of four trees $T_j, j = 1, 2, 3, 4$. Tree T_j has $6 - j$ nodes and $5j$ edges. Each edge represents the corresponding pair-copula and the label of the edge represents the subscript of the pair copula. Nodes in the figure are only used for finding the label of edges.

By using the decompositions shown in Figure 3.1, the joint density function of five random variables can be approximated as follows using a D-vine (Aas et al. [2009]):

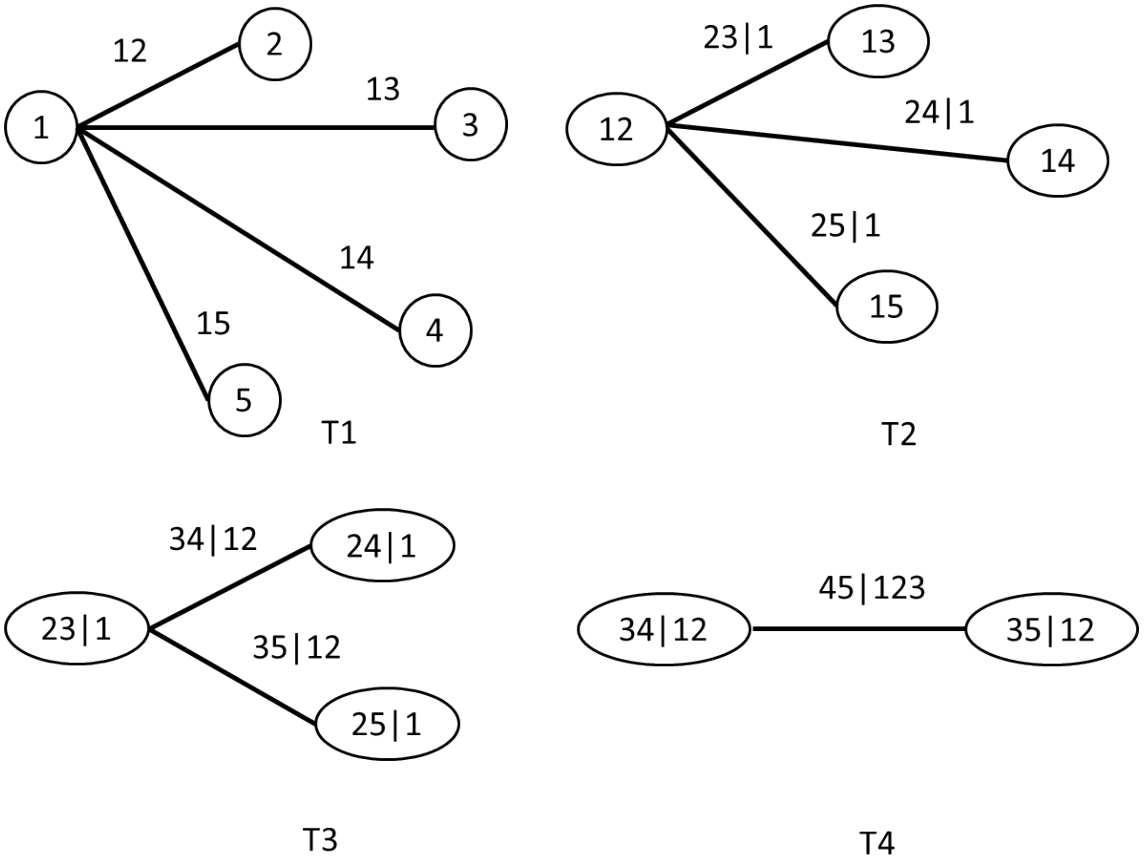


Figure 3.2: A canonical vine (5 variables).

$$
\begin{aligned}
&f_{12345}(x_1, x_2, x_3, x_4, x_5) = \\
&\quad f_1(x_1) \times f_2(x_2) \times f_3(x_3) \times f_4(x_3) \times f_5(x_3) \times \\
&\quad c_{12}(F_1(x_1), F_2(x_2)) \times c_{23}(F_2(x_2), F_3(x_3)) \times c_{34}(F_3(x_3), F_4(x_4)) \times \\
&\quad c_{45}(F_4(x_4), F_5(x_5)) \times c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \times \\
&\quad c_{24|3}(F_{2|3}(x_2|x_3), F_{4|3}(x_4|x_3)) \times c_{35|4}(F_{3|4}(x_3|x_4), F_{5|4}(x_5|x_4)) \times \\
&\quad c_{14|23}(F_{1|23}(x_1|x_2, x_3), F_{4|23}(x_4|x_2, x_3)) \times \\
&\quad c_{25|34}(F_{2|34}(x_2|x_3, x_4), F_{5|34}(x_5|x_3, x_4)) \times \\
&\quad c_{15|234}(F_{1|234}(x_1|x_2, x_3, x_4), F_{5|234}(x_5|x_2, x_3, x_4)).
\end{aligned}
$$

According to Figure 3.2, approximation of the joint density function of five random variables can be written as follows using a canonical vine (Aas et al. [2009]):

$$
\begin{aligned}
f_{12345}(x_1, x_2, x_3, x_4, x_5) = \\
& f_1(x_1) \times f_2(x_2) \times f_3(x_3) \times f_4(x_3) \times f_5(x_3) \times \\
& c_{12}(F_1(x_1), F_2(x_2)) \times c_{13}(F_1(x_1), F_3(x_3)) \times c_{14}(F_1(x_1), F_4(x_4)) \times \\
& c_{15}(F_1(x_1), F_5(x_5)) \times c_{23|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)) \times \\
& c_{24|1}(F_{2|1}(x_2|x_1), F_{4|1}(x_4|x_1)) \times c_{25|1}(F_{2|1}(x_2|x_1), F_{5|1}(x_5|x_1)) \times \\
& c_{34|12}(F_{3|12}(x_3|x_1, x_2), F_{4|12}(x_4|x_1, x_2)) \times \\
& c_{35|12}(F_{3|12}(x_3|x_1, x_2), F_{5|12}(x_5|x_1, x_2)) \times \\
& c_{45|123}(F_{4|123}(x_4|x_1, x_2, x_3), F_{5|123}(x_5|x_1, x_2, x_3)).
\end{aligned}
$$

3.3 Pair-copula Construction for Spatial Data

This section provides instruction for the application of pair-copula models to spatial data, as summarised from Gräler and Pebesma [2011] and Gräler [2014].

Gräler and Pebesma [2011] introduced pair-copula construction for spatial frameworks. This method allows modelling of complex spatial dependency in a fully flexible way. They used a canonical vine structure to construct a pair-copula for spatial data because this structure benefits spatial interpolation by giving higher priority to the interaction between the unobserved locations and nearby locations if unobserved locations are selected as the root element.

3.3.1 Assumptions for the copula based geostatistical model

As with conventional geostatistical models, some assumptions are required to apply the copula based model. Like conventional geostatistical models, copula based modelling assumes that the set of measured values of the variable of interest are realisations of a random function (Bárdossy and Li [2008]). However, when applying copula based models, a stationary random function (see the definition in Gaetan and Guyon [2010]) is assumed over the domain of interest. This

assumption is stronger than the conventional linear geostatistical assumption of a second order stationary random function over the domain of interest because the copula based model requires all the moments of the data generating process be unaffected by a change of spatial distance. However, copula based modelling has more advantages when compared to conventional geostatistical modelling, even though it requires a more limiting assumption, such as the ability to obtain the full conditional distribution, ability to remove the influences of marginal distributions when modelling the dependence structure and the ability to model non-linear spatial dependence (Haslauer et al. [2010]). Based on this strong stationarity assumption, the marginal distributions of the variable of interest for each location in the domain are identical, that is, $F_i(z_i) = F(z_i)$. The empirical bivariate copula can be used to explore the spatial variability. As with the variogram, it is assumed that the bivariate spatial copula C_s at any two locations only depends on the separation vector $\boldsymbol{h}$ and is independent of the locations x (Bárdossy and Li [2008], Bárdossy [2006]), that is

$$\begin{aligned} C_s(u,v) &= Pr(F(Z(x)) \leq u, F(Z(x+\boldsymbol{h})) \leq v) \\ &= C_{\boldsymbol{h}}(F(Z(x)), F(Z(x+\boldsymbol{h}))). \end{aligned}$$

All of the above mentioned assumptions are also applicable to spatial modelling based on the pair-copula model. To simplify application of the pair-copula model, spatial dependency is restricted to the isotropic case here. In isotropic situations, it is assumed that spatial dependence varies only with distance and not with direction. In this case the vector $\boldsymbol{h}$ becomes distance h.

3.3.2 Procedure for spatial interpolation using the pair-copula model

The general procedure for applying the pair-copula model for spatial interpolation, based on the details provided in Gräler and Pebesma [2011] and Gräler

[2014], is described step by step in detail as follows.

STEP 1: Empirical bivariate copula densities construction

As mentioned above, the marginal univariate distributions of the variable of interest for each location are identical (based on the stationarity assumption). Therefore, the empirical marginal distribution function $F(z)$ can be estimated using all the observations $z(x_1), \ldots, z(x_N)$ where N is the total number of sample locations. Then a unit interval transformation is applied to the observations using the estimated distribution function.

Distances between every pair $x_i - x_j = h; i \neq j, \forall i, j = 1, 2, \ldots, N$ are then calculated and, thereafter, each pair $\{F(z(x_i)), F(z(x_j))\}$ is placed into a relevant distance class from the following classes $[0, h_1), [h_1, h_2), \ldots, [h_{l-1}, h_l)$, where h_l is the maximum distance at which significant dependence is observed. The mean distance is considered as the representative value for each class.

The empirical bivariate copula densities can be calculated using a kernel density smoothing method if the number of pairs per distance class is considerably large enough, otherwise the empirical bivariate copula can be calculated by defining a regular grid on the unit square and calculating the cumulative frequency of values for each grid. The next step is to fit the theoretical copula model to the empirical copula densities. This is similar to fitting a theoretical model to the experimental variogram.

STEP 2: Theoretical bivariate copula densities and spatial copula construction

Even though it is possible to apply the maximum likelihood method for estimation of bivariate copulae, in the spatial setting, several copula families for each distance class need to be estimated in order to fit the most suitable spatial copula. As an example, if there are ten distance classes and nine copula families are to be compared for each distance class, altogether, ninety bivariate copulas need to be estimated in the first step of pair-copula construction. This may be computationally demanding and time consuming. In this kind of situation, it is simpler

and faster to calculate the inverse of Kendall's tau (or Spearman's rho) for each distance class and convert these values to estimates of the dependence parameter using the functional relationship between Kendall's tau and the dependence parameter of the copula families (Genest and Rivest [1993]). Following this, the copula that produces the maximum likelihood, amongst the copulas for a given distance class, is selected as the spatial copula for the corresponding class.

STEP 3: Pair-copula construction and spatial interpolation

The final aim of any spatial analysis method is to estimate the variable of interest $Z(x)$ at an unsampled location x. Although the kriging estimator is able to produce the expected value at the unsampled location as the estimator, the copula based methodology allows one to estimate the full conditional distribution of $Z(x)$, which is:

$$F(Z(x) \mid Z(x_1) = z(x_1), \ldots, Z(x_N) = z_N) = \\ Pr(Z(x) < z | Z(x_1) = z_1, \ldots, Z(x_N) = z_N)$$

where N is the total number of observations.
The full conditional distribution of the variable of interest at an unsampled location can be written using the corresponding conditional copula $C_{x,N}$:

$$F(Z(x) \mid Z(x_1) = z(x_1), \ldots, Z(x_N) = z_N) = \\ C_{x,N}(F(Z(x)) | u_1 = F(Z(x_1)), \ldots, u_N = F(Z(x_N))).$$

However, it may be computationally intensive to use all the observations in this process. Therefore, the conditional distribution is obtained based on the local nearby points. Bárdossy and Li [2008] explain the method of selecting a sufficient number of nearby locations. For a few randomly selected locations, the density

functions are estimated and plotted for different numbers of nearby locations. The number of nearby locations that produce nearly identical density functions for almost all considered locations can be selected as the sufficient number.
Let n be the nearby points to the unsampled location x, then

$$F(Z(x) \mid Z(x_1) = z(x_1), \ldots, Z(x_n) = z_n) = \\ C_{x,n}(F(Z(x))|u_1 = F(Z(x_1)), \ldots, u_n = F(Z(x_n))).$$

Therefore, the conditional density function can be derived as

$$\begin{aligned} f(z|z_1, ., z_n) &= \frac{\partial F(Z(x)|Z(x_1) = z_1, \ldots, Z(x_n) = z_n)}{\partial z} \\ &= \frac{\partial C(u|u_1 = F(Z(x_1)), \ldots, u_n = F(Z(x_1)))}{\partial z} \\ &= \frac{\partial C(u|u_1 = F(Z(x_1)), \ldots, u_n = F(Z(x_1)))}{\partial u} \times \frac{\partial F(z)}{\partial z} \end{aligned}$$

that is

$$f(z|z_1, \ldots, z_n) = c(u \mid u_1 = F(Z(x_1)), \ldots, u_n = F(Z(x_1))) \times f(z) \tag{3.7}$$

where $f(z)$ is the marginal density and $F(z)$ is its distribution function.

It is clear that in order to construct the conditional density function of the variable of interest at an unsampled location, constructing a conditional copula density is essential. The procedure for constructing the conditional copula density using the pair-copula construction is described using an example as follows.
Assume the value of the variable of interest at unobserved spatial locations is required to be estimated using four nearby locations. Figure 3.3, which is reproduced from Gräler and Pebesma [2011] , shows how the pair-copula decomposition should be carried out based on a canonical vine structure to obtain the full five dimensional pair-copula density. In the figure, an edge represent a bivariate copula and the two nodes connected to each edge represent the two arguments of the corresponding bivariate copula. The unobserved location is x_0 and x_1, x_2, x_3 and

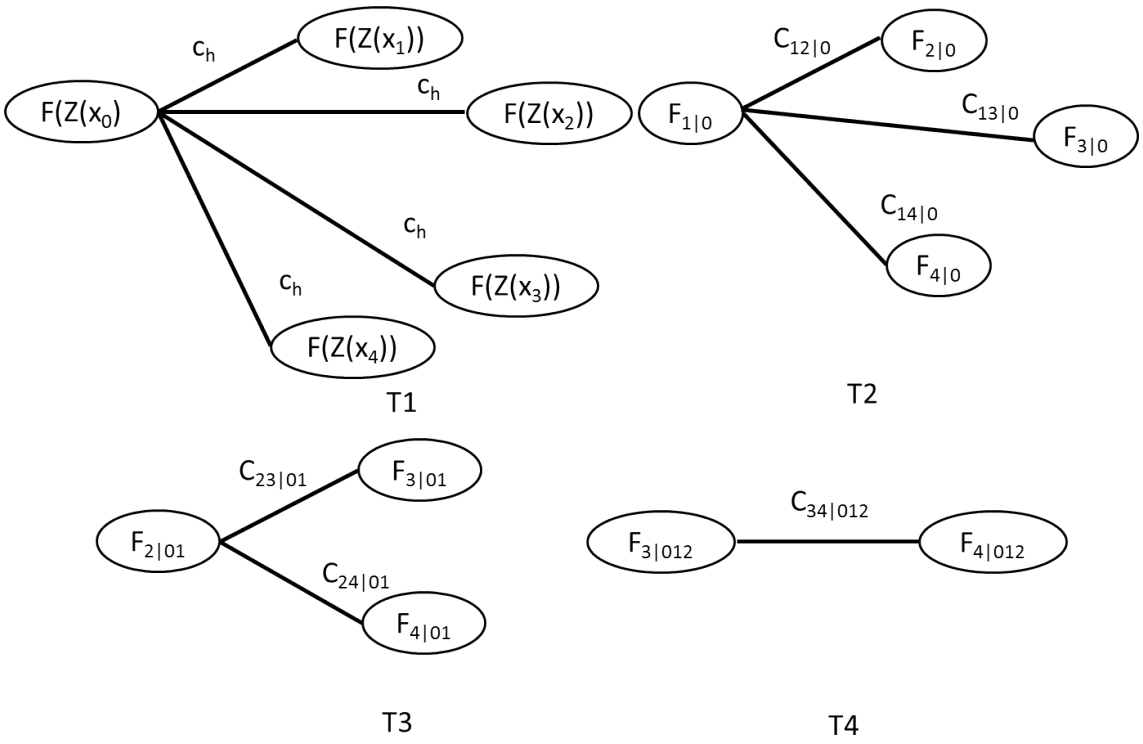


Figure 3.3: Five dimensional spatial vine.

x_4 are nearby locations.

The estimation process of the copulas in the first tree, T1, has already been discussed in STEP 2. By using these copulas, $F_{i|0}, i = 1, 2, 3, 4$, can be calculated according to Eq.(3.6).

Then the conditional pair-copula in the second tree can be estimated. The same procedure can be repeated to estimate the conditional copulas in other trees. However, one can see that these conditional copulae may be influenced not only by their conditional distribution function arguments but also by the value of the conditioning variable. For example, $c_{12|0}$ is influenced by its arguments $(F_{1|0}(z(x_1)|z(x_0)), F_{2|0}(z(x_2)|z(x_0)))$ and the value of $Z(x_0)$. But in pair-copula construction, estimation of a conditional pair-copula is simplified by ignoring the influence from the value of the conditioning variable to keep the construction process more practicable (Haff et al. [2010]). Moreover, Haff et al. [2010] showed that even though this simplified version has some limitations, it is a good approximation for the actual model.

Finally, according to the decomposition shown in Figure 3.3, the full five dimensional copula density can be written as

$$
\begin{aligned}
&c(u_0, u_1, \ldots, u_4) = \\
&\quad c_h(F(z(x_0)), F(z(x_1))) \times c_h(F(z(x_0)), F(z(x_2))) \times c_h(F(z(x_0)), F(z(x_3))) \times \\
&\quad c_h(F(z(x_0)), F(z(x_4))) \times c_{12|0}(F_{1|0}(z(x_1)|z(x_0)), F_{2|0}(z(x_2)|z(x_0))) \times \\
&\quad c_{13|0}(F_{1|0}(z(x_1)|z(x_0)), F_{3|0}(z(x_3)|z(x_0))) \times \\
&\quad c_{14|0}(F_{1|0}(z(x_1)|z(x_0)), F_{4|0}(z(x_4)|z(x_0))) \times \\
&\quad c_{23|01}(F_{2|01}(z(x_2)|z(x_0), z(x_1)), F_{3|01}(z(x_3)|z(x_0), z(x_1))) \times \\
&\quad c_{24|01}(F_{2|01}(z(x_2)|z(x_0), z(x_1)), F_{4|01}(z(x_4)|z(x_0), z(x_1))) \times \\
&\quad c_{34|012}(F_{3|012}(z(x_3)|z(x_0), z(x_1), z(x_2)), F_{4|012}(z(x_4)|z(x_0), z(x_1), z(x_2))).
\end{aligned}
$$

The conditional copula density of the variable of interest at the unsampled location can then be obtained as follows

$$c(u_0 \mid u_1, \ldots, u_4) = \frac{c(u_0, u_1, \ldots, u_4)}{\int_0^1 c(v, u_1, \ldots, u_4) dv}.$$

Finally, point estimates (mean and median) for the variable of interest at unobserved location x_0 can be obtained as follows (Bárdossy and Li [2008])

$$
\begin{aligned}
\hat{Z}_{mean}(x_0) &= \textstyle\int_0^1 F^{-1}(u) c(u|u_1, \ldots, u_n) du, \\
\hat{Z}_{median}(x_0) &= F^{-1}(u = C_n^{-1}(0.5|u_1, \ldots, u_n)).
\end{aligned}
$$

Since this method provides the full conditional distribution at an unsampled location, it is easy to obtain a more "complete" estimation of uncertainty, such as confidence intervals, when compared to the kriged model. Here "complete" is used to emphasise that the copula based model is fully capable of producing uncertainty estimation dependent on both the observations' configuration and values. This feature is very important for additional drilling campaigns, where a reduction of uncertainty is expected based on the influence of additional measurements.

3.4 Application

Confidential data on one particular metal from a real mine site are presented, in which there are nearly $80,000$ measurements from over $2,000$ drill holes. A

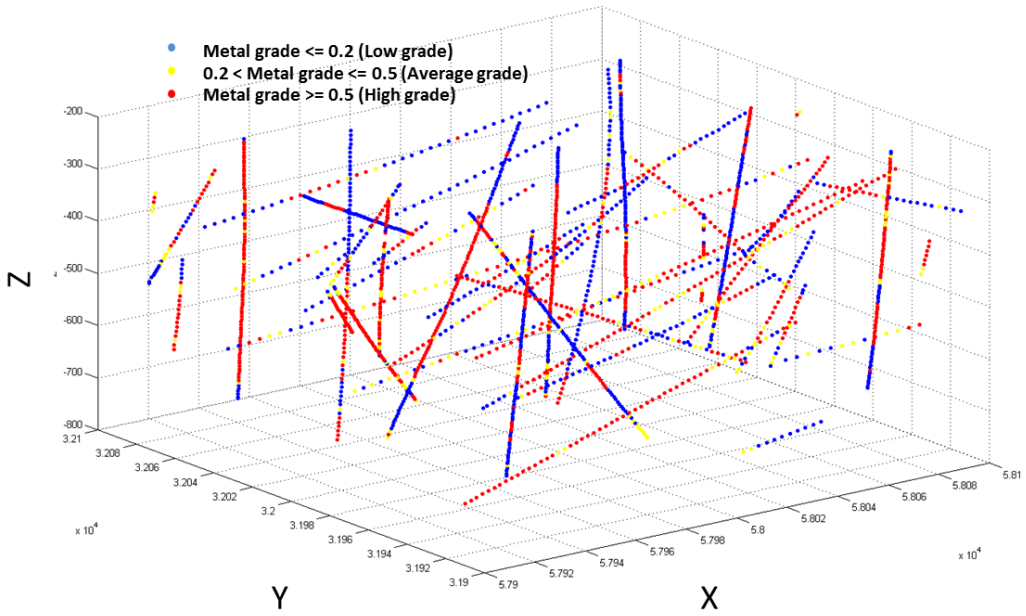


Figure 3.4: Spatial 3D plot of main metal grade.

small scale example is presented here based on a random subset of the spatial observations, and spatial analysis performed on this subset for the grade of the main metal. This subset consists of 2,086 measurements of grade of the main metal $z(x_i)$ at three dimensional locations $x_i = (x_{1i}, x_{2i}, x_{3i}), i = 1, \ldots, 2086$, as shown in the Figure 3.4. The following spatial statistical analysis was carried out using R software (R Core Team [2014]) and R-package "spcopula" of Gräler (see `http://r-forge.r-project.org/projects/spcopula/`).

Summary statistics of the grade of the main metal can be seen in Table 3.1 whilst a histogram of the main metal grade can be seen in Figure 3.5, from which positive skewness is clearly demonstrated. The blue, red and green curves on the histogram demonstrate the fitted gamma, generalised extreme value and lognormal distributions, respectively.

The first step of copula based spatial analysis is estimating the marginal distribution function of the variable of interest $Z(x)$. Two approaches have been used in this application for estimating the marginal distribution function. The first one is a purely empirical marginal function and the second is the best fitted parametric marginal function based on the maximum likelihood values. For the parametric marginal function, the log-likelihood values for the generalised

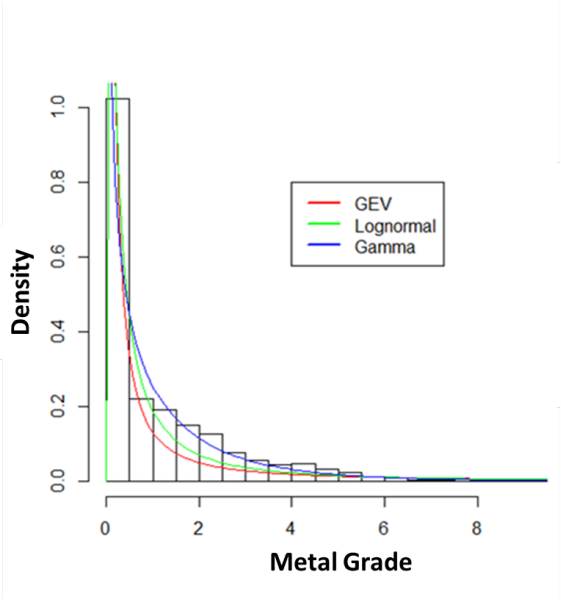


Figure 3.5: Histogram of main metal grade.

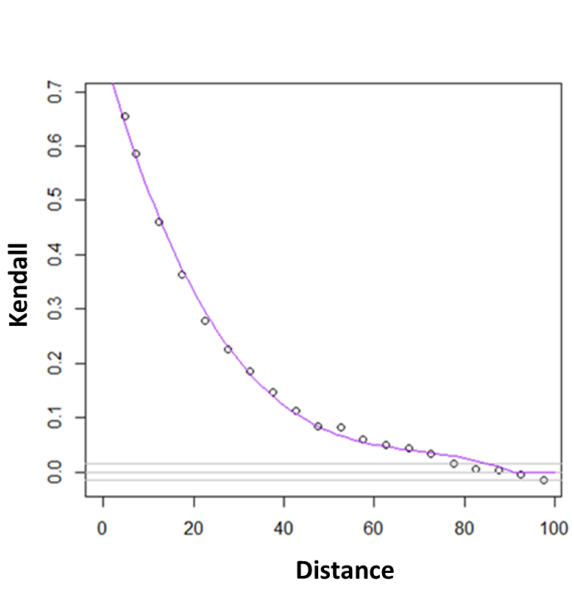


Figure 3.6: Kendall tau values against the mean of the distance classes.

Table 3.1: Summary statistics of the main metal grade.

Statistic	Value
n	2086
Mean	1.106
Standard deviation	1.340
Coefficient of variation	1.265
Min	0.009
First quartile Q1	0.064
Median	0.464
Third quartile Q3	1.728
Max	9.015

extreme value distribution, log-normal distribution and gamma distribution were -2082.0, -1971.9, and -1964.0, respectively. Therefore, the gamma distribution was selected as the best fitting distribution amongst the competing distribution functions. The maximum likelihood estimates are 0.544 and 2.033 for the shape and scale parameters respectively for the selected parametric distribution. Using the estimated marginal distribution function, observed measurements were then transformed to the unit interval in order to construct the empirical copula densities to explore the spatial dependency structure.

Five metre by five metre classes were constructed. Selecting this width for the classes ensures high flexibility in the pair-copula model. Moreover, this class width leads to accurate copula estimation since each class contains more than

100 pairs. Figure 3.6 presents the plot of the calculated Kendall tau (See the definition in Frahm et al. [2003]) values against the mean of the distance classes. The cubic function is a good fit for the relationship between Kendall tau values and means of the distance class values. Once the cubic function approaches the x-axis sufficiently closely, zero is assumed for the Kendall tau estimates. Here it can be seen that it is reasonable to assume spatial independence for measurement of the main metal at any two locations which are more than 85 metres apart.

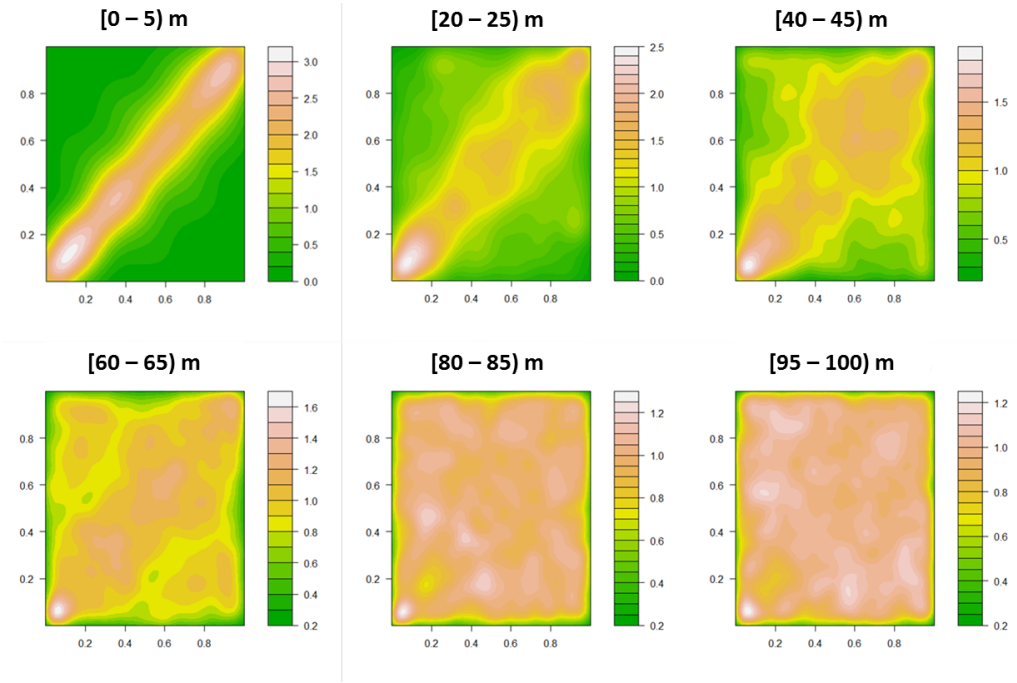


Figure 3.7: Empirical copula density of metal grade for 0-5 m, 20-25 m, 40-45 m, 60-65 m, 80-85 m and 95-100 m distance classes.

Figure 3.7 shows the empirical copula densities obtained for six of the twenty different distance classes. If the spatial dependency is linear then the empirical copula density plots should demonstrate a similar structure to that shown in Figure 3.8. Even though the distance class $[0, 5)$ m demonstrates a linear spatial structure, the other distance classes have more complex spatial structures than linearity. The empirical plots in Figure 3.7 confirm the spatial independency of any two locations which are more than 85 metres apart. Inversion of Kendall's tau was used to estimate the dependence parameter of the spatial copula and the copula with the highest log-likelihood value produced amongst the Gaussian, Student t, Frank, Clayton, Gumbel, Joe and survival version of the last three

copulas was fitted to each distance class. Table 3.2 shows the best fitting spatial copula for each distance class, while the fitted conditional pair-copulas are shown in Table 3.3.

Table 3.2: Best fit copulas for each distance class.

Class	Copula	Dependence Parameter	Degrees of freedom
0-5	Student t	0.634	4
5-10	Survival Joe	2.210	-
10-15	Survival Gumbel	1.520	-
15-20	Student t	0.453	4
20-25	Student t	0.388	4
25-30	Survival Joe	1.490	-
30-35	Survival Joe	1.400	-
35-40	Survival Joe	1.330	-
40-45	Survival Gumbel	1.150	-
45-50	Student t	0.170	4
50-55	Student t	0.143	4
55-60	Survival Joe	1.140	-
60-65	Student t	0.098	4
65-70	Survival Clayton	0.107	-
70-75	Joe	1.070	-
75-80	Frank	0.276	-
80-85	Survival Joe	1.040	-
85-90	Independent	-	-

The anisotropy (the directional effect on the spatial dependence structure) of the data set was evaluated mainly in two directions: horizontal and vertical. The Kendall tau plots show fairly similar dependence structures for these two directions. Hence, throughout the application, isotropic spatial dependency is assumed. This pair-copula model was applied to real world mine data and cross-validation was carried out to compare the performance of the model with ordinary kriging. Figure 3.9 presents the experimental variogram that was used for ordinary kriging where the exponential model was used to model spatial dependency. The estimated nugget, sill and the range of the exponential model are 0.898, 2.027 and 15.215 respectively. The same bin size as the pair-copula model was used when constructing the variogram model. The leave-one-out cross-validation technique was used, with ten nearby locations in the interpolation process. Unlike the kriged model, the copula based model has the ability to produce the full con-

Table 3.3: Fitted conditional bivariate copulas.

Notation	Copula	Dependence Parameter	Degrees of freedom
$C_{1,2\|0}$	Student t	0.359	3.906
$C_{1,3\|0}$	Survival Gumbel	1.210	-
$C_{1,4\|0}$	Survival Gumbel	1.090	-
$C_{1,5\|0}$	Survival Gumbel	1.200	-
$C_{1,6\|0}$	Survival Joe	1.220	-
$C_{1,7\|0}$	Frank	1.010	-
$C_{1,8\|0}$	Student t	0.218	4.978
$C_{1,9\|0}$	Tawn Type 1- Rotated 90	-1.711	0.040
$C_{1,10\|0}$	Frank	0.698	-
$C_{2,3\|0,1}$	Clayton	0.162	-
$C_{2,4\|0,1}$	Student t	0.256	3.569
$C_{2,5\|0,1}$	Survival Joe	1.160	-
$C_{2,6\|0,1}$	Frank	0.535	-
$C_{2,7\|0,1}$	Frank	0.569	-
$C_{2,8\|0,1}$	Joe- Rotated 270	-1.050	-
$C_{2,9\|0,1}$	Survival Joe	1.210	-
$C_{2,10\|0,1}$	Survival Gumbel	1.150	-
$C_{3,4\|0,1,2}$	Frank	0.868	-
$C_{3,5\|0,1,2}$	Survival Clayton	0.046	-
$C_{3,6\|0,1,2}$	Survival Tawn Type 2	1.373,	0.123
$C_{3,7\|0,1,2}$	Survival Tawn Type 1	1.436	0.231
$C_{3,8\|0,1,2}$	Survival Gumbel	1.060	-
$C_{3,9\|0,1,2}$	Student t 0.161,	4.837	-
$C_{3,10\|0,1,2}$	Survival Joe	1.110	-
$C_{4,5\|0,1,2,3}$	Survival Clayton	0.101	-
$C_{4,6\|0,1,2,3}$	Survival Joe	1.09 0	-
$C_{4,7\|0,1,2,3}$	Student t	0.024	3.490
$C_{4,8\|0,1,2,3}$	Joe	1.090	-
$C_{4,9\|0,1,2,3}$	Student t	0.177	5.032
$C_{4,10\|0,1,2,3}$	Survival Joe	1.180	-
$C_{5,6\|0,1,2,3,4}$	Survival Gumbel	1.070	-
$C_{5,7\|0,1,2,3,4}$	Clayton	0.122	-
$C_{5,8\|0,1,2,3,4}$	Frank	0.843	-
$C_{5,9\|0,1,2,3,4}$	Survival Joe	1.060	-
$C_{5,10\|0,1,2,3,4}$	Clayton	0.230	-
$C_{6,7\|0,1,2,3,4,5}$	Survival Gumbel	1.110	-
$C_{6,8\|0,1,2,3,4,5}$	Joe	1.070	-
$C_{6,9\|0,1,2,3,4,5}$	Clayton- Rotated 90	-0.098	-
$C_{6,10\|0,1,2,3,4,5}$	Tawn Type 2- Rotated 90	-1.660	0.032
$C_{7,8\|0,1,2,3,4,5,6}$	Frank	0.310	-
$C_{7,9\|0,1,2,3,4,5,6}$	Survival Joe	1.100	-
$C_{7,10\|0,1,2,3,4,5,6}$	Survival Clayton	0.053	-
$C_{8,9\|0,1,2,3,4,5,6,7}$	Survival Gumbel	1.090	-
$C_{8,10\|0,1,2,3,4,5,6,7}$	Survival Tawn Type 1	1.798	0.172
$C_{9,10\|0,1,2,3,4,5,6,7,8}$	Survival Joe	1.180	-

ditional distribution of the variable of interest at unsampled locations. Therefore any estimator can be obtained. Here, two estimators, the mean and median, were estimated from the copula model.

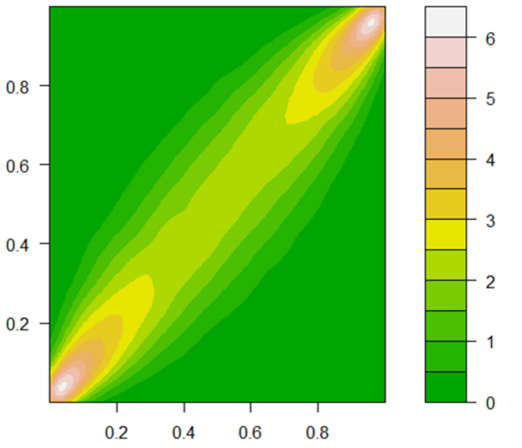


Figure 3.8: Gaussian copula density.

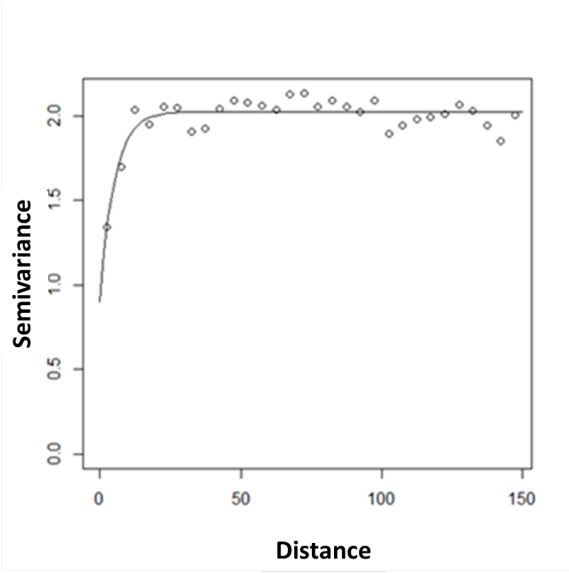


Figure 3.9: Empirical variogram with fitted theoritical model (Exponential).

The performances of the models were evaluated using different criteria: mean absolute error (MAE), bias (average difference between estimated and true values) and Pearson correlation coefficient of true and estimated values. Table 3.4 summarises these statistics. Figure 3.10 shows the bias against the true value. According to Figure 3.10, existence of conditional bias (lower values are overestimated and higher values are underestimated) can clearly be seen in all models. The main reason for the conditional bias in kriging and indicator kriging is the smoothing effect of the variance of the estimator. Conditional bias arising from smoothing is well-documented and understood in the literature Seo [2013], McLennan and Deutsch [2004]. Even though the smoothing effect does not directly apply to the pair-copula model, throughout the estimation process this model uses several approximations and numerical integrations. It can be conjectured that this might be the reasons for the existence of conditional bias in the estimators of the pair-copula model.

All the pair-copula approaches produce estimates with smaller MAE compared to kriging and the median estimator of the pair-copula with empirical margin

Table 3.4: Results of cross-validation.

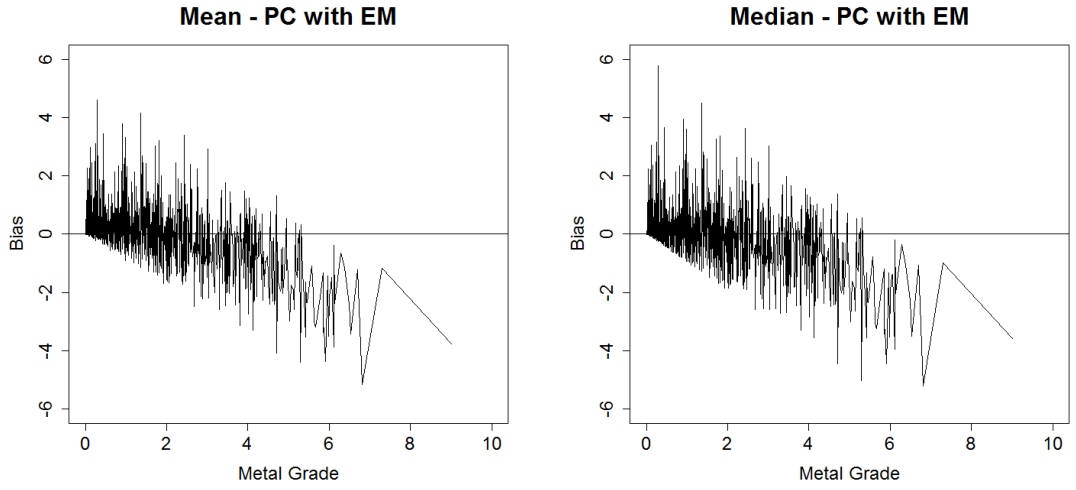
Margin	Approach	MAE	Bias	Correlation
Empirical	Pair-copula -Mean	0.455	0.024	0.831
	Pair-copula -Median	0.439	−0.048	0.826
Gamma	Pair-copula -Mean	0.466	−0.003	0.820
	Pair-copula -Median	0.457	−0.081	0.813
	Ordinary Kriging	0.508	−0.006	0.817

model produces the smallest MAE. On the other hand, the median estimator of the pair-copula empirical margin model has what may be considered, in this practical mining application, unacceptable large global bias. Hence the median estimator of the pair-copula model with empirical margin cannot be considered as the best overall estimator taking into consideration both MAE and bias. However, the mean estimator of the pair-copula model with gamma margin has the lowest global bias amongst all models considered and it produces smaller, if not at least comparable, results compared to the kriged model in terms of MAE. Moreover, Figure 3.10 indicates that all estimators from the pair-copula models perform better than the estimator of the kriged model over the right tail of the distribution of metal grade. Notice that the bias of the individual observations are generally larger as metal grade increases for the kriged model (Figure 3.10(e)), compared to the pair-copula models (Figures 3.10(a)-(d)).

3.5 Discussion and Conclusions

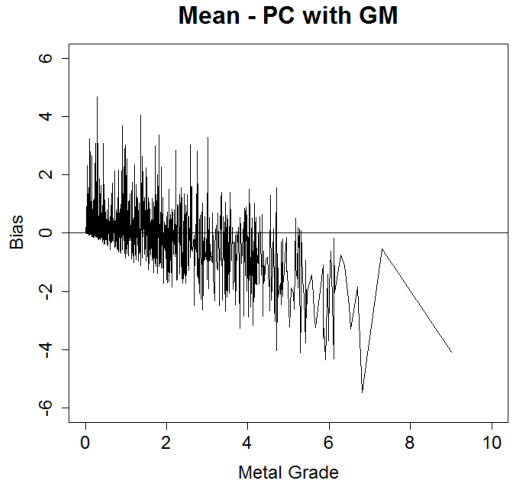
It should be noted that, in mining applications, the mean estimator is expected to perform well because it has the ability to produce unbiased estimates for total metal content. This requirement is satisfied by the pair-copula model with gamma margin for this application. The mean estimator of the pair-copula model with gamma margin has the ability to produce more accurate and less biased estimation for total metal grade compared to kriging. Even though this research focuses on modelling grade, this method can be used to model any geoscientific variable.

The pair-copula model has the potential to become a popular geostatistical model

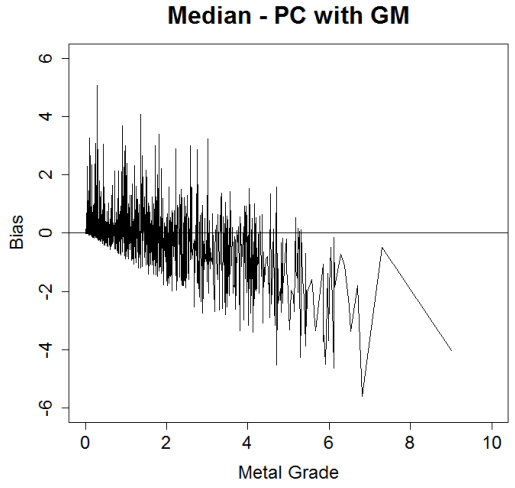


(a)

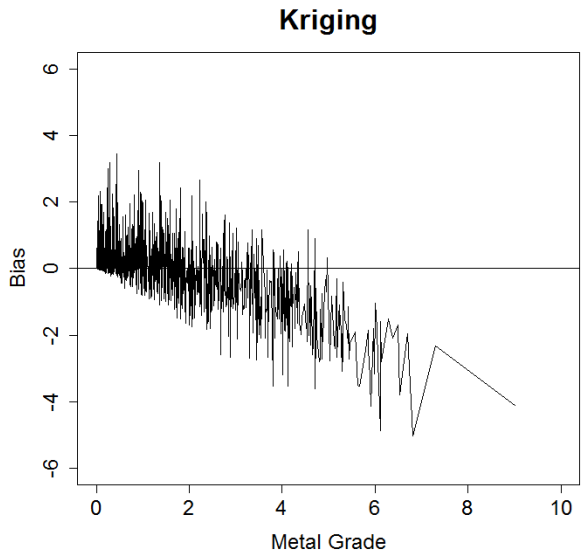
(b)



(c)



(d)



(e)

Figure 3.10: Bias against true metal grade for (a) mean estimate from pair-copula model with empirical margin, (b) median estimate from pair-copula model with empirical margin, (c) mean estimate from pair-copula model with gamma margin (d) median estimate from pair-copula model with gamma margin and (e) kriging.

because of the capability to fit the full conditional distribution, ability to remove the influences of marginal distributions when modelling the dependence structure and the ability to model non-linear spatial dependence and tail dependence. As a result, the copula based model is fully capable of producing uncertainty estimation dependent on both the observations configuration and values. Hence more complete uncertainty estimation can be used to obtain more precise optimal designs than optimal designs obtained using a kriged model for additional drillings.

However, more asymmetric copula families could be introduced in this model to capture the *in-situ* dependency structure. It is not only the correct pair-copula model but also the chosen marginal distribution that will affect the interpolation process, as can be seen from Table 3.4. This can be confirmed through the results. It is also worth mentioning that whilst the use of the empirical marginal distribution limits the range of the possible values for the estimates to be constrained between the minimum and maximum values of observed values, use of a gamma marginal allows any possible values for the estimates.

Finally, from these results, it has been demonstrated that, in our application, the pair-copula model is, overall, better than the kriged model. Moreover, the pair-copula model has the ability to reproduce the right tail of the skewed distribution more successfully than kriging.

Further improvements in the pair-copula model are expected to be gained through, for example, development of an efficient method for defining the lag distance classes, use of advanced search strategies, e.g., quadrant search, to remove the obvious cluster effects, and use of more families of copulas. These improvements are the focus of current research.

Chapter 4

Optimal Distance Classes for Spatial Pair-copulas

The research in this chapter has been submitted to Journal of Computer & Geosciences as detailed below.

- Musafer, G.N., and Thompson, M.H. (n.d). Determination of optimal lag distance classes in spatial pair-copula models. *Computer & Geosciences . Submitted.*

Abstract

An efficient algorithm for finding the optimal distance classes in spatial pair-copula models is presented based on the development of a new test for equality between two spatial copulas. The aim of optimal distance class determination is improvement in fit of the pair-copula model. There is currently no well-defined procedure for determination of distance classes in spatial pair-copula models even though the pair-copula model is based on distance classes. In determining optimal distance classes, a statistical test that is used to test the equality between dependence structures of two empirical copulas in the non-spatial framework is extended to the spatial framework. The test of equality between two spatial copulas is then used to develop an algorithm to determine optimal distance classes. The algorithm is applied to two data sets: data obtained from a real mine site

and the Meuse river bank data set. The results show an improvement in fit of the pair-copula model using the proposed algorithm compared to a pair-copula model with distance classes of equal width.

4.1 Introduction

The purpose of this chapter is the development of methodology for optimal determination of distance classes used in the spatial pair-copula model. The pair-copula model can be classified as a hierarchical model building concept. Aas et al. [2009] initially introduced this method to decompose high dimensional copula random variables using a set of bivariate copulas based on the work of Joe [1996], Bedford and Cooke [2002], and Kurowicka and Cooke [2006]. Gräler and Pebesma [2011] adapted this method to the spatial framework. The pair-copula decomposition of high dimensional spatial copulas allows modelling of complex spatial dependence in a fully flexible way by fitting, potentially, different copula families to different lag distance classes. Consequently, different dependence structures can be fitted to different lag distances. Gräler and Pebesma [2011] used a canonical vine structure to construct a pair-copula for spatial data because this structure benefits spatial interpolation by giving higher priority to the interaction between the unobserved locations and nearby locations, if unobserved locations are selected as the root element.

The first step in building a pair-copula model for a spatial framework is construction of the distance classes for a given data set [Gräler and Pebesma, 2011]. The distance between every data pair is calculated and each pair is placed into a relevant distance class. However, there is currently no well-defined procedure for distance class determination. For instance, two consecutive distance classes may show similar spatial dependence structures. In this situation, it may be more computationally efficient and parsimonious to fit a pair-copula model by combing the two classes than fitting a pair copula model by considering the classes as two separate classes. Moreover, it is more efficient and objective to compare the dependence structure between two distance classes using a statistical test than

by comparing empirical bivariate density plots visually. Since the pair-copula fits a copula to each class, the equality between two fitted copulas can be tested by testing the equality between the corresponding two dependence structures.

In the non-spatial setting, testing two copulas or, in other words, testing the equality between two dependence structures, is very little addressed in the literature. However, Rémillard and Scaillet [2009] introduced a non-parametric test to compare the equality between two copulas that measures similarity between the copulas using a Cramér-von Mises type distance between empirical estimations of the copulas. This can be used for samples coming from two independent populations and also samples coming from two paired populations in a non-spatial environment.

A full description of distance classes, including how these are used in spatial pair-copula models, can be found in Musafer et al. [2015]. The algorithm developed in this chapter is based on an extension of the test of equality between to non-spatial copulas proposed by Rémillard and Scaillet [2009] using the dependent wild bootstrap of Shao [2010] to introduce spatial dependence. Application of the algorithm to two data sets: data obtained from a real mine site and the Meuse data set, demonstrates an improved fit of the pair-copula model based on the proposed algorithm when compared to a pair-copula model that uses distance classes of width.

The following sections describe the test of copula equality for non-spatial data [Rémillard and Scaillet, 2009] and the dependent wild bootstrap [Shao, 2010]. New methodology for testing copula equality in a spatial framework, and its use in determining distance classes in the spatial pair-copula model, is subsequently presented. Finally, results from two applications are discussed, followed by concluding remarks.

4.1.1 Test of equality between non-spatial copulas

The test of copula equality between two non-spatial copulas, proposed by Rémillard and Scaillet [2009], aims to test the following hypothesis

$$H_0 : C = D \ vs \ H_1 : C \neq D$$

where C and D are the copulas associated with first sample, $X_1, \ldots, X_{n_1}$, and second sample, $Y_1, \ldots, Y_{n_2}$, respectively. Here X_i and Y_i are d dimensional real valued vectors. In the spatial framework, in Section 4.2, the first and second samples are two distance classes and $d = 2$.

Rémillard and Scaillet [2009] proposed the following test statistic S_{n_1,n_2}, which is a function of the difference between the empirical copulas C_{n_1} and D_{n_2}, to test the equality between two copulas

$$S_{n_1,n_2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \int_{[0,1]^d} \left(C_{n_1}(u) - D_{n_2}(u)\right)^2 du. \tag{4.1}$$

Rémillard and Scaillet [2009] proved that, under the null hypothesis, the test statistic can the written as

$$S_{n_1,n_2} = \int_{[0,1]^d} \varepsilon(u)^2 du \tag{4.2}$$

where

$$\varepsilon(u)^2 = \sqrt{(1-\lambda)}\mathbb{C}(u) - \sqrt{\lambda}\mathbb{D}(u)$$

with $\lambda = n_1/(n_1 + n_2)$. $\mathbb{C}(u)$ and $\mathbb{D}(u)$ are centred Gaussian processes that have the following representation

$$\mathbb{C}(u) = \alpha(u) - \sum_{l=1}^{d} \beta_l(u_l)\partial_{u_l} C(u), \tag{4.3}$$

$$\mathbb{D}(u) = \gamma(u) - \sum_{l=1}^{d} \delta_l(u_l)\partial_{u_l} D(u), \tag{4.4}$$

where

$$\alpha(u) = \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \{I(U_i \leq u) - C(u)\}, \tag{4.5}$$

$$\beta_l(u_l) = \alpha(1, \ldots, 1, u_l, 1, \ldots, 1), \tag{4.6}$$

$$\gamma(u) = \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} \{I(V_i \leq u) - D(u)\}, \tag{4.7}$$

$$\delta_l(u_l) = \gamma(1, \ldots, 1, u_l, 1, \ldots, 1). \tag{4.8}$$

Here, the distribution of the test statistic cannot be obtained directly due to the unknown C and D. However, Rémillard and Scaillet [2009]) approximated the random terms $\alpha(u)$ and $\gamma(u)$ using the multiplier central limit theorem, as given below. Based on this approximation, a simulation study can be conducted to obtain the p-value.

$$\hat{\alpha}(u) = \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \xi_i^{(k)} \{I(U_i \leq u) - C_{n_1}(u)\}, \tag{4.9}$$

$$\hat{\beta}_l(u_l) = \hat{\alpha}(1, \ldots, 1, u_l, 1, \ldots, 1), \tag{4.10}$$

$$\hat{\gamma}(u) = \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} \zeta_i^{(k)} \{I(V_i \leq u) - D_{n_2}(u)\}, \tag{4.11}$$

$$\hat{\delta}_l(u_l) = \hat{\gamma}(1, \ldots, 1, u_l, 1, \ldots, 1), \tag{4.12}$$

where $\xi_i^{(k)}$ and $\zeta_i^{(k)}$ are independently and identically distributed variables with mean zero and variance one. Here k denotes the simulation number; $k \in \{1, 2, \ldots, N\}$, where N is the number of simulations. The approximation of $\partial_{u_l} C(u)$ and $\partial_{u_l} D(u)$ in Eqs. 4.3 and 4.4 can be found in Rémillard and Scaillet [2009].

Finally, an approximate value $\hat{S}_{n_1,n_2}^{(k)}$ for S_{n_1,n_2} in Eq. 4.2, under the null hypothesis, can obtained by substituting the estimated values in Eqs. 4.9 to 4.12 for Eqs. 4.5 to 4.8, respectively. By doing this for the N simulations, the distribution of the test statistic can be obtained under the null hypothesis.

If two samples are independent, realisations $\xi_i^{(k)}$ and $\zeta_i^{(k)}$ are obtained independently of each other from the standard Guassian distribution. If two samples are paired, $\xi_i^{(k)}$ are obtained from the standard Guassian distribution and $\zeta_i^{(k)}$ is equal to $\xi_i^{(k)}$.

4.1.2 Dependent wild bootstrap

Shao [2010] introduced the dependent wild bootstrap (DWB) method to carry out statistical tests in time series. By using this DWB, a dependence structure can be injected into a time series to obtain the distribution of the Cramér-von Mises test statistic under the null hypothesis. Doukhan et al. [2015] discussed three variants of the DWB used by different authors in the literature. In this chapter, the second version of the DWB (DWB-2) is adopted due to its suitability and simplicity in defining a dependence structure for spatial data.

In the DWB-2, independent and identical realisations are obtained for each observation in the time series from the Gaussian distribution with mean zero and variance $1/l$, where l is the number of dependent lags. Then, for each observation, the following random component is calculated using those realisations

$$\varepsilon^*_{t,n} \quad = \quad \zeta^*_t + \ldots + \zeta^*_{t-l+1} \tag{4.13}$$

where $\zeta^*_t \sim N(0, 1/l)$ are independently and identically distributed, t is time lag and n is the number of observations.

Thereafter, the random component in Eq. 4.13 is used to obtain an approximation of the Cramér-von Mises test statistic $V_n = \frac{1}{n^2}\sum_{s,t=1}^{n} h(X_s, X_t)$ as follows

$$V^*_n \quad = \quad \frac{1}{n^2}\sum_{s,t=1}^{n} h(X_s, X_t)(\varepsilon^*_{s,n} - \bar{\varepsilon^*_n})(\varepsilon^*_{t,n} - \bar{\varepsilon^*_n})$$

where $X = \{X_1, X_2, \ldots, X_n\}$ is the time series of interest and $\bar{\varepsilon^*_n}$ is the average of the random components $\varepsilon^*_{i,n}$.

4.2 Methodology

Distance classes are the building blocks of the pair-copula model [Musafer et al., 2015]. However there is no well-defined procedure for defining the distance classes. Existing papers that apply the pair-copula model for spatial data use distance classes of equal width [Gräler and Pebesma, 2011, Gräler, 2014]. As discussed

previously, there may be situations where two consecutive distance classes have similar spatial dependence structures. This section describes a novel systematic algorithm for combining classes with similar dependence structures.

4.2.1 Test of equality between two spatial copulas

From Section 4.1.1, $\xi_i^{(k)}$ and $\zeta_i^{(k)}$ are the terms in the test statistic for a test of equality between to copulas that quantify the dependence, or otherwise, of the data. The test for equality between two non-spatial copulas is extended to the spatial framework here by replacing $\xi_i^{(k)}$ and $\zeta_i^{(k)}$ in Eqs. 4.9 and 4.11 with quantities that capture spatial dependence. A modification of the DWB random component, given in Eq. 4.13, is proposed to replace $\xi_i^{(k)}$ and $\zeta_i^{(k)}$.

Assume that $z_1, \ldots, z_n$ are n spatial observations obtained from the study domain and l number of neighbours are used for interpolation. As the first step, n independent realisations are obtained from the Gaussian distribution with mean zero and variance $1/(l+1)$, i.e., $e_i \sim N(0, 1/(l+1)); i = 1, \ldots, n$. Thereafter, a spatially dependent random component w_i for each observation can be obtained as follows

$$w_i \;=\; e_i + \sum_{t=1}^{l} e_{i,t} \tag{4.14}$$

where the $e_{i,t}$, $t = 1, \ldots, l$, represent the independent realisations that are obtained for the l locations neighbouring the i-th spatial location and w_i follows the standard Gaussian distribution.

Now, w_i can be considered a spatially dependent component similar to $\varepsilon_{t,n}^*$ in Eq. 4.13 for temporal dependence. With the w_i, it is possible to generate random components to replace $\xi_i^{(k)}$ and $\zeta_i^{(k)}$ in Eqs. 4.9 and 4.11, respectively, for a spatial framework as described below.

The first step in fitting a pair-copula model is the construction of the bivariate empirical copula for each distance class [Musafer et al., 2015]. Assume that C_{n_1} and D_{n_2} are the empirical copulas for the two distance classes of interest. Here n_1 and n_2 are the number of pairs of observations belonging to each distance class.

First, consider the empirical copula C_{n_1} for the first distance class. Any data pair in an empirical copula C_{n_1} for a distance class contains the information of the two spatial observations comprising the pair. For example, the s_1-th pair of C_{n_1} is $\{F(z_i), F(z_j)\}_{s_1}$; $i, j = 1, 2, \ldots, n$, $i \neq j$ and $s_1 = 1, \ldots, n_1$. Hence, there are two random components, w_i and w_j, associated with each pair.

However, in Eqs. 4.9 and 4.11, only one random component is used for each pair in the given empirical copula. Hence, the summation of the w_i and w_j is proposed as the random component in the spatial setting. However, even though the mean of the distribution of the summation of w_i and w_j is zero, the variance will not be equal to one. Thus, the random component $\xi_{s_1}^{(k)}$ for the s_1-th pair in C_{n_1} is replaced by $\xi_{s_1}^{*\,(k)}$, where

$$\xi_{s_1}^{*\,(k)} = \frac{w_i + w_j}{standard\ erorr(w_i + w_j)}. \tag{4.15}$$

The s_2-th pair of empirical copula D_{n_2} for the second distance class consists of two spatial observations $\{F(z_i'), F(z_j')\}_{s_2}$; $i', j' = 1, 2, \ldots, n$, $i' \neq j'$ and $s_2 = 1, \ldots, n_2$. Therefore, w_i' and w_j' are able to be generated similarly to w_i and w_j. Hence, as with the first distance class, the random component $\zeta_i^{(k)}$ for the s_2-th pair in D_{n_2} can be replaced by $\zeta_{s_2}^{*\,(k)}$, that is

$$\zeta_{s_2}^{*\,(k)} = \frac{w_i' + w_j'}{standard\ erorr(w_i' + w_j')}. \tag{4.16}$$

The random components defined in Eqs. 4.15 and 4.16 can be used in Eqs. 4.9 to 4.12 for the spatial framework as follows

$$\hat{\alpha}(u) = \frac{1}{\sqrt{n_1}} \sum_{s_i=1}^{n_1} \xi_{s_1}^{*\,(k)} \{I(U_i \leq u) - C_{n_1}(u)\}, \tag{4.17}$$

$$\hat{\beta}_l(u_l) = \hat{\alpha}(1, \ldots, 1, u_l, 1, \ldots, 1), \tag{4.18}$$

$$\hat{\gamma}(u) = \frac{1}{\sqrt{n_2}} \sum_{s_2=1}^{n_2} \zeta_{s_2}^{*\,(k)} \{I(V_i \leq u) - D_{n_2}(u)\}, \tag{4.19}$$

$$\hat{\delta}_l(u_l) = \hat{\gamma}(1, \ldots, 1, u_l, 1, \ldots, 1). \tag{4.20}$$

These values can be used to obtain an approximate value $\hat{S}_{n_1,n_2}^{(k)}$ for S_{n_1,n_2} in

Eq. 4.2, under the null hypothesis. Here k represents the k-th simulation. To obtain the distribution for S_{n_1,n_2} under the null hypothesis, this whole process is repeated N times.

After calculating the test statistic S_{n_1,n_2}, as in Eq. 4.1, using the empirical copulas for the distance classes, the p-value can be calculated as follows

$$p = \frac{\sum_{k=1}^{N} I(\hat{S}_{n_1,n_2}^{(k)} > S_{n_1,n_2})}{N}. \tag{4.21}$$

The following steps summarise the proposed test of equality between two spatial copulas.

1. Draw n independent e_i from the Gaussian distribution with mean zero and variance $1/(l+1)$.
2. Calculate n dependent random components w_i for the corresponding spatial observation using Eq. 4.14.
3. Using the w_i's obtained in step 2, obtain random components ${\xi_{s_1}^*}^{(k)}$, using Eq. 4.15, for each pair in the first class.
4. Using the w_i's obtained in step 2, obtain random components ${\zeta_{s_2}^*}^{(k)}$, using Eq. 4.16, for each pair in the second class .
5. Calculate the quantities in Eqs. 4.17 and 4.18 using the values obtained in step 3 and calculate the quantities in Eqs. 4.19 and 4.20 using the values obtained in step 4 .
6. Substitute the approximated values obtained in step 5 into the Eq. 4.2 to obtain an approximate value $\hat{S}_{n_1,n_2}^{(k)}$ for S_{n_1,n_2} under the null hypothesis.
7. Repeat steps 1 to 6, N times to obtain the distribution of the test statistic in Eq. 4.2 under the null hypothesis.
8. Calculate the test statistic S_{n_1,n_2} using observed values and Eq. 4.2.
9. Calculate the p-value using Eq. 4.21.
10. Reject H_0 if p-value $<$ significance level.

4.2.2 Defining distance classes

Generally, there may be more than two distance classes when developing a pair-copula model. Hence, multiple comparisons should be carried out to compare the dependence structure between pairs of distance classes to define the optimal classes for a given case study.

Constructing distance classes of equal width is the first step in the pair-copula modelling of Gräler and Pebesma [2011]. It is essential to have at least ten data pairs in each distance class for maximum likelihood estimation. If this requirement is not satisfied, the width of the distance classes could be increased. However, by using too wide a distance class, distance classes with different dependence structures may be combined.

After placing the data pairs in relevant distance classes, a plot of Kendall's tau against the mean distance of each class can be obtained. From this plot, the maximum distance (L) of any two locations that have a significant dependence structure can be determined. Hence, for any two points with distance greater than L, independence can be assumed. For distance classes with distance less than L, carry out pair-wise tests on consecutive distance classes using the test of equality between two spatial copulas described in Section 4.2.1. If the test produces a non-significant p-value, combine the consecutive distance classes into a new wider distance class. Then, depending on those results, further combine distance classes.

Let $d = (d_1, \ldots, d_k)$ denote the initial distance classes and C_i the copula corresponding to distance class $i; i = 1, \ldots, k$. An algorithm for combining classes is given in Algorithm 1, such that redundant pair-wise tests are not carried out. For example, if $d_1 \neq d_2$, that is, distance classes d_1 and d_2 are not combined, this implies that $d_1 \neq d_3$, since it is sensible, in the spatial setting, to combine only consecutive distance classes. Additionally, in combining multiple distance classes, all pair-wise comparisons of the distance classes comprising the combined distance classes must produce a non-significant p-value. For example, if the combined distance class is $d_1 + d_2 + d_3$, then $C_1 = C_2$, $C_2 = C_3$ and, importantly,

$C_1 = C_3$. This ensures that the first and last classes are actually similar.

Figure 4.1 shows the application of Algorithm 1 for a pair-copula with four equally spaced initial distance classes, $d_1, \ldots, d_4$.

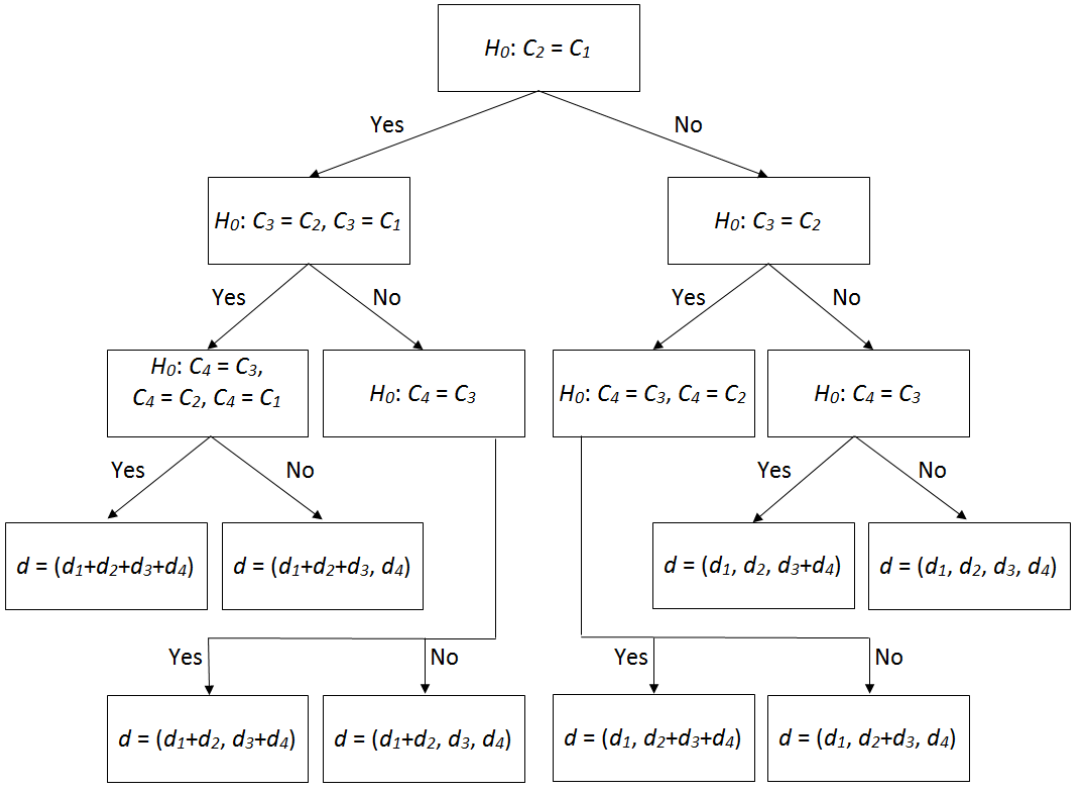


Figure 4.1: Application of Algorithm 1 for four distance classes.

4.3 Application

The following analysis was carried out using a computer with an Intel(R) Core(TM) i5 CPU (2.53GHz) processor and 4 GB memory. R statistical software [R Core Team, 2014] and its 'spcopula' package (see `http://r-forge.r-project.org/projects/spcopula/`) were used to carry out the analysis.

4.3.1 Data from a real mine

A small scale example using data from a real mine site is presented here. The data set consists of 200 spatial observations of the main metal at three dimensional locations.

As an initial step, the data points should be transformed to the unit interval

Algorithm 1: Algorithm for combining distance classes.

Definition:

$d = [d_1, \ldots, d_k]$ # Vector of initial distance classes of equal width

$k = length(d)$ # The number of initial distance classes

$newd = NULL$ # Vector to store the combined distance classes

$C = [C_1, \ldots, C_k]$ # Vector of bivariate empirical copulas for initial distance classes

$i \leftarrow 1$, $j \leftarrow 1$, $combine \leftarrow 1$

Notation:

$twocop(C_i, C_j)$ # Test of equality between spatial empirical copulas C_i and C_j using the test in Section 4.2.1. Output of this test is 0 (not equal) or 1 (equal).

Calculation:

while $(k > i - 1)$

while $(combine > 0)$

if $(k > i + j - 1)$

$l \leftarrow 1$, $m \leftarrow i + j - l$

while $(m > i - 1)$

if $(twocop(C_{i+j}, C_{i+j-l}) = 1)$

$l \leftarrow l + 1$, $m \leftarrow i + j - l$

else $m \leftarrow i - 1$

end if

end while

if $(l = j + 1)$

$j \leftarrow j + 1$

else $combine \leftarrow 0$

end if

else $combine \leftarrow 0$

end if

end while

add $sum(d(i)$ *to* $d(i + j - 1))$ *to* $newd$

$i \leftarrow i + j$, $j \leftarrow 1$, $combine \leftarrow 1$

end while

using a rank transformation or using the estimated marginal distribution of the data in order to construct the empirical copulas for each distance class.

In this application, the estimated marginal distribution was used in fitting a pair-copula. Then, 5 metre by 5 metre classes were constructed, which ensured a minimum of ten data pairs in each distance class. Data points were then assigned to relevant distance classes. Figure 4.2 is a plot of the calculated Kendall tau values against the mean of the distance classes. According to Figure 4.2, spatial independence can be assumed for the measurement of the main metal at any two locations which have more than a 95 metre separation distance. Hence, there are 19 distance classes to consider. A cubic relationship is appropriate in describing the relationship between the Kendall tau values and the distance for the first 19 classes. Thereafter, Algorithm 1 was applied to determine the optimal distance classes for the pair-copula model. Table 4.1 shows the original class boundaries, the best fitted theoretical copula with estimated Kendall tau values for the original class boundaries, the number of data pairs for the corresponding classes, the class boundaries for the combined classes from Algorithm 1, the best fitting theoretical copula for the classes after applying Algorithm 1 and their Kendall tau values.

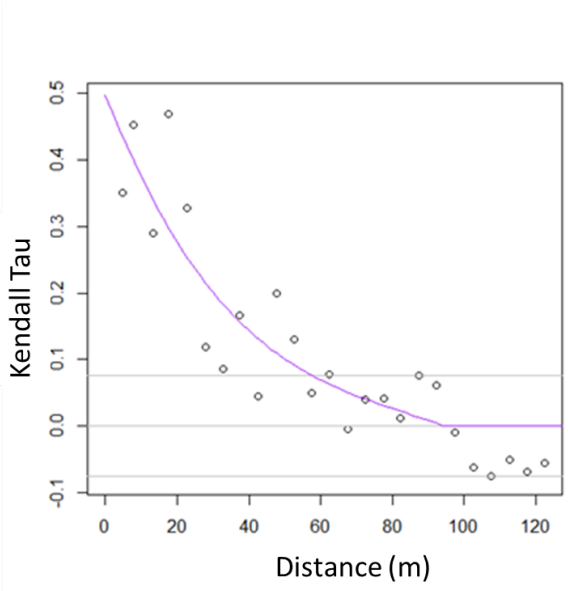


Figure 4.2: Mine data. Kendall tau values against the mean of the distance classes.

Finally, a pair-copula model was fitted to the original classes of equal width

Original distance classes				Algorithm 1 distance classes		
Boundaries	Best fitted copula	Kendall's tau	No. pairs	Boundaries	Best fitted copula	Kendall's tau
0–5	Student t	0.35	12	0–25	Survival Gumbel	0.38
5–10	Survival Joe	0.45	21			
10–15	Survival Gumbel	0.29	29			
15–20	Student t	0.47	35			
20–25	Student t	0.33	40			
25–30	Survival Joe	0.12	52	25–55	Survival Joe	0.13
30–35	Survival Joe	0.09	68			
35–40	Survival Joe	0.17	87			
40–45	Survival Gumbel	0.04	100			
45–50	Student t	0.20	120			
50–55	Student t	0.13	149			
55–60	Survival Joe	0.05	164	55–65	Student t	0.07
60–65	Student t	0.08	190			
65–70	Survival Clayton	−0.01	219	65–95	Survival Clayton	0.04
70–75	Joe	0.04	261			
75–80	Frank	0.04	263			
80–85	Survival Joe	0.01	281			
85–90	Frank	0.08	273			
90–95	Survival Gumbel	0.06	290			

Table 4.1: Mine data. Class boundaries using Algorithm 1.

and the combined classes using Algorithm 1. In fitting the pair-copula model, inversion of Kendall's tau was used to estimate the dependence parameter of the spatial copula and the copula with the highest log-likelihood value produced amongst the Gaussian, Student t, Clayton, Gumbel, Joe, Survival Clayton and Survival Gumbel copulas was fitted to the each distance class, since these copula families are able to capture different dependence structures, as explained in Trivedi and Zimmer [2007]. Thereafter, cross-validation was carried out to compare the performance of the pair-copula model with equal width classes and

the pair-copula model with combined classes using Algorithm 1. The leave-one-out cross validation technique was used, with ten nearby locations used for each location when constructing the conditional copula in the interpolation process. Gräler and Pebesma [2011] use four nearby locations. Increasing the number of nearby locations reduces mean absolute error (MAE) but at a cost in increased computational time. For this example, Figure 4.3 shows the reduction in MAE for an increasing number of neighbour locations and the corresponding increase in computational time. For the cross validation of the pair-copula model with equal width classes, using a sub-sample of the data, computational time increases sharply for more than 10 neighbour locations, and the reduction in MAE decreases for more than 10 neighbour locations. Hence, ten nearest neighbours were used in this example. Unlike conventional linear geostatistical models, the pair-copula model has the ability to produce the full conditional distribution of unsampled locations. Therefore, any estimator can be obtained. Here, two estimators, the mean and median, were estimated.

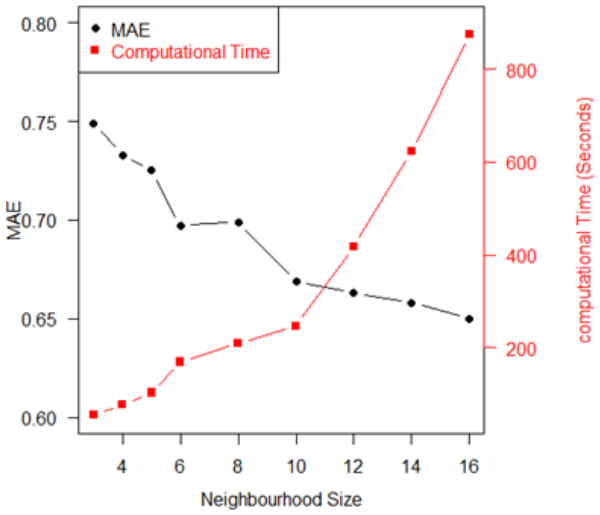


Figure 4.3: Mine data. MAE and computational time against number of neighbour locations.

The performance of the models was evaluated using different criteria: mean absolute error (MAE) and bias (average difference between estimated and true values). Other than these two criteria, goodness of fit of the fitted copula to each distance class should be evaluated between the equally spaced distance classes and the

combined classes from Algorithm 1. To that, the average of the mean square error of bivariate kernel density estimations (bivariate KDE MSE) over the distance classes was used. For a given distance class, the bivariate KDE MSE is calculated as the mean square difference of the KDE of the empirical and fitted theoretical copula.

The mean square error of the bivariate kernel density estimations of the empirical copula and fitted theoretical copula was calculated for each class (bivariate KDE MSE). The average value of the bivariate KDE MSE for all the classes was selected as the final value to represent the goodness of fit of the model. A smaller number for this statistic indicates a better fit. For ease of comparison, the mean bivariate KDE MSE values are divided by the smaller mean bivariate KDE MSE. As a result, the value of the statistic for the model that produces the smaller bivariate KDE MSE is equal to one and the statistic for the alternative model is the KDE MSE relative to the best model.

Table 4.2 presents a summary of these statistics. The model using Algorithm 1 produces the lowest bivariate KDE MSE and the pair-copula model with equal width classes has, approximately, a 40% increase in the mean bivariate KDE MSE compared to the Algorithm 1 model. When comparing pair-copula models, the estimator of the pair-copula model with combined classes using Algorithm 1 (PCA) produces the lowest MAE regardless of the estimator. However, the bias of the median estimator for the PCA model is slightly larger than the median estimator for the pair-copula model with original classes (PCO). However, in terms of KDE MSE, the PCA model produced a better fit.

Model	Boundaries	Relative KDE MSE	Mean		Median	
			MAE	Bias	MAE	Bias
PCO	[0,5,10,....,95]	1.39	0.878	−0.130	0.857	−0.349
PCA1	[0,25,55,65,95]	1.00	0.830	−0.101	0.806	−0.359

Table 4.2: Mine data. Results of cross-validation. PCO = pair-copula model with original distance classes and PCA = pair-copula model with distance classes from Algorithm 1.

4.3.2 Meuse data set

The Meuse river bank data set [Bivand et al., 2013], which is available in the R package [R Core Team, 2014], was used as a second application. This data set has spatial observations on four top soil heavy metal concentrations and seven other secondary variables at 155 different two dimensional locations. The top soil zinc concentration was selected as the variable of interest for this example.

As discussed in the previous application, it is essential to have at least ten data pairs in each class to apply Algorithm 1 using maximum likelihood estimation. The same methodology that was used for the mining data set was applied to the Meuse data set with 70 by 70 metre classes. A plot of the calculated Kendall tau values against the mean of the distance classes is displayed in Figure 4.4. Spatial independence can be assumed after approximately 600 metres. Hence, there are eight distance classes to consider. Moreover, a linear relationship best describes the relationship between the Kendall tau values and distance for the first eight distance classes. Table 4.3 presents the boundaries of the original classes of equal width and the combined classes using Algorithm 1. In addition, the best fitted copula and the estimated Kendall tau values for each distance class, for both the original classes and the combined classes using the Algorithm 1, can be also be found in Table 4.3. Leave-one-out cross validation was used, with eight nearby locations, to evaluate the performance of Algorithm 1. The choice of eight nearby locations was chosen in a similar fashion to the mining example and was chosen based on Figure 4.5. For the same data set, Gräler and Pebesma [2011] constructed a pair-copula model using four nearby locations. As per Figure 4.5, it can be seen that the use of eight nearby locations requires only a small increase in computational time for a marked reduction in MAE. Thebivariate KDE MSE, MAE and bias for the PCO and PCA models are presented in Table 4.4.

As with the mining application, the PCA model performed better than the PCO model in terms of bivariate KDE MSE. Whilst the bias of the median estimator of the PCA model is only slightly smaller than the PCO model, the bias of the mean estimator of the PCA model is approximately 50% of the bias in the PCO

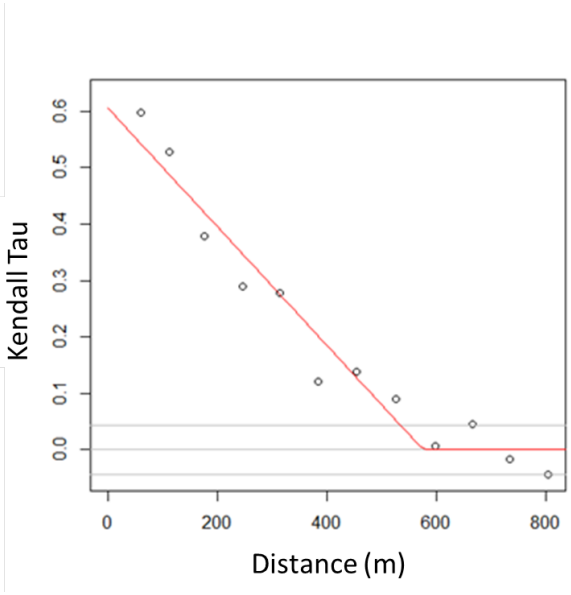


Figure 4.4: Meuse data. Kendall tau values against the mean of the distance classes.

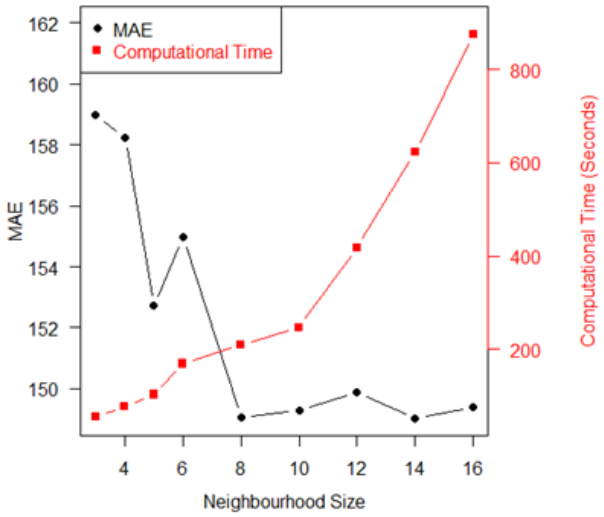


Figure 4.5: Meuse data: MAE and computational time against number of neighbour locations.

model. Moreover, in terms of MAE, the PCA model is better than the PCO model regardless of the estimator.

4.4 Conclusions

In this chapter, a new algorithm to determine the distance classes for the spatial pair-copula model is developed. In developing the algorithm, the test proposed

Original distance classes				Algorithm 1 distance classes		
Boundaries	Best fitted copula	Kendall's tau	No. pairs	Boundaries	Best fitted copula	Kendall's tau
0–70	Survival Joe	0.60	18	0–140	Survival t	0.53
70–140	Student t	0.53	112			
140–210	Survival Gumbel	0.38	217	140–210	Survival Gumbel	0.38
210–280	Survival Joe	0.29	263	210–350	Survival Joe	0.28
280–350	Survival Joe	0.27	282			
350–420	Survival Joe	0.12	322	350-560	Survival Joe	0.11
420–490	Survival Joe	0.14	346			
490–560	Survival Joe	0.10	371			

Table 4.3: Meuse data. Class boundaries using Algorithm 1.

Model	Boundaries	Relative KDE MSE	Mean		Median	
			MAE	Bias	MAE	Bias
PCO	[0,70,140,..,560]	1.13	153.036	8.814	147.130	−37.217
PCA	[0,140,210,350,560]	1.00	149.707	4.546	144.952	−36.854

Table 4.4: Meuse data. Results of cross-validation. PCO = pair-copula model with original distance classes and PCA = pair-copula model with distance classes from Algorithm 1 .

by Rémillard and Scaillet [2009] is extended to the spatial framework by using the dependent wild bootstrap [Shao, 2010].

In this research, improvement in the pair-copula model was expected to be gained through the development of an efficient method for defining the lag distance classes. In both applications, pair-copula models with classes constructed using Algorithm 1 show a significantly better fit to the data than the pair-copula model with classes constructed using equal widths. The expected improvement was successfully achieved as seen by the more accurate estimates for the pair-copula model using distance classes determined by Algorithm 1 than the pair-copula model with original classes.

From these results, it may be reasonable to assume that more accurate estimates can be obtained by using the pair-copula model with combined classes instead of using the pair-copula copula model with equal distance classes.

The proposed test of equality between spatial copulas should be evaluated based on a simulation study. However, for spatial data, it is difficult to simulate the distance classes with a specific dependence parameter using existing simulation tools. This perspective is the focus of future research.

Chapter 5

Multivariate Modelling

This chapter is in preparation of journal submission as below and was presented at the 10th Congress on Geostatistics for Environmental Applications 2014 , Paris, France.

- Musafer, G.N., Thompson, M.H., Wolff, R.C. and Kozan, E. (n.d). Non-linear multivariate spatial modelling using NLPCA and pair-copulas. *In preparation* .

Abstract

A novel geostatistical modelling approach is developed to model non-linear multivariate spatial dependence using non-linear principal components analysis (NLPCA) and pair-copulas. In spatial studies, multivariate measurements are frequently collected at each location. The dependence between such measurements can be complex. In this chapter a multivariate geostatistical model is developed that can capture both non-linear spatial dependence across locations and non-linear dependence between measurements at a particular location. Non-linear multivariate dependence between spatial variables is removed using NLPCA. Subsequently, a pair-copula based model is fitted to each transformed variable to model the univariate non-linear spatial dependencies. NLPCA and pair-copulas in the proposed model are compared with stepwise conditional transformation (SCT) and conventional kriging, respectively, using cross-validation. The results show that

the proposed model that uses NLPCA and pair-copulas reproduces non-linear multivariate structures and univariate distributions better than existing methods based on SCT and kriging.

5.1 Introduction

The focus of this chapter is on the modelling of non-linear multivariate spatial data. More specifically, interest is in modelling multiple non-linearly spatial variables where the relationship between variables is additionally non-linear.

In spatial studies, multivariate measurements are frequently collected at a given location. For example, environmental monitoring stations yield measurements on ozone, nitrous oxide, carbon monoxide, and so on. In geometallurgical modelling, measurements of rock hardness, mineral grade, geochemical attributes, and so on, are collected. The measurements for the different variables are unlikely to be spatially independent and dependence between these measurements can be non-linear. In addition, measurements of a specific variable are spatially correlated across the locations, and this correlation can also be non-linear. Ignorance of these non-linearities when modelling multivariate spatial data may consequently affect decisions based on the spatial model. For instance, mining projects carry out their financial evaluations based on estimates of potential ore reserves. Inadequate spatial modelling of an ore reserve can, thus, lead to potential project failure. Hence, it is essential to account for these non-linearities when performing spatial modelling.

Existing multivariate models for multiple spatial variables include co-regionalisation models, such as the co-regionalisation Markov model and the Markov-Bayes model. Both of these models ignore non-linear dependence between the variables and, additionally, fail to reproduce the within-variable spatial dependence successfully across locations (Leuangthong and Deutsch [2003]). Moreover, modelling of multiple spatial variables is complex and time consuming when compared to single variable modelling due to the requirement of the number of cross-variograms with an increasing number of variables (Bandarian et al. [2008]). As

a solution to this, multi-variables can be transformed to spatially uncorrelated variables (factors) by using a suitable transformation method. Univariate geostatistical modelling can then be subsequently performed on each factor separately. To restore the dependence structure of the original variables, the factors are appopriately back transformed (Rondon [2012]).

Principal components analysis (PCA) is the most popular method to obtain uncorrelated factors from linearly correlated variables (Wackernagel [2003]). Hence, this method is not an appropriate transformation for practical applications where non-linear dependence is present. Non-linear principal components analysis (NLPCA) is an extension to PCA that can be used to identify and remove any kind of non-linearity between variables (Kramer [1991]). This technique is widely used in different fields, such as micro-biology and image processing as an aid for dimension reduction (feature extraction), visualisation and exploratory data analysis (Kruger et al. [2008]). In this chapter, NLPCA is proposed for use in a spatial framework to identify and remove non-linear relationships between spatial variables. Other popular non-linear transformation techniques, such as stepwise conditional transformation (SCT) and projection pursuit multivariate transformation (PPMT) are competitive techniques to NLPCA. Whilst SCT accurately reproduces the distribution of the variable that is transformed first, the quality of the reproduction of distributions for the second, and subsequent, transformed variables decreases rapidly (Leuangthong [2003]). Thus, SCT is not suitable for application to higher dimensional data. However, as discussed in Barnett et al. [2014], PPMT can be successfully applied to higher dimensional data.

The drawback of NLPCA, SCT and PPMT is that these methods only remove cross-correlation at zero lag distance. If cross-correlation is present at lag distances greater than zero, additional transformation is required to remove the cross-correlation. Barnett and Deutsch [2012] carry out a modification of the minimum/maximum autocorrelation factors (MAF) transformation (Desbarats and Dimitrakopoulos [2000]) following SCT. Barnett et al. [2014] use the same modified MAF following PPMT to remove the remaining cross-correlation.

Once the multivariate spatial variables have been decomposed into uncorrelated factors at all lag distances, univariate geostatistical interpolation methods can be carried out on the uncorrelated factors. Most of the literature concerning non-linear multivariate decomposition techniques, such as those discussed above, employ traditional geostatistical interpolation methods, such as kriging, to model transformed independent factors (e.g., Leuangthong [2003], Barnett et al. [2014]). Since conventional geostatistical models use the covariance function to capture spatial dependence, they frequently fail to capture non-linear dependence. The pair-copula model for spatial data has the flexibility to capture more complex spatial dependence structures and will render more accurate results than traditional interpolation methods (Gräler and Pebesma [2011], Gräler [2014]). Additionally, unlike traditional geostatistical interpolation methods, the pair-copula does not require a Gaussian assumption on the marginal distribution. In this chapter, pair-copula based spatial interpolation is proposed for modelling the uncorrelated univariate factors. In doing so, this chapter introduces the pair-copula to the multivariate setting.

In summary, the non-linear multivariate spatial modelling approach considered involves transforming the multivariate variables into uncorrelated factors at all lag distances, fitting univariate geostatistical models to the factors separately, and back transforming to restore the properties of the observed data. SCT and PPMT are, currently, applied in practice to transform non-linear multivariate variables into uncorrelated factors. As a competitive approach, the use of non-linear principal components analysis (NLPCA) is proposed. The pair-copula approach is additionally proposed to model the univariate uncorrelated factors. Without loss of generality, the implementation of NLPCA and pair-copulas into the non-linear multivariate spatial modelling approach is illustrated using two two-dimenstional data sets, one real and the other artificial. Extension to higher dimensions merely requires additional computation. The accuracy and reliability of the proposed NLPCA and pair-copula implementations are evaluated via cross-validation and are compared to existing methods. Overall, the results indicate that, in the case

studies presented, non-linear multivariate spatial modelling based on transformation of the variables to uncorrelated factors is best implemented using NLPCA and pair-copulas.

5.2 Methodology

This section outlines the proposed methodology for modelling non-linear multivariate spatial data. The general algorithm, which is that used in Barnett and Deutsch [2012] and Barnett et al. [2014], is described below. A description of both existing methods used in the algorithm, and methods newly proposed for use in the algorithm, follows the discussion of the algorithm.

5.2.1 Algorithm

The three main components of the algorithm considered in Barnett and Deutsch [2012] and Barnett et al. [2014] are: forward transformation to transform non-linear multivariate spatial variables into uncorrelated univariate spatial factors, univariate spatial interpolation of the uncorrelated factors, and back transformation of the interpolated factors to the original variables. It should be noted that Barnett and Deutsch [2012] and Barnett et al. [2014] only consider data that are linearly spatial, they do not model non-linear spatial data in their spatial interpolation. This chapter extends the work of Barnett and Deutsch [2012] and Barnett et al. [2014] by inclusion of non-linear spatial interpolation in the spatial interpolation component of the algorithm. Consequently, a non-linear multivariate transformation that preserves the non-linear spatial dependence of the data is also introduced in the forward and backward transformation components of the algorithm.

Forward transformation

The first step in modelling non-linear multivariate spatial variables, considered in Barnett and Deutsch [2012] and Barnett et al. [2014], is multivariate decomposition of the variables into uncorrelated factors at zero lag distance. Barnett

and Deutsch [2012] and Barnett et al. [2014] propose the use of SCT (Rosenblatt [1952]) and PPMT (Friedman and Tukey [1974]), respectively, for the multivariate decorrelation. Whilst both methods are capable of decomposing both linear and non-linear multivariate data into uncorrelated factors, the resulting univariate factors are only linearly spatial. That is, SCT and PPMT do not preserve any non-linear spatial properties of the data that may be present. Here, the use of NLPCA (Kramer [1991]) is proposed as a suitable multivariate decorrelation method for data that are non-linearly spatial.

Application of non-linear transformation methods, such as SCT, PPMT and NLPCA, remove cross-correlation between spatial variables at zero lag distance. However, cross-correlation between spatial variables at lag distances greater than zero may remain. Fitting univariate geostatistical models to transformed uncorrelated factors separately requires that the factors be uncorrelated not only at zero lag distance but at all lag distances. It is commonly assumed that decorrelation of the variables at zero lag distance also decorrelates the variables at all lag distances (Goovaerts [1993], Leuangthong and Deutsch [2003]). Clearly, if this premise does not hold true, the subsequent univariate geostatistical models may not adequately fit the data. A commonly used method that has the ability to remove spatial cross-correlation between variables at all lag distances is MAF transformation (Switzer and Green [1984], Desbarats and Dimitrakopoulos [2000]). MAF, in its full form, is only able to be applied to linear multivariate data, since the first step of MAF transformation involves PCA. However, the second step of the MAF approach (Rondon [2012]), in which the MAF factors are derived, is useful for removing cross-correlation at a lag distance greater than zero. Consequently, the second step of MAF can be applied following non-linear multivariate decorrelation of the variables at zero lag distance, as demonstrated in Barnett and Deutsch [2012] and Barnett et al. [2014].

Spatial interpolation

After obtaining uncorrelated spatial factors, univariate geostatistical modelling can be performed on each factor separately and interpolation carried out at unsampled locations. One of the most important aspects of modelling spatial variables is spatial correlation. Spatial correlation describes the relationship between realisations of a spatial variable sampled at different locations. Any method used to model a spatial variable should be capable of accurately estimating the true spatial correlation. The traditional kriging method (Krige [1951]), as implemented by Barnett and Deutsch [2012] and Barnett et al. [2014], is only able to model data with linear spatial dependence. Therefore, standard kriging models are unlikely to produce accurate estimators of distributional properties at unsampled locations when non-linear dependence is present. Here, the use of the pair-copula model (Gräler and Pebesma [2011]) is proposed for appropriately modelling data with non-linear spatial dependence. Pair-copula models (Gräler and Pebesma [2011], Gräler [2014]) have, to date, only been applied in univariate non-linear spatial settings. This chapter presents the first application of pair-copula models in a multivariate spatial framework.

Back transformation

The final step in the algorithm is to back transform the interpolated values to their original scale, ensuring estimates retain the spatial dependence structure and non-linear multivariate relationships of the original variables. The back transformation should be carried out in the reverse order in which the forward transformation is applied.

5.2.2 Multivariate decorrelation at lag $h=0$

This section describes methods for non-linear multivariate decorrelation of spatial variables into uncorrelated spatial factors at zero lag distance, which occurs in step 1 of Algorithm 2. The corresponding back-transformation, which occurs in step 5 of Algorithm 2, is also discussed for each method. The methods considered

are SCT, used in the algorithm by Barnett and Deutsch [2012], PPMT, implemented in Barnett et al.s' (2014) version of the algorithm, and NLPCA, which is proposed for the new version of the algorithm to facilitate the modelling of non-linear spatial data.

Algorithm 2: General algorithm for modelling non-linear multivariate spatial data using transformation methods.

Forward Transformation

1. Multivariate decorrelation at lag $h = 0$: Apply a non-linear transformation to the multivariate data to produce uncorrelated factors at lag $h = 0$.

2. Multivariate decorrelation at lag $h > 0$: If spatial cross-correlation exists at lag $h > 0$, derive the MAF factors to produce uncorrelated factors at lag $h > 0$.

Spatial Interpolation

3. Fit univariate geostatistical models to each factor separately and interpolate at unsampled locations.

Back Transformation

4. Multivariate recorrelation at lag $h > 0$: If MAF factors were derived in step 2, apply the corresponding MAF back transformation to recorrelate the original variables for the interpolated data at lag $h > 0$.

5. Multivariate recorrelation at lag $h = 0$: Apply the corresponding back transformation of the non-linear transformation used in step 1 to recorrelate the original variables for the interpolated data at lag $h = 0$.

The algorithm described above is summarised in Algorithm 2. More generally, in what follows, $\boldsymbol{h}$ denotes a separation vector. To simplify application of the methodology, spatial dependence is restricted to the isotropic case here. In isotropic situations, it is assumed that spatial dependence varies only with dis-

tance and not with direction. In this case the vector $\boldsymbol{h}$ becomes distance h.

SCT

The SCT method transforms multivariate variables to multivariate Gaussian variables with no cross-relationship at zero lag distance (Leuangthong [2003]). The stepwise conditional transformation for the m-variate case can be illustrated as follows.

Let $Y_1, Y_2, \ldots, Y_m$ be spatially dependent variables and let $F_{i|1,2,\ldots,i-1}(y_i|y_1, y_2, \ldots, y_{i-1})$ be the distribution function of variable Y_i conditioned on first $i-1$ variables. Let g^{-1} be the inverse standard Gaussian distribution function. The transformed variables are then given by

$$\begin{aligned} T_1 =& g^{-1}(F_1(y_1)) \\ T_2 =& g^{-1}(F_{2|1}(y_2 \mid y_1)) \\ & \vdots \\ T_m =& g^{-1}(F_{n|1,2,\ldots,n-1}(y_n \mid y_1, y_2, \ldots, y_{m-1})) \end{aligned}$$

where $T_1, T_2, \ldots, T_m$ are multivariate Gaussian distributed with no cross-correlation at zero lag distance, i.e., $cov(T_i(u), T_j(u)) = 0$ where $i \neq j$ and $i, j = 1, 2, \ldots, m$. Back transformation to the original scale is carried out by applying the standard Gaussian distribution function g to $T_1, T_2, \ldots, T_m$ in the same order as the forward transformation. That is, estimates of T_1 are first transformed to the original scale, then estimates of T_2 are transformed, and so on. This ensures that the multivariate dependence structure of the original variables is retained by the estimates.

As discussed previously, cross-correlation between the transformed variables at lag $h > 0$ may be present. Additionally, the resultant transformed variables will be linearly spatial due to the Gaussian transformation applied to the data, which may be problematic if the original data is non-linearly spatial.

PPMT

Barnett et al. [2014] proposed PPMT as a non-parametric method for transforming complex and high dimensional geologic data to an uncorrelated multi-Gaussian distribution. PPMT is based on, and very closely resembles, the projection pursuit density estimation (PPDE) algorithm of Friedman and Tukey (Friedman and Tukey [1974]) and Friedman (Friedman [1987]).

Before applying the projection pursuit transformation, a Normal score transformation is applied to each variable. Data sphering is then carried out to obtain centred variables with unit variance and orthogonal covariance matrix. Finally the projection pursuit transformation is applied to obtain uncorrelated multivariate Gaussian data (Barnett et al. [2014]). A conceptual description of the projection pursuit transformation is given below.

Consider m-dimensional unit vector α and the projection of the data upon it, $p = \alpha^T X$, where X is the matrix of sphered variables. Any α should yield a p that is univariate Gaussian if X is multi-Gaussian. The projection pursuit process conducts an optimised search to find the α that produces the most non-Gaussian projection of the multivariate data. The multivariate data is then Gaussianised to remove this structure. By iterating this procedure, the data gradually transforms to multi-Gaussian data. Details of the optimised search method and Gaussianisation can be found in Hwang et al. [1994]. After completing the interpolation procedure, based on the transformed multi-Gaussian data, the interpolated values are back transformed to the original scale based on the distance between the interpolated value and its nearest neighbours in transformed space (Barnett et al. [2014]). Because PPMT, as with SCT, is based around Gaussianisation of the data, PPMT faces the same issues as SCT as a result of the Gaussianisation. However, in contrast to SCT, PPMT may be applied to any arbitrary number of variables due to its non-parametric nature. PPMT was applied to the data sets considered in this chapter. However, the resulting transformed variables did not retain any spatial properties of the original data, thus an application of the PPMT approach is not presented in this chapter.

NLPCA

NLPCA is a non-linear generalisation of standard (linear) PCA that reduces observed, possibly non-linearly correlated, variables to a number of uncorrelated factors (Kruger et al. [2008], Linting et al. [2007]).

In NLPCA, an arbitrary non-linear mapping function is used to obtain uncorrelated factors from observed variables through the following transformation $\mathbf{t}_i = g(\mathbf{y}_i)$, where $\mathbf{y}_i$ is the i-th row of $n \times m$ data matrix Y with n observations on m variables, g is a non-linear vector-valued function composed of f individual non-linear functions $g = (g_1, g_2 \ldots, g_f)$, and $\mathbf{t}_i$ is the corresponding row of the $n \times f$ matrix T containing the $f \leq m$ uncorrelated factors. The (i, j)-th element of T is given by $t_{ij} = g_j(\mathbf{y}_i)$.

The non-linear functions $g_1, g_2, \ldots, g_f$ are analogous to the columns of the loadings matrix in (linear) PCA; g_1 is referred to as the primary non-linear factor, and g_j is the j-th non-linear factor of Y. The inverse transformation, which restores the original dimensionality of the data, is obtained through a second non-linear vector-valued function $h = (h_1, h_2, \ldots, h_m)$: $y'_{ij} = h_j(\mathbf{t}_i)$, where y_{ij} is the (i, j)-th element of the reconstructed data matrix Y'. The functions g and h are chosen to minimise $||E||$, the Euclidean norm of the residual matrix $E = Y - Y'$:

$$\|E\| = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - y'_{ij})^2}. \tag{5.1}$$

AANN implementation of NLPCA

There are several methods that can be used to implement NLPCA, such as principal curve techniques (Hastie and Stuetzle [1989]), kernel PCA (Schölkopf et al. [1998]) and auto-associative neural networks (AANN) (Scholz et al. [2008], Kramer [1991]). Principal curve techniques demand computational power to extract the non-linear factors. Consequently, the number of extracted factors is typically limited to two (Scholz et al. [2008]). Kernel PCA is better used as a visualisation tool or noise reduction method, rather than as a technique for extracting the non-linear factors (Scholz et al. [2008]). The most common imple-

mentation of NLPCA is via AANNs. Hence, this chapter focusses on the AANN approach (Bishop [1995]) to extract non-linear factors from non-linear multivariate spatial data.

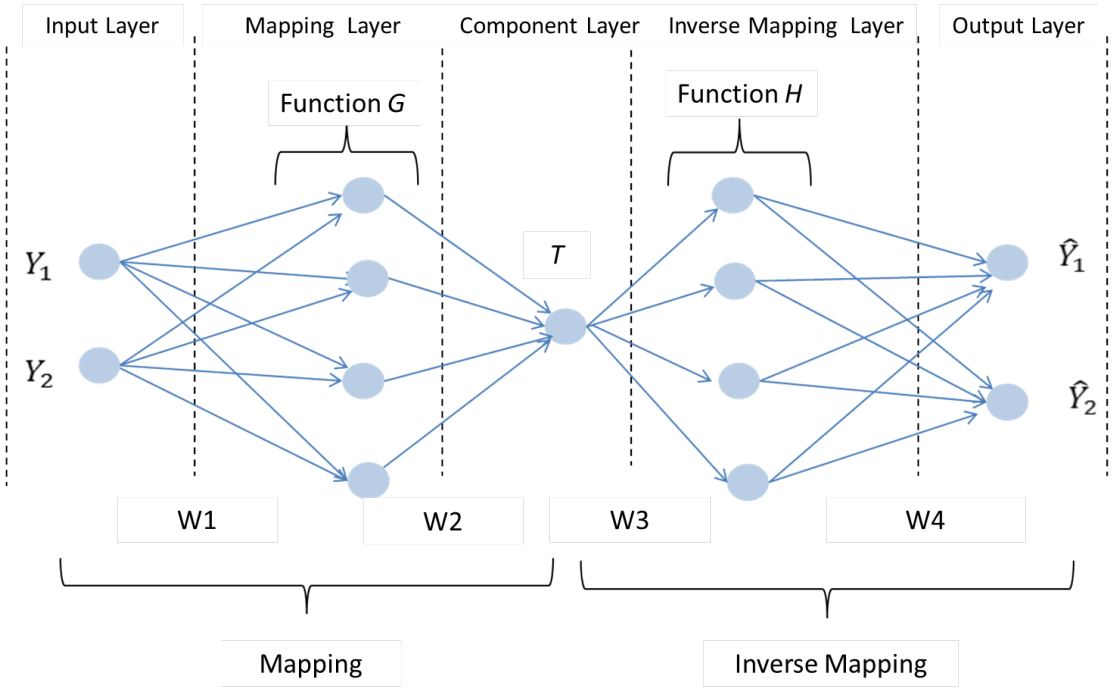


Figure 5.1: A standard AANN used to obtain a single non-linear factor.

The AANN used to implement NLPCA has five layers, as depicted in Figure 5.1. The layers include an input layer, three hidden layers (mapping layer, component layer and inverse mapping layer), and an output layer. The input and output layers represent Y and Y' respectively. The modelling of g and h is carried out in the mapping layer and the inverse mapping layer, respectively. The component layer, which is often refereed to as the bottleneck layer, represents T. The importance of the three hidden layers is discussed in Kramer [1991]. When the mapping and inverse mapping layers are absent, the AANN is equivalent to PCA.

As indicated in Figure 5.1, the AANN contains weights, $w_1, \ldots, w_4$, which are the parameters of the network. Training of the neural network is important in identifying suitable weights for the network. There are three different ways to train a neural network: supervised training, unsupervised training and reinforcement training. For NLPCA, the target of the network is to reproduce Y. In other words, the network needs to be trained as an identity mapping, $Y \rightarrow Y$. Since the desired output is well known, the network can be trained using supervised training.

Through supervised training, the weights of the network can be tweaked continuously until the desired outputs are achieved. This kind of network is known as "auto-associative" or "self-supervised back-propagation" (Kramer [1991]). Scholz et al. [2008] recommend the addition of a weight decay term to Equation 5.1 to penalise large network weights:

$$\|E\|_{total} = \|E\| + \vartheta \sum_{i=1}^{4} {w_i}^2. \tag{5.2}$$

For a wide range of cases, $\|E\|_{total}$ is minimised at around $\vartheta = 0.01$.

Figure 5.1 shows the structure of a standard AANN used to obtain a single nonlinear component from two variables. This network is generically denoted as a $2 - k - 1 - k - 2$ network architecture, corresponding to the two observed variables in the input layer, k nodes in the mapping layer ($k = 4$ in Figure 5.1), one component in the bottleneck layer, k nodes in the inverse mapping layer, and two reconstructed variables in the output layer. The number of nodes in the mapping and inverse mapping layers depend on the complexity of the nonlinearity of the data. A reasonable approach to decide the number of nodes in these layers is discussed in Kramer [1991].

This standard AANN can be modified to carry out NLPCA under alternative network architectures. Different data structures can also be be handled through modifications to this AANN. For example, h-NLPCA and circular NLPCA can be implemented for hierarchical and circular data, respectively. The h-NLPCA AANN can be constructed by using a constraint on the variance of the components or using a constraint on the reconstruction error given in Equation 5.2 (Scholz et al. [2008]). Circular NLPCA can be constructed using two components in the component layer of the standard AANN, whose outputs are constrained to project onto a circle (Kirby and Miranda [1996]). More details on these extended AANNs can be found in various articles (Scholz and Vigário [2002], Scholz [2007], Scholz et al. [2008]).

After development of the NLPCA, by training the AANN to perform an identity mapping on the data, the weights of the network can be estimated by minimising

the error function given in Equation 5.2. The weights and uncorrelated factors (components) are implicitly stored in the AANN. The factors stored in the network are the forward transformed variables in step 1 of Algorithm 2. Following spatial interpolation, using univariate geostatistical models based on the uncorrelated factors, the interpolated values can be back transformed through the same AANN used for the forward transformation. The dimensionality of the original data and non-linear correlation between the original variables is restored through this back-transformation.

The advantages of NLPCA for decomposing non-linear multivariate spatial data into uncorrelated factors are that it preserves both multivariate non-linearity and spatial non-linearity present in the observed data. Multivariate non-linear dependence can be destroyed or distorted when a Normal score transformation is applied to the data, such as in SCT and PPMT (Bandarian et al. [2008]). Additionally, Gaussinisation of multivariate data into multi-Gaussian space, as takes place in SCT and PPMT, can destroy or distort non-linear spatial dependence (Leuangthong and Deutsch [2003], Barnett et al. [2014]). The drawback of NLPCA implemented using an AANN is, however, the computational burden in decomposing higher dimensional data (> 10 dimensions). Also, as with SCT and PPMT, whilst transformed spatial variables are uncorrelated at zero lag distance, cross-correlation may remain at lag $h > 0$.

It should be noted that, if data are not available for all variables at all sampled locations, imputation of missing data is required to allow decomposition of the data into uncorrelated factors (Barnett et al. [2014]). The inverse network model for NLPCA can be used to deal with missing data for data with non-linear structures, as discussed in Scholz et al. [2005].

5.2.3 Multivariate decorrelation at lag $h > 0$

In this section, decorrelation of multivariate data at lag $h > 0$, which is step 2 of Algorithm 2, is discussed. The corresponding back transformation to recorrelate the data, which occurs in step 4 of Algorithm 2, is also described. The method

under consideration is the MAF transformation used by Barnett and Deutsch [2012] and Barnett et al. [2014] in their versions of Algorithm 2.

Initial decomposition of non-linear multivariate spatial data into uncorrelated factors using non-linear transformation methods, such as SCT, PPMT and NLPCA, generally decorrelates the factors at zero lag distance only. To fit univariate geostatistical models to the factors separately, the factors must be uncorrelated at all lag distances. Therefore, the correlation between the factors should be checked using, for example, a cross-semivariogram or cross-correlogram, to verify whether any correlation is present at lag distances greater than zero. In most situations, removal of correlation at lag $h = 0$ indirectly removes correlation at far away lag distances. However, correlation at shorter lag distances may be present. Such correlation should not be ignored, as subsequent interpolated or simulated data based on the non-linearly transformed variables may not successfully reproduce the dependence structure of the original variables (Barnett et al. [2014]). Thus, it is necessary to remove this remaining correlation if possible. The second step of the MAF approach can be applied to remove cross-correlation at lag distances greater than zero (Rondon and Tran [2008]).

MAF can be categorised into two techniques: model based and data driven. In the model based technique, direct and spatial cross-correlation is modelled using a specific linear model of co-regionalisation (LMC) (Desbarats and Dimitrakopoulos [2000]). These models are used to obtain factors that are independent at all lag distances. However, fitting a LMC to model the spatial cross-correlation is a difficult and time consuming task. Conversely, the data driven technique does not require any prior model to carry out the transformation (Switzer and Green [1984]). However, it is only capable of removing spatial cross-correlation at shorter lag distances. The data driven MAF approach is adopted in this chapter since, in most situations, spatial cross-correlation is likely only to remain at shorter lag distances after initial decomposition of the data into uncorrelated factors at zero lag distance.

After obtaining uncorrelated factors at lag distance $h = 0$, the second step of

data driven MAF can be applied to produce MAF factors that are uncorrelated at a short lag distance $h > 0$ using the following steps.

1. Select a suitable lag distance, $h = h_{max}$ ($h_{max} > 0$), at which cross-correlation is to be removed, where h_{max} is the distance at which non-negligible maximum cross-correlation exists.

2. Let the factors that are uncorrelated at $h = 0$ be denoted by $T = (T_1, T_2, \ldots, T_m)$. Obtain the variance-covariance matrix at lag distance $h = h_{max}$, $\Gamma(h_{max})$, for $T_1, T_2, \ldots, T_m$.

3. Obtain the spectral decomposition of $\Gamma(h_{max})$. Let Q be the matrix of eigenvectors.

4. Transform $T = (T_1, T_2, \ldots, T_m)$ to spatially independent factors $M = (M_1, M_2, \ldots, M_m)$ at lag $h = h_{max}$, where $M = QT$.

The back transformation to restore the correlation of the factors $T_1, T_2, \ldots, T_m$ into the data at lag distance $h > 0$ is the inverse transformation: $T = Q^{-1}M$, which uses the same matrix Q as the forward transformation.
Note that, whilst cross-correlation is removed at lag distance $h = h_{max}$, cross-correlation is often indirectly removed at all lag distances $0 < h < h_{max}$.

5.2.4 Spatial interpolation

This section describes the geostatistical models used to carry out spatial interpolation in step 3 of Algorithm 2. The methods considered are ordinary kriging, which is implemented by Barnett and Deutsch [2012] and Barnett et al. [2014] in their versions of the algorithm, and the pair-copula model, which is proposed for the new version of the algorithm to enable modelling of non-linear spatial data.

Ordinary Kriging (OK)

Since the ordinary kriging (OK) model was developed (Matheron [1970]), it has become popular in different spatial fields, such as mining, petroleum, hydrology,

meteorology, oceanography, environmental control, landscape ecology and agriculture (e.g., Cressie [1990], Venäläinen and Heikinheimo [2002], Mishra et al. [2009], Thomson and Emery [2014]). Simply, the OK estimator for the variable $Z(x_0)$ at unsampled location x_0 can be written as a linear combination of nearby samples:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} w_i Z(x_i).$$

The weights w_i are obtained by minimising the error variance σ_R^2 under the constraint $\sum_{i=1}^{n} w_i = 1$ to ensure the unbiased property of the estimator. The error variance σ_R^2 is

$$\begin{aligned}\sigma_R^2 &= Var[\hat{Z}(x_0) - Z(x_0)] \\ &= \sigma^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C_{ij} - 2\sum_{i=1}^{n} w_i C_{i0},\end{aligned}$$

where $\sigma^2 = Var[Z(x)]$, x is any sampled location, $C_{ij} = Cov[Z(x_i), Z(x_j)]$ and $C_{i0} = Cov[Z(x_i), Z(x_0)]$. Hence w_i can be calculated by solving the following system of equations (Isaaks and Srivastava [1989]): $\sum_{j=1}^{n} w_j C_{ij} + \mu = C_{i0}$ for all $i = 1, \ldots, n$ and $\sum_{i=1}^{n} w_i = 1$, where μ is the Lagrange multiplier. The aim of the Lagrange parameter is to obtain weights that produce minimum variance. Moreover, $Cov[Z(x_i), Z(x_j)]$ and $Cov[Z(x_i), Z(x_0)]$ are estimated using variogram modelling. Since the OK method employs the minimum variance concept, $\hat{Z}(x_0)$ is called the "best linear estimator" of the spatial variable of interest at unsampled locations x_0.

Pair-copula based geostatistical interpolation

The variogram and covariance function are the most common methods used to capture the spatial dependence structure of a spatial variable (Kazianka and Pilz [2010a], Gräler and Pebesma [2011]). These methods are only capable of providing one simple average measurement of dependence and also assume linear dependence over the distribution of the variable of interest. However, in reality, the spatial dependence structure may vary over the distribution of the variable

of interest (Journel and Alabert [1989]). Therefore, conventional geostatistical models, such as kriging, which uses the variogram to model spatial dependence, are unable to produce accurate estimators of distributional properties of the variable at unsampled locations when a complex dependence structure is present. Moreover, conventional linear kriging only produces optimal results when the random field is Gaussian. Even though non-linear kriged models, such as indicator kriging, are a solution for non-Gaussian random fields, indicator kriging has a loss in statistical power to detect the true relationship between the variables due to binary transformation of the data. Pair copula-based spatial models overcome these problems of conventional kriged models as they can deal with both non-Gaussian random fields and non-linear dependence structures.

A copula is a function that joins or "couples" a multivariate distribution function to its one-dimensional marginal distribution functions. The term "copula" was first introduced in a mathematical or statistical sense by Sklar [1959]. Following Sklar's theorem, any n-variate distribution function $H(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$ of the vector of random variables $(X_1, X_2, \ldots, X_n)$ with marginal distribution functions $F_i(x_i) = P(X_i \leq x_i)$ can be written as

$$H(x_1, x_2, \ldots, x_n) = C_n(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)),$$

where C_n is an n-dimensional copula. If the marginals $F_i(x_i)$ are continuous, then the copula is unique. Thus, a copula fully describes the dependence structure between random variables.

The main advantage of using a copula-based model for spatial data is that it has the ability to produce the full distribution of the variable of interest at unsampled locations which depend on both configuration of observations and their values (Bárdossy and Li [2008]).Consequently, estimators other than the mean are able to be estimated. However, modelling an n-variate distribution of the unsampled locations and their neighbouring locations requires an n-dimensional spatial copula, and the most readily available copulae in the literature are unable to be extended to higher dimensions. Additionally, some copulae that do have the

ability to be extended to higher dimensions do not provide good parameterisation for the dependence structure to appropriately reflect the spatial configuration of the data points (Bárdossy and Li [2008]). Even though the most popular copulae, such as Gaussian and Student t copulae, fulfil both requirements, these copulae cannot be used to model asymmetric dependence structures.

Unlike high dimensional copulae, bivariate copulae are well understood and readily estimated using maximum likelihood or moment based estimators. Fortunately, an $n-$dimensional copula can be decomposed into a set of $n(n-1)/2$ bivariate copulae using the pair-copula construction described by Aas et al. [2009]. Gräler and Pebesma [2011] adapted the pair-copula model to a spatial framework. The bivariate decomposition of a high-dimensional spatial copula provides a flexible way of using different types of copula families when modelling spatial dependence for different lag distances, and for higher order dependencies as well. However, the pair-copula decomposition is not unique. Each decomposition approximates the full copula density differently. Gräler and Pebesma [2011] used a canonical vine structure (Aas et al. [2009]) to construct a pair-copula for spatial data because this structure benefits spatial interpolation by giving higher priority to the interaction between the unobserved locations and nearby locations.

Interpolation of the spatial data is based on the conditional density of the pair-copula:

$$c_{n+1}(u_0 \mid u_1, \ldots, u_n) = \frac{c_{n+1}(u_0, u_1, \ldots, u_n)}{\int_0^1 c_{n+1}(v, u_1, \ldots, u_n)dv},$$

where n is the number of nearby locations, $u_i = F(Z(x_i))$ for $1 \leq i \leq n$, u_0 denotes the marginal distribution at the unsampled location x_0, and $c_{n+1}(u_0, u_1, \ldots, u_n)$ is the joint multivariate copula of the unsampled location and the nearby locations. The point estimates (mean and median) for the variable of interest at unobserved location x_0 based on n nearby locations can be obtained by calculating

$$\hat{Z}_{mean}(x_0) = \int_0^1 F^{-1}(u) \cdot c_{n+1}(u \mid u_1, \ldots, u_n)du,$$

$$\hat{Z}_{median}(x_0) = F^{-1}(u = C_{n+1}^{-1}(0.5 \mid u_1, \ldots, u_n)),$$

where C_{n+1} is the conditional copula distribution function and F is the estimated distribution function from the observed spatial data. A detailed description of the pair-copula construction of conditional copula densities in a spatial context can be found in Gräler and Pebesma [2011] and Gräler [2014] .

5.3 Data

Algorithm 2 was applied to two data sets: real data from the Bartlett Experimental Forrest (Finley et al. [2007]) and simulated artificial data. Whilst the forest data exhibit multivariate non-linearity, the artificial data were simulated to possess extreme multivariate non-linearity.

5.3.1 Bartlett Experimental Forest data

The real data set used in the application of Algorithm 2 is taken from georeferenced forest inventory plots on the United States Department of Agriculture Forest Service Bartlett Experimental Forest (BEF) in Bartlett, new Hampshire (Finley et al. [2007]). The BEF covers an area of 1,053 hectares. The data set consists of 437 measurements for more than 50 attributes at two dimensional locations $x_i = (x_{1i}, x_{2i}), i = 1, \ldots, 437$. Two attributes, generically labelled Z_1 and Z_2, have been selected to demonstrate the application of Algorithm 2. The extension to higher dimensions is trivial and merely requires additional computation.

Figures 5.2(a) and 5.2(b) show the spatial distribution of Z_1 and Z_2, respectively. It can be seen that Z_1 has a larger variation in attribute values in comparison to Z_2, and low attribute values tend to occur in similar locations for the two variables. The marginal distributions and joint distribution are illustrated in Figures 5.2(c) - 5.2(e), respectively. Strong skewness can be clearly seen in Figures 5.2(c) and 5.2(d), whilst the non-linear structure of the bivariate data can be clearly seen in Figure 5.2(e). Figure 5.3 indicates that there is univariate spatial dependence in Z_1 and Z_2, as well as multivariate spatial dependence between Z_1 and Z_2.

5.3.2 Artificial data

To investigate how well NLPCA can deal with different non-linear structures, artificial two dimensional spatial data were simulated with an extremely non-linear structure. These data comprise 2,304 simulated values.

Figures 5.4(a) and 5.4(b) show the spatial distribution of variables Z_1 and Z_2, respectively. Variable Z_1 has a larger variation in attribute values in comparison to Z_2, and Z_2 generally takes higher attribute values. From Figure 5.4(c), the marginal distribution of Z_1 appears to be approximately uniform, whilst Figure 5.4(d) indicates strong skewness in Z_2. Figure 5.4(e) clearly shows the circular relationship between the two variables. Figure 5.5 indicates some univariate spatial dependence in Z_1 and Z_2, as well as clear multivariate spatial dependence between Z_1 and Z_2.

5.4 Application

Seven versions of Algorithm 2 were implemented. NLPCA or SCT was selected for step 1 (and, consequently, step 5) of Algorithm 2, in which correlation between the bivariate variables is removed at lag distance $h = 0$. PPMT was not implemented in step 1, since, for the data sets considered in this chapter, PPMT removed all spatial properties of the original data. The second step of MAF was implemented in step 2 of Algorithm 2 (and, consequently, step 4) if spatial cross-correlation persisted at lag distance $h > 0$. The estimates from spatial interpolation, in step 3 of Algorithm 2, were obtained from kriging, the mean estimate from the pair-copula, or the median estimate from the pair-copula. Various combinations of these methods within Algorithm 2 resulted in seven competing models, which are summarised in Table 5.1.

Model 5 is comparable to the implementation of Algorithm 2 proposed by Barnett and Deutsch [2012]. The remaining models are newly proposed models. Models 3 and 4 are proposed as the models of preference for modelling non-linear multivariate spatial data, since pair-copulas are able to model spatial non-linearity

Table 5.1: Competing models for modelling non-linear multivariate spatial data.

Model	Transformation	Spatial Interpolation
1	NLPCA	Kriging
2	NLPCA+MAF	Kriging
3	NLPCA+MAF	Pair-copula mean
4	NLPCA+MAF	Pair-copula median
5	SCT	Kriging
6	SCT	Pair-copula mean
7	SCT	Pair-copula median

and NLPCA is able to model multivariate non-linearity whilst retaining spatial non-linearity in the transformed factors.

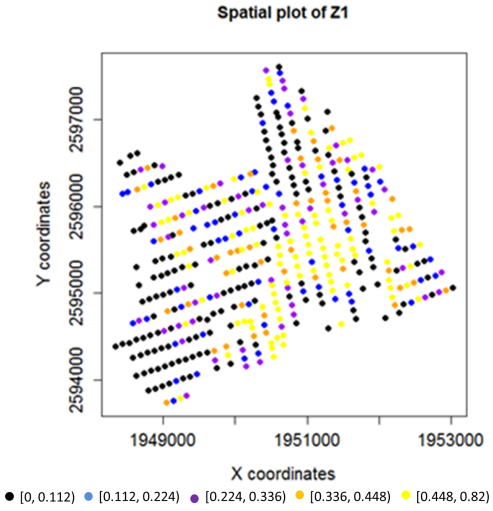
Leave-one-out cross-validation was used, with ten nearby locations in the interpolation process, to facilitate comparison of the models. The performance of the models was assessed based on reproduction of univariate and bivariate distributions. Reproduction of the univariate distributions was evaluated using the mean absolute error between estimated and original values (MAE), bias (average difference between estimated and original variables) and Pearson correlation coefficient of the original and estimated values. The absolute correlation error was used to evaluate the reproduction of bivariate relationships for the BEF data. This statistic is calculated by taking the absolute difference between the Spearmann correlation of the original data and associated estimates. Since the artificial data were simulated to have a circular structure, the circular correlation coefficient (Jammalamadaka and Sengupta [2001]) was used to assess the reproduction of the bivariate distribution. Similar to the Pearson correlation coefficient, the circular correlation coefficient takes values between -1 and 1, where higher values for this statistic represent better reproduction of the bivariate distribution.

5.4.1 Bartlett Experimental Forest data

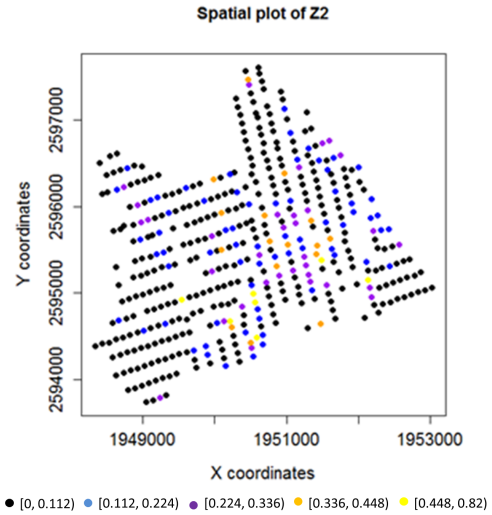
For NLPCA, the AANN mapping is based on a $2-6-2-6-2$ network (two input and output variables, two components, and six non-linear nodes in each mapping and inverse mapping layer). The two components extracted by the network for the BEF data are shown in Figure 5.6. Component A captures the non-linear structure of the data whilst component B captures the random variation of the data.

Figure 5.7(a) shows the scatterplot of the two extracted components using NLPCA. It is clear that the non-linear structure of the data was successfully removed by NLPCA in comparison to Figure 5.2(e). Whilst the cross-correlation between variables was removed at zero lag distance, and indirectly at large lag distances, a small negative cross-correlation remained at lag distances up to 400 km. The second step of the MAF transformation was subsequently carried out to remove spatial cross-correlation at lag $h = 400$ km, and indirectly at all lags up to 400 km. Figure 5.7(b) indicates that cross-correlation was removed at all lag distances after transformation with NLPCA followed by the second step of MAF.

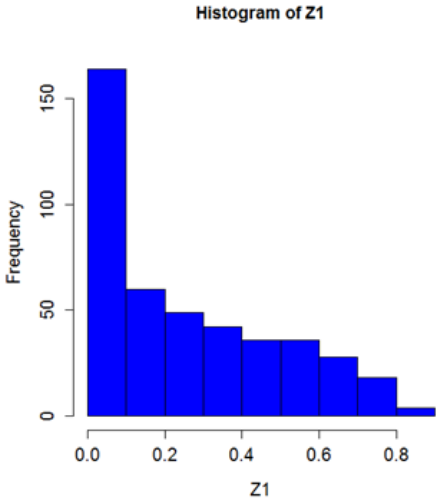
In Figure 5.7(c), the scatterplot of the transformed variables using SCT indicates no correlation at lag $h = 0$. The somewhat structured pattern that appears in Figure 5.7(c) is typical of the SCT method. Figure 5.7(d) indicates indirect removal of almost all spatial cross-correlation at all lag distances after applying SCT. Thus, there was no need to perform the second step of MAF following SCT. Univariate and bivariate statistics for models 2-7 fitted to the BEF data are presented in Table 5.2. Figure 5.8 displays the scatterplots of the estimated Z_1 values against the estimated Z_2 values for models 2-7.



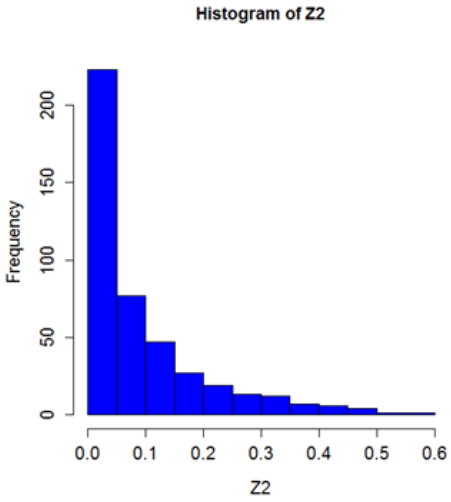
(a)



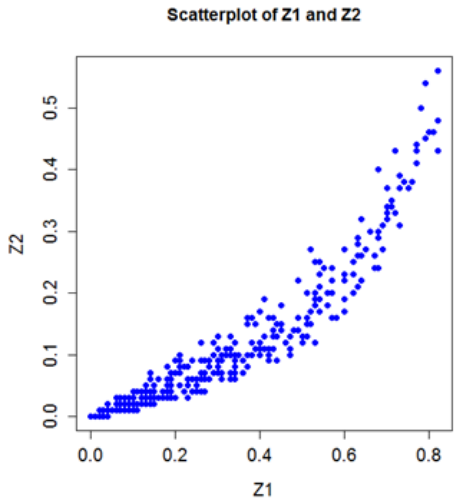
(b)



(c)



(d)



(e)

Figure 5.2: Data from Bartlett Experimental Forest – spatial distributions for (a) Z_1 and (b) Z_2, histograms for (c) Z_1 and (d) Z_2, and (e) scatterplot between Z_1 and Z_2.

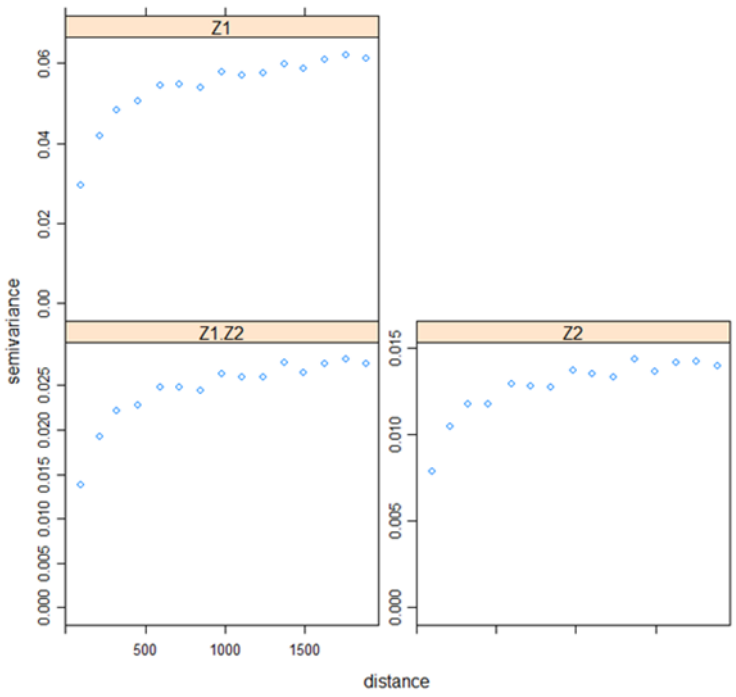


Figure 5.3: Semi-variograms and cross-variogram for variables Z_1 and Z_2 from the Bartlett Experimental Forest data.

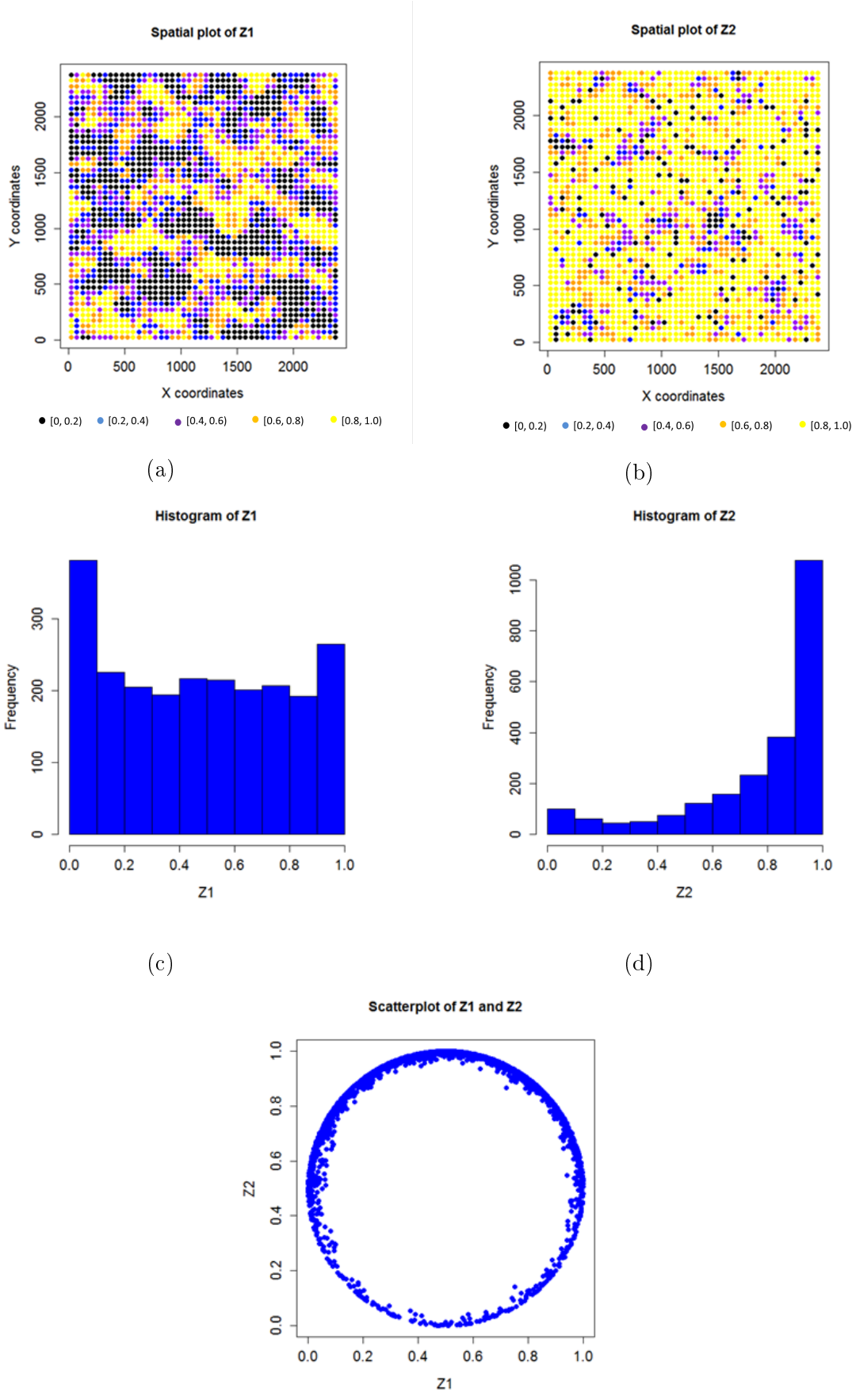


Figure 5.4: Artificial data – spatial distributions for (a) Z_1 and (b) Z_2, histograms for (c) Z_1 and (d) Z_2, and (e) scatterplot between Z_1 and Z_2.

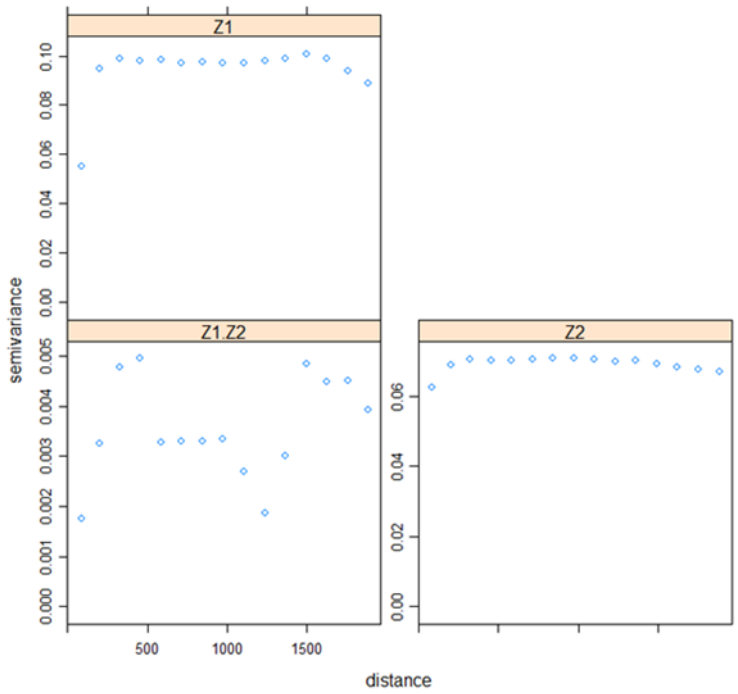


Figure 5.5: Semi-variograms and cross-variogram for variables Z_1 and Z_2 from the simulated artificial data set.

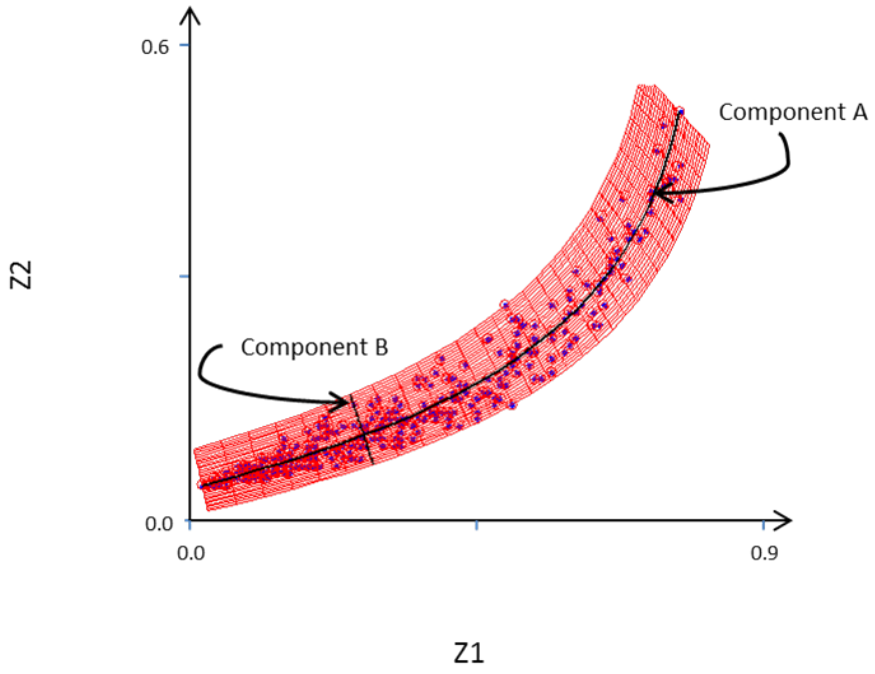


Figure 5.6: The two structures identified by the AANN for the BEF data. Solid dots represent the observed data and the curved line represents the non-linear structure present in the data.

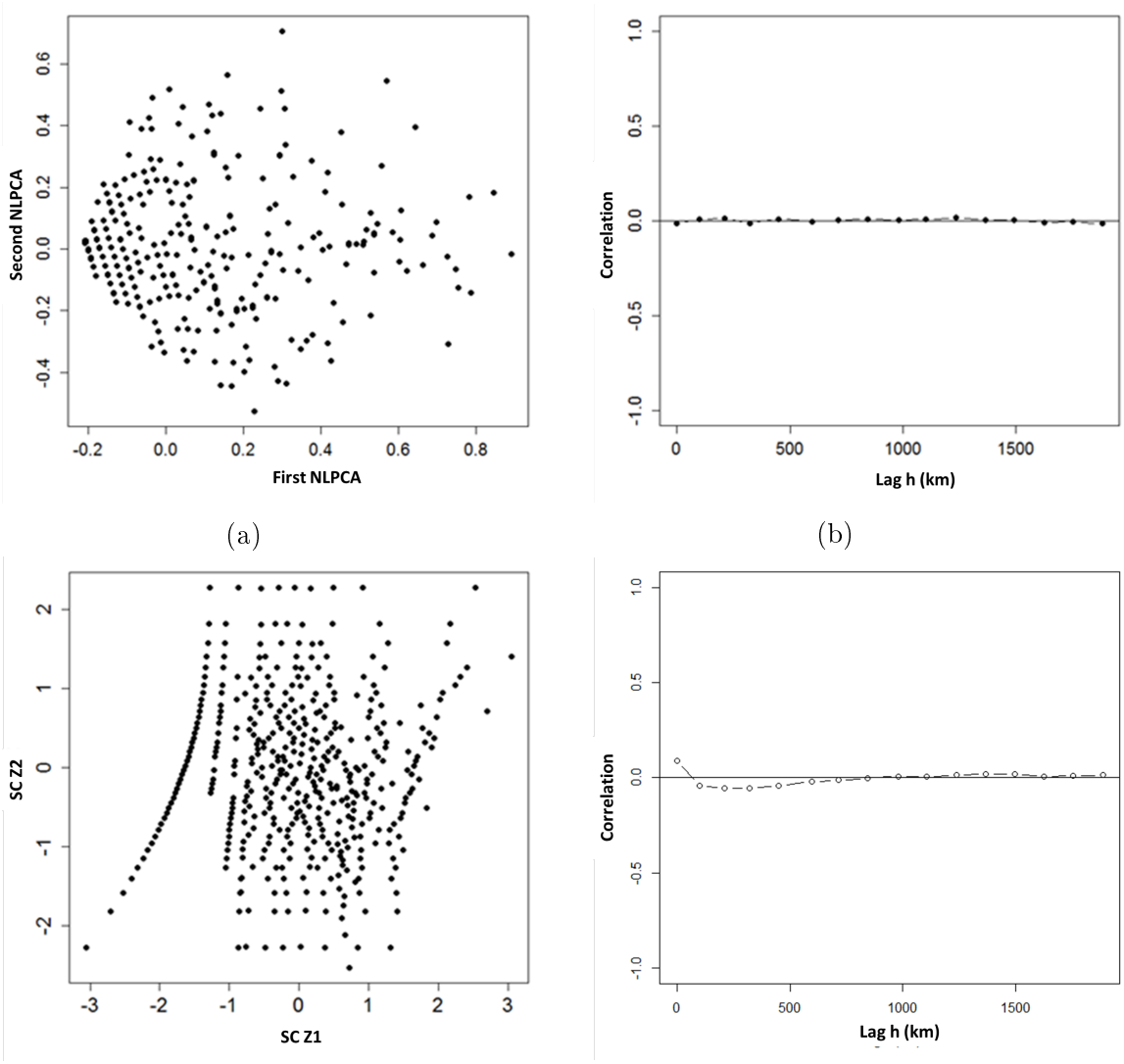


Figure 5.7: Bartlett Experimental Forest data – (a) scatterplot of extracted components from NLPCA, (b) correlogram of transformed variables from NLPCA+MAF, (c) scatterplot of transformed variables from SCT and (d) correlogram of transformed variables from SCT.

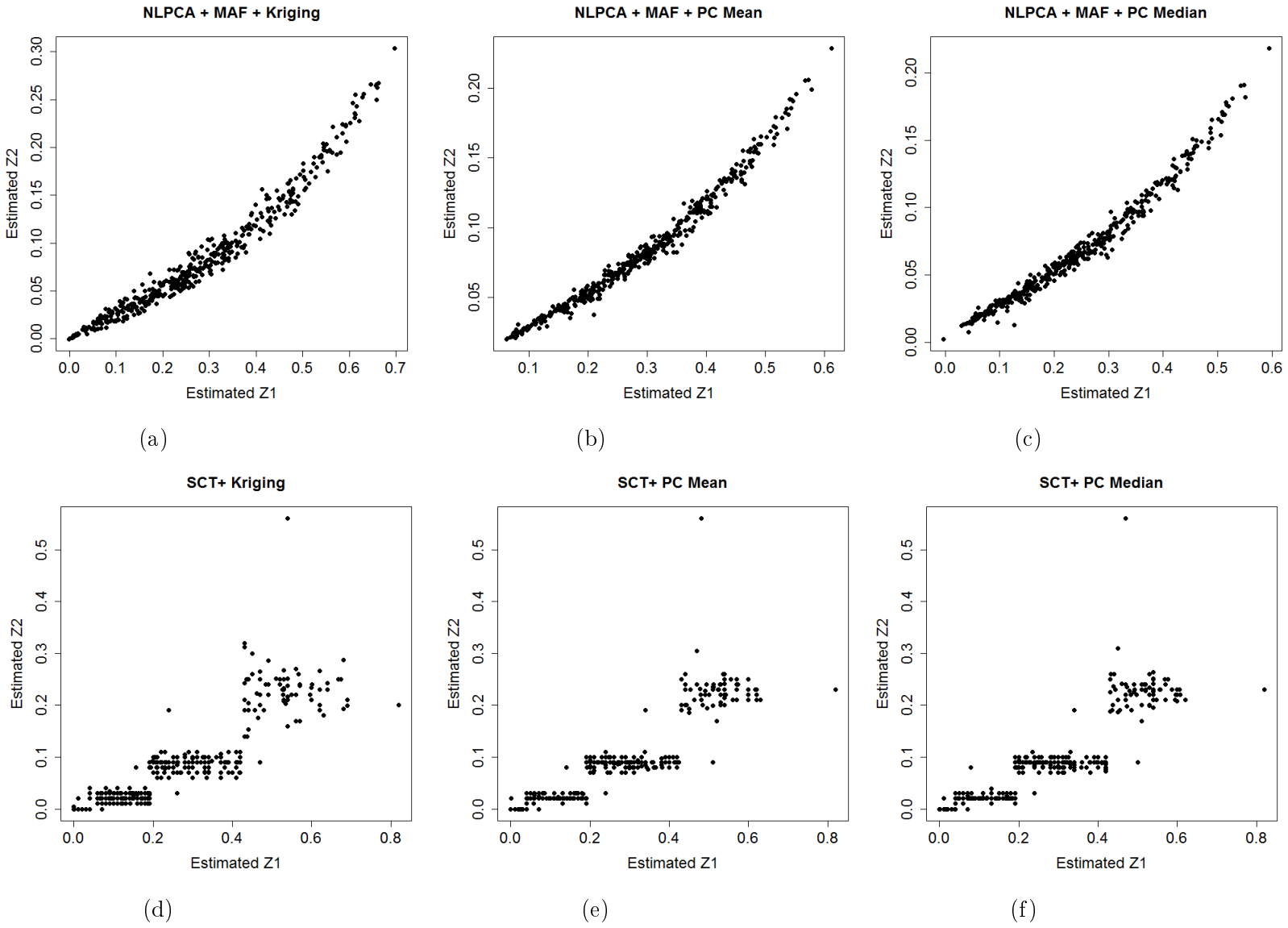


Figure 5.8: Reproduction of non-linear multivariate structure for Bartlett Experimental Forest data. Figures (a)-(f) are the estimated values for Z_1 versus estimated values for Z_2 for models 2-7, respectively.

The primary focus in this chapter is modelling multivariate non-linearity in spatial data. Figure 5.8 clearly demonstrates that models using NLPCA (Figures 5.8(a) - 5.8(c)) reproduce the non-linear bivariate structure of the original variables more successfully than models that use SCT (Figures 5.8(d) - 5.8(f)), irrespective of the interpolation method. Quantitatively, this can be confirmed by the absolute correlation error in Table 5.2, all of which are smaller for NLPCA compared to SCT.

Table 5.2 also indicates that the correlation between the original and estimated values was higher for NLPCA based models than SCT based models for variable Z_2, and perhaps slightly higher for NLPCA than SCT for Z_1. Whilst the multivariate modelling approaches do not focus on bias or MAE, the lowest bias, for both Z_1 and Z_2, was from NLPCA based models compared to SCT. MAE was similar between NLPCA and SCT for Z_2, and slightly worse for NLPCA for Z_1. Note that, whilst the correlation for Z_2 was worse than Z_1 for both NLPCA and SCT, the difference is exaggerated for SCT. In comparing interpolation methods, within the NLPCA based models, the pair-copula median (PC-Median) produced the best univariate results, compared with kriging, for all statistics except bias for Z_2. For SCT based models, the pair-copula mean (PC-Mean) produced the best results for all univariate statistics for both Z_1 and Z_2.

Table 5.2: Goodness of fit statistics for the BEF data, measuring the accuracy in reproduction of univariate and bivariate distributions.

Model	Transform	Interpolation	Z_1			Z_2			Abs. Corr. Error
			MAE	Bias	Corr.	MAE	Bias	Corr.	
2		Kriging	0.158	0.025	0.573	0.068	−0.006	0.519	0.006
3	NPLCA	PC-Mean	0.160	0.029	0.600	0.068	−0.009	0.550	0.013
4	+ MAF	PC-Median	0.152	−0.019	0.600	0.066	−0.024	0.553	0.012
5		Kriging	0.147	−0.025	0.583	0.070	−0.016	0.466	0.068
6	SCT	PC-Mean	0.145	−0.024	0.593	0.067	−0.015	0.494	0.063
7		PC-Median	0.145	−0.026	0.592	0.067	−0.016	0.488	0.069

5.4.2 Artificial data

For the artificial data, since the data have a circular structure, a circular AAAN architecture was used in the NLPCA. The two solid lines in Figure 5.9 show the

two components obtained from the NLPCA. Component A captures the circular structure while component B represents random variation in the data.

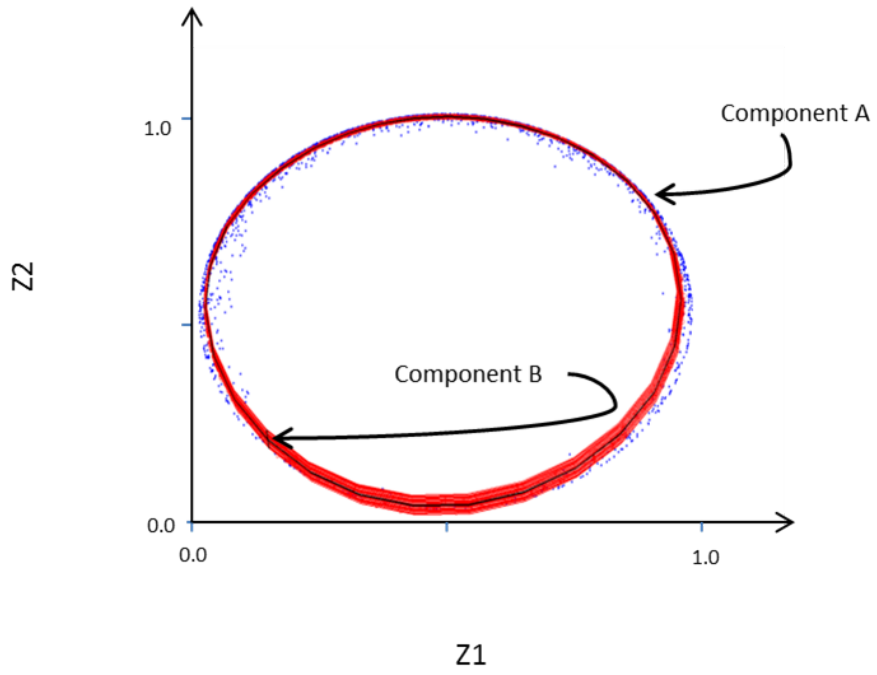


Figure 5.9: The two structures identified by the AANN for the artificial data. Solid dots represent the observed data and the circular line represents the circular structure present in the data.

Figures 5.10(a) and 5.10(c) are scatterplots of the two extracted components using NLPCA and SCT, respectively. The zero correlation structure between NLPCA transformed variables (extracted components) and SCT transformed variables is evident. The correlograms in Figure 5.10(b) and 5.10(d) indicate the removal of almost all cross-correlation at all lag distances for both NLPCA and SCT. Hence, the second step of MAF was not required following NLPCA or SCT.

Since the artificial data set was generated via a Gaussian random field, kriging interpolation will, generally, outperform interpolation based on pair-copulas. Consequently, only kriging was considered for spatial interpolation of the artificial data. Univariate and bivariate statistics for models 1 and 5 fitted to the artificial data are presented in Table 5.3. Figure 5.11 displays the scatterplots of the estimated Z_1 values against the estimated Z_2 values for models 1 and 5.

With regards to bivariate goodness of fit, Figure 5.11 clearly demonstrates that the NLPCA based model (Figure 5.11(a)) reproduced the bivariate structure of

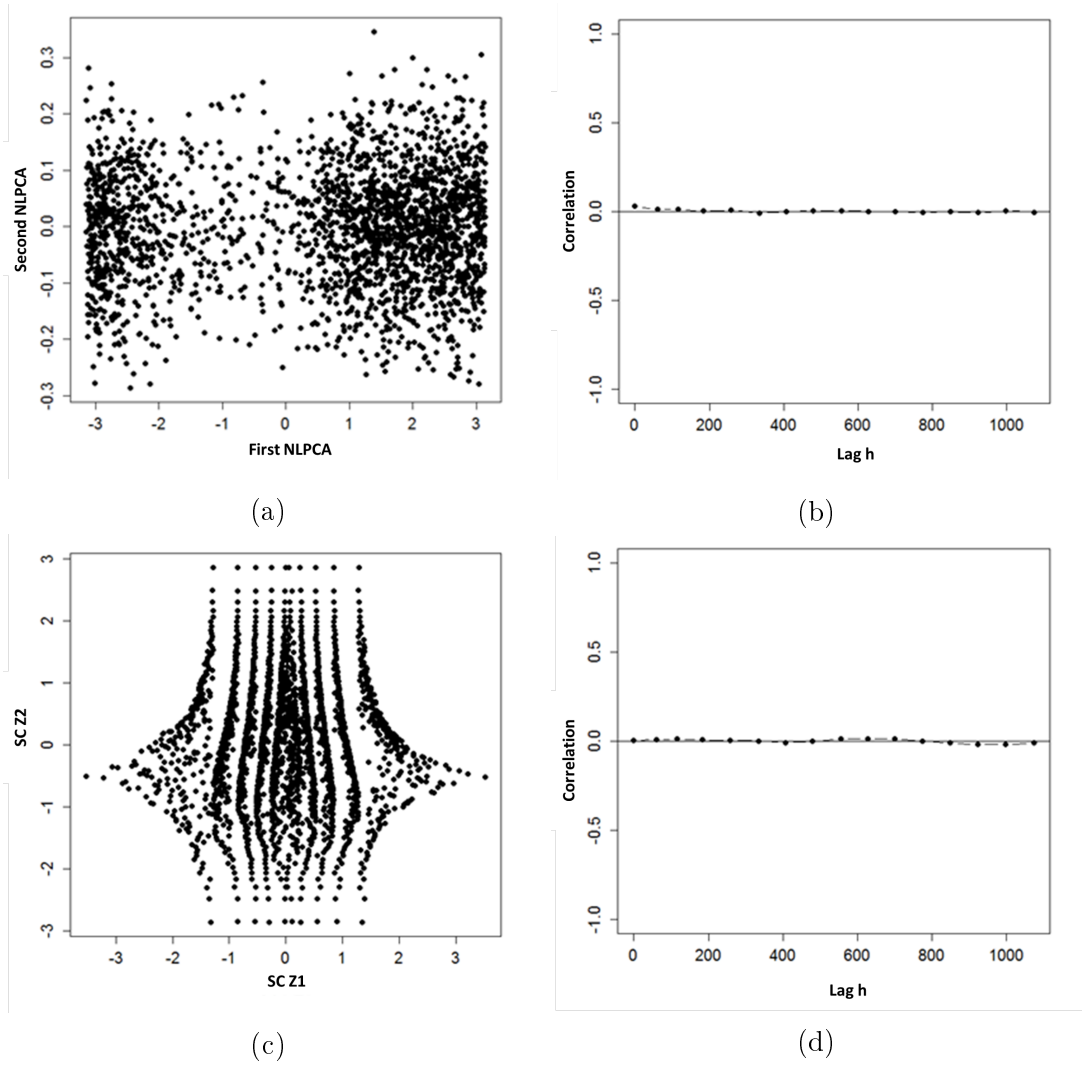


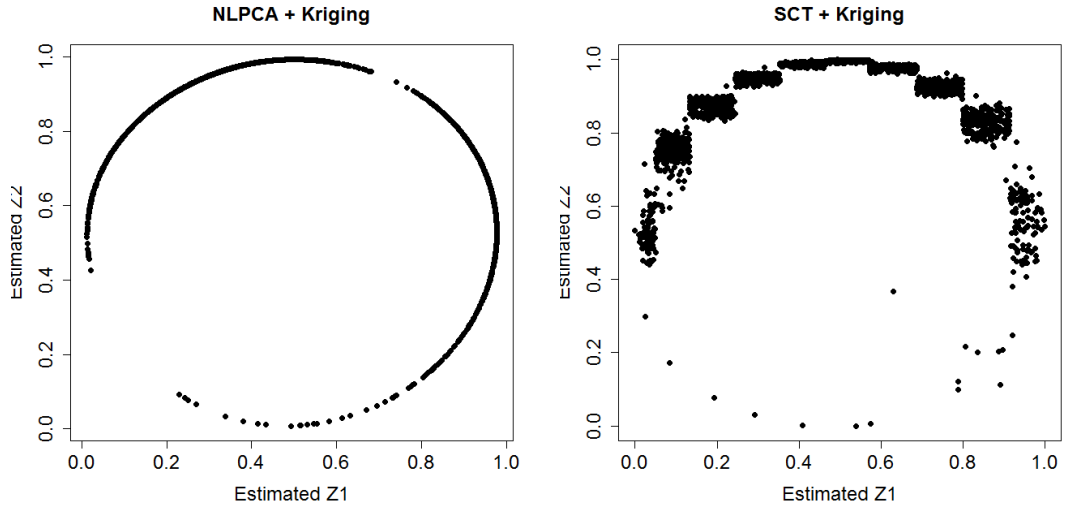
Figure 5.10: Artificial data – (a) scatterplot of extracted components from NLPCA, (b) correlogram of transformed variables from NLPCA, (c) scatterplot of transformed variables from SCT and (d) correlogram of transformed variables from SCT.

the original variables more successfully than the SCT based model (5.11(b)). This is confirmed by the larger estimate of circular correlation for NLPCA compared to SCT in Table 5.3.

In terms of univariate goodness of fit, NLPCA produced better univariate statistics (lower MAE and bias, and higher correlation) for Z_2. For Z_1, NLPC performed on par with SCT in all measures, except for bias, which was better for SCT. As with the BEF data, the correlation for Z_2 was worse than Z_1 for both NLPCA and SCT, but more so for SCT.

Table 5.3: Goodness of fit statistics for the artificial data, measuring the accuracy in reproduction of univariate and bivariate distributions

Model	Transform	Interpolation	Z_1			Z_2			Circ. Corr.
			MAE	Bias	Corr.	MAE	Bias	Corr.	
1	NLPCA	Kriging	0.134	0.021	0.832	0.142	0.032	0.569	0.853
5	SCT	Kriging	0.133	−0.004	0.835	0.157	0.100	0.368	0.677



(a) (b)

Figure 5.11: Reproduction of non-linear multivariate structure for artificial data. Figures (a) and (b) are the estimated values for Z_1 versus estimated values for Z_2 for models 1 and 5, respectively.

5.5 Discussion

Based on the two case studies, NLPCA transformation reproduced the non-linear bivariate structure of the data more successfully than the SCT transformation. For the univariate goodness of fit statistics, NLPCA performed on par with STC, if not better than STC, in the majority of models, in terms of the correlation between the original and estimated values. The lowest bias occurred in all NLPCA-based models compared to STC based models for both variables, except for Z_1 in the artificial data, where the lowest bias was for the SCT-based model. MAE was similar, or smaller, for variable Z_2 in the NLPCA-based models compared to STC. For Z_1, NLPCA and STC produced similar MAEs in the artificial data, but a slightly worse MAE in the BEF data.

In both case studies the correlation for Z_2 was worse than Z_1 for both NLPCA and SCT, with the difference being larger for SCT. This reflects the poor re-

production of the variable transformed second using SCT. In both case studies, variable $Z_1$was transformed first, then variable Z_2. In general, for SCT, the quality of the reproduction of univariate distributions declines for variables as their position in the order of transformation numerically increases. This is known as the "effect of ordering" in the literature (Leuangthong [2003]). That is, the variable transformed second will be worse than the variable transformed first, the variable transformed third will be worse than the variable transformed second, and so on. This ordering effect can be seen in Figures 5.8(d) - 5.8(f) for the BEF data and Figure 5.11(b) for the artificial data.

The interpolation methods cannot be compared between the case studies, since the artificial data was interpolated using kriging only. However, for the BEF data, the pair-copula model produced the best univariate results within the NLPCA based models and also within the SCT based models, compared to kriging, with the exception of the bias for Z_2 in the NLPCA based model.

5.6 Conclusions

Based on the results of the two case studies, NLPCA, followed by the second step of MAF, if required, and pair-copula based spatial interpolation is the recommended implementation of Algorithm 2 for modelling data that are both multivariately non-linear and spatially non-linear. The results demonstrate that NLPCA, in combination with MAF, when required, is effective in facilitating the modelling of non-linear multivariate spatial data, even in the presence of extreme multivariate non-linearity. In the case studies, NLPCA reproduced the bivariate distributions of the original data better than SCT, for all models considered.

The extent to which NLPCA can handle heteroscedasticity in non-linear data was not investigated, and this remains an open problem. We conjecture that NLPCA is not only able to capture non-linear structures among continuous spatial variables, but also among spatial variables with mixed types, such nominal and rank data. Extension of NLPCA to these types of variables will be considered in future research.

The results also indicate that, for the BEF data, the pair-copula model, generally, outperforms conventional kriging in terms of univariate goodness of fit statistics, regardless of the transformation method. This is most likely due to the ability of the pair-copula to more accurately reproduce tails of skewed distributions compared to kriging, which, being Gaussian-based, fails to capture asymmetric or heavy tails.

Further improvements to the pair-copula model are expected to be gained through, for example, development of an efficient method for defining lag distance classes, use of advanced search strategies (e.g., quadrant search to remove obvious cluster effects), and applications of wider classes of copulas. These developments can be incorporated , where pair-copulas are used in the non-linear multivariate modelling approach considered in this chapter.

Chapter 6

Univariate Optimal Spatial Design

The research in this chapter has been submitted to Geoderma for journal submission as detailed below.

- Musafer, G.N. and Thompson, M.H. (n.d). Pair-copula based optimal spatial design for additional samples. *Geoderma. Submitted.*

Abstract

A spatial sampling design that uses pair-copulas is presented that aims to reduce prediction uncertainly by selecting additional sampling locations based on both the spatial configuration of existing locations and the values of the observations at those locations. The novelty of the approach arises in the use of pair-copulas to estimate uncertainty at unsampled locations. Spatial pair-copulas are able to more accurately capture spatial dependence compared to other types of spatial copula models. Additionally, unlike traditional kriging variance, uncertainty estimates from the pair-copula account for influence from measurement values and not just the configuration of observations. This feature is beneficial, for example, for more accurate identification of soil contamination zones where high contamination measurements are located near measurements of varying contamination. The proposed design methodology is applied to a soil contamination example from the Swiss Jura region. A partial redesign of the original sampling configuration demonstrates the potential of the proposed methodology.

6.1 Introduction

The focus of this chapter is the development of a new optimal spatial design for additional sample locations using spatial pair-copulas in order to reduce uncertainty in spatial prediction that takes into account the configuration of observations and their measured values. The spatial variable is considered as a random field and the spatial dependence may be non-linear and non-Gaussian. Optimal design concepts are applied to determine the collection of additional samples in order to balance the benefit between additional information and reduction in prediction uncertainty. The optimal design will vary according to the scientific goal, such as parameter estimation of the model (Webster and Oliver, 1992, Lark, 2002, Zimmerman, 2006) and prediction using the geostatistical model (Zimmerman, 2006, Zhu and Stein, 2006, Diggle and Lophaven, 2006, Diggle and Ribeiro, 2007b). If prediction of the random field is the aim, then optimality of the sampling design is evaluated based on the maximum or average estimation of uncertainty of the predicted locations (Diggle and Ribeiro, 2007b).

The estimation of prediction uncertainty should be able to capture all types of variability present in the spatial random field. Variability occurs from the configuration of the data and variability in the measured values. In the literature, the majority of optimal spatial designs (e.g., Cressie, 1993, Journel, 1994, Van Groenigen et al., 1999, Zimmerman, 2006, Emery et al., 2008) aim to minimise the kriging variance. However, kriging variance is only dependent on the spatial configuration of observation locations and does not depend on the values of the observations under Gaussian assumption. The consequences of developing an optimal design that ignores the variability in measured values in an extreme spatial scenario is discussed in Chang et al. [2007]. Some spatial designs have attempted to capture the variability of sampled values in the uncertainty measurement by using conditional simulation (Pilger et al., 2001, Koppe et al., 2011). However, conditional simulation only uses one possible value for the additional samples from an infinite number of outcomes and so is unable to produce full

uncertainty estimation.

The need to incorporate the variability in measurement values into an optimal spatial sampling design motivates the use of copula based geostatistical models. Spatial copula models are capable of producing uncertainty estimation that is dependent on both the observations' configuration and values (e.g., Bárdossy, 2006, Bárdossy and Li, 2008, Haslauer et al., 2010, Gräler and Pebesma, 2011, Gräler, 2014). Moreover, spatial pair-copula models (Gräler and Pebesma, 2011, Gräler, 2014) are more able to accurately capture non-linear spatial dependence compared to less flexible copula based models (e.g, Bárdossy, 2006). This is because pair-copula models allow the use of different copula families when modelling spatial dependence for different separating vectors and for higher order dependencies whilst less flexible copula based models assume the same copula family for all separating vectors and for higher order dependencies.

Li et al. [2011] developed an observation network design based on a spatial copula model with the objective of maximising the expected gain defined by a utility function. The utility function constrains selection of the additional locations by taking into account estimation uncertainty, a critical threshold value that defines water quality and the gain-loss in the decision to sample or not. The research developed in this chapter builds on Li et al. [2011] through use of pair-copulas, rather than less flexible copula models, and considers unconstrained sampling design for the additional locations. Additional sampling locations are selected from those locations that produce the highest estimate of prediction uncertainty over the sampling region.

The proposed methodology is presented in Section 6.2. Section 6.3 provides a description of the two-dimensional Swiss Jura data set (Goovaerts, 1997). In Section 6.4, the proposed design methodology is applied to the Swiss Jura data set where the potential of the proposed method is demonstrated through a partial redesign of the existing sampling design for the Swiss Jura data. In addition, a design based on kriged model under Gaussian assumption and the design obtained from the proposed methodology are compared. Concluding remarks and future

research are discussed in Section 6.5.

6.2 Methodology

Let $Z(\boldsymbol{x})$ denote a univariate spatial random field where $\boldsymbol{x}$ is a two dimensional location belonging to the study domain $\mathcal{X}$. The set of existing sampled locations is denoted by $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. The objective of optimal spatial sampling here is to select additional measurement locations that reduce uncertainty over the random field. One such example arises in additional drill core sampling in which additional measurements are desired to reduce the uncertainty in the spatial distribution of an ore reserve. Of particular focus in this chapter is the reduction in the predictive quantile interval (PQI), that is, the difference between predictive quantiles of $Z(\boldsymbol{x})$. Let $X' = (\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m)$ be a set of candidate locations from which the additional new locations are chosen and $X' \subseteq \mathcal{X}$. The PQI at unsampled locations $X^* = (\boldsymbol{x}^*_1, \ldots, \boldsymbol{x}^*_N)$, can be estimated from the sampled observations by interpolation of the random field $Z(\boldsymbol{x})$. Note that, in practice, the study domain $\mathcal{X}$ is discretised into an interpolation grid so that the set of unsampled locations X^* are the nodes of the interpolation grid. The interpolation method of Gräler and Pebesma [2011], which uses spatial pair-copulas, is applied here. The PQI corresponding to the difference between the 95-th and 5-th predictive quantiles of $Z(\boldsymbol{x})$ at unsampled location $\boldsymbol{x}^*_j$, given the existing sampled observations, is

$$\begin{aligned} PQI(u^*_j|u_1, \ldots, u_n) &= F_Z^{-1}\left(C^{-1}_{\boldsymbol{x}^*_j,n}(0.95|u_1, \ldots, u_n)\right) \\ &\quad - F_Z^{-1}\left(C^{-1}_{\boldsymbol{x}^*_j,n}(0.05|u_1, \ldots, u_n)\right) \end{aligned}$$

where $C_{\boldsymbol{x}^*_j,n}(u^*_j|u_1, \ldots, u_n)$ is the conditional copula at unsampled location $\boldsymbol{x}^*_j$, conditioned on the n existing sampled observations. Note that u^*_j denotes the value of the Uniform random variable U^* (on $[0, 1]$) at the unsampled location $\boldsymbol{x}^*_j$, while $u_i = F_Z(z(\boldsymbol{x}_i))$ with F_Z denoting the estimated marginal cumulative

distribution function of the data.

The candidate location $\boldsymbol{x}'_i$, $i = 1, \ldots, m$, from the set X', selected as the new additional measurement location, is that which corresponds to the smallest total expected PQI summed over the study domain after it has been added to the existing sampled observations. Since pair-copulas are used in the spatial interpolation process, selection of an additional measurement location depends not only on the spatial location of the existing sampled observations and the spatial location of the new candidate but also on the values of the spatial variable at these locations. A pair-copula model that appropriately describes the spatial dependence has to be selected. In doing so, the values of the spatial variable $Z(\boldsymbol{x})$ have to be transformed to the probability space $[0, 1]$ using the estimated distribution function F_Z.

The conditional copula at the candidate location $\boldsymbol{x}'_i$, conditioned on the existing sampled observations, is

$$C_{\boldsymbol{x}'_i,n} \;=\; C_{\boldsymbol{x}'_i,n}(u'_i|u_1, \ldots, u_n) \tag{6.1}$$

where u'_i denotes the value of the Uniform random variable U' (on $[0, 1]$) at the candidate location $\boldsymbol{x}'_i$.

After adding a candidate to the set of existing observations, the values on the interpolation grid can be re-estimated. For any possible value u'_i of U' at the candidate location $\boldsymbol{x}'_i$, the conditional copula at unsampled interpolation location $\boldsymbol{x}^*_j$, conditioned on the existing sampled observations and the newly added candidate $\boldsymbol{x}'_i$, is

$$C_{\boldsymbol{x}^*_j,n+1} \;=\; C_{\boldsymbol{x}^*_j,n+1}(u^*_j|u'_i, u_1, \ldots, u_n) \tag{6.2}$$

where u^*_j is any possible value of Uniform random variable U^* at $\boldsymbol{x}^*_j$.

Using Eq. (6.2), any uncertainty measure, such as variance, coefficient of variation, interquartile range and PQI, can be estimated at all points on the interpolation grid after adding the candidate $\boldsymbol{x}'_i$. The PQI corresponding to the difference

between the 95-th and 5-th predictive quantiles at unsampled interpolation location $\boldsymbol{x}_j^*$ after adding the candidate $\boldsymbol{x}_i'$ with a proposed value u_i' as the assumed observed value is

$$\begin{aligned} PQI(u_j^*|u_i', u_1, \ldots, u_n) &= F_Z^{-1}\left(C^{-1}_{\boldsymbol{x}_j^*, n+1}(0.95|u_i', u_1, \ldots, u_n)\right) \\ &\quad - F_Z^{-1}\left(C^{-1}_{\boldsymbol{x}_j^*, n+1}(0.05|u_i', u_1, \ldots, u_n)\right). \end{aligned} \tag{6.3}$$

This is the PQI at $\boldsymbol{x}_j^*$ for one possible value of u_i'. The expected PQI at $\boldsymbol{x}_j^*$ is calculated as the integral of the PQI in Eq. (6.3) over the entire range of possible values of u_i' corresponding to candidate location $\boldsymbol{x}_i'$:

$$E\left[PQI(u_j^*|u_i', u_1, \ldots, u_n)\right] = \int_0^1 PQI(u_j^*|u_i', u_1, \ldots, u_n) dC_{\boldsymbol{x}_i', n} \tag{6.4}$$

where $C_{\boldsymbol{x}_i', n}$ is the conditional copula given in Eq. (6.1).

The total expected PQI of the entire interpolation grid after adding the candidate $\boldsymbol{x}_i'$ as a new observation is then the sum of the expected PQI at all interpolation points:

$$E_T(\boldsymbol{x}_i') = \sum_{j=1}^{N} \left(\int_0^1 PQI(u_j^*|u_i', u_1, \ldots, u_n) dC_{\boldsymbol{x}_i', n} \right). \tag{6.5}$$

Computational efficiency can be gained by interchanging the summation and integration in Eq. (6.5):

$$E_T(\boldsymbol{x}_i') = \int_0^1 \left(\sum_{j=1}^{N} PQI(u_j^*|u_i', u_1, \ldots, u_n) \right) dC_{\boldsymbol{x}_i', n}.$$

The candidate $\boldsymbol{x}_i'$ that produces the smallest total expected PQI is selected as the new sample location. Alternatively, minmax approach can be used here instead of averaging PQI over the study domain. A summary of the procedure is outlined in the following steps.

1. Transform the observations $z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_n)$ to the unit interval $[0, 1]$ us-

ing the estimated distribution function F_Z: $u_1 = F(Z(\boldsymbol{x}_1)), \ldots, u_n = F(Z(\boldsymbol{x}_n))$.

2. Use the transformed observations $u_1, \ldots, u_n$ to fit a spatial pair-copula $C_{\boldsymbol{x},n}(u|u_1, \ldots, u_n)$ using the method of Gräler and Pebesma [2011].

3. For each candidate location $\boldsymbol{x}'_i$, for all values of u'_i, calculate the conditional copula density $c_{\boldsymbol{x}'_i,n} = c_{\boldsymbol{x}'_i,n}(u'_i|u_1, \ldots, u_n)$ (Gräler and Pebesma, 2011). In practice, it is not possible to obtain the conditional copula density for all possible values of u'_i, hence the range of values of U', i.e., $[0, 1]$, is discretised and the conditional copula density is calculated for the midpoint of each interval.

4. For each interpolation grid point $\boldsymbol{x}^*_j$, calculate the conditional copula $C_{\boldsymbol{x}^*_j,n+1} = C_{\boldsymbol{x}^*_j,n+1}(u^*_j|u'_i, u_1, \ldots, u_n)$, conditioned on the existing observations $u_1, \ldots, u_n$ and the proposed value u'_i at the candidate location $\boldsymbol{x}'_i$. Use this conditional copula to calculate the predictive quantile interval $PQI(u^*_j|u'_i, u_1, \ldots, u_n)$ given in Eq. (6.3). Calculation of the conditional copula and, consequently, the predictive quantile interval is repeated for all discretised values of U' at the candidate location $\boldsymbol{x}'_i$.

5. For each interpolation grid point $\boldsymbol{x}^*_j$, calculate the expected PQI using Eq. (6.4). The integral in Eq. (6.4) can be approximated by

$$\int_0^1 PQI(u^*_j|u'_i, u_1, \ldots, u_n) dC_{\boldsymbol{x}'_i,n} = \sum_{l=1}^{M} PQI(u^*_j|u'_i = u'_{i,l}, u_1, \ldots, u_n) c_{\boldsymbol{x}'_i,n}(u'_i = u'_{i,l}|u_1, \ldots, u_n) \Delta u'_{i,l}$$

where $u'_{i,l}$ is the midpoint of the l-th discretised interval of U', $c_{\boldsymbol{x}'_i,n}(u'_i = u'_{i,l}|u_1, \ldots, u_n)$ is the conditional copula density calculated in step 3 at $u'_i = u'_{i,l}$ and $\Delta u'_{i,l}$ is the width of the l-th discretised interval.

6. For the candidate location $\boldsymbol{x}'_i$, calculate the total expected PQI of the entire interpolation grid using Eq. (6.5) by summing up the expected PQI calculated for all the interpolation grid points.

7. Repeat steps 2 to 6 for the remaining candidate points and select the candidate point that produces the smallest total expected PQI, $E_T(\boldsymbol{x}'_i)$, as the new sample location.

Note that, in step 2, whilst the spatial pair-copula of Gräler and Pebesma [2011] is used, alternative spatial copulas could be substituted into the procedure, such as the spatial copula of Bárdossy and Li [2008]. Moreover, minimisation of the maximum PQI over the study domain can be used as an alternative for using average over the study domain.

Additionally, there are some practical issues that require consideration in implementing the proposed design methodology. Firstly, the transformation applied in step 1 and the spatial copula fitted in step 2 are important, since it is assumed that the dependence of the random variable Z follows the selected copula model. For further details on interpolation using spatial pair-copulas, see Gräler and Pebesma [2011] and Gräler [2014]. For a more practical perspective, Musafer et al. [2015] provide detailed instructions on the steps involved in fitting, and interpolating from, a spatial pair-copula.

Secondly, the range of values of U' should be appropriately discretised in step 3 to provide a reasonable numerical approximation of the expected PQI calculated in step 5. Li et al. [2011] suggest a simple approximation of the expected PQI using a division of the $[0, 1]$ interval into hundreds of equally spaced intervals. For equally spaced intervals, if the width of the intervals is not sufficiently small, approximation of the expected PQI may be poor, consequently resulting in sub-optimal selection of additional sampling locations. Narrower intervals will result in more accurate approximation of the expected PQI, but at an increased cost in computational time. Even though more sophisticated deterministic quadratic scheme can be used for the integration, numerical approximation of the expected PQI, Monte Carlo integration (Shapiro, 2003) is used in step 5 for more computationally efficient .Consequently, the intervals of the discretised range of U' in step 3 are determined by Monte Carlo sampling and are not necessarily equally spaced. Thus, the expected PQI using Eq. (6.4) can be approximated using Monte

Carlo intergration as follows. For M Monte Carlo samples of U' intergration

$$\int_0^1 PQI(u_j^*|u_i', u_1, \ldots, u_n) dC_{\boldsymbol{x}_i', n} = \frac{1}{M} \sum_{l=1}^{M} PQI(u_j^*|u_i' = u_{i,l}', u_1, \ldots, u_n).$$

Algorithm 3 describes the Monte Carlo sampling of U' using the conditional copula density $c_{\boldsymbol{x}_i', n} = c_{\boldsymbol{x}_i', n}(u_i'|u_1, \ldots, u_n)$ at the i-th candidate location. Here, the uniform distribution is used as the envelope distribution.

Algorithm 3: Algorithm for Monte Carlo sampling of U', i.e., $[0, 1]$.

Definition:

\# Let M be the number of Monte Carlo samples

$sample \leftarrow NULL$ \# Vector of Monte Carlo sampling values

Calculation:

1. Calculate the conditional copula density $c_{\boldsymbol{x}_i', n}(u_i'|u_1, \ldots, u_n)$ at the i-th candidate location $\boldsymbol{x}_i'$.
2. Obtain the modal value u_{modal}' of $c_{\boldsymbol{x}_i', n}(u_i'|u_1, \ldots, u_n)$ and the corresponding density value $c_{\boldsymbol{x}_i' n}(u_i' = u_{modal}'|u_1, \ldots, u_n)$.
3. Obtain the Monte Carlo sampling values:
   while $(length(sample) < M)$

   $x \leftarrow random\ value \sim Uniform(0, 1)$

   $y \leftarrow random\ value \sim Uniform(0, c_{\boldsymbol{x}_i' n}(u_i' = u_{modal}'|u_1, \ldots, u_n))$

   if $(y \leq c_{\boldsymbol{x}_i' n}(u_i' = x|u_1, \ldots, u_n))$

   add x *value to sample*

   end if

end while

Finally, it may be computationally expensive to use all of the observations $u_1, \ldots, u_n$ in obtaining the conditional copula distributions $C_{\boldsymbol{x}_i', n}$, at the candidate location $\boldsymbol{x}_i'$, and $C_{\boldsymbol{x}_j^*, n+1}$, at the interpolation grid point $\boldsymbol{x}_j^*$. The conditional copula distribution based on nearby locations is a good approximation for the conditional copula distribution based on all of the observations if a sufficient

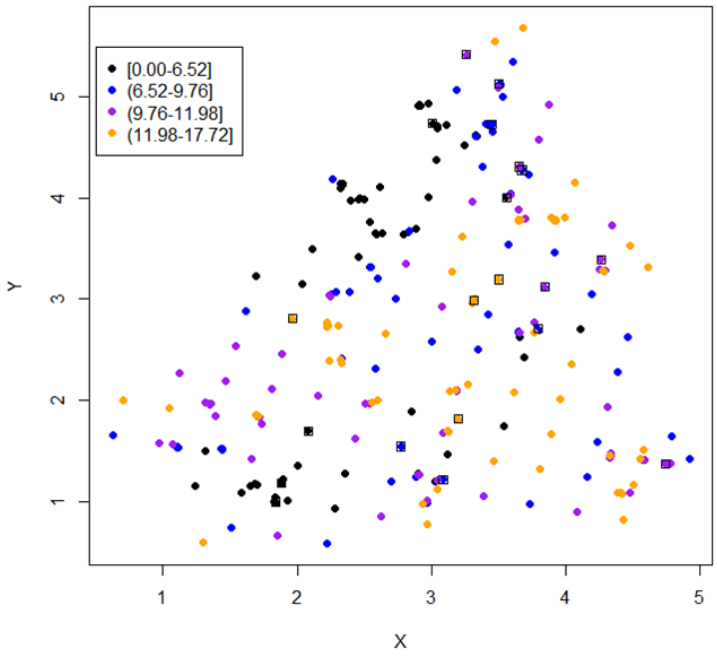
number of nearby locations is used (Bárdossy and Li, 2008).

The expected prediction uncertainty is sensitive to the number of nearby locations that use in prediction process. However, no significant difference can be observed if sufficient number of nearby locations was used. Number of nearby location was selected calculating expected prediction uncertainty for several randomly selected locations with different number of nearby locations. From that experiment, it was clear that after nine nearby locations no significant reduction can be seen in expected prediction uncertainty. However, computation time rapidly increased when number of nearby location increased. Hence nine nearby locations are used in this application.

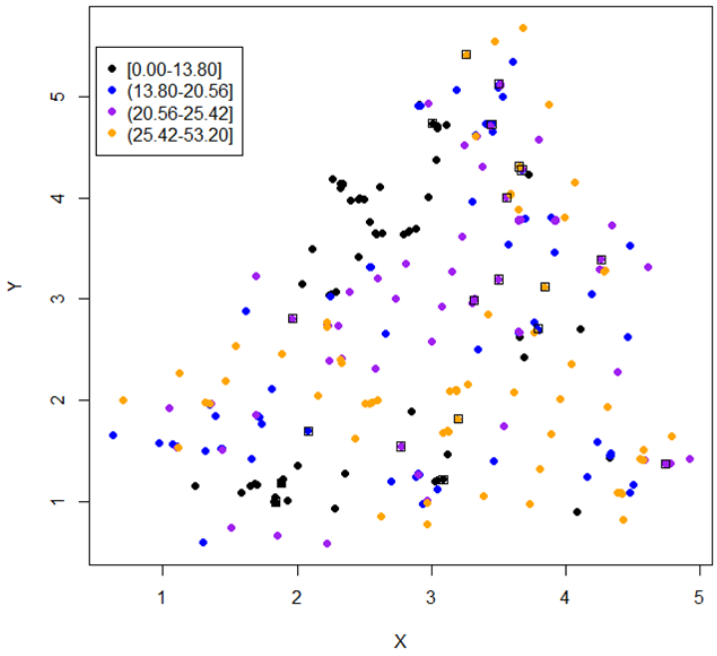
6.3 Data

The Swiss Jura data set (Goovaerts, 1997) was used in the application of the proposed sampling methodology. The data set contains 259 samples that were taken from the top soil of the region near La Chaux-de-Fonds in the Swiss Jura, which covers an area of 14.5 km^2. From these 259 top soil samples, seven toxic metals, namely, cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), nickel (Ni), lead (Pb), and zinc (Zn), were measured. The main intention of this survey was to identify contamination zones to restrict the use of those lands or apply remedies. In order to do this, prediction of the concentration of the metals must be carried out over the study domain and regions with high metal concentration identified. Hence, the reduction in prediction uncertainty, particularly in areas neighbouring high metal concentrations, is beneficial in the identification of contamination zones.

In this chapter, only two toxic metals, cobalt and nickel, were selected for application of the proposed methodology. Figures 6.1(a) and 6.1(b) are spatial plots of the concentrations of Co and Ni, respectively. For both metals, the more densely sampled areas tend to correspond to lower concentration values and the more sparsely sampled areas correspond to a mixture of moderate to high metal concentrations.



(a)



(b)

Figure 6.1: Spatial plots for (a) Co and (b) Ni.

6.4 Application

In this section, the proposed methodology is applied to Co and Ni separately. A 250m by 250m interpolation grid was defined over the study domain, as shown

in Figure 6.2. There are 196 grid points. The interpolation grid points are also considered as the potential candidates for the new samples. Performance of the design methodology is assessed through a partial redesign of the initial sampling. Twenty observations were removed randomly from an existing spatial design with 259 observations, based on the design in Atteia et al. [1994]. Subsequently, 20 design points were added back into the reduced data set from potential candidates using the proposed optimal design. The red squares in Figures 6.1 and 6.2 denote the 20 observations that were removed from the original 259 observations. Uncertainty measures of prediction over the interpolation grid are compared for the existing spatial design and the redesigned spatial design using the proposed methodology. In addition, a design based on kriged model under Gaussian assumption is compared with the redesigned spatial design under the proposed methodology.

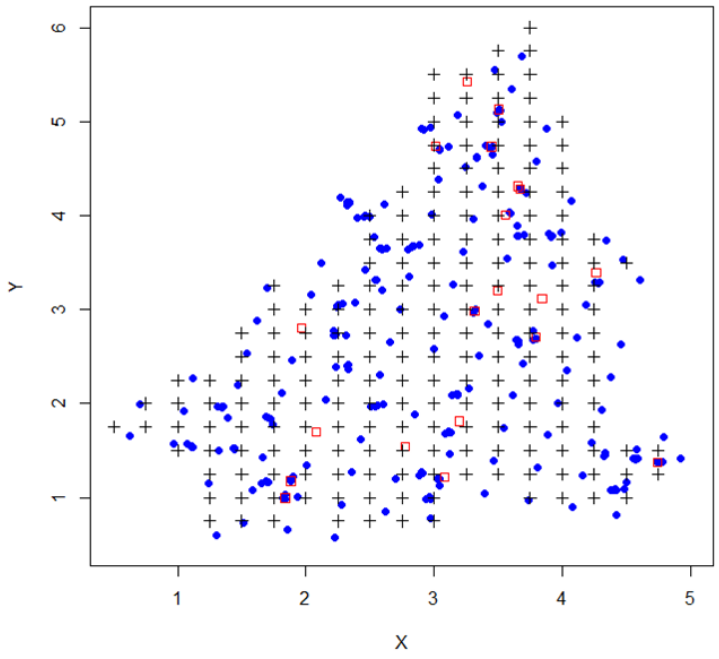


Figure 6.2: Study domain with retained old locations (blue dots) and removed locations (red squares) for both Co and Ni. Interpolation locations are denoted by black crosses.

6.4.1 Comparison of pair-copula and kriged models

Figures 6.3(a) and 6.3(b) give maps of the kriging variance under Gaussian assumptions for Co and Ni, respectively, while Figures 6.3(c) and 6.3(d) show the maps of the widths of the 90% prediction intervals from the pair-copula models for Co and Ni, respectively, for the reduced data set with 239 observations. The 90% prediction interval is calculated as the difference between the 95-th and 5-th predictive quantiles. These maps are overlaid with the retained old observations and the removed observations.

Figures 6.3(c) and 6.3(d) indicate that wider 90% prediction intervals under the pair-copula models correspond both to areas that are more sparsely sampled as well as areas with high variability in metal concentrations. Hence, the prediction intervals from the pair-copula models not only capture the spatial configuration of the data but also the variability in data values. The areas corresponding to wide prediction intervals differ between Co and Ni, due to the differing metal concentrations of Co and Ni at the observed locations. Hence, when the proposed design methodology is implemented, the locations for new observations will differ for Co and Ni, with the new locations occurring in areas with wide prediction intervals.

From Figures 6.3(a) and 6.3(b), it can be seen that the more sparsely sampled areas correspond to higher kriging variance and that the regions showing higher kriging variance are similar for both Co and Ni. However, unlike the pair-copula prediction intervals, the kriging variance doesn't capture the variability in metal concentrations. As a result, when a kriging based design is implemented, both Co and Ni will have very similar locations for new observations located in areas with high kriging variance.

6.4.2 Simulation study for non-sequential spatial redesign

Twenty new locations, out of the 196 potential candidate locations, were selected to replace the 20 removed locations. The performance of the proposed methodolgy is assessed by comparing the redesigned spatial design to the existing spatial

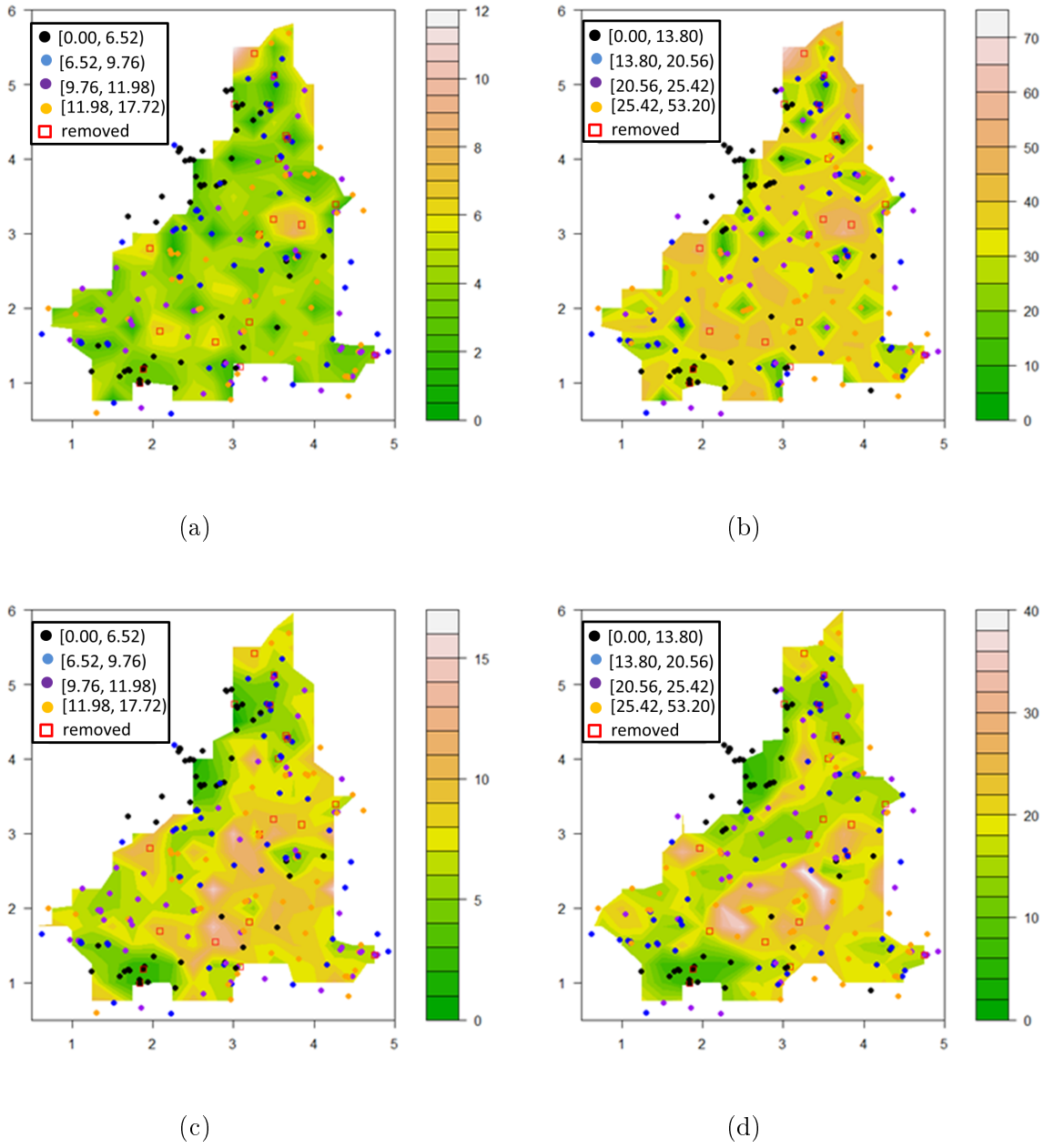


Figure 6.3: Maps for the (a) kriging variance of Co, (b) kriging variance of Ni, (c) 90% prediction interval widths based on the pair-copula for Co and (d) 90% prediction interval widths based on the pair-copula for Ni, overlaid with the retained old observations (dots) and removed observations (hollow red squares)

design through a simulation study similar to Li et al. [2011]. The procedure for the simulation study is outlined in the following steps.

1. Randomly remove 20 observed locations from the original observation set X^0, of 259 observations, to produce a reduced data set X, of 239 observations.

2. For each candidate point, calculate the total expected PQI over the inter-

polation grid after adding the candidate location as a possible new location to the reduced data set X. There will be 196 total expected PQIs obtained for the 196 candidate locations.

3. Select the 20 locations that produce the lowest total expected PQIs as the new sampling locations.

4. Randomly order the 20 new sampling locations and let this set of sampling locations be denoted by $S = (s_1, \ldots, s_{20})$. Sequentially simulate realisations for the locations. That is, simulate a value for s_1, then s_2, then s_3 and so on, up to s_{20}, as follows. For s_1, fit a conditional copula at s_1, conditioned on the reduced data set X. Obtain a random value from the conditional copula using Monte Carlo simulation and assign this value to the location s_1. Add s_1 to the reduced data set X. For s_i, $i = 2, \ldots, 20$, obtain a random value from the conditional copula at s_i, conditioned on the reduced data set and locations $s_1, \ldots, s_{i-1}$, using Monte Carlo simulation and assign this value to the location s_i. Add s_i to the data set containing X and the locations $s_1, \ldots, s_{i-1}$.

5. Repeat step 4, 100 times to obtain 100 sequential simulations. This results in 100 data sets, with each data set containing 259 observations.

6. For each simulated data set, calculate the total PQI over the interpolation grid. Sort the total PQIs in increasing order to form the set $PQI_T = (PQI_1, \ldots, PQI_{100})$, where $PQI_j < PQI_{j+1}$ for $j = 1, \ldots, 99$.

7. Calculate the total PQI for the original set of observations X^0 over the interpolation grid and let this be denoted by PQI_0.

8. Compare the total PQI from the original observations PQI_0 with the total PQIs from the simulated data sets $PQI_1, \ldots, PQI_{100}$ and observe the number of total PQIs from the simulated data sets that are less than the total PQI from the original observations. If $PQI_j < PQI_0 < PQI_{j+1}$, then the proportion of sequential simulations that have a lower total PQI than the

the total PQI of the original observations X^0 is $j/100$.

Figures 6.4(a) and 6.4(b) show the maps of the 196 total expected PQIs that are obtained for the 196 candidate locations, as detailed in step 2 above, for Co and Ni, respectively. As determined by the simulation procedure above, the new sampling locations (solid red squares) are located in regions corresponding to lower values of total expected PQI. Figures 6.4(c) and 6.4(d) are the maps of the 90% prediction interval widths from the pair-copula models for Co and Ni, respectively, for the reduced data set with 239 observations. As expected, comparing Figure 6.4(a) with Figure 6.4(c) for Co, and Figure 6.4(b) with Figure 6.4(d) for Ni, the areas with wide prediction intervals correspond to areas with low total expected PQI. It was commented previously that these are areas that are more sparsely sampled and with high variability in metal concentrations.

Figures 6.5(a) and 6.5(b) show the distributions of the total PQIs for the 100 different realisations of the redesigned spatial design for Co and Ni, respectively. The total PQI of the original 259 observations is represented by the value in bold on the x-axis. For Co, the redesigned spatial design outperforms the original spatial design, that is, the simulated total PQIs are less than the PQI of the original observations, in 98% of the simulations. For Ni, the redesigned spatial design outperforms the original design in 99% of the simulations.

6.4.3 Sequential spatial redesign

In the procedure for the simulation study, described above, the selection of the 20 new locations is not sequential. However, the proposed methodology specifies sequential addition of new locations, which means that the second optimal candidate location can only be determined after measurement at the first selected location. If measurement at the first additional location cannot be taken, the conditional copula density in step 3 of the proposed methodology can be obtained, not only for all possible values of u_i' but, for all possible values of the first additional location. Consequently, the integral in step 5 becomes a double integral. For the selection of the third optimal location, the integral becomes a triple inte-

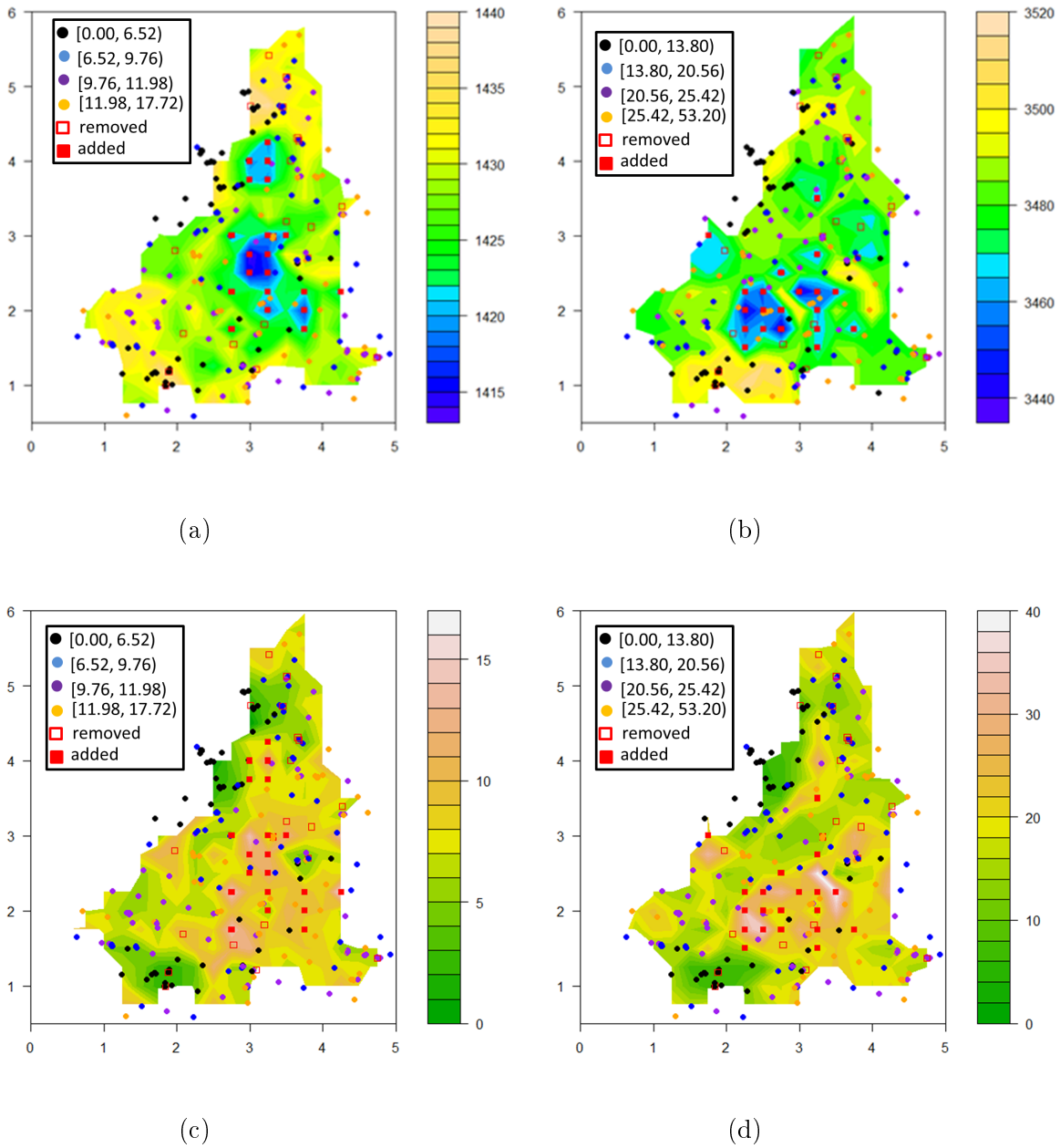
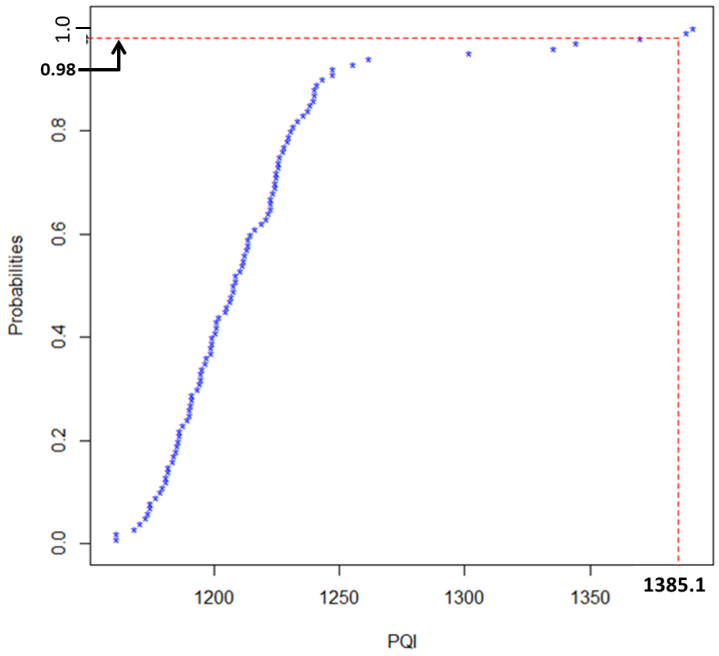


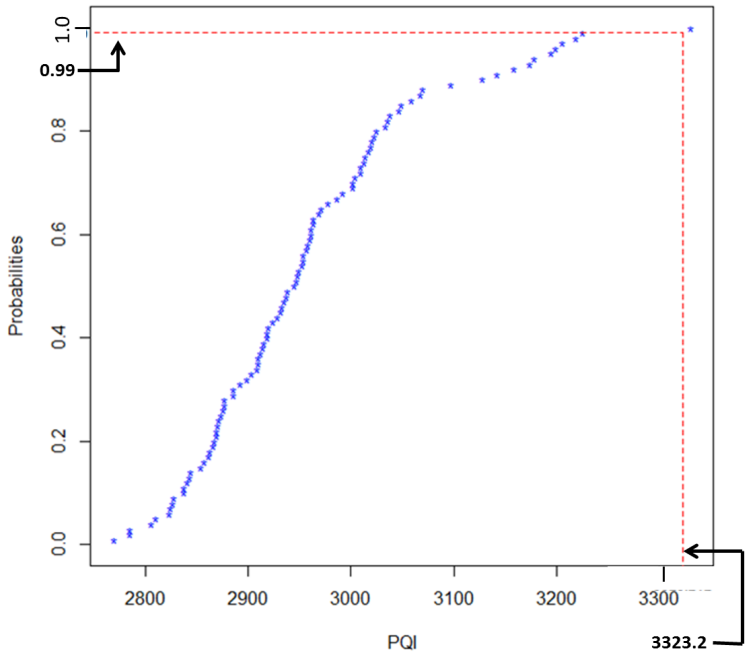
Figure 6.4: Maps of the total expected PQI for (a) Co and (b) Ni, and the 90% prediction interval widths based on the pair-copula for (c) Co and (d) Ni, overlaid with the retained old observations (dots), removed observations (hollow red squares) and new non-sequentially added observations (solid red squares).

gral. For selection of the n-th optimal location, an n-dimensional integral must be calculated. This approach is clearly computationally intensive.

To demonstrate the sequential design methodology, rather than considering all possible values of a newly selected location, the modal value from the conditional copula density, conditioned on the existing sampled observations and the previously added additional locations, is assigned as the observed value for the



(a)



(b)

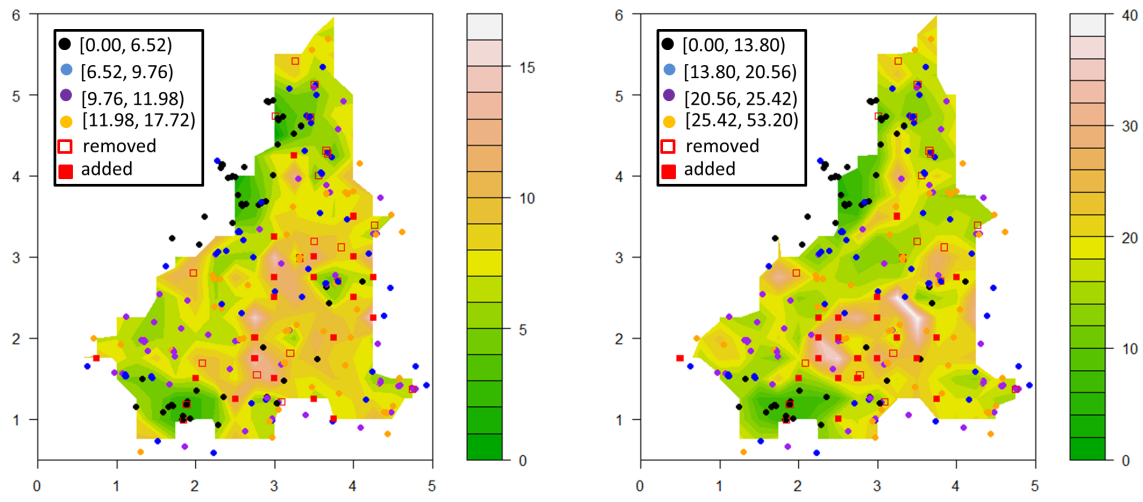
Figure 6.5: Distribution of total PQI for (a) Co and (b) Ni from 100 simulated data sets.

location. Moreover, as observed, in this application, the conditional distribution at given location is typically unimodal and very peaked. Therefore, modal value typically has a high probability of occurrence. Hence, it is reasonable to use

modal value when compared to computational intense that need to handle by using all the possible values.

Figures 6.6(a) and 6.6(b) show the maps of the 90% prediction interval widths from the pair-copula models for Co and Ni, respectively, for the reduced data set with 239 observations. The 20 new observations, obtained sequentially and assigned modal values, appear as solid red squares. Comparing Figure 6.4(c) with Figure 6.6(a), and Figure 6.4(d) with Figure 6.6(b), indicates that the areas where the 20 new observations are located are similar for the non-sequential and sequential designs, for both Co and Ni. The new locations for the sequential designs are still located in areas where the 90% predication intervals are wide. However, the selected locations are more scattered in the sequential design than the non-sequential design, due to updating of the total expected PQI after addition of each new location and the modal values assigned to the new locations in the sequential design. Since sequential design is only one possible realisation from an infinite number of designs, it is not possible to compare the sequential design and non-sequential design quantitatively. However, the similarity of the non-sequential and sequential designs suggests that the redesigned sequential designs are also likely to outperform the original designs in a large percentage of cases.

The green line in Figure 6.7 shows the total PQI after adding each selected location with the modal value assigned as the observation value for the location. For Co (Figure 6.7(a)) and Ni (Figure 6.7(b)), the total PQI decreased to less than the total PQI of the original 259 observations after adding just two new observations for Co and one observation for Ni. Hence, for this example, 18-19 less observations are required in the optimal redesign to achieve the total PQI, or less, of the original design. Note that the total PQI does not always decrease after adding a new observations due to the dependence of the total PQI on the values assigned to the new locations. However, any increase due to adding a new observation does not exceed the total PQI of the original observations.



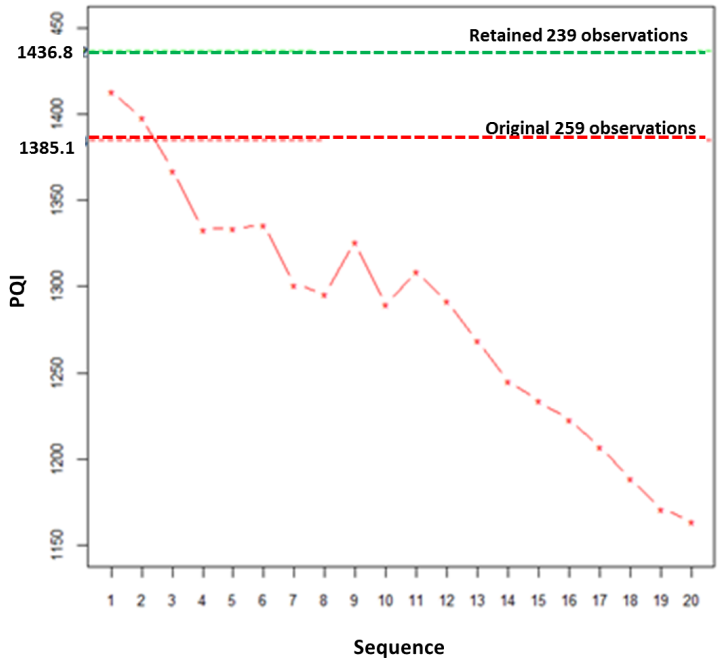
(a) (b)

Figure 6.6: Maps of the 90% prediction interval widths based on the pair-copula for (a) Co and (b) Ni, overlaid with the retained old observations (dots), removed observations (hollow red squares) and new sequentially added observations (solid red squares).

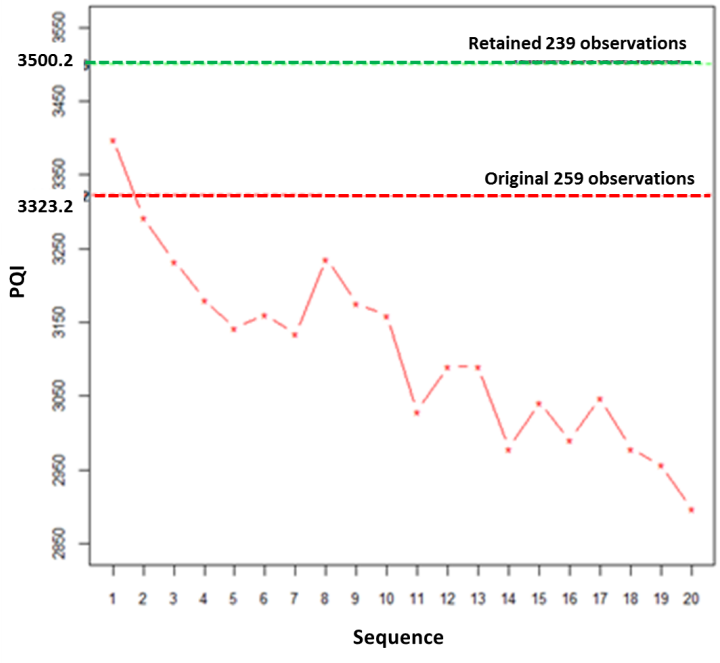
6.4.4 Kriging based design

Here, the performance of the proposed optimal design is evaluated against an optimal design based on kriged model under Gaussian assumption. Total kriging variance over the interpolation grid is used as the optimisation criterion for the kriged based design. A candidate location that produces the lowest total kriging variance is selected as the new observation. The variogram models that are discussed in Bandarian et al. [2008] were used to model the spatial dependency for Co and Ni.

From Figure 6.8, the new locations from the kriged based designs are located in areas with a lower density of observed points, as would be expected, since areas with less observations correspond to higher kriging variances. Unlike the designs based on the proposed methodology, which use pair-copulas, the new sampling locations for Co and Ni for the kriged based designs are identical for the non-sequential design and nearly identical for the sequential design. This is because designs based on kriged model under Gaussian assumption are dependent only on the spatial location of the observations, which are the same for Co and Ni,



(a)



(b)

Figure 6.7: Total PQI for sequentially selected candidate points for (a) Co and (b) Ni.

and not on the values of the observations. It is also worth noting that, as with the pair-copula based designs found using the proposed methodology, there is no notable difference between the sequential and non-sequential kriged based design

for both Co and Ni.

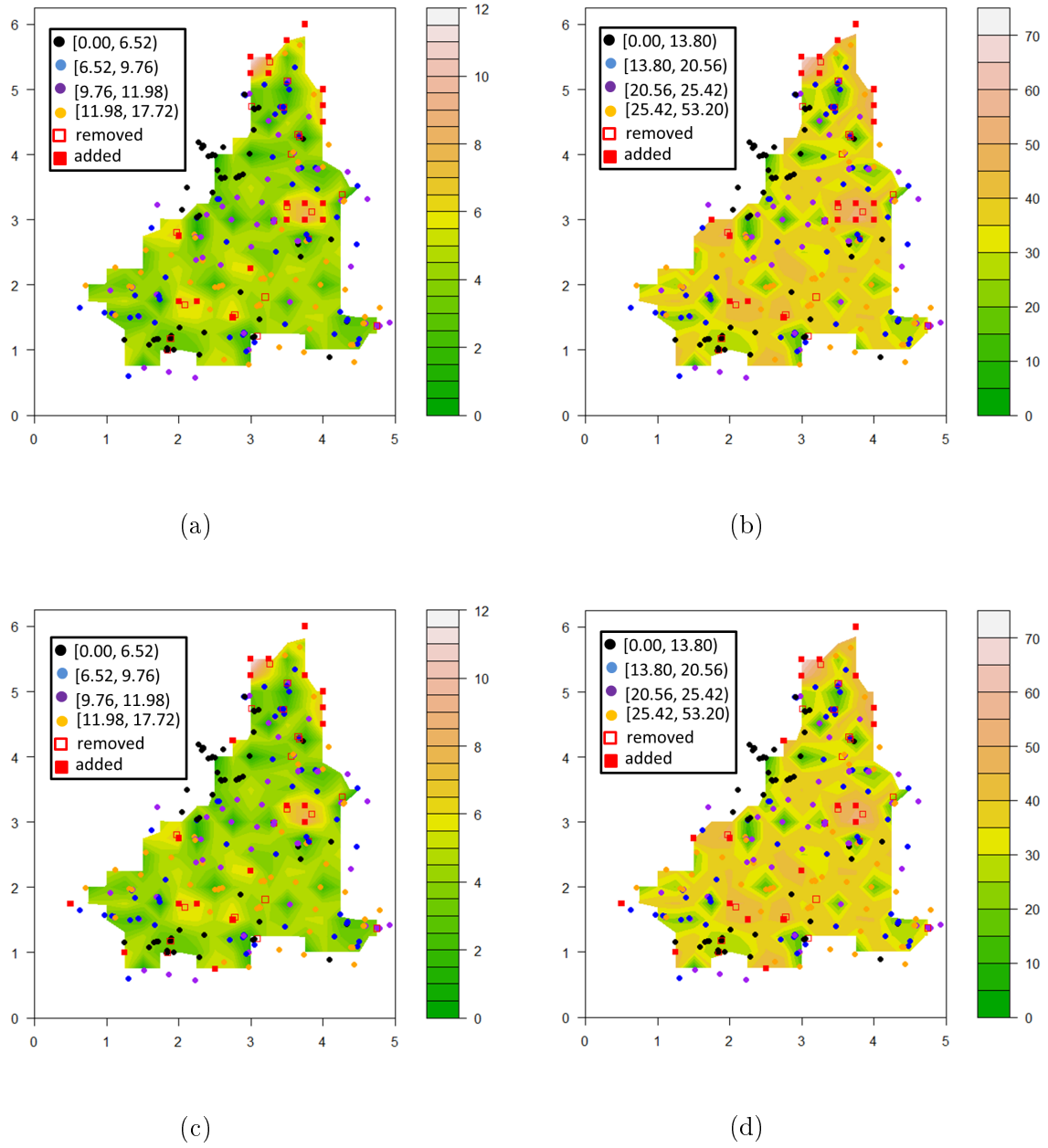
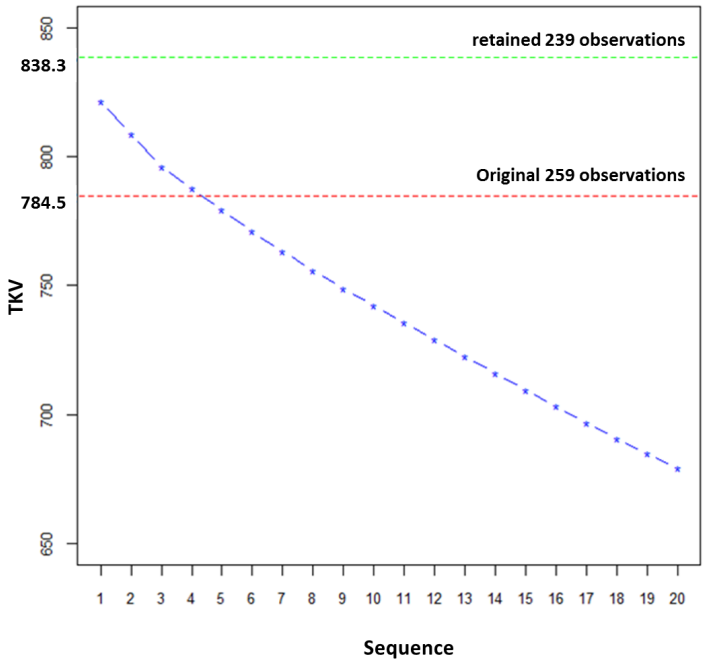
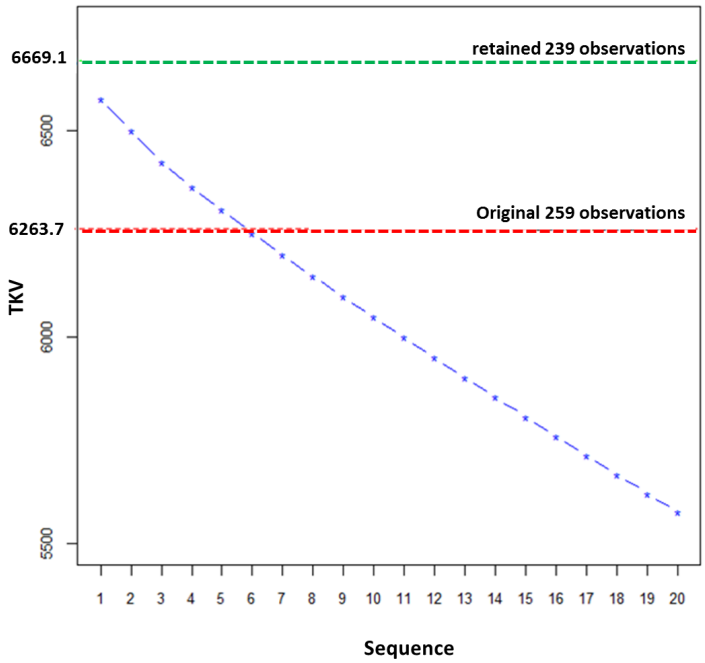


Figure 6.8: Kriging based non-sequential optimal design for (a) Co and (b) Ni, and sequential optimal design for (c) Co and (d) Ni.

As seen in Figure 6.9, the kriging variance for both variables decreases monotonically after adding a new observation. Moreover, to achieve the total kriging variance of the original 259 observations, 5-6 new observations are need, compared to only 1-2 for the pair-copula based designs.



(a)



(b)

Figure 6.9: Total kriging variance for sequentially selected candidate points for (a) Co and (b) Ni.

6.5 Conclusions

In this chapter, a new optimal design methodology based on the spatial pair-copula model is proposed. The optimal design methodology sequentially adds

new observations to an existing spatial design. The design is adaptive in that selection of a new location depends on the values observed for previously added locations.

Use of a copula-based spatial model in the proposed design methodology was motivated by the advantages of spatial copula models over other types of spatial models. Spatial copula-based models are capable of capturing both linear and non-linear spatial dependence and can additionally be used to study non-Gaussian processes. Specifically, the spatial pair-copula model more capably captures spatial dependence over other types of spatial copula models because it permits a different copula family, hence, different dependence structure, to be fitted to observations of differing distances.

In the application of the proposed methodology, the ability of the predictive quantile interval from the spatial pair-copula model to capture both the configuration and the variability of measured values, and the inability of the Gaussain based kriging variance to capture the variability of measured values, was demonstrated. Consequently, optimal designs based on spatial pair-copulas are likely to differ from optimal designs based on kriged model under Gaussian assumption. Pair-copula based spatial designs not only locate new measurements in areas that are more sparsely sampled, as do kriged based designs, but also add new measurements to areas where the measurement values vary. This feature is beneficial in the Swiss Jura application where identification of zones with high metal concentration is desired and where areas with varying levels of metal concentration are present. Moreover, it should be noted that the selected design is sensitive to the interpolation grid used on study domain as similar to prediction-based kriging designs.

In the simulation study, redesign of the spatial design using the proposed methodology outperformed the original design in more than 95% of simulations, even though the new sampling locations were not selected sequentially, and, hence, potentially selected sub-optimally. Whilst the sequential optimal design presented in the application is just one realisation of infinitely many possibly designs, it

is worth noting that the total predictive quantile interval for the sequential design is lower than all 100 simulated total predictive quantile intervals for the non-sequential design.

Determination of additional sampling locations in the sequential design, subsequent to the first additional location, ideally, requires measurement at previously added locations. If measurement is not possible, prior to determining the next sampling location, computation of a sequential optimal design increases to the n-th power for n additional locations. To overcome this computational challenge, an approach for selecting blocks should be developed, as discussed in Li et al. [2011].

In proposed design, cost contain didn't included. However cost constrains would be incorporated with proposed design methodology in future research.

Li et al. [2011] developed a sampling design based on a more simple copula based geostatistical model. The aim of their research was to add observation locations to an existing water observation network. A decision theoretic framework was used to constrain additional locations such that locations that were expected to fall below pollution thresholds for drinking water were more likely to be selected. As an extension to the methodology presented in this chapter, the method of Li et al. [2011] can be adapted to develop a decision theoretic design for additional samples based on the pair-copula mode based on a utility function. However, this is a topic of future research.

Chapter 7

Multivariate Optimal Spatial Design

The research in this chapter is in preparation for journal submission as detailed below.

- Musafer, G.N and Thompson, M.H. (n.d). Non-linear mutivariate optimal spatial design. *In preparation.*

Abstract

In this chapter, a new non-linear multivariate optimal spatial design methodology is proposed to simultaneously reduce the prediction uncertainty of multiple variables by selecting additional sampling locations based on the existing locations' configuration and their values. Novel aspects of the design methodology include the use of spatial pair-copulas to estimate the prediction uncertainty and the use of transformation methods for dimension reduction to model multivariate spatial dependence. Spatial pair-copulas are able to capture non-linear spatial dependence within variables better than other types of spatial copula models whilst a chained transformation that uses non-linear principal components captures the non-linear multivariate dependence between variables. The proposed design methodology is applied to two environmental case studies. Performance of the proposed methodology is evaluated through partial redesigns of the original

spatial designs. The first case study demonstrates the ability of the proposed design methodology to honour spatial non-linearity in the data. The second case study highlights the strength of the proposed design methodology in incorporating non-linear multivariate dependence into the design.

7.1 Introduction

Optimal spatial sampling design can be simply defined as optimal allocation of sampling points to spatial coordinates [Pilz and Spöck, 2008]. In most spatial processes, the first sampling campaign is conducted to obtain good geostatistical coverage and projection of the distribution of the variables of interest. However the decision of the sampling pattern for the next phase can be derived using the statistical information obtained from the first phase [Moon and Whateley, 2006]. This means that the information obtained from the first campaign can be used to develop an appropriate geostatistical model for prediction and additional samples can be used to improve precision of the prediction [Hassanipak and Sharafodin, 2004], which reduces the uncertainty of the prediction.

Some researchers are interested in objectives other than reducing the uncertainty of prediction. For example, Van Groenigen and Stein [1998] used a Monte Carlo method, such as simulated annealing, to maximise the spatial spread of the sample locations. This procedure is called a space-filling design [Royle and Nychka, 1998]. Others [Webster and Oliver, 1992, Zimmerman, 2006, Lark, 2002] are interested in improving the precision of the parameters of variograms. Also of interest is minimising both uncertainty of the prediction and minimising the uncertainty of parameter estimation of the variogram [Zimmerman, 2006, Diggle and Lophaven, 2006, Zhu and Stein, 2006]. In the mining field, Hassanipak and Sharafodin [2004] introduced another strategy to find the optimal design for additional samples with the aim of improving the reliability of resource classification and improving the estimates of grade and tonnage of the ore reserve. Li et al. [2011] maximised the expected gain defined by a utility function based on a simple spatial copula model in order to add observation locations to an existing water observation network.

Regardless of the objective, the target of these sampling design methodologies was one spatial variable. But in reality, measurements for multiple variables are frequently collected at a given location in spatial process and more than one variable may be of interest. If multiple variables are of interest, spatial design for additional samples should be optimal for all variables of interest under any objective. However, these variables are unlikely to be totally independent and dependence between these variables can be non-linear. In addition to this, in reality, the spatial dependence of individual variables is unlikely to be linear. Since optimal design is model dependent, an optimal design based on a spatial model that can capture the non-linear dependence structure between the spatial variables and non-linear dependence structure of individual variables is required. For instance, suppose that the actual relationships between spatial variables are non-linear, but one fits a model assuming a linear relationship between variables, and uses that model to develop an optimal sampling design for additional samples with the objective of reducing of prediction uncertainty and subsequently perform the prediction of variables of interest after adding new sample information. Under this scenario the final prediction will be inaccurate. There is no improvement that can be gained with an optimal deign without a valid model.

As far as the author is aware, very little work exists for spatial designs for multivariate settings [Vašát et al., 2010, Brown et al., 1994, Bueso et al., 1999, Li and Zimmerman, 2015]. Most of the existing multivariate spatial designs in the literature were developed based on co-kriging with the objective of reducing the uncertainty of simultaneous prediction where co-kriging is only capable of accounting for the linear relationship between the spatial variables. Moreover, co-kriging assumes a linear spatial dependence structure (spatial autocorrelation) of individual variables by employing variogram in modelling the spatial dependence. Hence, improvements cannot be expected in spatial prediction by adding new samples based on designs that are obtained through co-kriging if any dependence structure (between variables or within variables) is non-linear. In this research a new methodology for optimal spatial design for more than one variable

with the objective of reducing the uncertainty of the prediction of all variables simultaneously based on the spatial modelling approach is developed that can capture any non-linearity between variables and within variables.

Modelling of multiple spatial variables jointly is time consuming and complex as the number of variables increases. Hence, most researchers use different modelling approaches for prediction and simulation by transforming spatial variables into spatially uncorrelated variables (factors) using a suitable transformation method or combination of transformation methods [Leuangthong and Deutsch, 2003, Barnett et al., 2014, Barnett and Deutsch, 2012]. In this chapter, the multivariate modelling approach developed in Chapter 5 is used. In the multivariate modelling, spatial pair-copulas are used for the spatial interpolation. The pair-copula model, which was introduced by Gräler and Pebesma [2011], among other copula-based models, has more flexibility to capture the non-linear dependence structure of individual variables. This means that the uncertainty estimation for prediction produced by a pair-copula model has the capability to capture the variability of both the observations' configuration and its measured values [Bárdossy and Li, 2008, Haslauer et al., 2010]. Moreover, the pair-copula model can produce the conditional distribution of the variable of interest at unsampled locations.

The theory behind the proposed design methodology is discussed in detail in the next section. The implementation of the proposed design methodology is illustrated using the two case studies. The first case study has linear related variables whilst the variables of the second case study show non-linearity. The validity of the proposed methodology is evaluated by redesigning the existing design of two case studies. In addition to the implementation of the proposed methodology, these case studies are used to compare the different optimal design methodologies with the objective of reducing uncertainty. The first case study is used to investigate the difference between designs based on a model that can cater for only linear dependence of individual variables and designs based on a model that can cater for non-linear dependence of individual variables. Investigation of the difference between designs based on a model that captures the non-linearity

between variables and a design based on a model that ignores the non-linearity between variables is carried out in the second case study. Overall, the results demonstrate, in the case studies presented, the potentialities of the methodology. Li et al. [2011] proposed a sampling design based on a spatial copula. The aim of their research was to add observation locations to an existing water observation network with the aim of maximising the expected gain defined through a utility function. In Chapter 6, the methodology of Li et al. [2011] was extended by defining a statistical criterion in order to tally with the aim of reduction of uncertainty in univariate predictions and used a more flexible pair-copula model. In this research, the univariate optimal design of Chapter 6 is extended to multivariate optimal spatial design. This methodology enables optimal sampling design for more than one variable by reducing the uncertainty of prediction of all variables simultaneously.

Without loss of generality, the proposed multivariate methodology is described using a bivariate spatial design.

7.2 Methodology

Let $Z(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), Z_2(\boldsymbol{x})]$ denote a bivariate spatial random field. Here $\boldsymbol{x}$ is a two dimensional location belonging to the study domain $\mathcal{X}$. The set of existing sampled locations is denoted by $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. The objective of optimal spatial sampling here is to select additional measurement locations that reduce the combined uncertainty of two spatial random variables over the random field. One such example arises in additional sampling to monitor soil quality, where additional measurements are desired to reduce the combined uncertainty of toxic metal concentrations for the purpose of determining contamination zones. Of particular focus in this chapter is the reduction in the weighted average of the predictive quantile intervals (PQI) for the two variables, where the PQI is the difference between predictive quantiles of $Z_k(\boldsymbol{x})$, $k = 1, 2$. Let $X' = (\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m)$ be a set of candidate locations from which the additional new locations are chosen and $X' \subseteq \mathcal{X}$. The PQI at unsampled locations $X^* = (\boldsymbol{x}^*_1, \ldots, \boldsymbol{x}^*_N)$ can be esti-

mated from the sampled observations by interpolation of the random field $Z_k(\boldsymbol{x})$. Note that, in practice, the study domain $\mathcal{X}$ is discretised into an interpolation grid so that the set of unsampled locations X^* are the nodes of the interpolation grid. The interpolation method of Gräler and Pebesma [2011], which uses spatial pair-copulas, is applied here.

Throughout, superscript Z is used to denote quantities associated with $Z(\boldsymbol{x})$. The predictive quantile of $Z_k(\boldsymbol{x})$ at unsampled location $\boldsymbol{x}_j^*$, given the existing sampled observations is,

$$PQ^Z_{k,\boldsymbol{x}_j^*,n;q} \;=\; F^{-1}_{Z_k}\left({C^Z_{k,\boldsymbol{x}_j^*,n}}^{-1}(q|\boldsymbol{u}_k^Z)\right)$$

where $C^Z_{k,\boldsymbol{x}_j^*,n}(u^{*Z}_{kj}|\boldsymbol{u}_k^Z)$ is the conditional copula for variable $Z_k(\boldsymbol{x})$ at unsampled location $\boldsymbol{x}_j^*$, conditioned on the n existing sampled observations. Note that u^{*Z}_{kj} denotes the value of the k-th Uniform random variable U_k^{*Z} (on $[0,1]$) at the unsampled location $\boldsymbol{x}_j^*$, and $\boldsymbol{u}_k^Z = (u^Z_{k1},\ldots,u^Z_{kn})$ where $u^Z_{ki} = F_{Z_k}(z_k(\boldsymbol{x}_i))$ with F_{Z_k} denoting the estimated marginal cumulative distribution function of the data for variable Z_k.

The PQI corresponding to the difference between the 95-th and 5-th predictive quantiles of $Z_k(\boldsymbol{x})$ at unsampled location $\boldsymbol{x}_j^*$, given the existing sampled observations, is

$$PQI^Z_k(u^{*Z}_{kj}|\boldsymbol{u}_k^Z) \;=\; PQ^Z_{k,\boldsymbol{x}_j^*,n;0.95} - PQ^Z_{k,\boldsymbol{x}_j^*,n;0.05}.$$

The objective of the optimal design is to reduce the combined PQI for $Z_1(\boldsymbol{x})$ and $Z_2(\boldsymbol{x})$. Following Vašát et al. [2010], the combined PQI is taken as the weighted average of the PQIs for the two variables:

$$PQI^Z(u^{*Z}_{1j},u^{*Z}_{2j}|\boldsymbol{u}_1^Z,\boldsymbol{u}_2^Z) \;=\; \sum_{k=1}^{2}\frac{w_k}{\sigma_k}PQI^Z_k(u^{*Z}_{kj}|\boldsymbol{u}_k^Z)$$

where w_k are weights that are assigned depending on the relative importance of each variable with $\sum_{i=k}^{2} w_k = 1$. For spatial variables that have different measurement units, the PQIs can be standardised by division with the corresponding

standard deviation σ_k of the data. The candidate location $\boldsymbol{x}'_i$, $i = 1, \ldots, m$, from the set X', selected as the new location for measurement, is that which corresponds to the smallest total expected weighted PQI over the study domain after it has been added to the existing observations. In this chapter weighted PQI is used as statically criterion for as the PQI is used as a measure of predictive uncertainty in most of the spatial applications and less computationally expensive. However any arbitrary measure of predictive uncertainty such as spatial variance can be used in this proposed design methodology.

To obtain the weighted PQI for $Z(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), Z_2(\boldsymbol{x})]$ that takes account of the bivariate dependence between the variables, the variables are first transformed into uncorrelated factors $M(\boldsymbol{x}) = [M_1(\boldsymbol{x}), M_2(\boldsymbol{x})]$. The transformation method of Chapter 4 is applied here with principal components analysis (PCA) used in the transformation for linear bivariate relationships and non-linear PCA (NLPCA) used for non-linear relationships.

The transformation of variables $Z(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), Z_2(\boldsymbol{x})]$ into uncorrelated factors $M(\boldsymbol{x}) = [M_1(\boldsymbol{x}), M_2(\boldsymbol{x})]$ is given by

$$M(\boldsymbol{x}) \;=\; G[T\,(Z(\boldsymbol{x}))] \tag{7.1}$$

where T is the transformation used to remove cross-correlation at zero lag distance and G is the transformation used to decorrelate the variables at distances greater than lag zero. Note that, for higher dimensions, transformation of $Z(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), \ldots, Z_K(\boldsymbol{x})]$, $K > 2$, results in decomposition of the variables into $L \leq K$ uncorrelated factors $M(\boldsymbol{x}) = [M_1(\boldsymbol{x}), \ldots, M_L(\boldsymbol{x})]$. Henceforth, superscript M is used to denote quantities associated $M(\boldsymbol{x})$.

Predictive quantiles can be calculated for each factor after fitting independent pair-copula models to each factor. The conditional copula for $M_l(\boldsymbol{x})$, $l = 1, 2$, at the candidate location $\boldsymbol{x}'_i$, conditioned on the existing observations, is

$$C^M_{l,\boldsymbol{x}'_i,n} \;=\; C^M_{l,\boldsymbol{x}'_i,n}(u'^M_{li}|\boldsymbol{u}^M_l) \tag{7.2}$$

where $u_{li}'^M$ denotes the value of the l-th Uniform random variable $U_l'^M$ at the candidate location $\boldsymbol{x}_i'$, and $\boldsymbol{u}_l^M = (u_{l1}^M, \ldots, u_{ln}^M)$ where $u_{li}^M = F_{M_l}(m_l(\boldsymbol{x}_i))$ with F_{M_l} denoting the estimated marginal distribution function of the transformed data corresponding to factor M_l.

After addition of a candidate to the set of existing observations, the conditional copula for each factor can be re-estimated at the interpolation grid points. For any possible value $u_{li}'^M$ of $U_l'^M$ at the candidate location $\boldsymbol{x}_i'$, the conditional copula for $M_l(\boldsymbol{x})$ at interpolation location $\boldsymbol{x}_j^*$, conditioned on the existing observations and the newly added candidate $\boldsymbol{x}_i'$, is

$$C^M_{l,\boldsymbol{x}_j^*,n+1} = C^M_{l,\boldsymbol{x}_j^*,n+1}(u_{lj}^{*M}|u_{li}'^M, \boldsymbol{u}_l^M) \tag{7.3}$$

where u_{lj}^{*M} is any possible value of Uniform random variable U_l^{*M} at $\boldsymbol{x}_j^*$.

Using Eq. (7.3), any predictive quantile of $M_l(\boldsymbol{x})$ can be estimated at all points on the interpolation grid after adding the candidate $\boldsymbol{x}_i'$. The q^M-th predictive quantile, $0 < q^M < 1$, at interpolation location $\boldsymbol{x}_j^*$ after adding the candidate $\boldsymbol{x}_i'$ with a proposed value $u_{li}'^M$ as the assumed observed value is

$$PQ^M_{l,\boldsymbol{x}_j^*,n+1;q^M} = F^{-1}_{M_l}\left({C^M_{l,\boldsymbol{x}_j^*,n+1}}^{-1}(q^M|u_{li}'^M, \boldsymbol{u}_l^M)\right). \tag{7.4}$$

In total, predictive quantiles are calculated for N_q values of q^M.

The predictive quantiles for variables $Z_k(\boldsymbol{x})$, which incorporate the bivariate dependence, are then obtained by applying the corresponding back transformation of Eq. (7.1) to the predictive quantiles for factors $M_l(\boldsymbol{x})$:

$$\left[PQ^Z_{1,\boldsymbol{x}_j^*,n+1;q^Z}, PQ^Z_{2,\boldsymbol{x}_j^*,n+1;q^Z}\right] = T^{-1}\left[G^{-1}\left(PQ^M_{1,\boldsymbol{x}_j^*,n+1;q^M}, PQ^M_{2,\boldsymbol{x}_j^*,n+1;q^M}\right)\right] \tag{7.5}$$

where $PQ^Z_{k,\boldsymbol{x}_j^*,n+1;q^Z}$ is the q^Z-th predictive quantile of $Z_k(\boldsymbol{x})$ at $\boldsymbol{x}_j^*$ after adding $\boldsymbol{x}_i'$ to the set of observations with a proposed value $u_{ki}'^Z$. Note that q^Z, $0 < q^Z < 1$, does not necessarily equally q^M and is unknown.

To estimate the q^Z-th predictive quantile of $Z_k(\boldsymbol{x})$, the N_q predictive quantiles $PQ^Z_{k,\boldsymbol{x}^*_j,n+1;q^Z}$ are sorted in ascending order and the average of the $\lfloor q^Z N_q + 1/2 \rfloor$ and $\lceil q^Z N_q + 1/2 \rceil$ values in the ordered set is taken as the q^Z-th predictive quantile. For example, for $N_q = 100$ quantiles, the $q^Z = 0.05$ quantile is the average of the 5-th and 6-th values in the ordered set. The number of quantiles N_q should be sufficiently large for accurate estimation of the q^Z-th predictive quantile of $Z_k(\boldsymbol{x})$. From Eq. (7.5), the PQI corresponding to the difference between the 0.95 and 0.05 predictive quantiles of $Z_k(\boldsymbol{x})$ at unsampled interpolation location $\boldsymbol{x}^*_j$ after adding the candidate $\boldsymbol{x}'_i$ is

$$PQI^Z_k(u^{*M}_{1j}, u^{*M}_{2j} | u'^M_{1i}, u'^M_{2i}, \boldsymbol{u}^M_1, \boldsymbol{u}^M_2) \;\; = \;\; PQ^Z_{k,\boldsymbol{x}^*_j,n+1;0.95} - PQ^Z_{k,\boldsymbol{x}^*_j,n+1;0.05}. \tag{7.6}$$

Hence, using Eq. (7.6), the weighted average of the PQIs for $Z_1(\boldsymbol{x})$ and $Z_2(\boldsymbol{x})$ is

$$\begin{aligned} &PQI^Z(u^{*M}_{1j}, u^{*M}_{2j} | u'^M_{1i}, u'^M_{2i}, \boldsymbol{u}^M_1, \boldsymbol{u}^M_2) = \\ &\qquad \sum_{k=1}^{2} \frac{w_k}{b_k} PQI^Z_k(u^{*M}_{1j}, u^{*M}_{2j} | u'^M_{1i}, u'^M_{2i}, \boldsymbol{u}^M_1, \boldsymbol{u}^M_2). \end{aligned} \tag{7.7}$$

This is the weighted PQI at $\boldsymbol{x}^*_j$ for one possible combination of values for u'^M_{1i} and u'^M_{2i}.

The expected weighted PQI at $\boldsymbol{x}^*_j$ is calculated as the integral of the weighted PQI in Eq. (7.7) over the entire range of possible values of u'^M_{1i} and u'^M_{2i} corresponding to candidate location $\boldsymbol{x}'_i$:

$$\begin{aligned} &E\left[PQI^Z(u^{*M}_{1j}, u^{*M}_{2j} | u'^M_{1i}, u'^M_{2i}, \boldsymbol{u}^M_1, \boldsymbol{u}^M_2)\right] = \\ &\qquad \int_0^1 \int_0^1 PQI^Z(u^{*M}_{1j}, u^{*M}_{2j} | u'^M_{1i}, u'^M_{2i}, \boldsymbol{u}^M_1, \boldsymbol{u}^M_2) dC^M_{1,\boldsymbol{x}'_i,n} dC^M_{2,\boldsymbol{x}'_i,n} \end{aligned} \tag{7.8}$$

where $C^M_{l,\boldsymbol{x}'_i,n}$ is the conditional copula given in Eq. (7.2).

The total expected weighted PQI of the entire interpolation grid after adding the candidate $\boldsymbol{x}'_i$ as a new observation is then the sum of the expected weighted PQI

at all interpolation points:

$$E_T(\boldsymbol{x}_i') = \sum_{j=1}^{N} \left(\int_0^1 \int_0^1 PQI^Z(u_{1j}^{*M}, u_{2j}^{*M} | u_{1i}'^M, u_{2i}'^M, \boldsymbol{u}_1^M, \boldsymbol{u}_2^M) dC^M_{1\boldsymbol{x}_i',n} dC^M_{2\boldsymbol{x}_i',n} \right). \tag{7.9}$$

Computational efficiency can be gained by interchanging the summation and integration in Eq. (7.9):

$$E_T(\boldsymbol{x}_i') = \int_0^1 \int_0^1 \left(\sum_{j=1}^{N} PQI^Z(u_{1j}^{*M}, u_{2j}^{*M} | u_{1i}'^M, u_{2i}'^M, \boldsymbol{u}_1^M, \boldsymbol{u}_2^M) \right) dC^M_{1,\boldsymbol{x}_i',n} dC^M_{2,\boldsymbol{x}_i',n}.$$

The candidate $\boldsymbol{x}_i'$ that produces the smallest total expected weighted PQI is selected as the new sample location. A summary of the procedure is outlined in the following steps.

1. Use the transformation method in Chapter 4 to transform $Z(\boldsymbol{x}) = [Z_1(\boldsymbol{x}), Z_2(\boldsymbol{x})]$ into uncorrelated factors $M(\boldsymbol{x}) = [M_1(\boldsymbol{x}), M_2(\boldsymbol{x})]$ using $M(\boldsymbol{x}) = G[T(Z(\boldsymbol{x}))]$, where transformation T decorrelates the variables at zero lag distance and transformation G decorrelates the variables at lag distances greater than zero.

2. For each factor $M_l(\boldsymbol{x})$, $l = 1, 2$:

   (a) Transform values $m_l(\boldsymbol{x}_1), \ldots, m_l(\boldsymbol{x}_n)$ to the unit interval $[0, 1]$ using the estimated distribution function F_{M_l}: $u_{l1}^M = F_{M_l}(m_l(\boldsymbol{x}_1)), \ldots, u_{ln}^M = F_{M_l}(m_l(\boldsymbol{x}_n))$.

   (b) Use the transformed values $u_{l1}^M, \ldots, u_{ln}^M$ to fit a spatial pair-copula $C^M_{l,\boldsymbol{x},n}(u_l^M | u_{l1}^M, \ldots, u_{ln}^M)$ using the method of Gräler and Pebesma [2011].

   (c) For each candidate location $\boldsymbol{x}_i'$, for all values of $u_{li}'^M$, calculate the conditional copula density $c^M_{l,\boldsymbol{x}_i',n} = c^M_{l,\boldsymbol{x}_i',n}(u_{li}'^M | u_{l1}^M, \ldots, u_{ln}^M)$ (Gräler and Pebesma [2011]). In practice, it is not possible to obtain the conditional copula density for all possible values of $u_{li}'^M$, hence the range of values of $U_l'^M$, i.e., $[0, 1]$, is discretised and the conditional copula density is calculated for the midpoint of each interval.

3. For each interpolation grid point $\boldsymbol{x}_j^*$:

   (a) For each factor $M_l(\boldsymbol{x})$, calculate the conditional copula $C^M_{l,\boldsymbol{x}_j^*,n+1} = C^M_{l,\boldsymbol{x}_j^*,n+1}(u^{*M}_{lj}|u'^M_{li}, u^M_{l1}, \ldots, u^M_{ln})$, conditioned on the existing values $u^M_{l1}, \ldots, u^M_{ln}$ and the proposed value u'^M_{li} at the candidate location $\boldsymbol{x}'_i$. Use this conditional copula to calculate predictive quantiles $PQ^M_{l,\boldsymbol{x}_j^*,n+1;q^M}$, given in Eq. (7.4). Calculate predictive quantiles for N_q values of q^M, $0 < q^M < 1$.

   (b) For each of the N_q values of q^M, obtain predictive quantiles $PQ^Z_{1,\boldsymbol{x}_j^*,n+1;q^Z}$ and $PQ^Z_{2,\boldsymbol{x}_j^*,n+1;q^Z}$ for $Z_1(\boldsymbol{x})$ and $Z_2(\boldsymbol{x})$, respectively, by applying the corresponding back transformation of step 1 to the predictive quantiles $PQ^M_{1,\boldsymbol{x}_j^*,n+1;q^M}$ and $PQ^M_{2,\boldsymbol{x}_j^*,n+1;q^M}$.

   (c) For each variable $Z_k(\boldsymbol{x})$, sort the N_q predictive quantiles $PQ^Z_{k,\boldsymbol{x}_j^*,n+1;q^Z}$ in ascending order. Estimate the predictive quantile $PQ^Z_{k,\boldsymbol{x}_j^*,n+1;0.95}$ using the average of the $\lfloor 0.95N_q + 1/2 \rfloor$ and $\lceil 0.95N_q + 1/2 \rceil$ values from the ordered set. Similarly, estimate $PQ^Z_{k,\boldsymbol{x}_j^*,n+1;0.05}$ using the average of the $\lfloor 0.05N_q + 1/2 \rfloor$ and $\lceil 0.05N_q + 1/2 \rceil$ values from the ordered set. Subtract the 0.05 quantile from the 0.95 quantile to obtain the predictive quantile interval $PQI^Z_k(u^{*M}_{1j}, u^{*M}_{2j}|u'^M_{1i}, u'^M_{2i}, \boldsymbol{u}^M_1, \boldsymbol{u}^M_2)$, given in Eq. (7.6). Use these predictive quantile intervals to calculate the weighted predictive quantile interval $PQI^Z(u^{*M}_{1j}, u^{*M}_{2j}|u'^M_{1i}, u'^M_{2i}, \boldsymbol{u}^M_1, \boldsymbol{u}^M_2)$, given in Eq. (7.7).

   (d) Calculation of the conditional copulas and, consequently, the weighted predictive quantile interval is repeated for all discretised values of U'^M_1 and U'^M_2 at the candidate location $\boldsymbol{x}'_i$.

4. For each interpolation grid point $\boldsymbol{x}_j^*$, calculate the expected weighted PQI

using Eq. (7.8). The double integral in Eq. (7.8) can be approximated by

$$\int_0^1 \int_0^1 PQI^Z(u_{1j}^{*M}, u_{2j}^{*M} | u_{1i}'^M, u_{2i}'^M, \boldsymbol{u}_1^M, \boldsymbol{u}_2^M) dC_{1,\boldsymbol{x}_i',n}^M dC_{2,\boldsymbol{x}_i',n}^M =$$
$$\sum_{d_1=1}^{D_1} \sum_{d_2=1}^{D_2} \Big[PQI^Z(u_{1j}^{*M}, u_{2j}^{*M} | u_{1i}'^M = u_{1i,d_1}'^M, u_{2i}'^M = u_{2i,d_2}'^M, \boldsymbol{u}_1^M, \boldsymbol{u}_2^M)$$
$$\times c_{1,\boldsymbol{x}_i',n}^M (u_{1i}'^M = u_{1i,d_1}'^M | \boldsymbol{u}_1^M) \Delta u_{1i,d_1}'^M \times c_{2,\boldsymbol{x}_i',n}^M (u_{2i}'^M = u_{2i,d_2}'^M | \boldsymbol{u}_2^M) \Delta u_{2i,d_2}'^M \Big]$$

where $u_{li,d_l}'^M$ is the midpoint of the d_l-th discretised interval of $U_l'^M$, $c_{l,\boldsymbol{x}_i',n}^M(u_{li}'^M = u_{li,d_l}' | \boldsymbol{u}_l^M)$ is the conditional copula density for $M_l(\boldsymbol{x})$ calculated in step 2(c) at $u_{li}'^M = u_{li,d_l}'^M$ and $\Delta u_{li,d_l}'^M$ is the width of the d_l-th discretised interval.

5. For the candidate location $\boldsymbol{x}_i'$, calculate the total expected weighted PQI of the entire interpolation grid using Eq. (7.9) by summing up the expected weighted PQI calculated for all the interpolation grid points.

6. Repeat steps 2(b) to 5 for the remaining candidate points and select the candidate point that produces the smallest total expected weighted PQI, $E_T(\boldsymbol{x}_i')$, as the new sample location.

As with the univariate design methodology in Chapter 6, whilst the spatial pair-copula of Gräler and Pebesma [2011] is used in step 2(b), alternative spatial copulas could be substituted into the procedure, such as the spatial copula of Bárdossy and Li [2008].

Additionally, some practical issues that require consideration in implementing the proposed design methodology follow. Firstly, as with the univariate design methodology, the transformation applied in step 4 and the spatial copula fitted in step 2(b) are important, since it is assumed that the dependence of the random variable M_l follows the selected copula model.

Secondly, the range of values of $U_l'^M$ should be discretised in step 2(c) to provide a good numerical approximation of the expected weighted PQI calculated in step 4. Monte Carlo integration (Shapiro, 2003) is used in the numerical approximation of the expected weighted PQI. Hence, the intervals are determined by Monte

Carlo sampling. Therefore, the expected weighted PQI using Eq. (7.8) can be approximated by using Monte Carlo intergration as follows. For D_1 Monte Carlo samples of $U_1'^M$ and D_2 Monte Carlo samples of $U_2'^M$

$$\int_0^1 \int_0^1 PQI^Z(u_{1j}^{*M}, u_{2j}^{*M} | u_{1i}'^M, u_{2i}'^M, \boldsymbol{u}_1^M, \boldsymbol{u}_2^M) dC_{1,\boldsymbol{x}_i',n}^M dC_{2,\boldsymbol{x}_i',n}^M =$$
$$\frac{1}{D_1 \times D_2} \sum_{d_1=1}^{D_1} \sum_{d_2=1}^{D_2} \left[PQI^Z(u_{1j}^{*M}, u_{2j}^{*M} | u_{1i}'^M = u_{1i,d_1}'^M, u_{2i}'^M = u_{2i,d_2}'^M, \boldsymbol{u}_1^M, \boldsymbol{u}_2^M) \right].$$

Algorithm 3 in Chapter 5 can be used to carry out the Monte Carlo sampling of $U_l'^M$.

Thirdly, the predictive quantiles for M_l in step 3(a) are calculated for N_q values of q^M. Discretisation of the range of q^M, $0 < q^M < 1$, into N_q equally spaced points, such that N_q is sufficiently large to produce accurate estimation of the 0.95 and 0.05 predictive quantiles of Z_k, may be computationally intensive. Instead, Monte Carlo sampling is used to determine N_q values of the predictive quantiles for M_l by sampling values of $C_{l,\boldsymbol{x}_j^*,n+1}^M$. Algorithm 4 describes the estimation of the 0.95 and 0.05 predictive quantiles of Z_k using Monte Carlo sampling.

Finally, as discussed in Chapter 5, it may be computationally expensive to use all of the observations $u_{l1}^M, \ldots, u_{ln}^M$ in obtaining the conditional copula distributions $C_{l,\boldsymbol{x}_i',n}^M$, at the candidate location $\boldsymbol{x}_i'$, and $C_{l,\boldsymbol{x}_j^*,n+1}^M$, at the interpolation grid point $\boldsymbol{x}_j^*$. Therefore, the conditional copula distribution is calculated based on nearby locations. Ten nearby locations are used in the application. By using only the nearby neighbours, the the value of the conditional copula, used in calculation of the predictive quantile interval, at an interpolation grid point changes only if the newly added candidate location is a neighbour of the grid point. Hence, computation can be significantly reduced by use of an algorithm to find the grid points that are affected by newly added locations.

Algorithm 4: Algorithm for estimation of the 0.95 and 0.05 predictive quantiles for Z_1 and Z_2 using Monte Carlo sampling.

Definition:

\# Let N_q be the number of Monte Carlo samples

$sample_1 \leftarrow matrix(NA, N_q, 1)$ \# Vector of Monte Carlo sampling values for M_1

$sample_2 \leftarrow matrix(NA, N_q, 1)$ \# Vector of Monte Carlo sampling values for M_2

$Z_1 \leftarrow matrix(NA, N_q, 1)$ \# Vector of back transformed Z_1 values

$Z_2 \leftarrow matrix(NA, N_q, 1)$ \# Vector of back transformed Z_2 values

$PQ0.95_{Z_1} \leftarrow NULL$ \# 0.95 predictive quantile for Z_1

$PQ0.05_{Z_1} \leftarrow NULL$ \# 0.05 predictive quantile for Z_1

$PQ0.95_{Z_2} \leftarrow NULL$ \# 0.95 predictive quantile for Z_2

$PQ0.05_{Z_2} \leftarrow NULL$ \# 0.05 predictive quantile for Z_2

Calculation:

for l in 1 *to* 2

(a) Calculate the conditional copula density $c^M_{l,\boldsymbol{x}^*_j,n+1}(u^{*M}_{lj}|u'^M_{li}, \boldsymbol{u}^M_l)$ of M_l at interpolation location $\boldsymbol{x}^*_j$ for an assigned value of u'^M_{li} at candidate location $\boldsymbol{x}'_i$.

(b) Obtain the modal value $u^{*M}_{l,modal}$ of $c^M_{l,\boldsymbol{x}^*_j,n+1}(u^{*M}_{lj}|u'^M_{li}, \boldsymbol{u}^M_l)$ and the corresponding density value $c^M_{l,\boldsymbol{x}^*_j,n+1}(u^{*M}_{lj} = u^{*M}_{l,modal}|u'^M_{li}, \boldsymbol{u}^M_l)$.

(c) Obtain the Monte Carlo sampling values for M_l:
while $(length(sample_l) < N_q)$

$x \leftarrow$ *random value* $\sim Uniform(0, 1)$

$y \leftarrow$ *random value* $\sim Uniform(0, c^M_{l,\boldsymbol{x}^*_j,n+1}(u^{*M}_{lj} = u^{*M}_{l,modal}|u'^M_{li}, \boldsymbol{u}^M_l)$

if $(y \leq c^M_{l,\boldsymbol{x}^*_j,n+1}(u^{*M}_{lj} = x|u'^M_{li}, \boldsymbol{u}^M_l)$

add x *value to sample*

end if

end while

$sample_l \leftarrow F^{-1}_{M_l}(sample_l)$

$sample_l \leftarrow sort(sample_l)$ \# Sort vector in ascending order

end for

1. Back transform the Monte Carlo sampling values for M_1 and M_2:
$[Z_1, Z_2] \leftarrow T^{-1}(G^{-1}[sample_1, sample_2])$

2. Calculate the 0.95 and 0.05 predictive quantiles for Z_k:
for k in 1 *to* 2

$Z_k \leftarrow sort(Z_k)$ \# Sort vector in ascending order

$PQ0.95_{Z_k} \leftarrow (Z_k[\lfloor 0.95N_q + 1/2 \rfloor] + Z_k[\lceil 0.95N_q + 1/2 \rceil])/2$

$PQ0.05_{Z_k} \leftarrow (Z_k[\lfloor 0.05N_q + 1/2 \rfloor] + Z_k[\lceil 0.05N_q + 1/2 \rceil])/2$

end for

7.3 Data

The proposed optimal design methodology was applied to two data sets. The first application uses the Swiss Jura data set [Goovaerts, 1997]. In the Swiss Jura application, the bivariate relationship between the two variables investigated is linear. Data from the Bartlett Experimental Forrest (BEF) [Finley et al., 2007] was used in the second application, where the two variables investigated have a non-linear bivariate relationship. The variables from both data sets possess non-Gaussian and non-linear spatial dependence.

The purpose of the Swiss Jura application is to elucidate the features and advantages of multivariate pair-copula based sampling designs for data that are non-linearly spatial. The BEF application additionally demonstrates how a non-linear bivariate relationship impacts the sampling design.

7.3.1 Swiss Jura

A description of the Swiss Jura data set can be found in Chapter 6. The data set contains measurements of metal concentrations for seven toxic metals. In identifying contamination zones, that is, areas with high metal concentrations, simultaneous prediction of the metal concentrations should be carried out over the study domain. Prediction based on just one metal, as was done in Chapter 6, can be used to define regions that are contaminated with that particular metal, but may exclude areas with high concentrations of other metals. Hence, the simultaneous reduction in prediction uncertainty of all metals of concern, particularly in areas near high metal concentrations, is beneficial in the identification of regions with high metal concentrations of one or more metals.

For the purposes of illustrating the proposed design methodology, only two toxic metals, cobalt (Co) and nickel (Ni), were considered. Spatial plots of the concentrations of Co and Ni were given in Figures 6.1(a) and 6.1(b) of Chapter 6 and are repeated here in Figures 7.1(a) and 7.1(b) for ease of reference. It was noted in Chapter 6 that, for both metals, the more densely sampled areas tend

to correspond to lower concentration values and the more sparsely sampled areas correspond to metal concentrations with moderate to high values.

The scatter plot of the Co and Ni measurements in Figure 7.1(c) exhibits a strong linear relationship between Co and Ni at zero lag distance.

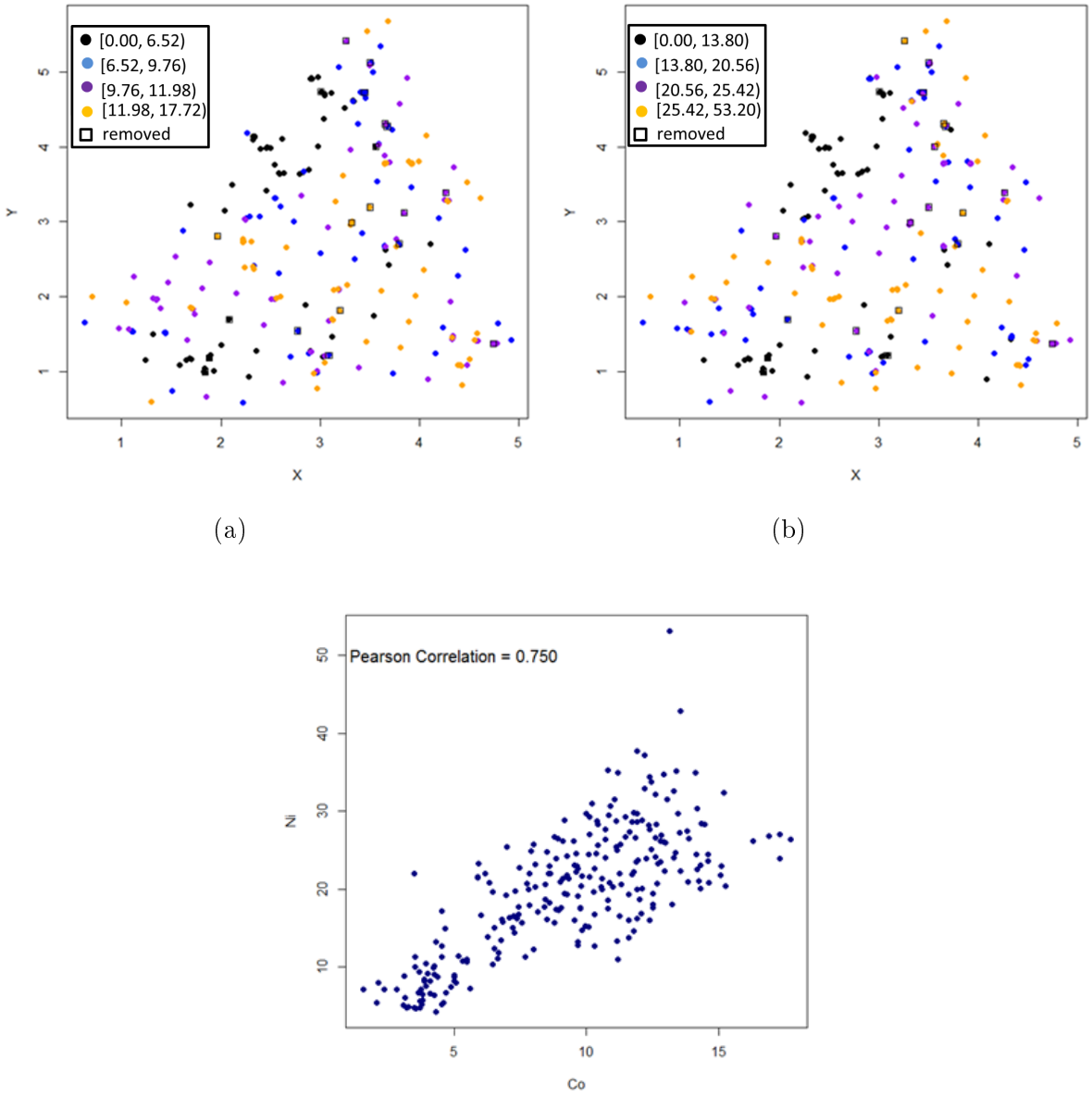


Figure 7.1: Swiss Jura data. Spatial plots for (a) Co and (b) Ni, and (c) scatter plot between Co and Ni.

7.3.2 Bartlett Experimental Forest

The BEF data set consists of 437 measurements at two dimensional locations for more than 50 attributes from georeferenced forest inventory plots on the United States Department of Agriculture Forest Service Bartlett Experimental Forest in Bartlett, new Hampshire [Finley et al., 2007]. The BEF covers an area of 1,053 hectares. Here, we are interested in two attributes that are non-linearly related. These attributes are generically labelled Z_1 and Z_2.

Figures 7.2(a) and 7.2(b) show the spatial distribution of Z_1 and Z_2, respectively. It can be seen that Z_1 has a larger variation in attribute values in comparison to Z_2, and low attribute values occur in similar locations for the two variables.

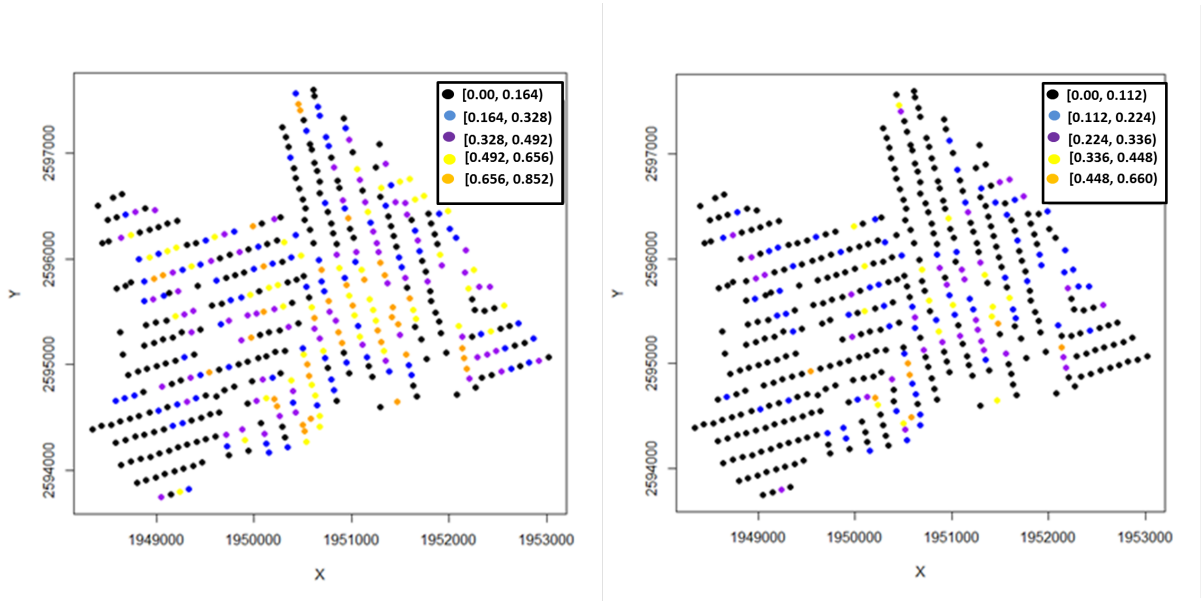
The non-linear structure of the bivariate data at zero lag distance can be clearly seen in the scatter plot of Figure 7.2(c).

7.4 Application

In this section, the proposed methodology is applied to the Swiss Jura and BEF data. Grids are defined over each study domain for interpolation. The interpolation grid points are also considered as potential candidates for the new samples. As with Chapter 6, performance of the design methodology is assessed through a partial redesign of the initial sampling for each data set. For the purposes of demonstrating the methodology, a random subset of points was removed from the existing spatial design. Subsequently, the same number of points were added back into the reduced data set from potential candidates using the proposed optimal design. Expected prediction quantile intervals over the interpolation grid are compared for the existing spatial design and the redesigned spatial design using the proposed methodology.

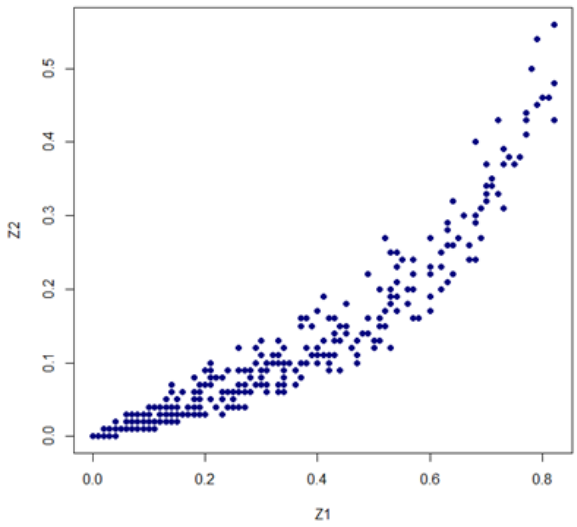
7.4.1 Simulation study for spatial redesign

The performance of the proposed methodology is assessed by comparing the redesigned spatial design to the existing spatial design through a simulation study



(a)

(b)



(c)

Figure 7.2: BEF data. Spatial plots for (a) Z_1 and (b) Z_2, and (c) scatter plot between Z_1 and Z_2.

similar to Chapter 6, which is based on the approach of Li et al. [2011]. The procedure for the simulation study for the bivariate context is outlined in the following steps.

1. Randomly remove p observed locations from the original observation set X^0 to produce a reduced data set X.

2. For each of the m candidate points, calculate the total expected weighted PQI over the interpolation grid after adding the candidate location as a

possible new location to the reduced data set X.

3. Select the p locations that produce the lowest total expected weighted PQIs as the new sampling locations. Let the p new sampling locations be denoted by $S = (s_1, \ldots, s_p)$.

4. For the reduced data set X, transform $Z = [Z_1, Z_2]$ into uncorrelated factors $M = [M_1, M_2]$ using the transformation described in Eq. (7.1).

5. From the p new sampling locations, randomly select one location s_i without replacement. For each factor M_1 and M_2, separately, fit a conditional copula at s_i, conditioned on the reduced data set X. From each conditional copula for M_1 and M_2, obtain a random value using Monte Carlo simulation and assign this value to the location s_i. Apply the back transformation of the transformation used in step 4 to obtain the corresponding Z_1 and Z_2 values. Add location s_i, with assigned values for M_1 and M_2, and their corresponding Z_1 and Z_2 values, to the reduced data set X.

6. Repeat steps 4 to 5 a further $p - 1$ times to obtain p simulated values for each new sampling location.

7. Repeat steps 4 to 6, 100 times to obtain 100 sequential simulations. This results in 100 data sets for M_1 and M_2.

8. For each simulated data set, sum the weighted PQI for Z_1 and Z_2, given in Eq. (7.7), over the interpolation grid to give the total weighted PQI. Sort the total weighted PQIs in increasing order to form the set $PQI_T = (PQI_1, \ldots, PQI_{100})$, where $PQI_j < PQI_{j+1}$ for $j = 1, \ldots, 99$.

9. Calculate the total weighted PQI for Z_1 and Z_2 over the interpolation grid using the original set of observations X^0 and let this be denoted by PQI_0. In order to account for the bivariate relationship between Z_1 and Z_2, the total weighted PQI for the original set of observations should be calculated by first transforming $Z = [Z_1, Z_2]$ to $M = [M_1, M_2]$ using the same transformation method as in step 4. Thereafter, the predictive quantiles for Z_1

and Z_2 can be obtained using the back transformation of the predictive quantiles for M_1 and M_2.

10. Compare the total weighted PQI from the original observations PQI_0 with the total PQIs from the simulated data sets $PQI_1, \ldots, PQI_{100}$ and observe the number of total PQIs from the simulated data sets that are less than the total PQI from the original observations. If $PQI_j < PQI_0 < PQI_{j+1}$, then the proportion of sequential simulations that have a lower total weighted PQI than the total PQI of the original observations X^0 is $j/100$.

As in the simulation study in Chapter 6, the selection of p new locations in not sequential. However, the proposed methodology specifies sequential addition of new locations. The approach for sequential addition of new locations in a simulation study discussed in Chapter 6 is computationally intensive for univariate designs. This is more so the case for multivariate designs. However, Chapter 6 also demonstrated that, in the univariate setting, the sequential design is similar to the non-sequential design. Hence assessment of the multivariate design methodology is conducted using the non-sequential design approach in this chapter.

7.4.2 Swiss Jura data

Figure 7.3 shows the 250m by 250m interpolation grid that was defined over the study domain. This is a replication of Figure 6.2 from Chapter 6, which appears here for ease of reference. There are 196 grid points. The interpolation grid points are also considered as the potential candidates for the new samples. Twenty observations were removed randomly from an existing spatial design with 259 observations, based on the design in Atteia et al. [1994]. Subsequently, 20 design points were added back into the reduced data set from potential candidates using the proposed optimal design. The hollow red squares in Figure 7.3 denote the 20 observations that were removed from the original 259 observations.

As seen in Figure 7.1(c), Co and Ni have a strong linear relationship. Hence, PCA was applied to remove the correlation between the variables at zero lag distance.

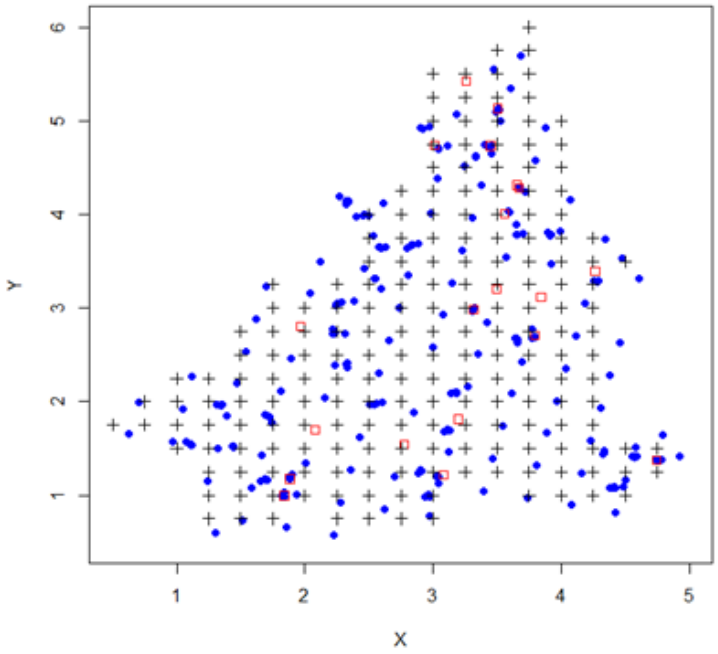


Figure 7.3: Swiss Jura data. Study domain with retained old locations (blue dots) and removed locations (hollow red squares) for both Co and Ni. Interpolation locations are denoted by black crosses.

Thereafter, the second rotation of MAF was used to remove cross-correlation between variables at lag distances greater than zero. Here, model based MAF was used. To perform model based MAF, the variogram and co-variogram should be modelled by the two structure linear model of coregionalisation (LMC). Thus, the LMC used by Bandarian et al. [2008] was used to model the dependence of Co and Ni here.

Since the concentration of Co and Ni are measured using the same units and both variables are equally important in the study, the standard deviation σ_k is not used in Eq. (7.7) and Co and Ni are assigned equal weights w_k.

Comparison of linear multivariate co-kriged and pair-copula models

The weighted co-kriging variance maps of Co and Ni concentrations at each interpolation grid point, for the reduced data set with 239 observations, are presented in Figures 7.4(a) and Figures 7.4(b). The weighted co-kriging variance is the weighted average of the co-kriging variance for Co and Ni. Figure 7.4(a) is overlaid with the spatial distribution of Co while Figure 7.4(b) is overlaid with the

spatial distribution of Ni.

Figures 7.4(c) and 7.4(d) show the maps for the widths of the weighted 90% prediction interval for Co and Ni. The weighted 90% prediction interval is the weighted average of the 90% PQIs for Co and Ni, for the reduced data set with 239 observations. A 90% PQI corresponds to the difference between the 95-th and 5-th predictive quantiles. Figure 7.4(c) is overlaid with the spatial distribution of Co while Figure 7.4(d) is overlaid with the spatial distribution of Ni.

Similar to the univariate case in Chapter 6, Figures 7.4(c) and 7.4(d) indicate that wider weighted 90% prediction intervals from the pair-copula models under the multivariate framework correspond both to areas that are more sparsely sampled as well as areas with high variability in metal concentrations for both variables. Note that the areas of high variability in metal concentrations are similar for Co and Ni. This is because low values of Co and Ni occur together and high values occur together. This is also apparent from Figure 7.1(c), which additionally indicates a linear relationship between Co and Ni concentrations. Hence, when the proposed design methodology is implemented, the locations for new observations will occur in areas that are sparsely sampled and areas with with high variability in both metal concentrations. This tallies with the aim of reducing the variability of both variables simultaneously.

From Figures 7.4(a) and 7.4(b), weighted co-kriging variances are higher in more sparsely sampled areas. However, unlike the weighted prediction intervals from the pair-copula model, the weighted co-kriging variance doesn't capture the variability in metal concentrations. As a result, when a multivariate co-kriging based design is implemented, new observations will be located in areas that are sparsely sampled, regardless of the metal concentration values.

Simulation study for linear Swiss Jura data

Figures 7.5(a) and 7.5(b) show the map of the 196 total expected weighted PQIs for Co and Ni that are obtained for the 196 candidate locations, as detailed in step 2 of the simulation study procedure. Figure 7.5(a) is overlaid with the spatial

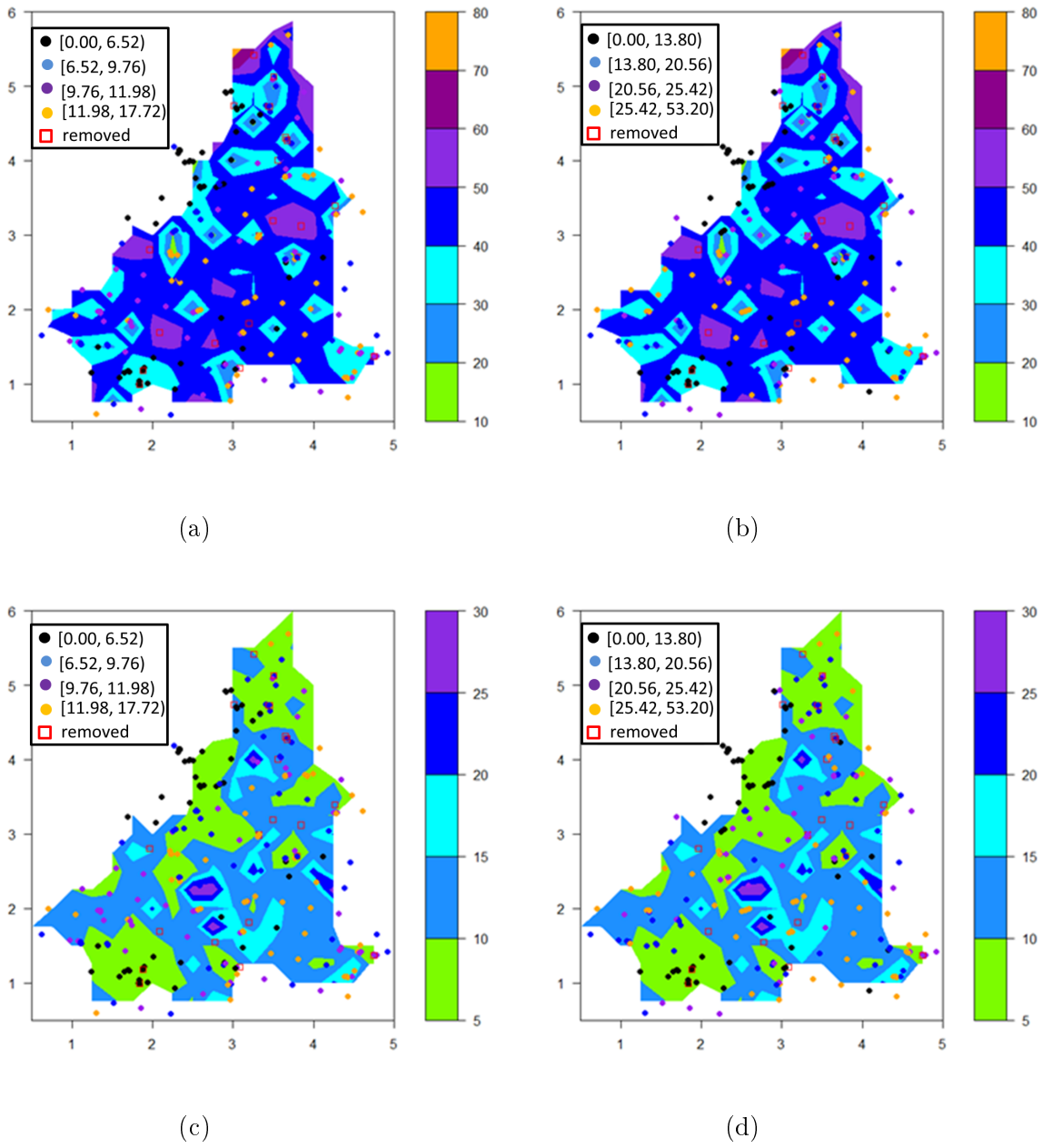


Figure 7.4: Maps for the (a) weighted co-kriging variance of Co and Ni overlaid with the spatial distribution of Co, (b) weighted co-kriging variance of Co and Ni overlaid with the spatial distribution of Ni, (c) widths of the weighted 90% prediction intervals for Co and Ni overlaid with the spatial distribution of Co and (d) widths of the weighted 90% prediction intervals for Co and Ni overlaid with the spatial distribution of Ni. Retained observations are displayed as dots and removed observations are hollow red squares.

distribution of Co while Figure 7.5(b) is overlaid with the spatial distribution of Ni. As determined by the simulation procedure above, the new sampling locations (solid red squares) are located in regions corresponding to lower values of total expected weighted PQI.

Figures 7.5(c) and 7.5(d) are the maps for the widths of the weighted 90% prediction intervals for Co and Ni, for the reduced data set with 239 observations. Figure 7.5(c) is overlaid with the spatial distribution of Co while Figure 7.5(d) is overlaid with the spatial distribution of Ni.

As expected, comparing Figures 7.5(a) and 7.5(b) with Figures 7.5(c) and 7.5(d), the areas with low total expected weighted PQI correspond to areas with wide weighted prediction intervals. It was commented previously that these are areas that are more sparsely sampled and with high variability in both metal concentrations.

Figure 7.6(a) shows the distribution of the total weighted PQIs for Co and Ni, which were obtained by applying steps 4 to 8 of the simulation study procedure, for the 100 different realisations of the redesigned spatial design.

The total weighted PQI of the original 259 observations is represented by the value in bold on the x-axis. The redesigned spatial design outperforms the original spatial design, that is, the simulated total weighted PQIs are less than the PQI of the original observations, in 99% of the simulations.

The proposed multivariate optimal design methodology reduces prediction uncertainty simultaneously for all variables by minimising the weighted average of the PQIs. However, the resultant design may be sub-optimal for an individual variable, where interest is in minimising the PQI just for that variable.

To assess whether the design points from the multivariate design are optimal for reduction in prediction uncertainty of Co alone, the total PQI for Co was obtained using steps 4 to 8 of the simulation study procedure, for the 100 different realisations of the redesigned spatial design. The total PQI for Co is simply the PQI for Co summed over the interpolation grid, that is, it is the total weighted PQI with weight $w_k = 0$ for Ni. The total PQI for Co for the original 259 observations was also obtained by setting $w_k = 0$ for Ni. Figure 7.6(b) shows the distribution of the total PQIs for Co for the 100 different realisations of the redesigned multivariate spatial design, with the total PQI for Co of the original 259 observations represented on the x-axis in bold. The simulated total PQIs

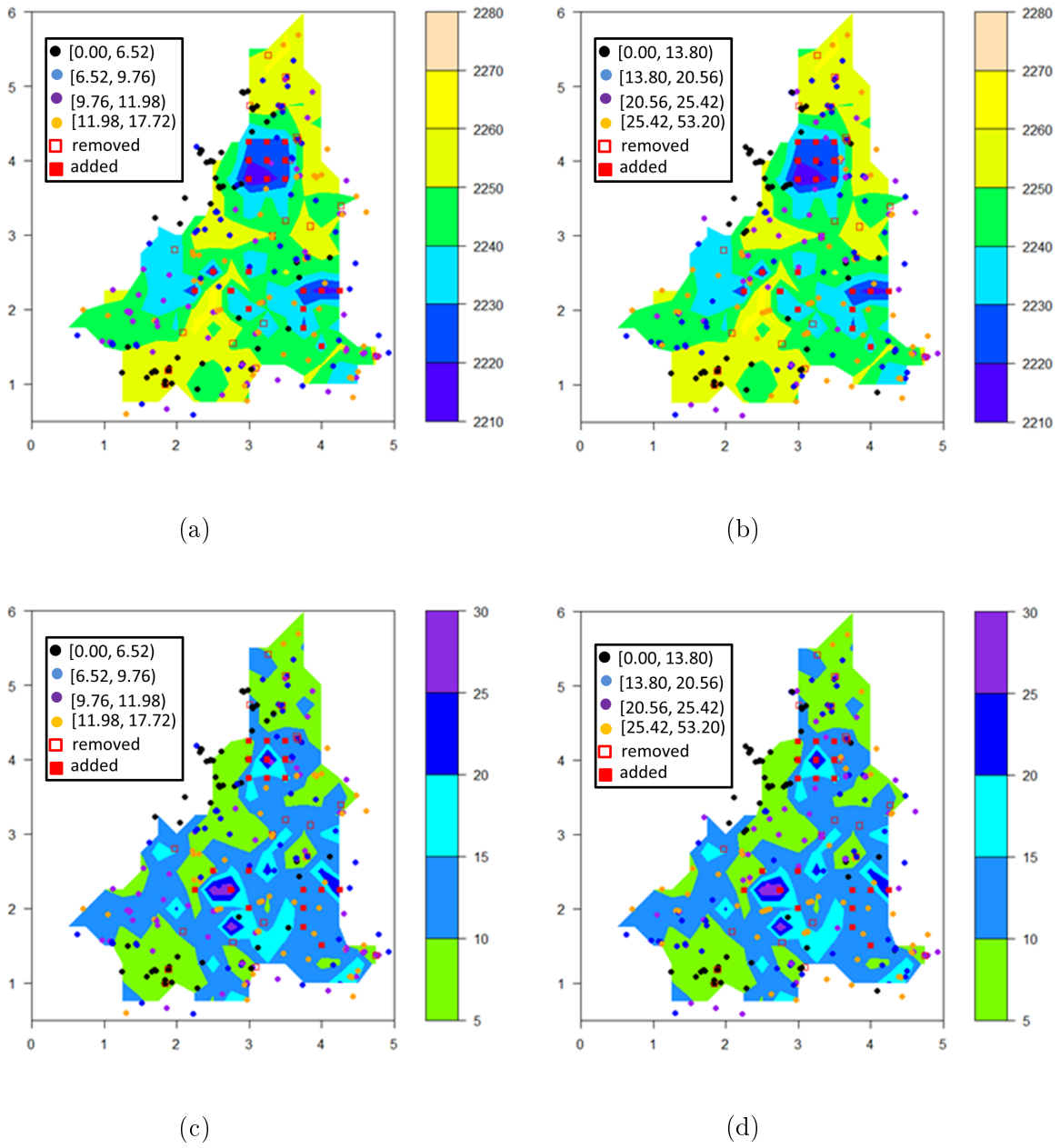


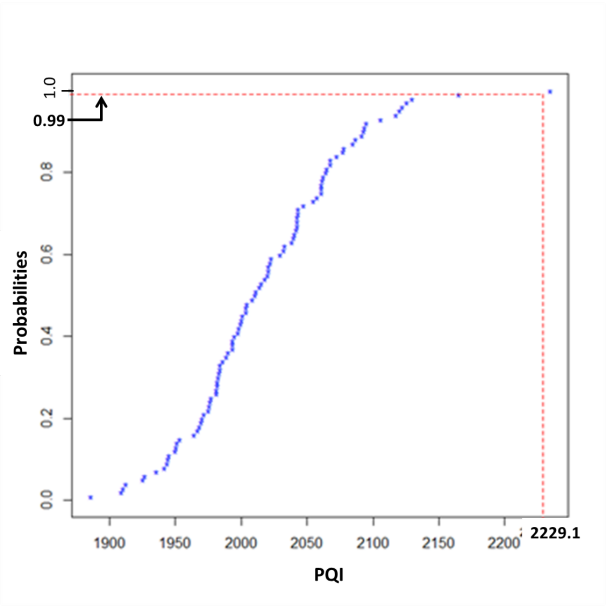
Figure 7.5: Maps for the (a) total expected weighted PQI for Co and Ni overlaid with the spatial distribution of Co, (b) total expected weighted PQI for Co and Ni overlaid with the spatial distribution of Ni, (c) widths of the weighted 90% prediction intervals for Co and Ni overlaid with the spatial distribution of Co and (d) widths of the weighted 90% prediction intervals for Co and Ni overlaid with the spatial distribution of Ni. Retained observations are displayed as dots, removed observations are hollow red squares and newly added locations are solid red squares.

for Co are less than the PQI for Co of the original observations in 99% of the simulations.

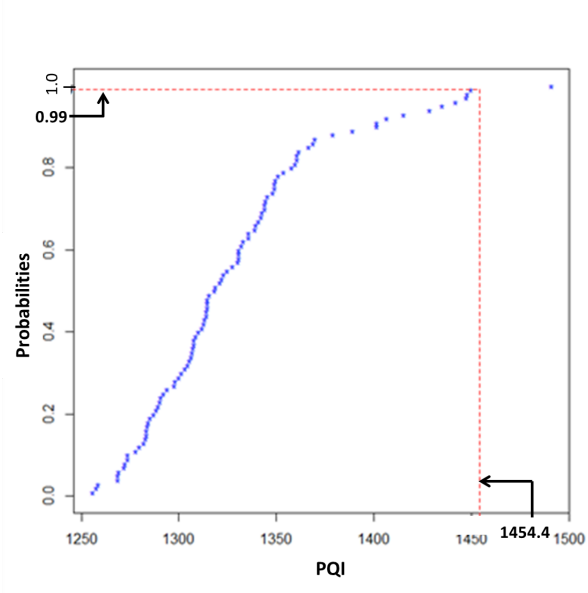
The total PQIs for Ni for the 100 different realisations of the redsigned multivariate spatial design and the PQI for Ni of the original 259 observations can be

found in a similar way to Co, by setting $w_k = 0$ for Co. Figure 7.6(c) indicates that the simulated total PQIs for Ni are less than the PQI for Ni of the original observations in 98% of the simulations.

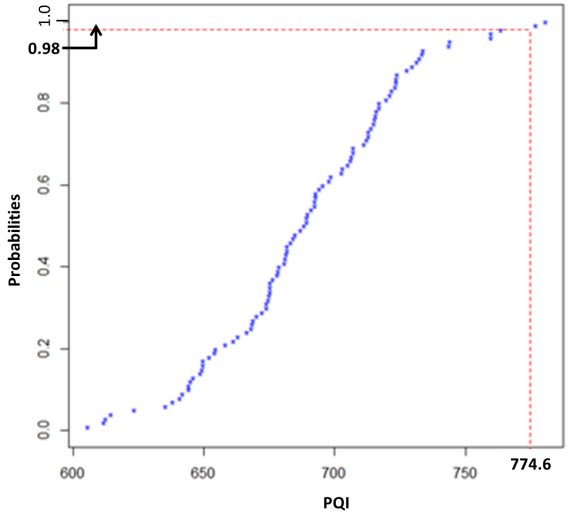
Hence, for this application, the optimal design points obtained in order reduce the prediction uncertainty simultaneously for Co and Ni are also optimal for Co and Ni separately. This is because the areas of high variability in metal concentrations are similar for Co and Ni, that is, because the Co and Ni concentrations have a positive linear relationship.



(a)



(b)



(c)

Figure 7.6: Distribution of (a) total weighted PQI for Co and Ni, (b) total PQI for Co and (c) total PQI for Ni, from 100 simulated data sets.

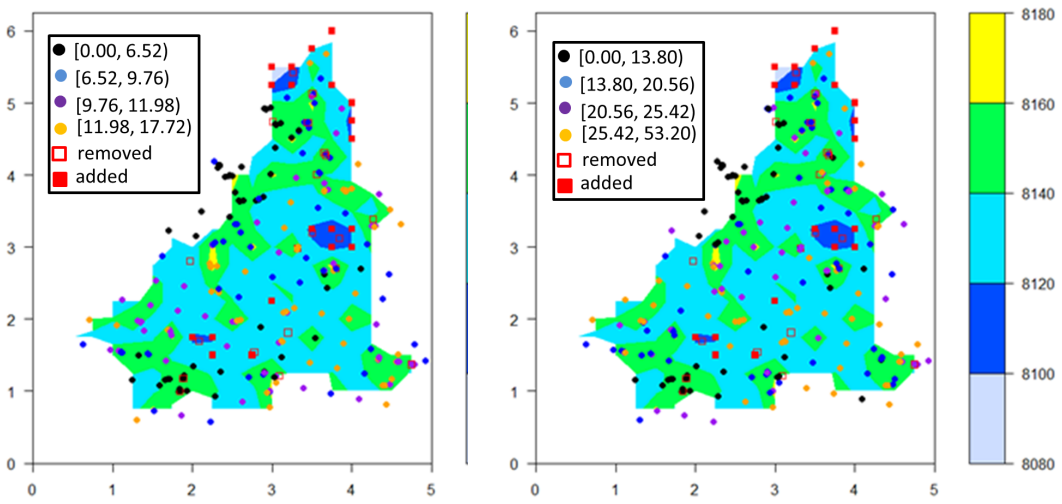
Co-kriging based multivariate design

Here, optimal bivariate designs, for variables that are bivariately linear, are compared for the design based on pair-copula models and the design based on co-kriging models. The purpose of such a comparison is to investigate how the optimal bivariate designs vary depending on the ability, or lack thereof, of the modelling approach to capture spatial non-linearity within individual variables. Total weighted co-kriging variance over the interpolation grid is used as the optimisation criterion for the co-kriged based design such that a candidate location that produces the smallest total weighted co-kriging variance is selected as the new sampling location. The LMC used by Bandarian et al. [2008] was used to model the spatial dependence of Co and Ni.

Figure 7.7(a) and 7.7(b) show the 196 weighted co-kriging variances obtained for the 196 candidate locations. Figure 7.7(a) is overlaid with the spatial distribution of Co while Figure 7.7(b) is overlaid with the spatial distribution of Ni. The solid red squares are the 20 new locations from the co-kriged based design. From Figures 7.7(a) and 7.7(b), the new locations are located in areas with a lower density of observed points, as would be expected, since areas with less observations correspond to larger weighted co-kriging variances. Unlike the design based on the proposed methodology, which uses pair copulas, the new sampling locations for Co and Ni for the co-kriged based design do not depend on the values of the observations for Co and Ni.

7.4.3 Bartlett Experimental Forest data

Figure 7.8 shows the 125km by 125km interpolation grid that was defined over the study domain. There are 492 grid points that are also considered as the potential candidates for the new sampling locations. Forty-eight observations were removed randomly from the existing spatial design with 437 observations. Subsequently, 48 design points were added back into the reduced data set from potential candidates using the proposed optimal multivariate design. The hollow red squares denote the 48 observations that were removed from the original 437



(a) (b)

Figure 7.7: Co-kriging based optimal bivariate design for Co and Ni overlaid with the spatial distribution of (a) Co and (b) Ni. Retained observations are displayed as dots, removed observations are hollow red squares and newly added locations are solid red squares.

observations.

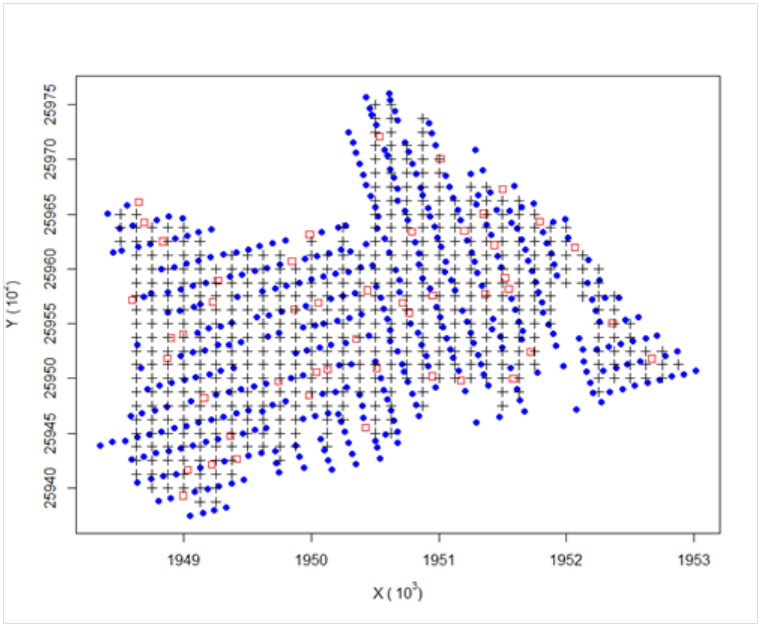


Figure 7.8: Bartlett Experimental Forest data. Study domain with retained old locations (blue dots) and removed locations (hollow red squares) for both Z_1 and Z_2. Interpolation locations are denoted by black crosses.

As seen in Figure 7.2(c), Z_1 and Z_2 are non-linearly related. To obtain uncorrelated factors at all lag distances, NLPCA and partial MAF (the second rotation of MAF) transformations were applied to the data for Z_1 and Z_2. NLPCA,

which was developed using an artificial neural network, was applied to remove the correlation between the variables at zero lag distance. Cross-correlation for lag distances greater than 150km were also removed indirectly. Thereafter, the second rotation of data driven MAF was used to remove cross-correlation between variables at a lag distance of 150km, which indirectly removed all remaining cross-correlation below this lag distance. Details on these transformation methods can be found in Chapter 4.

Since the values of Z_1 and Z_2 are measured using the same units and both variables are equally important in the study, the standard deviation σ_k is not used in Eq. (7.7) and Z_1 and Z_2 are assigned equal weights w_k.

Comparison of linear and non-linear multivariate pair-copula models

To fit a linear multivariate pair-copula model to Z_1 and Z_2, which ignores the fact that the relationship between Z_1 and Z_2 is actually non-linear, PCA and partial MAF (the second rotation of MAF) transformations were applied to the data for Z_1 and Z_2.

The maps for the widths of the weighted 90% prediction intervals for Z_1 and Z_2, based on the linear multivariate pair copula model, at each interpolation grid point are presented in Figures 7.9(a) and 7.9(b). The weighted 90% prediction interval is the weighted average of the 90% PQIs for Z_1 and Z_2, for the reduced data set with 389 observations. A 90% PQI corresponds to the difference between the 95-th and 5-th predictive quantiles. Figure 7.9(a) is overlaid with the spatial distribution of Z_1 while Figure 7.9(b) is overlaid with the spatial distribution of Z_2.

Figures 7.9(c) and 7.9(d) show the maps for the widths of the weighted 90% prediction intervals for Z_1 and Z_2 based on the non-linear multivariate pair-copula model, which uses NLPCA and partial MAF transformations.

The difference between the weighted 90% prediction interval map in Figures 7.9(a) and 7.9(b) and the weighted 90% prediction interval map in Figures 7.9(c) and 7.9(d) arises from the assumed linear, or non-linear, relationship between

Z_1 and Z_2.

Figure 7.2(c) shows the actual non-linear relationship between Z_1 and Z_2. Whilst low values of Z_1 and Z_2 occur together, and high values of Z_1 and Z_2 occur together, the non-linear relationship indicates that moderate values of Z_1 occur with moderately low values of Z_2. This can be seen in comparing Figures 7.9(c) and 7.9(d).

However, in Figures 7.9(a) and 7.9(b), moderate values of Z_1 occur with moderate values of Z_2 because of the assumed linear relationship between Z_1 and Z_2.

Since the weighted prediction intervals from pair-copula models depend on the values of Z_1 and Z_2, it is expected that weighted prediction interval maps will differ for different assumed bivariate relationships.

Hence, when the proposed design methodology is implemented, the locations for the new observations will occur in areas that are sparsely sampled and where the Z_1 and Z_2 values correspond to wide weighted prediction intervals. The Z_1 and Z_2 values in these areas depend on the relationship between Z_1 and Z_2.

Simulation study for BEF data under a non-linear bivariate model

Figures 7.10(a) and 7.10(b) show the map of the 492 total expected weighted PQIs for Z_1 and Z_2 that are obtained for the 492 candidate locations, as detailed in step 2 of the simulation study procedure. The total expected weighted PQIs are based on the non-linear multivariate pair-copula model. Figure 7.10(a) is overlaid with the spatial distribution of Z_1 while Figure 7.10(b) is overlaid with the spatial distribution of Z_2. As determined by the simulation procedure, the new sampling locations (sold red squares) are located in regions corresponding to lower values of total expected weighted PQI.

Figures 7.10(c) and 7.10(d) show the maps for the widths of the weighted 90% prediction intervals for Z_1 and Z_2, for the reduced data set with 389 observations. The weighted 90% prediction intervals are also based on the non-linear multivariate pair-copula model. Figure 7.10(c) is overlaid with the spatial distribution of Z_1 while Figure 7.10(d) is overlaid with the spatial distribution of Z_2.

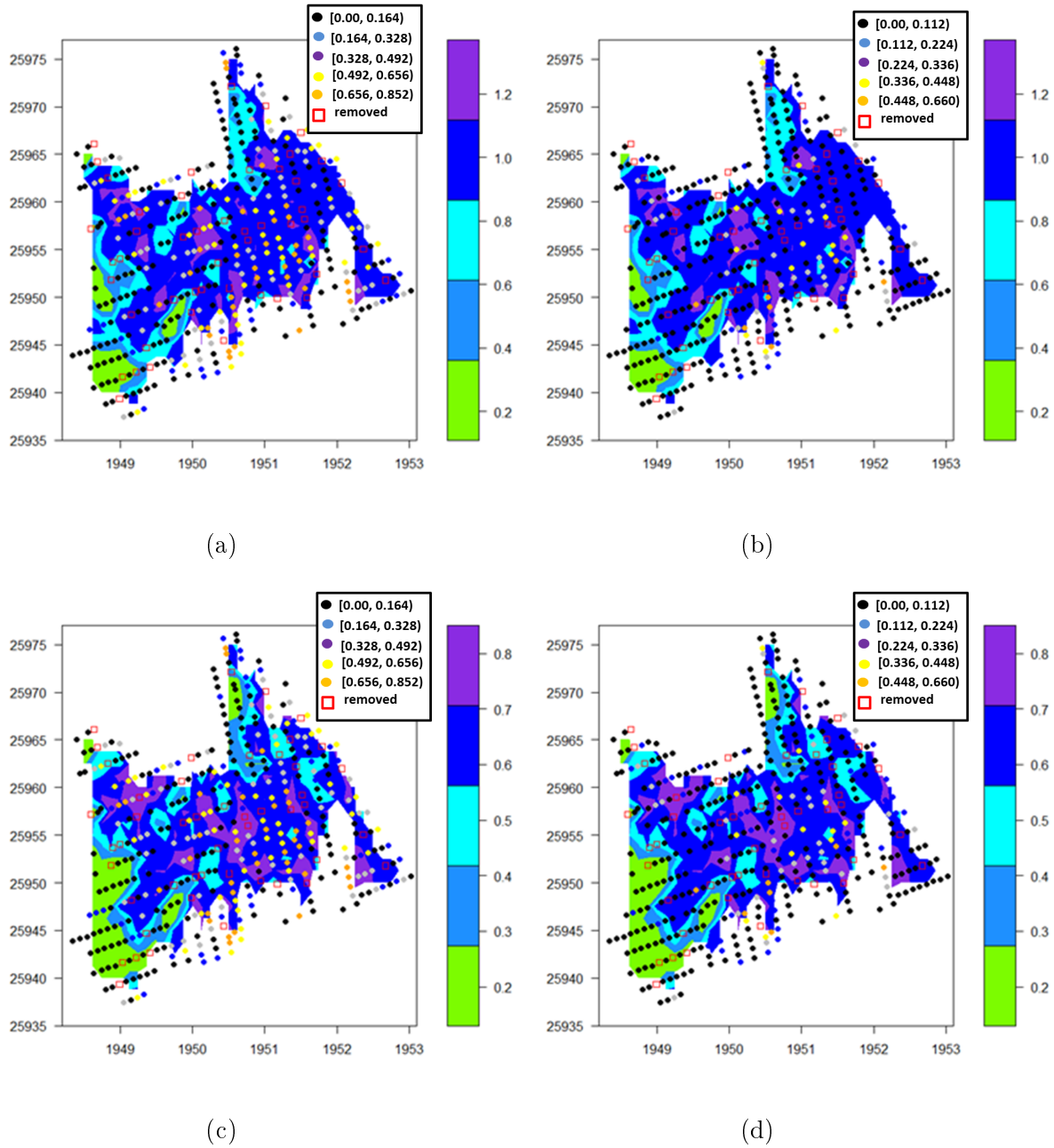


Figure 7.9: Maps for the widths of the weighted 90% prediction intervals for Z_1 and Z_2 based on the linear multivariate pair copula model overlaid with the spatial distribution of (a) Z_1 and (b) Z_2, and the widths of the weighted 90% prediction intervals for Z_1 and Z_2 based on the non-linear multivariate pair-copula model overlaid with the spatial distribution of (a) Z_1 and (b) Z_2. Retained observations are displayed as dots and removed observations are hollow red squares.

As expected, comparing Figures 7.10(a) and 7.10(b) with Figures 7.10(c) and 7.10(d), the areas with low total expected weighted PQI correspond to areas with wide weighted prediction intervals. It was commented previously that these are areas that are more sparsely sampled and with Z_1 and Z_2 values that depend on the relationship between Z_1 and Z_2.

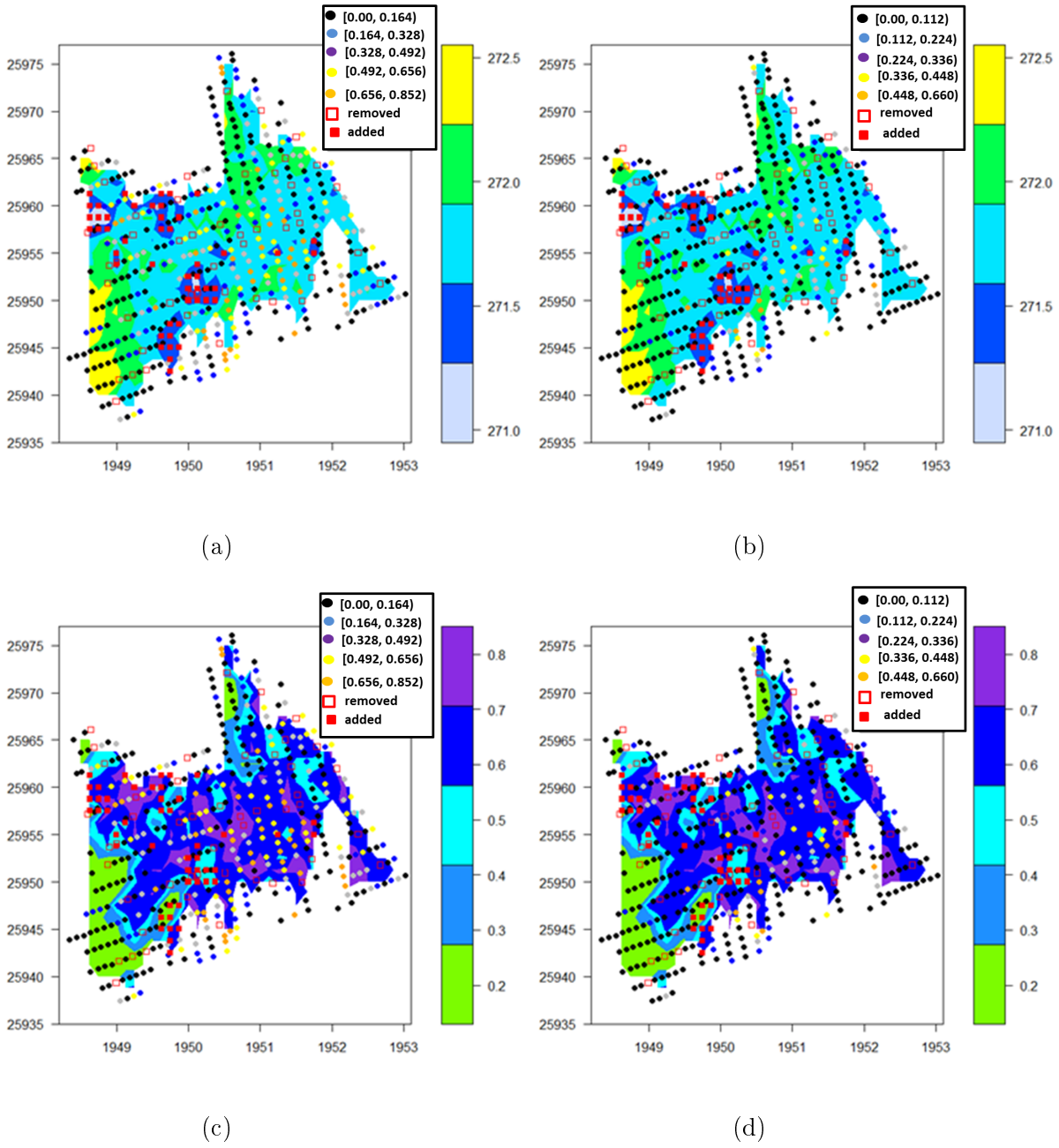


Figure 7.10: Maps for the total expected weighted PQI for Z_1 and Z_2 based on the non-linear multivariate pair copula model overlaid with the spatial distribution of (a) Z_1 and (b) Z_2, and the widths of the weighted 90% prediction intervals for Z_1 and Z_2 based on the non-linear multivariate pair copula model overlaid with the spatial distribution of (c) Z_1 and (d) Z_2. Retained observations are displayed as dots, removed observations are hollow red squares and newly added locations are solid red squares.

Figure 7.11(a) shows the distribution of the total weighted PQIs for Z_1 and Z_2, which were obtained by applying steps 4 to 8 of the simulation study procedure, for the 100 different realisations of the redesigned spatial design.

The total weighted PQI of the original 389 observations is represented by the

value in bold on the x-axis. The redesigned spatial design outperforms the original spatial design, that is, the simulated total weighted PQIs are less than the PQI of the original observations, in 92% of the simulations.

Additionally, the multivariate design for simultaneous reduction in prediction uncertainty of Z_1 and Z_2 was assessed to see if it was also optimal for Z_1 alone and Z_2 alone using a similar approach to that of the Swiss Jura application.

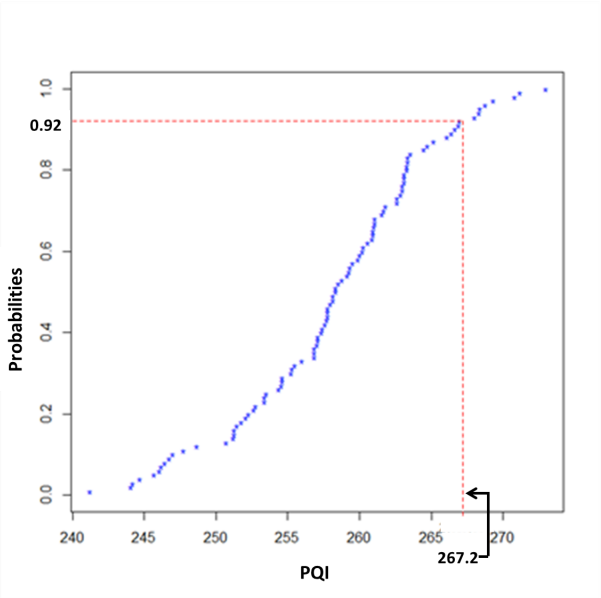
Figures 7.11(b) and 7.11(c) indicate that the redesigned spatial design outperforms the original spatial design in 93% of simulations for Z_1 and 90% of simulations for Z_2. Hence, for this application, the optimal design points obtained in order reduce the prediction uncertainty simultaneously for Z_1 and Z_2 are, in most simulations, optimal for Z_1 and Z_2 separately.

Simulation study for BEF data under a linear bivariate model

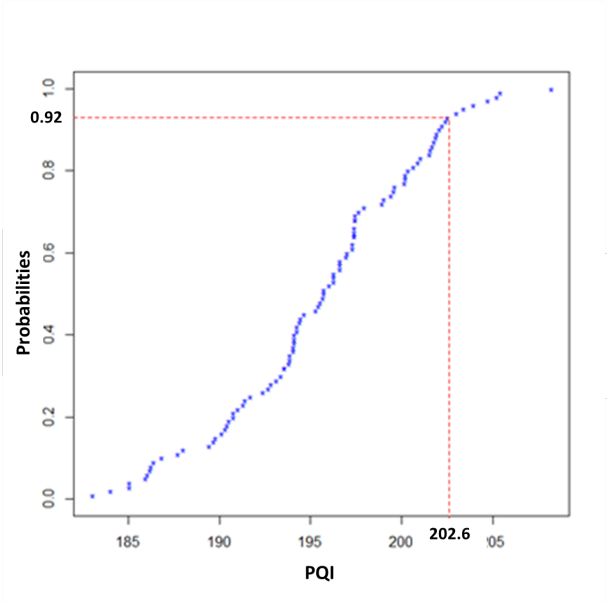
Here, optimal bivariate designs, for variables that are bivariately non-linear, are compared for the design based on the non-linear bivariate copula model and the design based on the linear bivariate copula model. The purpose of such a comparison is to investigate how the optimal bivariate designs vary depending on the ability, or lack thereof, of the modelling approach to capture the non-linear bivariate relationship between the variables.

Figures 7.12(a) and 7.12(b) show the map of the total expected weighted PQIs for Z_1 and Z_2 based on the linear multivariate pair-copula model. Figures 7.12(c) and 7.12(d) show the map for the widths of the weighted 90% prediction intervals for Z_1 and Z_2 based on the linear multivariate pair-copula model. Figures 7.12(a) and 7.12(c) are overlaid with the spatial distribution of Z_1 while Figures 7.12(b) and 7.12(d) is overlaid with the spatial distribution of Z_2. As with the design based on the non-linear multivariate pair-copula model, the new sampling locations (solid red squares) are located in regions corresponding to lower values of total expected weighted PQI. These regions correspond to areas with wide prediction intervals.

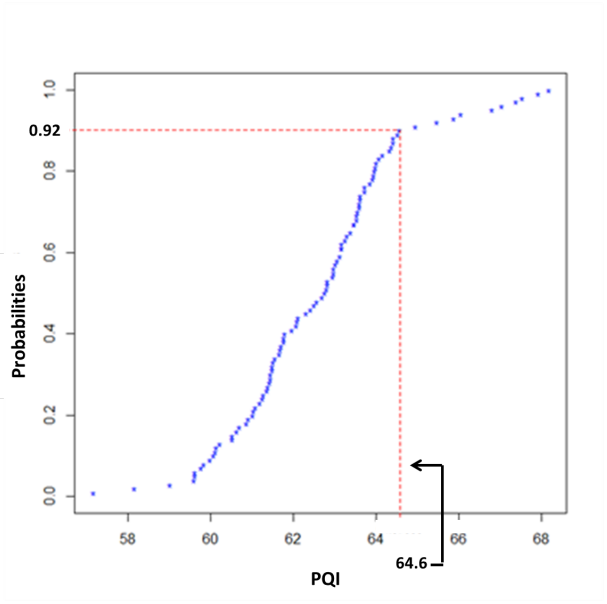
However, the design based on the non-linear multivariate copula model (Fig-



(a)



(b)



(c)

Figure 7.11: Distribution of (a) total weighted PQI for Z_1 and Z_2, (b) total PQI for Z_1 and (c) total PQI for Z_2, from 100 simulated data sets.

ure 7.10) differs to the design based on the linear multivariate copula model (Figure 7.12). This is because the new sampling locations for designs based on multivariate copula models depend on the observed values of Z_1 and Z_2 and hence differs if the relationship between Z_1 and Z_2 differs. That is, the design is model dependent and is optimal for the model used.

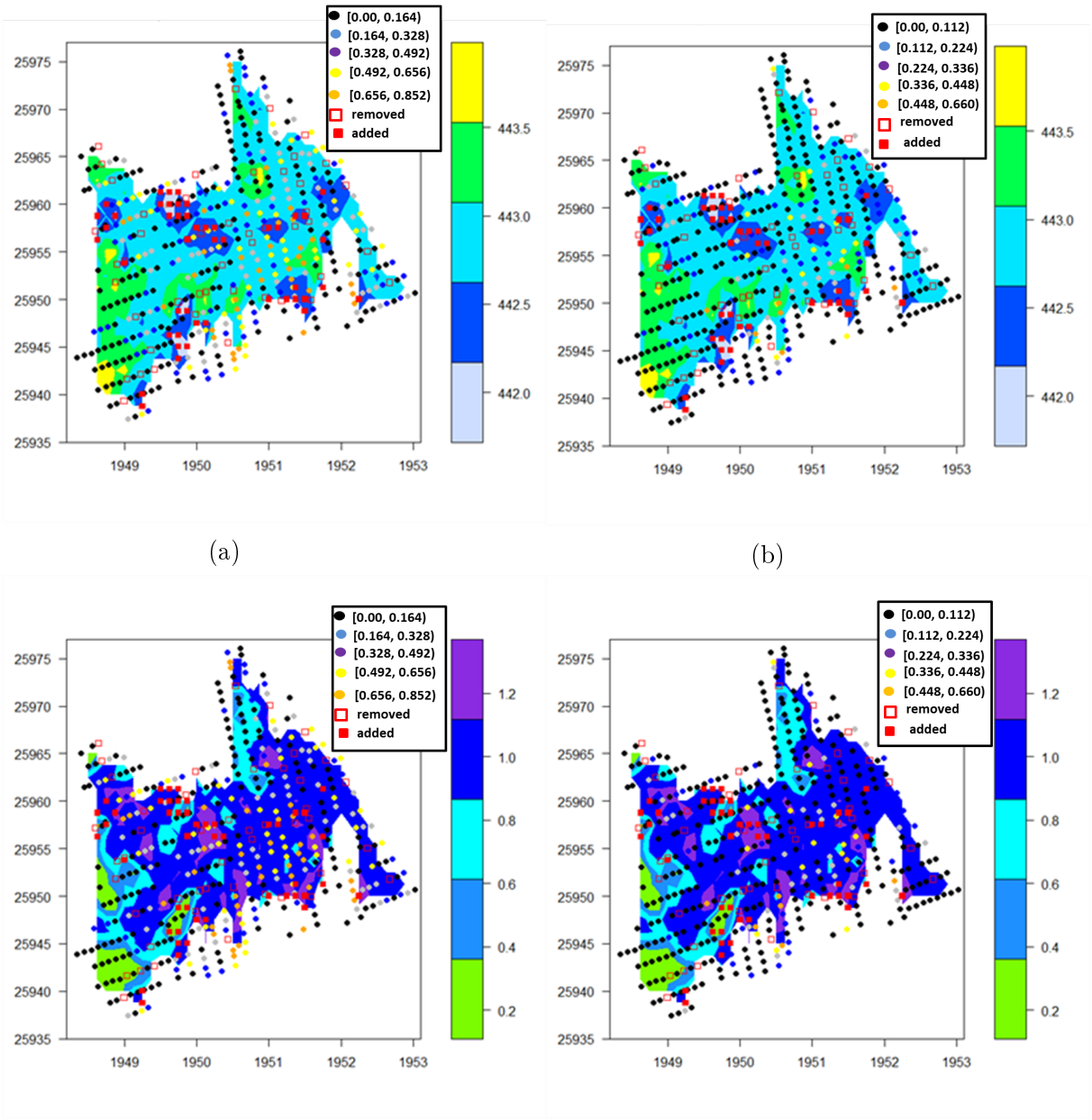
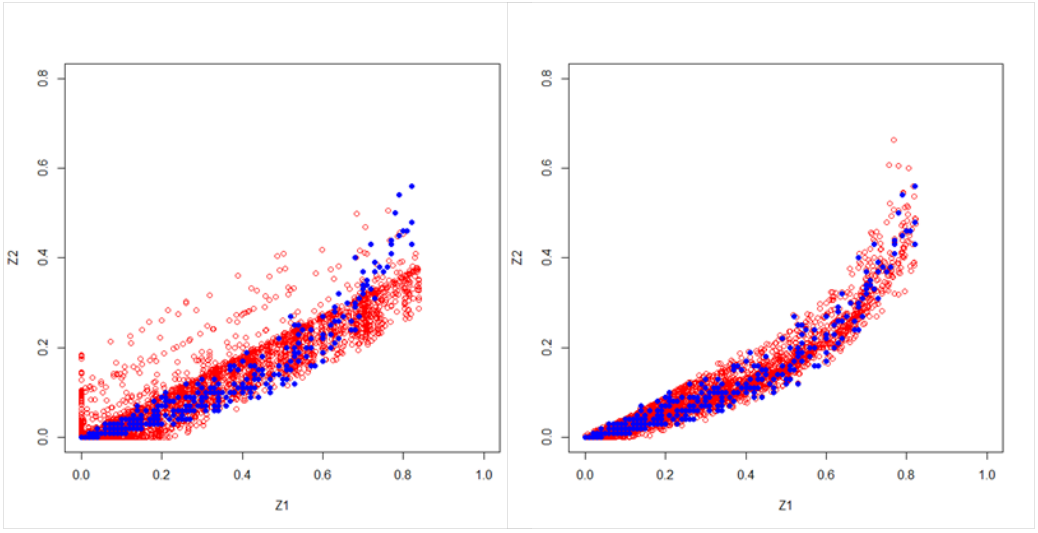


Figure 7.12: Maps for the total expected weighted PQI for Z_1 and Z_2 based on the linear multivariate pair copula model overlaid with the spatial distribution of (a) Z_1 and (b) Z_2, and the widths of the weighted 90% prediction intervals for Z_1 and Z_2 based on the linear multivariate pair copula model overlaid with the spatial distribution of (c) Z_1 and (d) Z_2. Retained observations are displayed as dots, removed observations are hollow red squares and newly added locations are solid red squares

Comparison of linear and non-linear multivariate designs

Figure 7.13(a) displays the scatter plot of Z_1 against Z_2 for the 100 simulated data sets obtained from the design based on a linear multivariate pair-copula model (red circles) overlaid with the original data (blue dots). It is clear that the

linear multivariate pair-copula model ignores the non-linearity present between the variables. Figure 7.13(b) shows the scatter plot for the 100 simulated data sets obtained from the design based on a non-linear multivariate pair-copula model (red circles) overlaid with the original data (blue dots). The non-linearity is captured reasonably well by the non-linear multivariate pair-copula model.

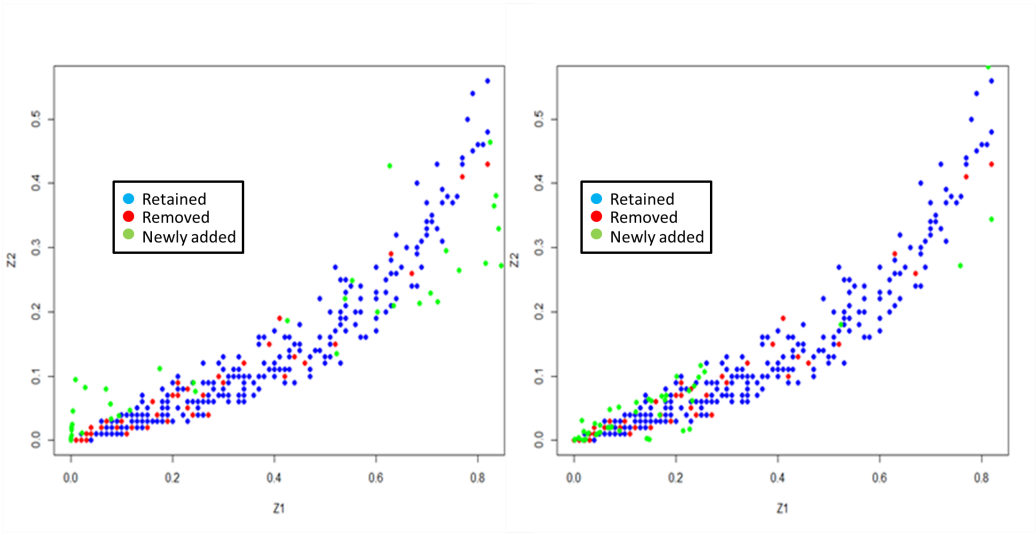


(a) (b)

Figure 7.13: Scatter plot of Z_1 against Z_2 for the 100 simulated data sets obtained from the (a) linear multivariate design and (b) non-linear multivariate design. Red circles denote simulated data and blue dots are the original data.

From the simulation studies, the average simulated observed value, over the 100 simulations, for each of the 48 newly added points was calculated. Figures 7.14(a) and 7.14(b) show the scatter plots of Z_1 against Z_2 for the retained observations (blue dots), removed observations (red dots) and the average values of the newly added locations, for the designs based on the linear and non-linear multivariate pair-copula models respectively.

From Figure 7.14(b), the newly added locations tend to take Z_1 and Z_2 values that correspond to the non-linear relationship of the observed data. Additionally, the number of newly added locations is greater where the relationship between Z_1 and Z_2 changes the most. Hence the locations of the new design points based on the non-linear multivariate pair-copula model appear to correspond to values of Z_1 and Z_2 that contribute to more accurate estimation of the true non-linear relationship between Z_1 and Z_2.



(a) (b)

Figure 7.14: Scatter plot of Z_1 against Z_2 obtained from the (a) linear multivariate design and (be) non-linear multivariate design. Blue dots are the retained observations, red dots are the removed observations and green dots are the simulated values of the newly added locations averaged over the 100 simulated data sets.

In Figure 7.14(a), the newly added locations tend to take Z_1 and Z_2 values that correspond to a linear relationship of the observed data, and the number of new locations is evenly spread across this linear relationship. Hence the locations of the new design points based on the linear multivariate pair-copula model appear to correspond to values of Z_1 and Z_2 that contribute to more accurate estimation of a linear relationship between Z_1 and Z_2, even though the observed relationship is clearly non-linear.

Using a design based on a non-linear multivariate pair-copula model when the multivariate relationship is non-linear has the advantage of being able capture the non-linear relationship and select new locations that contribute to more accurate estimation of the non-linear relationship. If the non-linear relationship is ignored in determining new sampling locations, the simultaneous reduction in the prediction uncertainty of the variables may be largely unaffected but the uncertainty of the multivariate relationship may possibly increase.

7.5 Discussion

In chapter 6, the ability of the univariate pair-copula model in capturing the variability of measured values of individual variables was demonstrated. This feature is also apparent in the multivariate modelling approach due to the use of spatial pair-copulas, as can be seen in the 90% predictive interval maps. These same maps also show a decrease in prediction uncertainty of individual variables when jointly predicting all variables independently. That is, use of a multivariate model allows more information to be utilised, thus reducing prediction uncertainty.

The simulation studies show that the proposed optimal design outperforms the original design in more than 90% of simulations. Additionally, whilst the objective of the proposed design is to reduce prediction uncertainty for both variables at the same time in both case studies, the proposed design points are also good design points for reduction of prediction uncertainty of the individual variables separately.

Furthermore, through this analysis, sensitivity of the proposed design to the non-linearity between variables and within variables was also demonstrated. The first case study shows the difference between design points obtained based on a model which ignores the non-linearity in individual variables (co-kriging) and design points obtained based on the model that can capture the non-linearity in individual variables (pair-copula model with MAF transformation). The second case study shows the difference between the design based on a model that can capture the non-linearity between variables compared to the design based on a model that cannot capture the non-linearity between variables.

In both case studies, it can be seen that most of the newly suggested sampling points are clustered together. One may think, by using this kind of design, information may be redundant. It should be noted that, in both case studies, the optimal sampling points were selected within one optimisation run. In other words, a non-sequential design approach was used to obtain the new sampling points. The non-sequential designs presented are purely to demonstrate the ap-

plication of the methodology and to show the potential of the methodology even when a non-sequential design is used. Practically, to obtain the optimal design, the proposed methodology should be applied in a sequential manner. Sequential addition of new observations is likely to produce designs with less clustered configurations. The sequential design is not applied here, because of the inability to obtain measurements for new sampling points.

7.6 Conclusions

In this chapter a new non-linear multivariate optimal spatial design methodology was proposed to reduce the prediction uncertainty of more one than variable simultaneously based on a pair-copula model and use of dimension reducing transformations. Even though bivariate case studies are used here for demonstration purposes, this methodology can be applied to ten dimensions. However this methodology would not be feasible to apply for more than ten dimensions due to increase of computational rapidly.

From the results, it can be concluded that, more precise predictions for variables under study can be obtained by adding additional samples that are determined by an optimal design based on a modelling approach that can honour the dependencies in the data. Furthermore, in this research a finite number of candidate locations was used to obtain the optimal design points. By using an optimisation technique, such as spatial simulated annealing, in the proposed methodology, any point in the study domain can be treated as a potential candidate point.

The proposed design approach cannot be directly applied when direct measurements are unable to be obtained after adding a new sample point. Even though a simulation-based sequential stochastic procedure can be applied in these kinds of situations, it would be very computationally expensive. However, as discussed in Li et al. [2011], an approach for selecting spatial blocks should be developed to overcome this problem. Moreover, here the optimal design was obtained by assuming the collocation of measurements, that is, measurements for all variables of the interests are obtained at each sampling point. There may be some situations

where collecting these kind of measurements is not feasible. For simulation-based design, this poses a problem that is to be addressed in future research. Moreover, in proposed design, cost contain didn't included. However cost constrains would be incorporated with proposed design methodology in future research as well.

Chapter 8

Discussion

8.1 Comparison of Univariate and Multivariate Pair-copula Modelling

Figures 8.1(a) and 8.1(c) show the maps for the widths of the 90% prediction intervals for Co and Ni respectively using the Swiss Jura data, discussed in Chapter 6, by applying the pair-copula model separately to Co and Ni, while Figures 8.1(b) and 8.1(d) represents the corresponding maps obtained by applying the pair-copula model in the multivariate setting with MAF used in the decorrelation transformation.

Figures 8.1(a) and 8.1(c) also appear in Chapter 6 but are repeated here for ease of reference. According to these two figures, areas with high variability in measured values and areas that are sparsely sampled have higher estimates of uncertainty from the pair-copula model. Hence, these figures demostrate that the uncertainty estimation produced by the pair-copula depends on both the observations' configuration and their values. That is, the pair-copula model has the ability to capture non-linear spatial dependence. A multivariate pair-copula model should also be able to capture non-linear dependence of spatial variables. The maps for the widths of the 90% prediction intervals for Co based on the univariate (Figure 8.1(a)) and multivariate (Figure 8.1(b)) pair-copula models are very similar. However, from Figures 8.1(c) and 8.1(d), it can be clearly

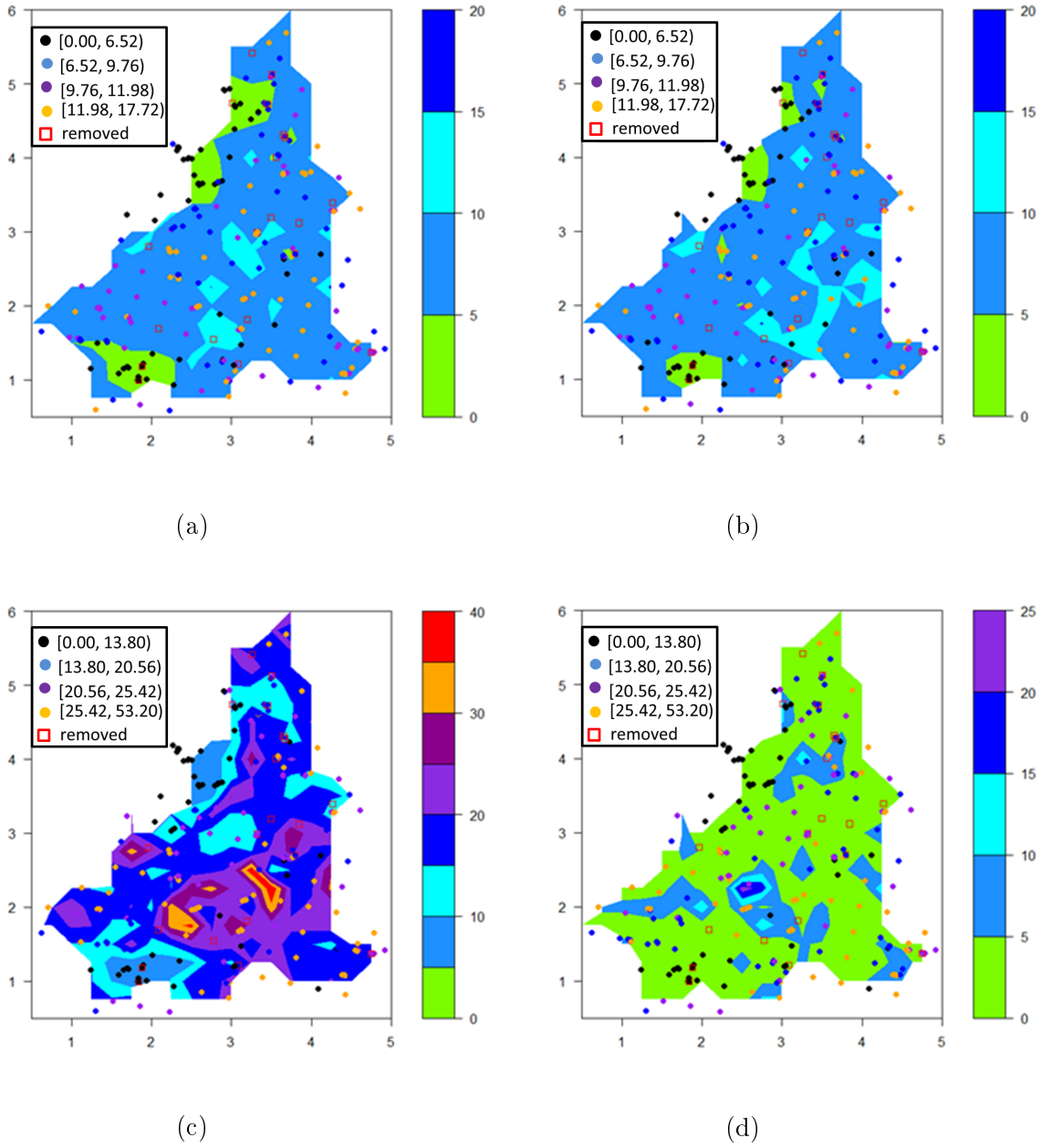


Figure 8.1: Maps for the widths of the 90% prediction intervasl for (a) Co based on the univariate pair-copula, (b) Co based on the multivariate pair-copula (c) Ni based on the univariate pair-copula and (d) Ni based on multivariate pair-copula.

seen that the range of the uncertainty estimation of the prediction for Ni is significantly reduced when the multivariate modelling approached is used. Also, uncertainty estimation of Ni based on the multivariate model also captures the non-linear dependence by producing higher uncertinity estimation for areas with high variability in measured values.

Figures 8.2(a) and 8.2(c) show the maps for the widths of the 90% prediction intervals for Z_1 and Z_2 respectively using the BEF data discussed in Chapter 5 by

applying pair-copula models separately for Z_1 and Z_2, while Figures 8.2(b) and 8.2(d) represent the corresponding maps obtained by applying the pair-copula model in the multivariate setting with NLPCA used in the decorrelation transformation.

According to Figures 8.2(a) and 8.2(b), the distribution of the 90% prediction interval widths for the univariate and multivarite pair-copula models are similar. However the range of the 90% predicton interval widths is less for the multivariate pair-copula when compared to the univariate pair-copula model. According to Figures 8.2(c) and 8.2(d) these features can also be observed for variable Z_2.

The maps also demonstrate the decrease in prediction uncertainty of the individual variables when jointly predicting all the variables using a multivariate modelling approach compared to univariate modelling.

Moreover, the multivariate maps in Figures 8.1 and 8.2 confirm the ability of the multivariate pair-copula model to capture non-linear dependence of spatial variables.

8.2 Comparison of Univariate and Multivariate Design

The univariate optimal design, with the objective of reduction of prediction uncertainty of individual variables based on the univariate pair-copula model, was presented in Chapter 6. The multivariate optimal design, with the objective of reduction of prediction uncertainty for more than two variables simultaneously based on the multivariate pair-copula model was presented in Chapter 7. In this section, the difference between the univariate and multivariate optimal designs is analysed.

Figures 8.3(a), 8.3(c) and 8.3(e) are non-sequential optimal designs based on the univariate pair-copula model, while non-sequential optimal designs based on the pair-copula model in the multivariate setting are demonstrated in Figures 8.3(b), 8.3(d) and 8.3(f). The optimal additional locations, indicated by red squares

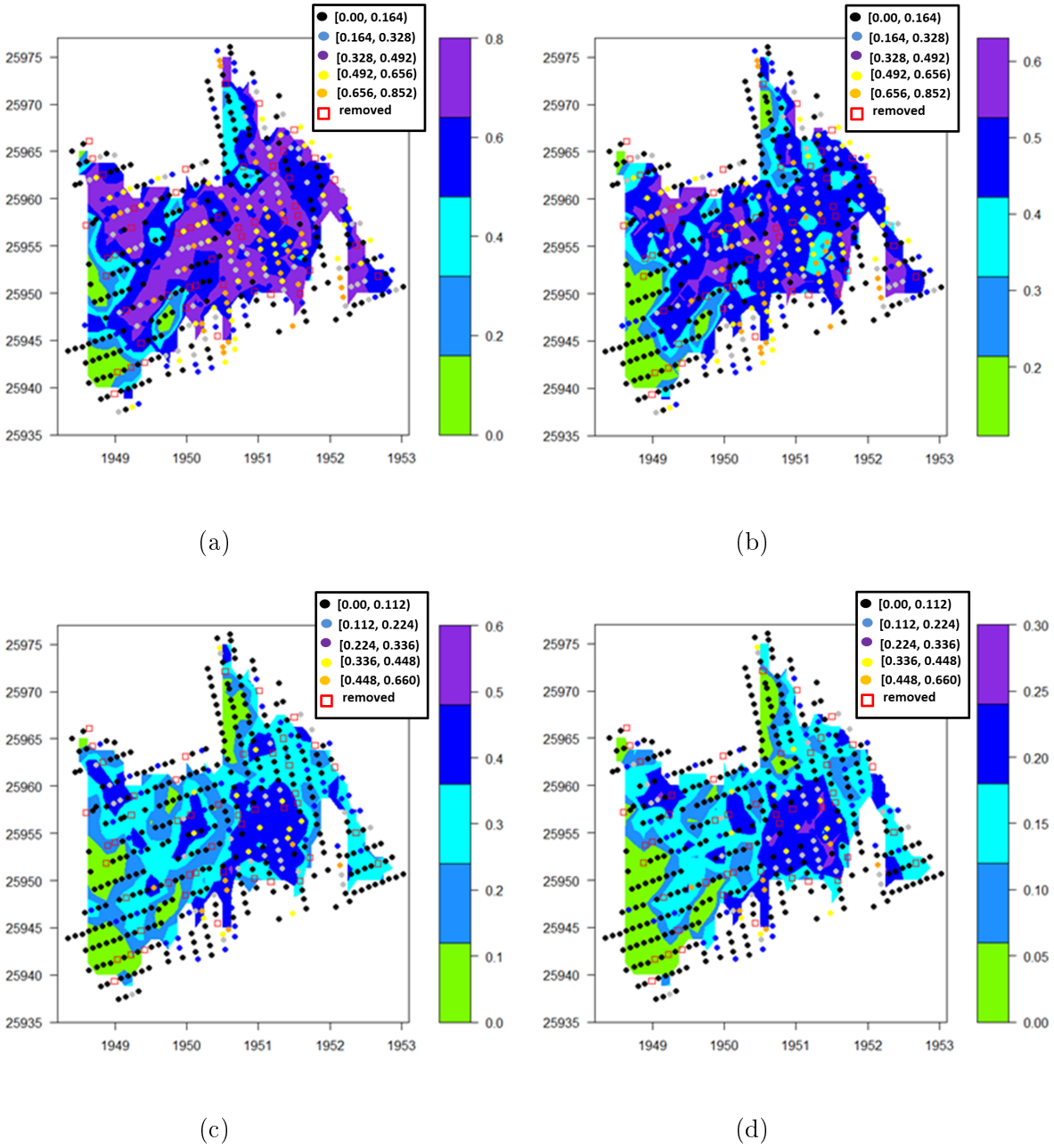


Figure 8.2: Maps for the widths of the 90% prediction intervals for (a) Z_1 based on the univariate pair-copula, (b) Z_1 based on the multivariate pair-copula (c) Z_2 based on the univariate pair-copula and (d) Z_2 based on multivariate pair-copula.

in Figures 8.3(a) and 8.3(b) are obtained by reducing the individual prediction uncertainty over the interpolation grid for Co. The additional locations in Figures 8.3(c) and 8.3(d) are obtained by reducing the prediction uncertainty of Ni only. The resultant optimal locations found by reducing the prediction uncertainty of both variables simultaneously are presented in Figures 8.3(e) and 8.3(f) for the univariate and multivariate pair copula models respectively.

When comparing Figures 8.3(a) and 8.3(b), it can be clearly seen that the non-

sequential univarite optimal design for Co based on the univariate and multivariate pair-copula models are similar. However, the non-sequential univariate optimal design for Ni based on the univariate and multivariate pair-copula models differ significantly. This is because of the reduction in prediction uncertainty of Ni in the multivariate pair-copula model (Figure 8.1(d)) compared to the univariate model (Figure 8.1(c)).

The multivariate optimal design for Co and Ni based on the univariate pair-copula model (Figure 8.3(e)), is most similar to the univariate design for Ni (Figure 8.3(c)). This is due to the high variability present in Ni (see Figure 8.3(c)) based on the univariate model. However, in the multivariate model, the variability of Ni is significantly decreased. Hence, the multivariate optimal design for Co and Ni based on the multivariate model (Figure 8.3(f)) can be considered a mixture of the univariate designs for Co (Figure 8.3(b)) and Ni (Figure 8.3(d)). Based on these results, it can be concluded that, a multivariate design based on univariate pair-copula models tends to be dominated by the points that reduce prediction uncertainty for the variable with highest variability. However, a multivariate design based on a multivariate pair-copula model produces design points that reduce prediction uncertainty in both variables. If one needs to obtain an optimal design in order to reduce the prediction uncertainty for all variables of interest simultaneously, the design approach based on the multivariate model is preferred.

8.3 Summary of the Contributions

In this thesis, the main aim was to develop general methodology for the optimal design of additional sampling based on a geostatistical model that can preserve both multivariate non-linearity and spatial non-linearity present in spatial variables. It has been indentified through analysing the literature that, without a valid model, no improvement can be gained with an optimal design. Hence, novel multivariate geostatistical modelling that can capture both multivariate non-linearity and spatial non-linearity was developed first, before developing the

methodology for optimal design. In this thesis, focus was mainly on copula based geostatistical models since they offer a solution to modelling the non-linear dependence structure in individual variables. In other words, the uncertainty estimation for predictions produced by copula based models capture not only the variation of the spatial configuration but also the variation in measured data values. Specifically, interest was in the pair-copula based geostatistical model (Gräler and Pebesma [2011]), since it has more flexibility to capture non-linear spatial dependence structures over a simple copula based model (Bárdossy [2006]).
Since this pair-copula based approach is relatively new to geostatistics, it has been used in few spatial fields and hasn't been used in the field of mining. The pair-copula based geostatistical model was introduced to the mining field for the first time in Chapter 3. That chapter also gave a step by step guideline for the use of pair-copula models in any practical application. Analysis of empirical copula density plots for the different distance classes revealed the non-linear dependence structure present in mining data. This result emphasised how the use of pair-copulas is able to capture realistic dependence structures compared to the use of the variogram, which ignores non-linearity. For the mining application used in the Chapter 3, better cross validation results were obtained by the pair-copula model compared to ordinary kriging, which is commonly used in the field of mining.
Improvement in the pair-copula model was gained by developing an algorithm to determine the distance classes of the pair-copula model in Chapter 4. In the literature, there is no well defined procedure for distance class determination of the pair-copula model even though the pair-copula model is based on distance classes. As the first part of the algorithm, a test used in the non-spatial setting to compare the equality between two copulas (Rémillard and Scaillet [2009]) was extended to the spatial setting by use of the dependent wild bootstrap. Based on the new test, Algorithm 1 was developed to define the distance classes for a pair-copula model. The application of the algorithm to the two dimensional Meuse data set and the three dimensional mining data set demonstrated a significant improvement in pair-copula fit when compared to the fit of a pair-copula with

equal distance classes.

In Chapter 5, a novel geostatistical multivariate modelling approach was developed to model the non-linear dependence between variables and the non-linear spatial dependence structure of the individual variables using NLPCA and pair-copulas. In addition, the pair-copula model was also introduced to the multivariate spatial setting for the first time. NLPCA was implemented to remove non-linear dependence between spatial variables at lag distance zero and if dependence between variables exist for lag distances larger than zero, then the second step of MAF was used to remove that dependence. Subsequently, the pair-copula model was used to individually model the uncorrelated transformed variables to capture the non-linear spatial dependence. The use of NLPCA was evaluated against the common non-linear transformation method SCT using two case studies. In both case studies, NLPCA reproduced the non-linear relationship between variables better than the SCT transformation. Moreover, the modelling approach with the pair-copula outperformed the modelling approach with conventional kriged model, regardless of the transformation method, in terms of reproduction of univariate statistics. In summary, based on the results obtained for the case studies, it can be concluded that use of NLPCA and pair-copulas has potential to improve modelling of non-linear multivariate data compared to existing non-linear modelling approaches.

A novel adaptive spatial design for additional samples based on the pair-copula model in order to reduce prediction uncertainty was proposed in Chapter 6. Introduction of the pair-copula model to spatial design is the main novelty of the proposed design approach. The uncertainty estimates from the pair-copula can capture not only the variation that comes from the spatial configuration of observations but also the variation that comes from the measured values from spatial observations. Hence, unlike traditional design approaches, pair-copulas are able to select optimal locations for additional samples based on both spatial configuration and values of the observations. Expected prediction uncertainty was used as the statistical criterion for selecting optimal locations, since the statistical

criterion should represent the effects of different values of a potential candidate location. This proposed design approach was applied to a two dimensional soil based application and the performance of the proposed approach was evaluated by partially redesigning the existing spatial design. The resulting redesign outperformed the existing spatial design. In addition, the efficiency of proposed design was compared with a conventional design approach based on a kriged model. Overall, the results demonstrate the potential of the proposed design.

In Chapter 7, a novel adaptive multivariate spatial design was proposed based on the model developed in the Chapter 5. The main objective of the proposed design is to reduce the uncertainty of the prediction of multiple spatial variables simultaneously. The novelty of this proposed design approach is the use of the model developed in Chapter 5. The uncertainty estimation from the model in Chapter 5 is able to capture both spatial and non-spatial non-linearity. Hence, the new sampling locations obtained through the proposed methodology were selected based on the relationship between variables, the spatial configuration of the observations and the measured values of the observations. Moreover, by using a case study with linear multivariate spatial variables, the difference between the spatial design based on the model that honours the non-linear spatial dependence of individual variables and the spatial design based on the model that doesn't was investigated. Based on this investigation, it can be conjectured that selecting optimal locations for new samples based on the correct model that honours the *in-situ* dependence of the spatial data will improve the precision of multivariate prediction in the spatial random field.

8.4 Limitations and Future Work

For simplicity, an isotropic dependence structure was assumed for all applications used in this thesis. However, anisotropy should be evaluated for different directions. Evaluation of the Kendall tau plots for different directions would not be sufficient to evaluate the anisotropy when fitting the spatial pair-copula model. The empirical copula density of each distance class for different directions should

be compared. This will be addressed in future research.

In Chapter 4, the test introduced by Rémillard and Scaillet [2009] to test the equality between two copulas is extended to the spatial framework. This test should at least be assessed by using random field simulation with known strength of dependence for each lag distance class. However, this kind of random simulation is not possible with existing spatial simulation tools. Hence, future research should focus on this perspective.

The proposed modelling approach based on NLPCA only investigated the non-linearity present in multivariate spatial data. However, both non-linearity and heteroscedasticity may be present in multivariate spatial data. Thus, the proposed methodology should be extended to deal with heteroscedasticity in future research. Moreover, we conjecture that NLPCA is not only able to capture non-linear structures among continuous spatial variables, but also among spatial variables with mixed types, such as nominal and rank data. Extension of NLPCA to these types of variables will also be considered in future research. Moreover, in Chapter 5, we only considered the non-linearity between variables at zero lag distance and linearity was assumed between variables at other lag distances when using the second step of MAF to remove cross-correlations. Hence, this issue should be investigated and solved in future research.

The objective of the proposed design is to reduce prediction uncertainty only. However, in practical applications, a campaign for additional samples should be carried out under a given budget. Thus, the process of finding the optimal number of additional samples should be included in the design methodology. Therefore, the proposed design methodology needs to be extended to find an optimum number of samples and their optimum locations in order to obtain the maximal knowledge about the spatial process under budget constraints in future research. In addition to that, the different location may have different cost associated to them. This constrain should be also included in the future research when considering the budget constrain.

Moreover, the limited number of potential candidates was pre-defined in the ap-

plications, which were used to demonstrate the proposed design methodology. However, there are an enormous number of candidate locations over the study domain. It would be computationally expensive to use exhaustive search over the study domain to find optimal locations. An efficient search algorithm, such as direct search simulated annealing can be integrated with the proposed algorithm to do this within a reasonable amount of computational time.

Even though the proposed design methodology can be applied to any two dimensional spatial application, it cannot be directly applied to three dimensional spatial applications. For instance, finding an optimal design for additional drillings in mining cannot be done directly based on the proposed design methodology. For each selected location, the optimal direction of drilling and optimum dip should be defined. Soltani and Hezarkhani (2011) proposed that the optimality of directional drilling should be evaluated by minimising the length of drill holes that lie on the outside of the ore body and maximising the length of drill holes that lie inside the ore body. However, the algorithm proposed by Soltani and Hezarkhani (2011) is only capable of optimising the dip angle. This algorithm should be extended to find the optimum azimuth of drilling and optimum dip for a given drilling location. The proposed methodology should be integrated with this algorithm for application in mining applications.

Li et al. [2011] developed an optimal sampling design methodology for an environmental observation network in order to increase expected gain defined by a utility function based on a more simple copula based geostatistical model. This method can be adopted and extended using pair-copula models. Since the pair-copula has more flexibility to capture the non-linear dependence structure, it can be conjectured that the design produced by pair-copulas would produce more precise estimates than the design proposed by the simple pair-copula model. Moreover, by applying this method to spatial applications, it would not only reduce the prediction uncertainty, but also significantly reduce the losses of making incorrect decisions and increase the gain of making correct decision. For example, in mining, this methodology can be applied to obtain optimal designs for addi-

tional drillings in order to maximise the expected return based on cut-off grade by minimising the loss of wrong decisions (deciding not to mine blocks with higher grade and deciding to mine blocks with lower grade) and by maximising the gain of correct decisions (deciding to mine blocks with higher grade and deciding to not mine blocks with lower grade). Here, it can be introduced an utility function based on decision theory to give positive value for making correct decision and negative value for making wrong decision. For each candidate location, it can be calculated expected utility over the study domain. The location which produce the maximum expected utility can be selected as the optimal point among the candidate location. Finally, implementing this optimal design for additional drilling would increase the precision of ore reserve estimation and also reduced expenses in making wrong decisions in the mining planning stage.

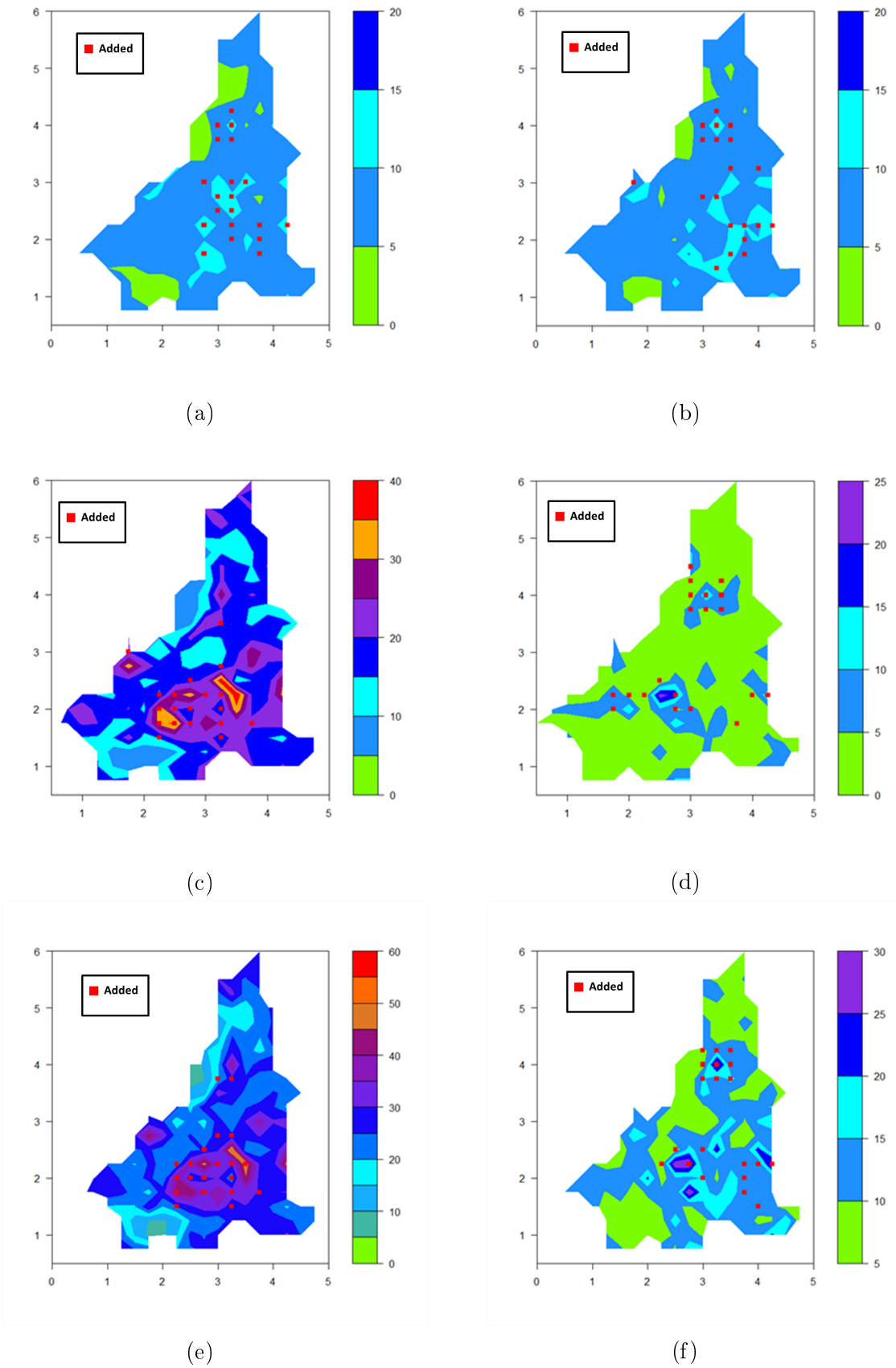


Figure 8.3: Non-sequential optimal design for (a) Co based on the univariate pair-copula (b) Co based on the multivariate pair-copula (c) Ni based on the univariate pair-copula (d) Ni based on the multivariate pair-copula model (e) Co+Ni based on the univariate pair-copula and (f) Co+ Ni based on the multivariate pair-copulal. Red squares represent the propsed optimal locations from each design approach.

Appendix A

Random field

A random field (or stochastic field), $X(s,\omega), s \in D, \omega \in \Omega$, is a random function specified by its finite-dimensional joint distribution.

$$F(y_1,\ldots,y_n;s_1,\ldots,s_n) = P(X(s_1) \leq y_1,\ldots,X(s_n) \leq y_n)$$

for every finite n and every collection $s_1,...,s_n$ of locations in D. The set D is usually a subset of $R^d, d \in N$ and for the special case $d = 1, X(s,\omega)$ is called a random process (or stochastic process). At every location $s \in D, X(s,\omega)$ is a random variable where the event ω lies in some abstract sample space Ω.

Strong stationary random field

Let $Z(x)$ be a random field and $P(Z(x_1) \leq z_1,\ldots,Z(x_n) \leq z_n)$ be the cumulative distribution function of the joint distribution of $Z(x)$ at the locations $x_1,\ldots,x_n \in D$. $Z(x)$ is said to be a strong stationary random field if, for all n and all vectors h that satisfy $x_1+h,\ldots,x_n+h \in D$,

$$P(Z(x_1) \leq z_1,\ldots,Z(x_n) \leq z_n) = P(Z(x_1+h)z_1,\ldots,Z(x_n+h) \leq z_n).$$

This implies that the cumulative distribution function is not a function of h.

Second order stationary random field

Let $Z(x)$ be a random field. $Z(x)$ is said to be a second order stationary random field if:

1. it has a constant mean over all spatial locations i.e., $E[Z(x)] = \mu$;

2. the auto-covariance of the data generating process depends only on distance h, i.e., $Cov(Z(x+h), Z(x)) = \gamma(h)$, where γ is the covariance function of $Z(x)$.

Spatial dependence

The relationship between realisations of a spatial variable sampled at different locations is described by the spatial dependence. High spatial dependence can be observed between samples that are close to the each other in space.

Linear spatial dependence

If spatial dependence of spatial data can be described by linear relationship, then spatial variable has linear spatial dependence. For an instance, Figure A.1 shows the kernel density plot of the unit transformation values of all the data pairs which are five meters apart in a spatial study. It can be clearly seen that regardless of the value of the data points, strength of the relationship between data pairs are constant. Hence, for this particular example, it can be mentioned that spatial dependence at lag five meters is linear.

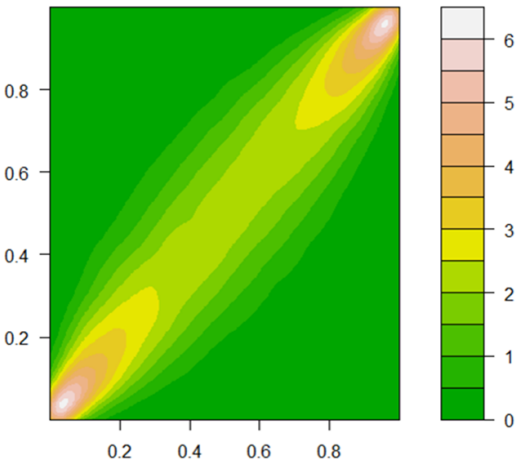


Figure A.1: Linear spatial dependence

Non-linear spatial dependence

If spatial dependence of spatial data cannot be described by a linear relationship,

then spatial variable has non- linear spatial dependence. Figure A.2 demonstrates the non-linear spatial dependence at lag ten meters. It can be seen strength of the relationship vary over the distribution values.

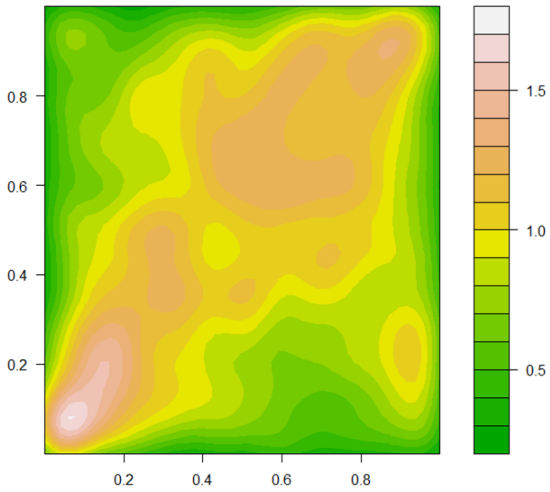


Figure A.2: Non-linear spatial dependence

Univariate spatial study

In univariate spatial study, only one spatial variable is considered from data collection to spatial analysis.

Multivariate spatial study

In multivariate spatial study, more than one spatial variable variable is considered from data collection to spatial analysis.

Multivariate dependence

The dependency between spatial variables at lag distance zero ($h = 0$) is defined as multivariate dependence. In other words, dependence of the measurements of the different variables at a particular location is defined as multivariate dependence.

Multivariate spatial dependence

The dependency between spatial variable at lag dstance greater than zero ($h > 0$) is defined as multivariate spatial dependence. In other words, dependence of the measurements of the different variables cross different locations defined as multivariate spatial dependence.

Cross variogram

The variogram discussed in the section above is only capable of dealing with the spatial dependency structure of a single variable (e.g., comparing percentage concentration of copper to other nearby percentage of copper concentration). To quantify the spatial relationship between two or more variables, a tool called the cross-variogram is used. The theoretical cross-variogram function can be defined as

$$\gamma^*_{jk}(h) = \frac{1}{2}Cov[\{Z_j(x) - Z_j(x+h)\}\{Z_k(x) - Z_k(x+h)\}].$$

where $Z_i(x)$ is i^{th} spatial variable at location x.

This can be estimated using the empirical cross variogram

$$\widehat{\gamma^*_{jk}(h)} = \frac{1}{2N}\sum_N \{z_j(x) - z_j(x+h)\}\{z_k(x) - z_k(x+h)\}].$$

This tool is only able of capturing the linear cross-relationship between the variables.

Bibliography

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics, 44(2):182–198.

Atteia, O., Dubois, J.-P., and Webster, R. (1994). Geostatistical analysis of soil contamination in the Swiss Jura. Environmental Pollution, 86(3):315–327.

Bandarian, E. M., Bloom, L. M., and Mueller, U. A. (2008). Direct minimum/maximum autocorrelation factors within the framework of a two structure linear model of coregionalisation. Computers & Geosciences, 34(3):190–200.

Bandarian, E. M., Mueller, U. A., Ferreira, J., and Richardson, S. (2010). Transformation methods for multivariate geostatistical simulation – minimum/maximum autocorrelation factors and alternating columns diagonal centres. In Dimitrakopoulos, R., editor, Old and New Dimensions in a Changing World, volume 17, pages 79–90. The Australasian Institute of Mining and Metallurgy (The AusIMM).

Bárdossy, A. (2006). Copula-based geostatistical models for groundwater quality parameters. Water Resources Research, 42(11):W11416.

Bárdossy, A. and Li, J. (2008). Geostatistical interpolation using copulas. Water Resources Research, 44(7):W07412.

Barnett, R. and Deutsch, C. (2012). Practical implementation of non–linear transforms for modeling geometallurgical variables. In Abrahamsen, P., Hauge,

R., and Kolbjørnsen, O., editors, Geostatistics Oslo 2012, volume 17 of Quantitative Geology and Geostatistics, pages 409–422. Springer, Netherlands.

Barnett, R., Manchuk, J., and Deutsch, C. (2014). Projection pursuit multivariate transform. Mathematical Geosciences, 46(3):337–359.

Bedford, T. and Cooke, R. M. (2002). Vines: A new graphical model for dependent random variables. The Annals of Statistics, 30(4):1031–1068.

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford University Press, New York.

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R. Springer, New York, 2nd edition.

Boardman, R. and Vann, J. (2011). A review of the application of copulas to improve modelling of non-bigaussian bivariate relationships (with an example using geological data). In Chan, F, M. D. and Anderssen, R. S., editors, Proceedings of the 19th International Congress on Modelling and Simulation, pages 627–633.

Brown, P. J., Le, N. D., and Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air pollutants. Canadian Journal of Statistics, 22(4):489–509.

Bueso, M., Angulo, J., Cruz-Sanjulián, J., and García-Aróstegui, J. (1999). Optimal spatial sampling design in a multivariate framework. Mathematical Geology, 31(5):507–525.

Chang, H., Fu, A., Le, N., and Zidek, J. (2007). Designing environmental monitoring networks to measure extremes. Environmental and Ecological Statistics, 14(3):301–321.

Cressie, N. (1990). The origins of kriging. Mathematical Geology, 22(3):239–252.

Cressie, N. A. C. (1993). Statistics for spatial data. Wiley, New York.

De Souza, L., Costa, J., and Koppe, J. (2004). Uncertainty estimate in resources assessment: A geostatistical contribution. Natural Resources Research, 13(1):1–15.

De-Vitry, C., Vann, J., and Arvidson, H. (2007). A guide to selecting the optimal method of resource estimation for multivariate iron ore deposits. In The AusIMM, editor, Iron Ore Conference 2007, pages 67–77. The Australasian Institute of Mining and Metallurgy (The AusIMM).

Desbarats, A. J. and Dimitrakopoulos, R. (2000). Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. Mathematical Geology, 32(8):919–942.

Deutsch, C. (1993). Kriging in a finite domain. Mathematical Geology, 25(1):41–52.

Diggle, P. and Lophaven, S. (2006). Bayesian geostatistical design. Scandinavian Journal of Statistics, 33(1):53–64.

Diggle, P. and Ribeiro, P. (2007a). Classical parameter estimation. In Model-based Geostatistics, pages 99–133. Springer, New York.

Diggle, P. and Ribeiro, P. (2007b). Geostatistical design. In Model-based Geostatistics, pages 199–212. Springer, New York.

Diggle, P. J., Ribeiro Jr, P. J., and Christensen, O. F. (2003). An introduction to model-based geostatistics. In Spatial statistics and computational methods, pages 43–86. Springer.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics), 47(3):299–350.

Doukhan, P., Lang, G., Leucht, A., and Neumann, M. H. (2015). Dependent wild bootstrap for the empirical process. Journal of Time Series Analysis, 36(3):290–314.

Dunham, S. and Vann, J. (2007). Geometallurgy, geostatistics and project value – does your block model tell you what you need to know? In Project Evaluation 2007, pages 19–20. The Australian Institute of Mining and Metallurgy (AusIMM).

Emery, X., Hernández, J., Corvalán, P., and Montaner, D. (2008). Developing a cost-effective sampling design for forest inventory. In Ortiz, J. and Emery, X., editors, Geostats 2008 – Proceedings of the Eighth International Geostatistical Congress, Santiago, Chile.

Erhardt, T. M., Czado, C., and Schepsmeier, U. (2015a). R-vine models for spatial time series with an application to daily mean temperature. Biometrics, 71:323–332.

Erhardt, T. M., Czado, C., and Schepsmeier, U. (2015b). Spatial composite likelihood inference using local c-vines. Journal of Multivariate Analysis, 138:74–88.

Fedorov, V. V. and Hackl, P. (2012). Model-oriented design of experiments. Springer Science & Business Media.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spbayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. Journal of Statistical Software, 19(4):1–24.

Frahm, G., Junker, M., and Szimayer, A. (2003). Elliptical copulas: applicability and limitations. Statistics & Probability Letters, 63(3):275–286.

Friedman, J. H. (1987). Exploratory projection pursuit. Journal of the American Statistical Association, 82(397):249–266.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers, C-23(9):881–890.

Gaetan, C. and Guyon, X. (2010). Second-order spatial models and geostatistics. In Spatial Statistics and Modeling, pages 1–52. Springer, New York.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedean copulas. Journal of the American Statistical Association, 88(423):1034–1043.

Getis, A. (2007). Reflections on spatial autocorrelation. Regional Science and Urban Economics, 37(4):491–496.

Goodchild, M. F. (1992). Geographical information science. International Journal of Geographical Information Systems, 6(1):31–45.

Goovaerts, P. (1993). Spatial orthogonality of the principal components computed from coregionalized variables. Mathematical Geology, 25(3):281–302.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford University Press, New York.

Gräler, B. (2014). Modelling skewed spatial random fields through the spatial vine copula. Spatial Statistics, 10:87–02.

Gräler, B. and Pebesma, E. (2011). The pair-copula construction for spatial data: a new approach to model spatial dependency. Procedia Environmental Sciences, 7:206–211.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed gaussian process models with an application to computer modeling. Journal of the American Statistical Association, 103(483):1119–1130.

Haff, I. H., Aas, K., and Frigessi, A. (2010). On the simplified pair-copula construction – simply useful or too simplistic? Journal of Multivariate Analysis, 101(5):1296–1310.

Haslauer, C., Li, J., and Bárdossy, A. (2010). Application of copulas in geostatistics. In Atkinson, P. M. and Lloyd, C. D., editors, geoENV VII – Geostatistics for Environmental Applications, pages 395–404. Springer, Netherlands.

Hassanipak, A. and Sharafodin, M. (2004). GET: A function for preferential site selection of additional borehole drilling. Exploration and Mining Geology, 13(1–4):139–146.

Hastie, T. and Stuetzle, W. (1989). Principal curves. Journal of the American Statistical Association, 84(406):502–516.

Hwang, J., Lay, S., and Lippman, A. (1994). Nonparametric multivariate density estimation: a comparative study. IEEE Transactions on Signal Processing, 42(10):2795–2810.

Isaaks, E. H. and Srivastava, R. M. (1989). An Introduction to Applied Geostatistics. Oxford University Press, New York.

Jammalamadaka, S. R. and Sengupta, A. (2001). Topics in Circular Statistics. World Scientific.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. In Rüschendorf, L., Schweizer, B., and Taylor, M. D., editors, Distributions with fixed marginals and related topics, pages 120–141. Institute of Mathematical Statistics, California.

Journel, A. (1994). Resampling from stochastic simulations. Environmental and Ecological Statistics, 1(1):63–91.

Journel, A. and Alabert, F. (1989). Non-Gaussian data expansion in the Earth Sciences. Terra Nova, 1(2):123–134.

Kazianka, H. and Pilz, J. (2010a). Copula-based geostatistical modeling of continuous and discrete data including covariates. Stochastic Environmental Research and Risk Assessment, 24(5):661–673.

Kazianka, H. and Pilz, J. (2010b). Spatial interpolation using copula–based geostatistical models. In Atkinson, P. M. and Lloyd, C. D., editors, geoENV VII – Geostatistics for Environmental Applications, volume 16 of Quantitative Geology and Geostatistics, pages 307–319. Springer, Netherlands.

Kazianka, H. and Pilz, J. (2011). Bayesian spatial modeling and interpolation using copulas. Computers & Geosciences, 37(3):310–319.

Khosrowshahi, S. and Shaw, W. (2001). Conditional simulation for resource characterization and grade control. In Edwards, A. C., editor, Mineral Resource and Ore Reserve Estimation - The AusIMM Guide to Good Practice: (Monograph 23), pages 285–292. Australasian Institute of Mining and Metallurgy (AusIMM).

King, P. (2011). Ore-body sampling and metallurgical testing. In Darling, P., editor, SME Mining Engineering Handbook. Society for Mining, Metallurgy, and Exploration (SME), 3rd edition.

Kirby, M. J. and Miranda, R. (1996). Circular nodes in neural networks. Neural Computation, 8(2):390–402.

Koppe, V., Costa, J., de Lemos Peroni, R., and Koppe, J. (2011). Choosing between two kinds of sampling patterns using geostatistical simulation: regularly spaced or at high uncertainty locations? Natural Resources Research, 20(2):131–142.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal, 37(2):233–243.

Krige, D. G. (1951). A statistical approach to some mine valuation and allied problems on the Witwatersrand. MSc dissertation, University of the Witwatersrand, Johannesburg.

Kruger, U., Zhang, J., and Xie, L. (2008). Developments and applications of nonlinear principal component analysis – a review. In Gorban, A., Kégl, B., Wunsch, D., and Zinovyev, A., editors, Principal Manifolds for Data Visualization and Dimension Reduction, pages 1–43. Springer, New York.

Kurowicka, D. and Cooke, R. (2006). Uncertainty analysis with high dimensional dependence modelling. Wiley, England.

Lark, R. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma, 105(1–2):49–80.

Larocque, G., Dutilleul, P., Pelletier, B., and Fyles, J. W. (2006). Conditional gaussian co-simulation of regionalized components of soil variation. Geoderma, 134(1–2):1–16.

Leuangthong, O. (2003). Stepwise conditional transformation for multivariate geostatistical simulation. PhD dissertation, University of Alberta, Canada.

Leuangthong, O. and Deutsch, C. V. (2003). Stepwise conditional transformation for simulation of multiple variables. Mathematical Geology, 35(2):155–173.

Li, J. (2010). Application of copulas as a new geostatistical tool. PhD thesis, Institute for Water and Environmental System, University of Stuttgart.

Li, J., Brárdossy, A., Guenni, L., and Liu, M. (2011). A copula based observation network design approach. Environmental Modelling & Software, 26(11):1349–1357.

Li, J. and Zimmerman, D. L. (2015). Model-based sampling design for multivariate geostatistics. Technometrics, 57(1):75–86.

Linting, M., Meulman, J. J., Groenen, P. J., and van der Koojj, A. J. (2007). Nonlinear principal components analysis: introduction and application. Psychological Methods, 12(3):336–358.

Malevergne, Y. and Sornette, D. (2003). Testing the gaussian copula hypothesis for financial assets dependences. Quantitative Finance, 3(4):231–250.

Marchant, B., McBratney, A., Lark, R., and Minasny, B. (2013). Optimized multi-phase sampling for soil remediation surveys. Spatial Statistics, 4:1–13.

Marchant, B., Saby, N., Jolivet, C., Arrouays, D., and Lark, R. (2011). Spatial prediction of soil properties with copulas. Geoderma, 162(3–4):327–334.

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

Matheron, G. (1970). La Théorie des variables régionalisées, et ses applications. Les Cahiers du Centre de morphologie mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines de Paris.

McLennan, J. A. and Deutsch, C. V. (2004). Conditional non-bias of geostatistical simulation for estimation of recoverable reserves. CIM Bulletin, 97(1080):68–72.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., and Van Meirvenne, M. (2009). Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Science Society of America Journal, 73(2):614–621.

Moon, C. and Whateley, M. (2006). From prospect to prefeasibility. In Evans, A. M., Whateley, M. K. G., and Moon, C. J., editors, Introduction to Mineral Exploration, pages 70–103. Wiley-Blackwell, US, 2nd edition.

Muller, W. G. (2007). Designs for spatial trend estimation. In Collecting Spatial Data, pages 101–139. Springer, Berlin Heidelberg.

Musafer, G. N., Thompson, M. H., Kozan, E., and Wolff, R. C. (2013). Copula-based spatial modelling of geometalurgical variables. In Dominy, S., editor, Proceedings of the Second AusIMM International Geometallurgy Conference, pages 239–246, Brisbane, Queensland. The Australian Institute of Mining and Metallurgy (AusIMM).

Musafer, G. N., Thompson, M. H., Kozan, E., and Wolff, R. C. (2015). Pair-copula modelling of grade in ore bodies. Technical report, Queensland University of Technology.

Nelsen, R. (2006). An Introduction to Copulas. Springer Series in Statistics. Springer, New York.

Noppé, M. (1994). Practical geostatistics for on-site analysis – a coal example. In Proceedings of the Mining Geostatistics Conference, page 14. Geostatistical Association of South Africa, Kruger National Park, South Africa.

Pilger, G., Costa, J., and Koppe, J. (2001). Additional samples: where they should be located. Natural Resources Research, 10(3):197–207.

Pilz, J. and Spöck, G. (2008). Bayesian spatial sampling design. In Ortiz, J. and Emery, X., editors, Proc. 8th International Geostatistics Congress , Gecamin Ltd., Santiago de Chile, pages 21–30.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rémillard, B. and Scaillet, O. (2009). Testing for equality between two copulas. Journal of Multivariate Analysis, 100(3):377–386.

Rondon, O. (2012). Teaching aid: Minimum/maximum autocorrelation factors for joint simulation of attributes. Mathematical Geosciences, 44(4):469–504.

Rondon, O. and Tran, T. T. (2008). Multivariate simulation using min/max autocorrelation factors: practical aspects and case studies in the mining industry. In Ortiz, J. M. and Emery, X., editors, Proceedings of the Eighth International Geostatistics Congress (GEOSTATS 2008), Santiago, Chile, pages 269–278.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. Ann. Math. Statist., 23(3):470–472.

Roth, C. (1998). Is lognormal kriging suitable for local estimation? Mathematical Geology, 30(8):999–1009.

Royle, J. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Computers & Geosciences, 24(5):479–488.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. Stat.Med., 25(1):127–141.

Saikia, K. and Sarkar, B. (2006). Exploration drilling optimisation using geostatistics: a case in Jharia coalfield, India. Applied Earth Science: Transactions of the Institutions of Mining and Metallurgy: Section B, 115(1):13–22.

Saito, H. and Goovaerts, P. (2000). Geostatistical interpolation of positively skewed and censored data in a dioxin-contaminated site. Environmental Science & Technology, 34(19):4228–4235.

Scheck, D. and Chou, D.-R. (1983). Optimum locations for exploratory drill holes. International Journal of Mining Engineering, 1(4):343–355.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5):1299–1319.

Scholz, M. (2007). Analysing periodic phenomena by circular PCA. In Hochreiter, S. and Wagner, R., editors, Bioinformatics Research and Development, pages 38–47. Springer, New York.

Scholz, M., Fraunholz, M., and Selbig, J. (2008). Nonlinear principal component analysis: Neural network models and applications. In Gorban, A. N., Kégl, B., Wunsch, D. C., and Zinovyev, A. Y., editors, Principal Manifolds for Data Visualization and Dimension Reduction, pages 44–67. Springer, New York.

Scholz, M., Kaplan, F., Guy, C. L., Kopka, J., and Selbig, J. (2005). Non-linear PCA: a missing data approach. Bioinformatics, 21(20):3887–3895.

Scholz, M. and Vigário, R. (2002). Nonlinear PCA: a new hierarchical approach. In Verleysen, M., editor, Proceedings of the 10th European Symposium on Artificial Neural Networks (ESANN 2002), pages 439–444.

Seo, D.-J. (2013). Conditional bias-penalized kriging (cbpk). Stochastic Environmental Research and Risk Assessment, 27(1):43–58.

Shao, X. (2010). The dependent wild bootstrap. Journal of the American Statistical Association, 105(489):218–235.

Shapiro, A. (2003). Monte carlo sampling methods. In Ruszczyriski, A. and Shapiro, A., editors, Stochastic Programming, volume 10 of Handbooks in Operations Research and Management Science, pages 353–425. Elsevier.

Sidler, R. (2003). Kriging and conditional geostatistical simulation based on scale-invariant covariance models. Master's thesis, Swiss Federal Institute of Technology, Zurich.

Sklar, A. (1959). Fonctions de répartition á n dimensions et leurs marges. Publ. Inst. Statist. Université Paris 8.

Soltani, S. and Hezarkhani, A. (2013). Proposed algorithm for optimization of directional additional exploratory drill holes and computer coding. Arabian Journal of Geosciences, 6(2):455–462.

Switzer, P. and Green, A. A. (1984). Min/max autocorrelation factors for multivariate spatial imagery. Technical report, Department of Statistics, Stanford University, California, `https://statistics.stanford.edu/sites/default/files/SWI%20NSF%2006.pdf`.

Thomson, R. E. and Emery, W. J. (2014). Data Analysis Methods in Physical Oceanography. pages 313–414. Elsevier, USA, 3rd edition.

Triantafilis, J., Odeh, I., Warr, B., and Ahmed, M. (2004). Mapping of salinity risk in the lower Namoi Valley using non-linear kriging methods. Agricultural Water Management, 69(3):203–231.

Trivedi, P. K. and Zimmer, D. M. (2007). Copula modeling: an introduction for practitioners. Now, Boston.

Van Groenigen, J., Siderius, W., and Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma, 87(3–4):239–259.

Van Groenigen, J. and Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality, 27(5):1078–1086.

Vann, J. and Guibal, D. (2001). Beyond ordinary kriging – an overview of non-linear estimation. In Edwards, A. C., editor, Mineral Resource and Ore Reserve

Estimation - The AusIMM Guide to Good Practice: (Monograph 23). The Australasian Institute of Mining and Metallurgy (The AusIMM).

Vašát, R., Heuvelink, G., and Borůvka, L. (2010). Sampling design optimization for multivariate soil mapping. Geoderma, 155(3–4):147–153.

Venäläinen, A. and Heikinheimo, M. (2002). Meteorological data for agricultural applications. Physics and Chemistry of the Earth, Parts A/B/C, 27(23):1045–1050.

Wackernagel, H. (2003). Multivariate Geostatistics. Springer, London, 3rd edition.

Walton, D. and Kauffman, P. (1982). Some practical considerations in applying geostatistics to coal reserve estimation. SME-AIME, Dallas.

Webster, R. and Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. Journal of Soil Science, 43(1):177–192.

Zhu, Z. and Stein, M. (2006). Spatial sampling design for prediction with estimated parameters. Journal of Agricultural, Biological, and Environmental Statistics, 11(1):24–44.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics, 17(6):635–652.