Multivariate Modelling of Geological and Geometallurgical Variables

Emmanuel ADDO JUNIOR

BSc. (Hons) Geological Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

MSc. Geology, University of Ghana, Legon, Ghana

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the

Faculty of Engineering, Computer and Mathematical Sciences

School of Civil, Environmental and Mining Engineering

The University of Adelaide

- January 2019-

Declaration of Authorship

I, Emmanuel ADDO JUNIOR, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: _____

Date: 04/03/2019

"*Better is the end of a thing than the beginning thereof*"

Ecclesiastes 7:8a (KJV)

"*If you really believe in what you're doing, work hard, take nothing personally and if something blocks one route, find another. Never give up*!"

Laurie Notaro

UNIVERSITY OF ADELAIDE

Abstract

Faculty of Engineering, Computer and Mathematical Sciences

School of Civil, Environmental and Mining Engineering

Doctor of Philosophy

Multivariate Modelling of Geological and Geometallurgical Variables

by Emmanuel ADDO JUNIOR

The mining and minerals industry is confronted with several challenges that were not common some decades ago. Deep-seated and complex orebodies, low metal grades, and fluctuating commodity prices have an increasingly high impact on the mining industry, potentially reducing profit margins. It follows that accurate modelling of geological and geometallurgical variables is needed to reduce risks associated with most mineral prospects. This modelling needs to include uncertainty in predictions, as well as outcomes, so that mining companies can use value at risk, for example, to make informed business decisions.

This thesis, comprises three journal papers and one conference paper. Several mathematical formulations have been used to model geological and geometallurgical variables. These novel modelling methodologies provides more versatile modelling techniques to traditional modelling techniques which are currently employed in the mining and minerals industry.

In Chapter 2 (Paper 1) and Chapter 3 (Paper 2), spatial pair-copula models are used to predict the geological grades of an anisotropic gold deposit within and outside a main field. These models are compared with a traditional kriging approach and results show that pair-copulas

models provide improved modelling of error structure than kriging. In Chapter 4 (Paper 3), different trivariate copulas were used to model and predict geological variables from a drill core. The models provided better estimates and prediction intervals of geological variables. In general, geological variables have a large number of outlying values and also exhibit tail dependence. Copulas provide a means of dealing with these practical issues. D-vine copula models, which are able to address the massive multivariate nature and non-linear bivariate relationships of geometallurgical variables, are employed to model geometallurgical variables. In most cases, geometallurgical variables have several missing data, which makes modelling and prediction of these variables difficult. Chapter 5 (Paper 4), a novel data imputation algorithm is developed as part of this thesis to address the issue of missing data of geometallurgical variables. The outcomes of this thesis are an improved geostatistical modelling framework and novel data imputation algorithm techniques, providing better estimates and prediction intervals for geological and geometallurgical variables, and with demonstrated application to practical mining case studies.

Acknowledgements

I would like to express my sincere gratitude to my supervisory team, Associate Professor Emmanuel Knox Chanda, Associate Professor Andrew Viggo Metcalfe and Dr Michael Leonard. Thanks for all the encouragement and support in the last 3.5 years of my research life, your love for research and expertise have been invaluable. I would also like to acknowledge my previous supervisory team, Associate Professor Chaoshui Xu and Professor Peter Dowd for pushing me hard through those difficult times of this research journey, it refined my thinking and brought out the best in me.

Many thanks to my family, my charming wife Afua and adorable children Elyssa, Elyana and Erel. Your love and support got me through this research, you tolerated all the late nights and weekends' work on this thesis during the most demanding and gruelling times. My richest appreciation also goes to my extended family, my prayerful mum (Alice) and sisters (Agnes, Becky, Lydia, Mary and Martha) who stood with me in prayers through the strenuous times of this research journey. My unfathomed gratitude further goes to the Pentecost International Worship Centre (PIWC) Adelaide family, in particular Overseer Goldwyn Baiden-Assan and wife, Elder Emmanuel Joel Aikins Abakah and Deaconess Becky Obeng. I couldn't have asked for a better family than this.

I am very much indebted to the late *Joseph Djan Mamphey* who gave me the opportunity to fall in love with resource modelling and geostatistics. I know you will be smiling in heaven knowing the seed you sowed some ten years ago has finally bore fruits. Also, to Dr Winfred Assibey-Bonsu who saw the potential in me and natured it, I am forever grateful. My exceptional gratitude also goes to Assistant Professor Eugene Ben-Awuah of Laurentian University whose advice and encouragement got me into graduate school.

I don’t want to forget all the old and new friends I made on this journey, you’ve all played an active role in this achievement. And to the school staff, professors and other PhD students you’ve all supported in many diverse ways to achieve this dream, I say thanks to you all.

Finally, I owe all this to **YAWEH**, he made everything beautiful and possible in his own time (*Eccl., 3:11a*).

Table of Contents

List of Figures

List of Tables

Dedicated to my charming wife Afua and children
Elyssa, Elyana and Erel

Chapter 1

1 Introduction

This research is motivated by the many challenges that arise when modelling spatial univariate/multivariate geological and geometallurgical variables in the field of mining. The modelling challenges faced in the mining context can generally be extended to any field that deals with spatial data. Examples are environmental modelling, geotechnical engineering and hydrology. Subsurface spatial resource and reserve modelling is a critical task for mining, petroleum and many environmental projects, as this has a high impact on business, technical and operational decision making. A mining context and data are used in this thesis, but most of the key concepts and modelling methodologies highlighted here can be applied to any subsurface spatial resource and reserve modelling.

Major modelling decisions are normally based on the geological and geometallurgical sampling of orebodies/deposits e.g., soil samples, geophysical surveys, trenching, auger drill holes, rotary air blast (RAB) holes, reverse circulation (RC) holes and diamond drilling (DD) holes. These measurements are used to inform modelling of critical variables across the entire orebody/deposit. For example, metal content (i.e., the grade) which is the average proportion contained in the ore and geology (i.e., lithology), which is typically summarized by expert assessment of visible characteristics such as texture of a mineral resource are modelled to evaluate the global resource and reserve for economic feasibility studies.

Until recently, orebody models were based almost exclusively on geological grade variables, but variables that are relevant to processing (i.e., geometallurgical variables) can also play a key role in the evaluation of the final cost of mining and mineral projects. Examples include: hardness, bond mill work index (BWi), resistance to abrasion and breakage (A*b), the semi-autogenous grinding power index (Spi), recovery, the degree of weathering and floatability. These processing or metallurgical variables, however, have some complex issues that need to

be preserved and captured when modelling. For instance, they are normally under-sampled compared to traditional geological grade variables, which hinders their integration into resource models using existing geostatistical methods (Hunt, Kojovic, and Berry, 2013). In addition, these variables are non-additive, difficult to scale up (i.e., from sample scale to plant scale) and finally have a lot of missing data (Deutsch, 2013). Furthermore, geological and geometallurgical variables have complex univariate and multivariate relationships, which are the result of multiple chaotic non-linear natural processes that occurred during their formation. More often than not, it is required to have an optimal geostatistical modelling tool that can model these non-linear spatial dependence structures and other complexities that exist with or within-variable of these geological and geometallurgical variables.

The concept of geometallurgy has been used in mining industry for over five decades and several definitions have been proposed by different authors. Johnson et al. (2007) used the term geometallurgy to imply the impact of ore quality on mine planning, plant performance and product quality. Furthermore, Coward et al. (2009) also explains that the value of geometallurgy is in the improvement of mining and ore treatment in the design and operational phases, by further improving the general understanding of rock properties. Lishchuk et al. (2015) define geometallurgy as a multidisciplinary approach that integrates geology, mineralogy, mineral processing, and metallurgy to create spatial models for production and operational decisions. Finally, Dunham and Vann (2007) define geometallurgy as a cross-disciplinary approach which aims to improve resource economics by integrating geology, mining planning, operational design, mineral processing and metallurgy. All the definitions posited above agree on the importance of integrating the available geological and geometallurgical variables for improved mine planning and operations.

The standard procedure to model a geological variable is a well understood problem in current geostatistical literature. However, modelling geological and geometallurgical variables

simultaneously, where these variables are correlated, is not well documented. Furthermore, the dependence between these variables (i.e. geological and geometallurgical) can be non-linear. Moreover, the spatial dependence of the individual geological and geometallurgical variables can be non-linear (Musafer et al 2016). If a geostatistician is tasked to model these variables, and that model fails to capture and preserve the non-linearity, the final predictions from the model may be inaccurate. However, when the geostatistical model captures the non- linearity in the dependence structure of individual variables, a much more "*detailed*" estimate of uncertainty can be generated realistically in the spatial field. The "*detailed*" uncertainty estimation here means, a model that can capture not only the individual variation in the configuration of the spatial locations but also variations in the measured values for those spatial locations (Kazianka and Pilz 2010). Therefore, an optimal geostatistical model is required to capture the non-linear spatial dependence in order to improve estimates.

Most univariate and multivariate geostatistical models use the variogram to model spatial dependence (Goovaerts 1998; Kazianka and Pilz 2010; Rossi et al. 1992). Standard geostatistical techniques are equivalent to using a Gaussian copula, where the multivariate Gaussian distribution has the property of linear regression. The variogram measures the linear dependence over the distribution of the variable for a given spatial distance. Hence, any geostatistical model that uses the variogram in an estimation routine may not be able to provide accurate and precise predictions when non-linearity is present. While standard geostatistical methods are versatile, they have the limitation that they do not model the co-movement of extreme values, which is popularly known as tail-dependence. Lastly, the variogram is sensitive to outlying values, which are common in geological and geometallurgical datasets.

Geological and geometallurgical datasets normally require the characterization of multiple variables. That is, modelling the relationships between the geological (i.e., grade, lithology etc.) variables and geometallurgical (i.e., recovery, hardness, BWi, Spi, impedance, magnetic

susceptibility etc.) variables can have a large impact on the final process forecasting. Geostatisticians most often aim to reproduce the complex relationships in these variables through multivariate geostatistical modelling. Existing traditional multivariate models for multiple spatial variables generally include a co-regionalisation model. These models, ignore the non-linear dependence between the variables and also fail to reproduce the within-variable spatial dependence at all sample locations. In addition, modeling these multiple spatial variables is time consuming and difficult, because of the requirements of modelling cross variograms, which normally increases with the number of variables (Bandarian et al 2008).

Alternatively, spatial multivariate variables can be assumed to be multivariate Gaussian, in order to fully parameterize the covariance matrix (Chiles and Delfiner, 2012; Isaaks, 1990; Journel and Huijbregts, 1978; Verly, 1983). In reality however, these geological and geometallurgical variables exhibit complex features such as a skewed distribution, heteroscedasticity, non-linearity and constraints. When faced with these problems the common practice usually involves transforming the variables to a univariate Gaussian, before assuming the variables to be multiGaussian and applying conventional geostatistical modelling techniques (Leuangthong and Deutsch 2003). Examples of transformations include the normal score transformation (Chiles and Delfiner 2012; Verly 1983), principal component analysis (Pearson 1901) and min/max autocorrelation factors (Switzer and Green 1984). After transformation to the univariate Gaussian, traditional geostatistical modelling tools which assume a multiGaussian are then used for estimation and/or simulation, before back-transforming the data to the original space. The back-transformation to the original space can worsen the problem when it fails to reproduce the complex multivariate relationships that existed prior to modelling of the multivariate variables. This approach of modelling ignores the non-linear dependence between the variables and also fails to reproduce the within-variable spatial dependence at all locations (Leuangthong and Deutsch 2003).

Another problem with most geological and geometallurgical datasets is the issue of under-sampling (i.e., unequal sampling at all locations). This normally occurs with legacy or metallurgical samples, which are mostly expensive to sample and analyse. It is not uncommon for a geological and geometallurgical database to have large differences between the number of samples recorded (i.e., lithology, grades etc.) yet relatively few metallurgical test-work samples (i.e., recover, hardness, BWi, Spi, etc.). Retaining only data where all variables are sampled could normally result in removing a large amount of data from the final geological and geometallurgical model (Deutsch 2013), which can lead to poor geostatistical modeling in areas where more data (of some variables) are sampled. Data exclusion is problematic for a number of reasons, including the introduction of bias and loss of information (Enders 2010; Little and Rubin 2002). In general, data imputation is the preferred method, however, there are only a few suitable imputation methods for geological and geometallurgical variables. Most of the available imputation methods normally do not integrate the statistical information of the complete geological and geometallurgical variables in the imputation (Barnett 2015).

The challenges and problems discussed above motivated this research to contribute a new data imputation algorithm and geostatistical modeling framework for spatial univariate/multivariate geological and geometallurgical variables. The main objective of this research is to develop a geostatistical modelling technique and novel data imputation algorithm that captures the non-linearity of geological and geometallurgical variables. In this thesis, copula based geostatistical models have been adapted to address the spatial non-linearity issue of geological and geometallurgical variables. These models offer a more "detailed" solution to modelling the non-linearity and spatial dependence in individual spatial geological and geometallurgical variables. Furthermore, a novel data imputation algorithm is developed to address the problem of under-sampling (i.e., missing data) that normally comes with geometallurgical datasets. This new algorithm uses statistical information of the complete variables for imputation. Practical

mining related case studies are presented in this thesis that employ the proposed geostatistical modelling tools and data imputation algorithm developed.

1.1 Research Aims and Objectives

The overall aim of this research is to develop advanced statistical (i.e., copula) models for prediction of spatial geological and geometallurgical variables. A novel geostatistical modelling approach is used that will preserve both the univariate and multivariate non-linearity and spatial non-linearity present in geological and geometallurgical variables. The overall aims can be achieved through specific aims related to univariate/multivariate modelling and data imputation as follows.

a) Univariate model: Improve on the existing pair-copula based models (Gräler and Pebesma 2011; Musafer and Thompson 2013) to estimate single geological grade variable with non-linear spatial dependence and anisotropy.
b) Multivariate model: To extend the copula based models in multivariate setting to model and predict three geological variables whilst capturing the non-linear spatial dependence, complexity and tail dependence.
c) Data imputation: Develop an optimal data imputation algorithm for missing geometallurgical variables that seeks to preserve critical statistics of the variables.

In order to achieve the overarching overall aim, specific objectives have been outlined which relate to the univariate and multivariate modelling and geostatistical data imputation of spatial variables as follows.

a) Univariate model: Extend on the existing copula-based models (Gräler and Pebesma 2011; Musafer and Thompson 2013) and develop a novel copula based anisotropy model to estimate geological variables that exhibits skewness and non-linear spatial dependence.

b) Multivariate model: To apply the copula-based spatial model in a multivariate setting to model and predict geometallurgical variables whilst capturing non-linear spatial dependence and complex multivariate relationships. In addition, develop a novel prediction framework using the copula based multivariate model.
c) Data Imputation: To develop a novel geostatistical data imputation algorithm for missing geometallurgical datasets, this new algorithm seeks to preserve the individual histograms and bivariate correlation among the geometallurgical variables.

The expected outcome of this research is to develop a novel geostatistical modelling framework to model anisotropic geological variables. In addition, a novel data imputation algorithm is developed and applied to geometallurgical variables. Several practical mining case studies are also presented using the proposed modelling methodologies.

1.2 Research Questions

Three principal research questions related to spatial geostatistical modeling and data imputation for geological and geometallurgical variables are addressed in this research.

a) Univariate modelling: How can a univariate geostatistical technique be formulated and extended to model univariate non-linearity and spatial non-linearity present in spatial anisotropy geological and geometallurgical variables based on pair-copulas?
b) Multivariate modelling: How can a multivariate geostatistical model be developed to capture the complex multivariate non-linearity, tail dependence and spatial non-linearity present in spatial geological and geometallurgical variables based on pair-copulas?
c) Data Imputation: How can an optimum data imputation algorithm be developed to preserve the statistical characteristics of missing geometallurgical and geological variables?

1.3 Significance of the Research

The three main contributions to the field of spatial statistics and geostatistics which have been addressed in this thesis are:

a) There is no procedure in current literature to define distance classes and directions required in pair-copula geostatistical models when dealing with anisotropic datasets. The first aspects of this research develops an optimum algorithm to define distance classes and directions required in pair-copula modelling when modelling anisotropic geological datasets. The extension of this new algorithm that captures the spatial variability by both distance and direction is a new contribution to the area of geostatistics.

b) The second contribution to research is the extension and application of pair-copula geostatistical models to be able to model multivariate non-linear, tail dependence and complex geological variables. A novel framework is developed using a copula-based geostatistical model of dependence to make predictions and generate prediction intervals of geological variables at further unknown depths of a drillhole.

c) The final contribution to research is the development of a new geostatistical data imputation algorithm. This new algorithm seeks to preserve individual histograms and the bivariate correlation among the geometallurgical variables in order to have a complete datasets to build a robust and reliable predictive model. Moreover, the D-vine copula-based regression which is not affected by quantile crossing, due to the monotonic increasing of Kendall's tau, is used to model and predict geometallurgical variables.

1.4 Thesis Overview and Outline

The scope of this thesis is to improve spatial univariate/multivariate modelling and spatial interpolation using different copulas and pair-copula based geostatistical model with the overall objective of increasing the accuracy in prediction of geological and geometallurgical variables.

Chapter 2 (Paper 1) presents a detailed description of the spatial pair-copula based geostatistical model, which includes its strength and weakness. The pair-copula model is applied to spatial, univariate, non-linear, anisotropic gold deposits. There is no procedure in literature to define distance classes required in pair-copula modelling for anisotropic datasets. This paper seeks to develop an efficient algorithm to define distance classes required in pair-copula modelling for anisotropic datasets. A novel improvement to the spatial pair-copula model was made by introducing an efficient algorithm to define lags by distance and direction, and is applied to anisotropic geological grade datasets. The pair-copula model developed within the main field is compared with traditional geostatistical log-normal kriging. Results show that the pair-copula model gives 5% better predictive performance on mean absolute error compared with traditional log-normal kriging.

Chapter 3 (Paper 2) presents a framework for using pair-copula models to make predictions at unknown locations outside the main geological field. This work is an extension from Chapter 2 (Paper 1). Customarily, geostatisticians are faced with the challenging issues of predicting samples at unknown locations to make informed business and mining/geological decisions. This paper uses a spatial pair-copula model to make predictions outside the main geological field for future exploration and drilling campaign. A total of twenty unknown points (10 apiece at the North-East and South-West) side of the main field were predicted and results were compared with traditional geostatistical log-normal kriging. Results from the cross-validation of both models, suggested that the pair-copula model slightly outperformed traditional geostatistical log-normal kriging for making predictions outside the main geological field. In

addition, results from the pair-copula model also suggested that the mining company were more likely to intersect high grade mineralisation if they concentrated on exploring the South-Western part of the main geological field.

Chapter 4 (Paper 3) presents a framework and methodology for modelling and predicting multivariate geological variables using copulas. In this chapter, the Gaussian, Student-*t* and Vine copulas were used to model the complex relationships which are often associated with geological variables. The non-linear spatial dependence structure at all lags and tail dependence of all variables were modelled with Gaussian, Student-*t* and Vine copulas. A novel framework is developed using reliable copula-based model of dependence to make estimates of geological variables at further depth. Prediction intervals of the estimate were generated for all three copula models using Monte-Carlo simulations and results were cross-validated and compared. In this application the Vine copula slightly outperforms the Student-*t* and Gaussian copulas for estimating points and predicting intervals at further depths. The Vine copula generally allows for tail dependence when modelling skewed and non-linear geological variables. In general, copulas provides a better and flexible model of the error structure and most importantly gives a much detailed modelling of geological variables compared with traditional geostatistical approaches.

Chapter 5 (Paper 4) addresses the problems of missing data, non-additivity and non-linear relationships that exist in geometallurgical datasets. A novel data imputation algorithm is developed and applied to geometallurgical variables. The data imputation is developed as an optimisation problem that seeks to preserve the individual histograms and the bivariate correlations of the geometallurgical variables. Moreover, applying the primary-response rock property framework and using quantitative and qualitative primary properties, metallurgical response prediction model was built using a D-vine copula quantile regression. In the D-vine copula quantile regression, because the Kendall's tau calculated from the bivariate copulas are

monotonically increasing, crossing quantiles corresponding to different quantile levels is not possible. Generally, traditional ordinary linear and generalised least squares regression models quantiles functions cross when modelling non-Gaussian geometallurgical datasets. D-vine quantile regression copula is applied to four non-Gaussian geometallurgical variables and predictions of one variable was made. Results show a better improvement compared with traditional ordinary linear and generalised least squares regression models.

Finally, chapter 6 addresses some of the key findings and limitations of the proposed methodologies and modelling framework developed in this thesis. In paper 1 and 2 of this thesis, the soil sample database used enabled a comprehensive methodology when using spatial pair-copula to model anisotropic dependence structures. However, resource and reserve modelling in mining context uses three dimensional drillhole database, future work should aim at using anisotropic 3-D database for the pair-copula modelling. In paper 3, the copula based models preserved and modelled the non-linearity of the multivariate geological database. However, some geological database may have further complex features, future work should aim at using the copula based approach to model these complexities. Finally, in paper 4 the GA algorithm used was designed to reproduce only two statistics, future work should aim at improving the algorithm to capture more statistics.

References

Bandarian, E, Mueller, UA, Fereira, J & Richardson, S 2018, 'Transformation methods for multivariate geostatistical simulation—Minimum/Maximum autocorrelation factors and alternating columns diagonal centres', in A
dvances in Applied Strategic Mine Planning, Springer, pp. 371-393.

Barnett, RM 2015, 'Managing Complex Multivariate Relations in the Presence of Incomplete Spatial Data', PhD Thesis, 2015.

Chilès, JP & Delfiner, P 2012, 'Wiley Series in Probability and Statistics', Geostatistics: Modeling Spatial Uncertainty, Second Edition, pp. 705-714.

Coward, S, Vann, J, Dunham, S & Stewart, M 2009, 'The Primary-Response framework for geometallurgical variables', in 7 th international Mining Geology Conference, pp. 109-113.

Deutsch, CV 2013, 'Geostatistical Modelling of Geometallurgical Variables - Problems and Solutions', THE SECOND AUSIMM INTERNATIONAL GEOMETALLURGY CONFERENCE / BRISBANE, QLD, 30 SEPTEMBER - 2 OCTOBER 2013.

Dunham, S & Vann, J 19 - 20 June 2007, 'Geometallurgy, Geostatistics and Project Value — Does Your Block Model Tell You What You Need to Know?', Project Evaluation Conference - Melbourne, 19 - 20 June 2007.

Enders, CK 2010, Applied missing data analysis, Guilford press.

Goovaerts, P 1998, 'Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties', Biology and Fertility of soils, vol. 27, no. 4, pp. 315-334.

Gräler, B & Pebesma, E 2011, 'The pair-copula construction for spatial data: a new approach to model spatial dependency', Procedia Environmental Sciences, vol. 7, pp. 206-211.

Hunt, J, Kojovic, T & Berry, R 2013, 'Estimating Comminution Indices from Ore Mineralogy, Chemistry and Drill Core Logging', THE SECOND AUSIMM INTERNATIONAL GEOMETALLURGY CONFERENCE / BRISBANE, QLD., 30 SEPTEMBER - 2 OCTOBER 2013.

Isaaks, EH & Srivastava, RM 1989, An introduction to applied geostatistics, Oxford university press.

Johnson, R, Scott, G & Lukey, H 2007, 'Implications of Mineralogy, Grain Size and Texture on Liberation and Pellet Quality of Great Lakes Iron ore', Iron Ore, vol. 2007, pp. 109-111.

Journel, AG & Huijbregts, CJ 1978, Mining geostatistics, Academic press.

Kazianka, H & Pilz, J 2010, 'Copula-based geostatistical modeling of continuous and discrete data including covariates', Stochastic Environmental Research and Risk Assessment, vol. 24, no. 5, pp. 661-673.

Leuangthong, O & Deutsch, CV 2003, 'Stepwise Conditional Transformation for Simulation of Multiple Variables', Mathematical Geology, vol. 35.

Lishchuk, V, Koch, P-H, Lund, C & Lamberg, P 2015, 'The Geometallurgical Framework. MALMBERGET AND MIKHEEVSKOYE CASE STUDIES.', Minerals and Metallurgical Engineering division Lulea University of Technology.

Little, RJ & Rubin, DB 2014, Statistical analysis with missing data, vol. 333, John Wiley & Sons.

Musafer, GN & Thompson, MH 2016, 'Non-linear optimal multivariate spatial design using spatial vine copulas', Stochastic Environmental Research and Risk Assessment, pp. 1-20.

Musafer, GN, Thompson, MH, Kozan, E & Wolff, RC 2013, 'Copula-Based Spatial Modelling of Geometallurgical Variables.pdf', THE SECOND AUSIMM INTERNATIONAL GEOMETALLURGY CONFERENCE / BRISBANE, QLD, 30 SEPTEMBER - 2 OCTOBER 2013.

Pearson, K 1901, 'Principal components analysis', The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 6, no. 2, p. 559.

Rossi, RE, Mulla, DJ, Journel, AG & Franz, EH 1992, 'Geostatistical tools for modeling and interpreting ecological spatial dependence', Ecological monographs, vol. 62, no. 2, pp. 277-314.

Switzer, P & Green, A 1984, 'Min/max autocorrelation factors for multivariate spatial imagery: Dept', Statistics, Stanford Univ. Tech. Rep.

Verly, G 1983, 'The multigaussian approach and its applications to the estimation of local reserves', Mathematical Geology, vol. 15, no. 2, pp. 259-286.

Chapter 2

Spatial Pair-Copula Model of Grade for an Anisotropic Gold Deposit (*Paper 1*)

Emmanuel Addo Jr, Emmanuel K. Chanda and Andrew V. Metcalfe

Mathematical Geoscience, published - July 2018

Statement of Authorship

Title of Paper	**Spatial Pair-Copula Model of Grade for an Anisotropic Gold Deposit**
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	E Addo Jr, EK Chanda, and AV Metcalfe. "Spatial Pair-Copula Model of Grade for an Anisotropic Gold Deposit". In: Mathematical Geosciences (2018), pp. 1–26. doi: 10.1007/s11004-018-9757-7

Principal Author

Name of Principal Author (Candidate)	Emmanuel ADDO JUNIOR
Contribution to the Paper	Developed methodology, conducted programming and execution of methods. Wrote the manuscripts and acted as the corresponding author.
Overall percentage (%)	80%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/0/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Emmanuel KNOX CHANDA
Contribution to the Paper	Supervised the development of work and assisted in reviewing the manuscript
Signature	Date 10/1/2019

Name of Co-Author	Andrew VIGGO METCALFE
Contribution to the Paper	Supervised the development of work and assisted in reviewing the manuscript
Signature	Date 14/1/2019

Abstract

Copulas provide a convenient way to express multivariate distribution. In this study pair-copula modelling is applied to a gold deposit in Western Ghana. The dataset for the gold deposit has 1,500 surface soil samples on an area of 7,810 hectares. The distribution of the grade appears to be anisotropic and a realization of a non-stationary random process. The objectives of the analysis are to use spatial pair-copula to model the anisotropic gold grade and to determine regions of highest gold value within the field for a future drilling campaign. In the analysis, possible transformations of the data were compared to reduce the influence of outliers, and in the case of copulas achieving a marginal uniform distribution. The anisotropy of the gold grades is described with empirical copula contour plots for each distance class and for two orthogonal directions. The non-stationarity is modelled by regression methods. The residuals from the regression are modelled with both spatial pair copulas and kriging. The different approaches are compared in terms of the mean absolute error (MAE) and root mean square error (RMSE), using different proportions of the data for training and validation. The pair-copula median with kernel margins had MAE of 17.4 compared to 18.3 for log-normal kriging. An advantage of copulas is that they provide a more accurate model, than kriging, for the uncertainty associated with predictions.

KEYWORDS: copula, geostatistical modelling, kriging, mining

2.1 Introduction

The minerals industry is now confronted with several challenges, which were not common some few decades ago. Low ore grades, more refractory ores, deep-seated orebodies, increased variability within orebodies and fluctuating commodity prices have an increasingly high impact on the industry and require more accurate modelling of all related variables to optimise the mining operations. This modelling needs to include uncertainty in predictions as well as expected outcome, so that mining companies can use continuous value at risk (Rockafellar and Uryasev 2002), for example, to make business decisions.

The standard procedure for modelling spatially distributed grade variables is to assume that some transformation of grade has a multivariate Gaussian distribution (MVG). Variogram models, or equivalently covariance functions, have widely been used to capture the spatial dependence structures within the orebody (Goovaerts 1998; Gräler and Pebesma 2011; Kazianka and Pilz 2010; Rossi et al. 1992). While the standard procedure is quite versatile the MVG has a limitation that it does not model co-movement of extreme values, known as tail-dependence. Random variables that generally show little correlation can show tail dependence in extreme deviations, as seen in financial derivatives.

In geostatistical applications, the spatial dependence structures can vary over different percentile values of the distribution of the variable of interest (Journel and Alabert 1989); and in particular the extremes can have different spatial dependence structure from the central values. One goal of geostatistics is to find a way to describe such spatial dependence. The variogram represents expected values of squared differences between the values of the variable measured at different locations. Empirical variograms are mostly sensitive to outliers, as has been demonstrated by Bárdossy and Kundzewicz (1990). However, these effects can be reduced by using the indicator approach, at a cost of calculating a large number of indicator variograms. Journel and Deutsch (1997), suggested another approach using the ranks for

interpolation. Indicator variograms can be used to define the difference in dependence as a function of the observed values, and Journel and Alabert (1990) noted that by using indicator variograms the dependence between variables enables departure from a Gaussian model. However, for interpolation and simulation purposes theoretical variograms will have to be fitted for each selected threshold. The amount of work involved in fitting the theoretical variograms may be reduced by using an automatic algorithm as suggested by Goovaerts et al. (2005). The indicator approach is highly empirical and is not based on any stochastic model, so it requires large data sets for adequate precisions. In particular, indicators are fitted for each threshold separately, which can lead to problems with monotonicity of the estimation.

An alternative strategy for modelling the multivariate distribution of grades is to use copula models, which encompass all multivariate distributions including the MVG (Bárdossy and Li 2008; Kazianka and Pilz 2011; Marchant et al. 2011). Copulas have been widely used in the financial and actuarial sectors to model tail dependence structures (Cherubini et al. 2004; Chollete et al. 2009; Embrechts et al. 2001; Hu 2006; Rodriguez 2007). Copulas have also been applied in the field of hydrology (AghaKouchak 2014; De Michele and Salvadori 2003; Favre et al. 2004; Genest and Favre 2007), and other spatial applications. For example: Bárdossy and Li (2008) used spatial copulas to model groundwater parameters; Marchant et al. (2011) used copulas to model soil properties; Kazianka and Pilz (2011) used copulas to model air pollutants; and Musafer et al. (2013, 2016) applied copulas to mining applications. The pair-copula model has been used in only a few spatial applications for example Gräler and Pebesma (2011) and Gräler (2014), and spatial temporal applications for example (Erhardt et al. 2015). Musafer et al. (2016) used the spatial pair-copula in modelling the grade of an isotropic gold orebody.

In this paper, spatial pair-copula is compared with kriging for surface soil samples from an anisotropic ore body. The objectives of the analysis are to model the anisotropic gold grade and to determine regions of highest gold value within the field for a future drilling campaign.

The distribution of gold grade is non-stationary and anisotropic, as well as being highly skewed. The analysis is based on the logarithms of grade, as is customary in the mining industry. The non-stationarity is modelled by a quadratic regression, and the residuals from the regression are modelled with both spatial pair-copulas and lognormal kriging. Although the rationale for taking logarithms of grades is to obtain an approximate normal distribution, the distribution of the residuals from the regression remains noticeably skewed and there are some substantial outliers. The copula approach allows for the skewness by using an empirical transform to a uniform distribution on (0, 1).The anisotropy of the gold grades is shown with empirical copula contour plots for each distance class and for two orthogonal directions. The pair-copulas and kriging will be compared in terms of mean absolute error (MAE) and root mean square error (RMSE), using different proportions of the data for training and validation, and in terms of the difference in predicted regions of highest gold yield.

This paper comprises three main sections. The "Method" section, together with the appendices, describes the theory of copulas, pair-copulas and vine copula constructions model for spatial data. The "Application" section describes the fitting of models from gold surface soil samples data from an operating mine in Western Ghana. The final section is a discussion and conclusion.

2.2 Methods

This section summarizes the principles of copulas, pair-copula and pair- copula construction for anisotropic spatial data. More details about the concept of copula can be found in Joe (1996) and Nelsen (2006). Further details about the pair-copulas and vine copulas model can also be found in Aas et al. (2009); Bedford and Cooke (2002) and Kurowicka and Cooke (2006). Applications to geostatistics can be found in Gräler and Pebesma (2011) and Gräler (2014).

2.2.1 Definition of Copulas

Copulas are multivariate uniform distribution. It follows from the definition of the cumulative distribution function (cdf) that the cdf of any continuous random variables has a uniform distribution (the probability integral transform). So any multivariate distribution has a copula form. Furthermore, given a copula the uniform margins can be transformed to any continuous distributions which can all be distinct. Therefore copulas provide a very flexible approach for modelling multivariate data.

The cdf of a continuous multivariate random variable $(Z_1, \dots, Z_d)$ is defined by

$$F(z_1, z_2, \dots, z_d) = P[Z_1 < z_1, Z_2 < z_2, \dots, Z_d < z_d] ,$$

and using the probability integral transforms the right hand side can be written in an equivalent form

$$= P\big(F_1(Z_1) < F_1(z_1), F_2(Z_2) < F_2(z_2), \dots, F_d(Z_d) < F_d(z_d)\big) . \qquad (1)$$

where $F_{i,\dots,d}(\,)$ are the marginal cdfs. The latter form is defined as a copula cdf $C(F_1(z_1), F_2(z_2), .., F_d(z_d))$. So, a copula is a uniform multivariate distribution. Sklar's Theorem (1959) is a formal proof that not only can any continuous multivariate cdf be expressed as a copula but also expanding a copula with any set of inverse cdfs $F_1^{-1}(\,), F_2^{-1}(\,), \dots., F_d^{-1}(\,)$ results in a valid multivariate distribution. The practical importance of the result is that the fitting of marginal distributions is separated from the modelling of the covariance structure. Any multivariate distribution can be modelled by fitting marginal distributions, which can all be quite different, and then fitting a suitable copula to the probability integral transforms of the marginal distributions.

To summarize, a copula of dimension d is a multivariate uniform distribution defined on the unit hypercube $[0,1]^d$. Copula is defined in terms of its cdf $C(u_1, u_2, \ldots., u_d)$ and its corresponding probability density function (pdf) is

$$c(u_1, u_2, \ldots., u_d) = \frac{\partial C(u_1, u_2, \ldots., u_d)}{\partial u_1 \, \partial u_2 \ldots \partial u_d} . \quad (2)$$

The copula pdf links the marginal pdfs to the multivariate pdf

$$f(z_1, \ldots, z_d) = c(u_1, \ldots, u_d) f(z_1) \ldots . f(z_d) . \quad (3)$$

There is a wide variety of copulas that are commonly used to model bivariate probability distributions, and some are given in Appendix B. In particular the Gaussian copula is equivalent to a bivariate Gaussian distribution.

More generally, multivariate distributions of more than two variables are required. The Gaussian copula, and copulas based on student-*t* distributions, readily extend to more than two variables, but this does not generally hold for other bivariate copulas. A flexible approach to modelling such multivariate distribution is pair-copula D-vines and canonical vines, as described by Gräler (2014).

2.2.2 Pair – Copula

It follows from the multiplicative rule of probability that any multivariate distribution can be factorised in many ways using conditional distributions. In particular any copula can be factorised as a product of marginal distributions and bivariate conditional copulas. Such a factorization is known as a pair-copula model. The pair-copula model assumes that all joint, marginal and conditional distributions are absolutely continuous with corresponding densities. Joe (1996) gave the first pair-copula construction for a multivariate copula based on the

distribution functions. Following Joe's work Bedford and Cooke (2002) gave a construction in terms of densities. They organized the constructions in a graphical way involving a sequence of nested trees, which they call regular vines. They further defined two popular sub-classes of pair-copula construction (PCC) models, which they called D-vines and canonical vines. Their work is developed by Kurowicka and Cooke (2006). The derivation of canonical vines, which are used in this application is outlined below.

2.2.2.1 Canonical Vine Distributions

Aas et al. (2009), published a construction principle of vine copulas as pair-copula construction based on the work by Bedford and Cooke (2002). The idea behind pair-copula construction is that multivariate copula can be estimated with a cascade of bivariate copula building blocks (Gräler 2014). The pair-copula can also be seen as a multivariate copula which aims to approximate the target copula because not all copulas can be re-build as a vine copula as discussed by Haff et al. (2010). The pair-copula decomposition is however not unique, five dimensional density can have about 240 different constructions. Each decomposition expresses the full copula density differently Aas et al. (2009) and Bedford and Cooke (2002) used a graphical model called the regular vine copula to arrange the large number of pair-copula constructions. This regular vine copula are made up of special cases of D-vine and canonical vines (Czado 2010).

Canonical vines are used when one can find a key variable that controls relationship of the dataset. Figures 2.1(a) and 2.1(b) is reproduced from Aas et al. (2009), it shows the graphical model used to demonstrate the canonical vine for three and five variables respectively. These figures consists of two trees $T_j, j = 1,2$ and four trees $T_j, j = 1,2,3,4$ respectively. Figures 2.1(a) and 2.1(b), Tree T_j has $4 - j$ nodes, $3 - j$ edges and Tree T_j has $6 - j$ nodes, $5 - j$ edges respectively. In general the number of trees for m variables is $m - 1$. T_j has $m + 1 - j$

nodes. An edge represents the corresponding pair-copula and the label of the edge represents the subscript of the pair-copula (Aas et al. 2009). The nodes in the (Fig. (2.1a)) and (Fig. (2.1b)) are used in deciding the labels of the edges. Figures 2.1(a) and 2.1(b) illustrates the canonical vine for three and five variables respectively, using these decomposition the joint density function of three and five random variables can be expressed using the C-vine as follows in Eq. (4a) and (4b) respectively (Aas et al. 2009)

$$f_{123}(z_1, z_2, z_3) = f_1(z_1).f_2(z_2).f_3(z_3).c_{12}\big(F_1(z_1), F_2(z_2)\big).$$
$$c_{13}\big(F_1(z_1), F_3(z_3)\big).c_{23|1}\left(F_{2|1}(z_2|z_1), F_{3|1}(z_3|z_1)\right). \qquad (4a)$$

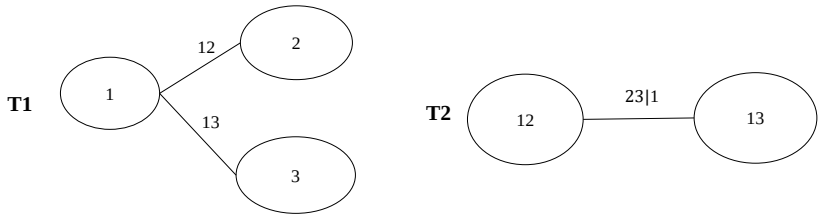


Fig. 2.1(a) Canonical vine for three variables

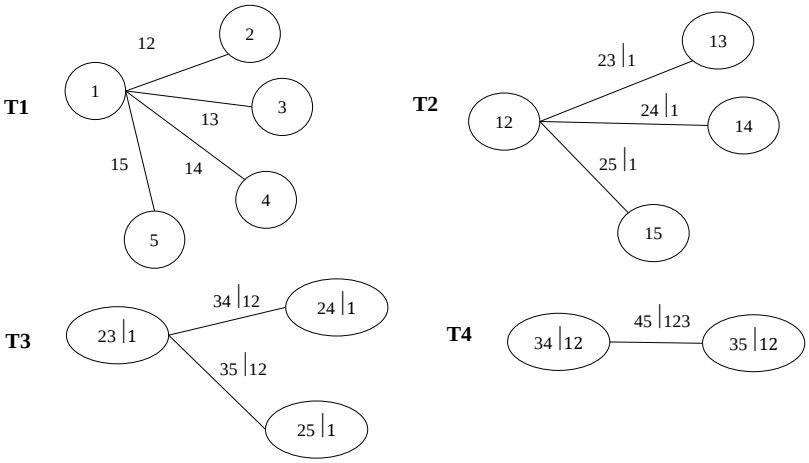


Fig. 2. 1(b) Canonical vine for five variables

The rationale for the correspondence between Eq. (4a) and (Fig. 2.1(a)) is explained in Appendix A. The joint density function of five variables is estimated using the C vines as follows

$$
\begin{aligned}
&f_{12345}(z_1, z_2, z_3, z_4, z_5) \\
&= f_1(z_1) \cdot f_2(z_2) \cdot f_3(z_3) \cdot f_4(z_4) \cdot f_5(z_5) \cdot c_{12}\big(F_1(z_1), F_2(z_2)\big) \\
&\cdot c_{13}\left(F_1(z_1), F_3(z_3)\right) \cdot c_{14}\big(F_1(z_1), F_4(z_4)\big) \cdot c_{15}\big(F_1(z_1), F_5(z_5)\big) \\
&\cdot c_{23|1}\left(F_{2|1}(z_2|z_1), F_{3|1}(z_3|z_1)\right) \cdot c_{24|1}\left(F_{2|1}(z_2|z_1), F_{4|1}(z_4|z_1)\right) \\
&\cdot c_{25|1}\left(F_{2|1}(z_2|z_1), F_{5|1}(z_5|z_1)\right) \cdot c_{34|12}\left(F_{3|12}(z_3|z_1, z_2), F_{4|12}(z_4|z_1, z_2)\right) \\
&\cdot c_{35|12}\left(F_{3|12}(z_3|z_1, z_2), F_{5|12}(z_5|z_1, z_2)\right) \\
&\cdot c_{45|123}\left(F_{4|123}(z_4|z_1, z_2, z_3)\right), F_{5|123}(z_5|z_1, z_2, z_3)\big) . \qquad (4b)
\end{aligned}
$$

2.2.3 Pair – Copula Construction for Anisotropic Spatial Data

In general the pair-copula construction explained by Aas et al. (2009) gives the procedure to disintegrate n-dimensional copula C_n into a set of $n(n-1)/2$ bivariate copulas. The density of the full pair-copula is a product of all the bivariate copula densities following the decomposition structure.

Gräler (2014), proposes a canonical vine copula for modelling spatial random fields. As in conventional linear geostatistics, an assumption of stationarity is essential for the pair-copula construction. This assumption of stationarity allows us to use the same marginal cumulative distribution function F for all locations in the domain, which is $F_i(z_i) = F(z_i)$. In the isotropic case, the bivariate spatial copula C_s at any two locations within a domain of interest only depends on the separation vector h and is independent of the location x (Bardossy 2006; Bardossy and Li 2008). That is

$$
\begin{aligned}
C_s(h, u, v) &= Pr\big(F\big(Z(x)\big) \le u, F\big(Z(x+h)\big) \le v\big) \\
&= C\left(F\big(Z(x)\big), F\big(Z(x+h)\big)\right). \qquad (5)
\end{aligned}
$$

In this paper, modifications have been made to be able to model the anisotropic cases, in which spatial dependence is assumed to vary both with distance and direction. The steps for the analysis are as follows:

1. *Checking Stationarity*

Fit a regression model for the field variable on location. In general, the location will be specified by the coordinates(x, y), and the predictor variables will be $x, y, x^2, y^2\ and\ xy$ for a quadratic surface. The fitting should be by generalized least squares, and a variogram model can be used to estimate the approximate covariance structure of the errors. If there is evidence against a hypotheses that all coefficients in the model are 0, then there is evidence of non-stationarity. In this case the residuals from the fitted regression model can be considered as a realization of a stationary random field.

2. *Empirical Bivariate Copula Contour or Densities for Orthogonal Directions*

Denote the observed value of the field variable at a point i, with coordinates (x_i, y_i), as $z_i(x_i, y_i)$ or more succinctly as z_i. A consequence of the stationarity assumptions is that the distribution of field variable is identical at all locations. So, $F(z)$ can be estimated from all the observations $(z_1, \dots\dots, z_N)$ where N is the total number of samples in the domain. If the distribution function $F(z)$ is estimated by some parametric form, $\hat{F}(z)$, then the probability integral transformation, $u_i = \hat{F}(z_i)$, can be used to transform all the observations to a unit interval (0,1). Alternatively, an empirical transformation $u_i = \frac{rank\ order\ (z_i)}{n+1}$ can be used to transform the observation to (0, 1). The distances between every pair of observations $z_i(x_i, y_i)\ and\ z_j(x_j, y_j); i \neq j, \forall i, j = 1,2, \dots\dots, N$ are calculated as $h_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. A set of distance classes $\{(0, h_1), (h_1, h_2), \dots, (h_{l-1}, h_l)\}$ is defined, where h_l is taken as the distance beyond which there is no significant dependence

between the observations (range). For every pair of observation (z_i, z_j) the probability integral transform of the two components were calculated to obtain $\left(\hat{F}(z_i), \hat{F}(z_j)\right)$, and associate this pair with the distance class which includes h_{ij}.

A rotation matrix is used to select pairs of points that are orientated along or nearly along a specific direction, θ say. If the two points have coordinates (x_i, y_i) and (x_j, y_j) the transformed points (x_i', y_i') and (x_j', y_j') are given by

$$\begin{pmatrix} x_i' & x_j' \\ y_i' & y_j' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_i & x_j \\ y_i & y_j \end{pmatrix}. \quad (6)$$

If the original points lie along a line with direction θ, the transformed points lie along the horizontal. Then the slope of the line segment joining the transformed points $\frac{y_j' - y_i'}{x_j' - x_i'} = 0$. Points are considered to be nearly orientated along the direction θ if the absolute value of the slope of the line segment joining the transformed points is less than a small defined tolerance (tol)

$$\left|\frac{y_j' - y_i'}{x_j' - x_i'}\right| < tol\,. \quad (7)$$

Given a defined tolerance (tol) distance and an angle(θ), all pairs of points that satisfy Eq. (7) are selected as the pairs of points for that directional angle(θ). The Kendall tau correlation between the field variable for all pairs in each distance class is calculated and plotted against the mean distance between the points in each distance class. In general, this is repeated for angle θ from 0° to 90° in steps 10° and the principal direction is taken as the direction with the greatest range. The orthogonal direction, is at 90° to the principal direction. However, in some cases the principal direction may be defined by a geological study.

Figure 2.2 illustrates a schematic diagram showing how the weighted correlogram is estimated. A weighted correlogram at an angle θ to the principal axis is defined by Eq. (8). Hence, the weighted correlogram is calculated as

$$\rho_{(\theta)} = \rho_{(principal)}\left(\frac{90-\theta}{90}\right) + \rho_{(minor)}\left(\frac{\theta}{90}\right). \quad (8)$$

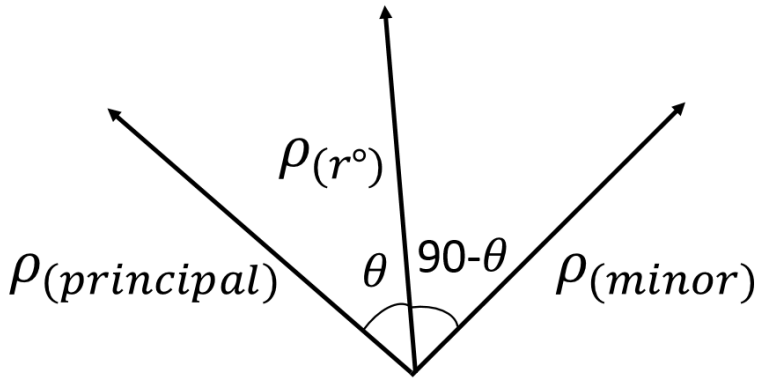


Fig. 2.2 Schematic diagram showing how weighted correlogram is estimated

3. Bivariate Copula Densities and Spatial Copula Construction

The pair copula construction is fitted to pairs of points that are orientated approximately along an axis that is at 45° to the principal axis. So the pair copula relates to a notional average correlogram. A bivariate copula is fitted to the pairs in each distance class. The simplest method of fitting a bivariate copula is to equate the sample values of Kendall's tau to the value of Kendall's tau implied by the dependence parameter. This is known as the methods of moments and requires the inverse function $\tau = \tau(\theta)$ of the function $\theta = \theta(\tau)$ given in appendix B. A limitation of the method of moments is that it does not lead easily to a criterion for choosing between copula forms. Therefore, the method of maximum likelihood is preferred for choosing between copula forms, the form with the highest likelihood being chosen. The R software

function spcopula (Gräler 2014) allows the user to define a set of candidate copulas for each distance class and selects the copulas that maximize the likelihood. Different copula forms are generally selected for the different distance classes. The selected copulas for all the distance classes are used to define convex combination of copulas as given in Eq. (1) of Gräler (2014).

Using a convex combinations ensures consistency between distance classes.

4. Pair-Copula Construction and Spatial Interpolation

Finally, the random variable Z representing the field variable at unknown location x_0 follows a distribution $F(z|z_1, \dots\dots\dots, z_k)$ which is conditioned on the known values of grade at the nearest neighbouring points $x_1, \dots\dots, x_k$, where k is the total number of observations. The conditional distribution is expressed in terms of the conditional pair-copula (Musafer et al. 2016).

Point estimates at unknown locations x_0 can then be obtained by computing the mean or the median of the conditional distribution (Bardossy and Li 2008), using the Metropolis-Hastings (Chibb and Greenberg 1995) algorithm as described in Appendix A.

The pair-copula modeling approach provides the full conditional distribution of the realisations at unknown locations. This full distribution can be used to generate the confidence intervals of the predicted point mean or median estimates, and also prediction intervals associated with point predictions. These prediction intervals are likely to be more accurate than those based on kriging because there is no assumption of normality.

2.3 Application

Surface soil samples of gold grade data from an operating open pit mine in Western Ghana have been used to investigate the feasibility and possible benefits of the copula based approach for geostatistical modelling. The geology of the project area is predominantly made up of the

Birimian and Tarkwaian rocks of the West-Africa region. These rocks host over 60 million ounces of gold in Ghana alone, and there are several operating mines and large green and brownfield exploration targets. A total of 1,500 surface soil samples (two dimensional data) were sampled over an area of 7,810 hectares. The soil gold grade measurements are in parts per million (ppm).

Mineralisation pods based on the average orientation of the quartz veins were digitized after initial exploration phase (i.e., outcrop mapping and quartz vein sampling) were completed. The average orientation of these mineralisation pods were along the 135° direction. These mineralisation pods were used as base template in planning the surface soil sampling campaign. Anisotropy of the grades was expected with substantial differences between the 135° and its orthogonal direction, (i.e., 45°) direction. The correlation of the surface soil samples is relatively low due to the mineralization style and the distances between samples. Based on the modelling results from the surface soil samples data, the mining company will consider investing large capital in a drilling campaign.

In the local grid system of the mine, the project extends from an easting of 243,258.6E to 233,183.8E and a northing of 210,425.9N to 202,674.5N. Figure 2.3(a) and 2.3(b) show a two dimensional plot of the project area with sampling locations of the surface soil samples and a contour map of the gold grade in the project area, respectively.

Summary statistics of the surface soil samples grades are given in Table 2.1. All statistical analysis and fitting of spatial pair-copula models were conducted using R software (Gräler & Appel 2015 and R Core Team 2016). In addition, variogram and kriging interpolations were conducted using Geostatistics for Windows (*GeostatWIN*) software (Chaoshui Xu, personal communication).

A histogram and box plot of the grades is shown in Fig. 2.4 (upper panel), and the extremely high positive skewness is evident. A histogram and box plot of the natural logarithms of grades are shown in Fig. 2.4 (lower panel). The skewness is much reduced, but not removed. The spike below 0 corresponds to samples with grades, which were recorded to two decimal places, below 1 ppm. As is common in mining applications, the following analysis is in terms of logarithms of grades. The steps in the subsequent analysis are shown in (Fig. 2.5).

The hypothesis of stationarity was tested by fitting a quadratic regression surface to the grades. The natural logarithm of the grade (w) was regressed on the standardized eastings (x), standardized northings (y), x^2, y^2 and the cross product xy in Eq. (9)

$$w = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy + \varepsilon\,. \qquad (9)$$

Where ε is random error, which is expected to be spatially correlated, with mean of 0 and standard deviation of σ_o. Generalized least squares was used for the fitting based on spherical variogram using the *nlme* function (Pinheiro and DebRoy 2016) in R. Standardization refers to scaling, by subtracting the mean and dividing by the standard deviation, to avoid the excessively large values of the quadratic terms and ill-conditioned matrices if original eastings and northings are used. The estimated coefficients in the model are shown in Table 2.2. A histogram of the residuals from the model is shown in (Fig. (2.6b)). The histogram of the residuals is bi-modal as a consequence of large number of grades below 1.

The estimated standard deviation of the residuals is 1.420 on 1,494 degrees of freedom, which is slightly smaller than the standard deviation of the logarithm grade (1.480). Whilst the reduction is small, the samples size is large and three of the coefficient in the fitted quadratic surface are highly statistically significant. An assumption of the residuals of the regression as a realisation of a stationary spatial process was made.

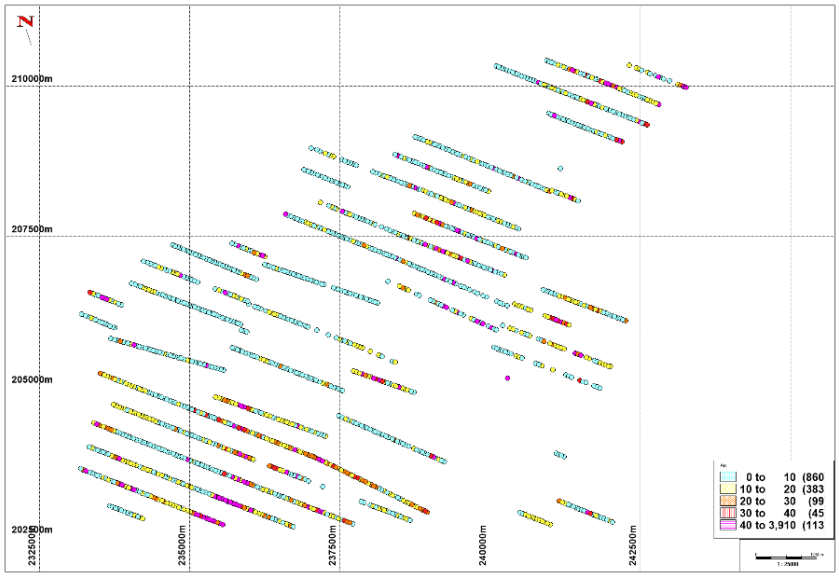


Fig. 2.3a Soil sampling locations of the project area (closed circle). Bottom right: Gold grades, Au in ppm.

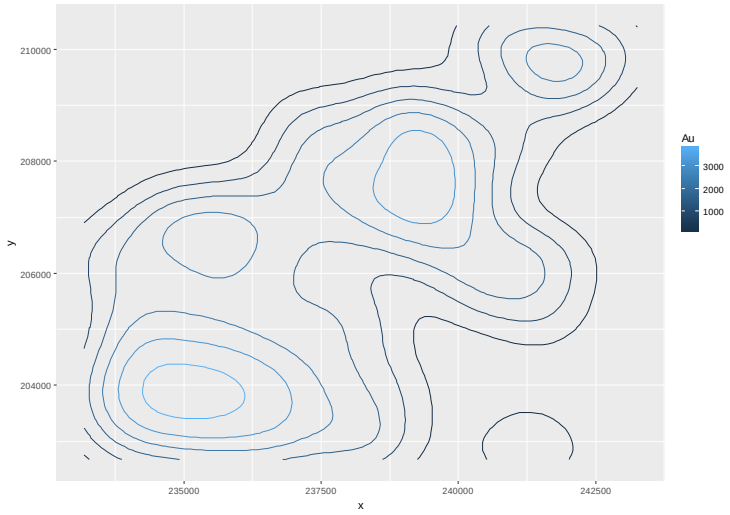


Fig. 2.3b Contour map showing areas of highest gold (Au in ppm) values

The contour plot for the fitted surface is shown in (Fig. 2.6(a)). It is qualitatively similar to the contour plot in (Fig. (2.3b)). Both suggest the higher average gold grades are in north-east and south-west.

Table 2.1 Summary statistics of the grades

Statistics value	
Number of Samples	1500
Minimum Value (ppm)	0.05
Maximum Value (ppm)	3906
First Quartile (25%)	1
Median (ppm)	8
Third Quartile (75%)	14
Mean Absolute Deviation from Median	18.24
Mean (ppm)	20.81
Mean Absolute Deviation from Mean	23.91
Standard Deviation (ppm)	114.66
Kurtosis	891.82
Skewness	27.19

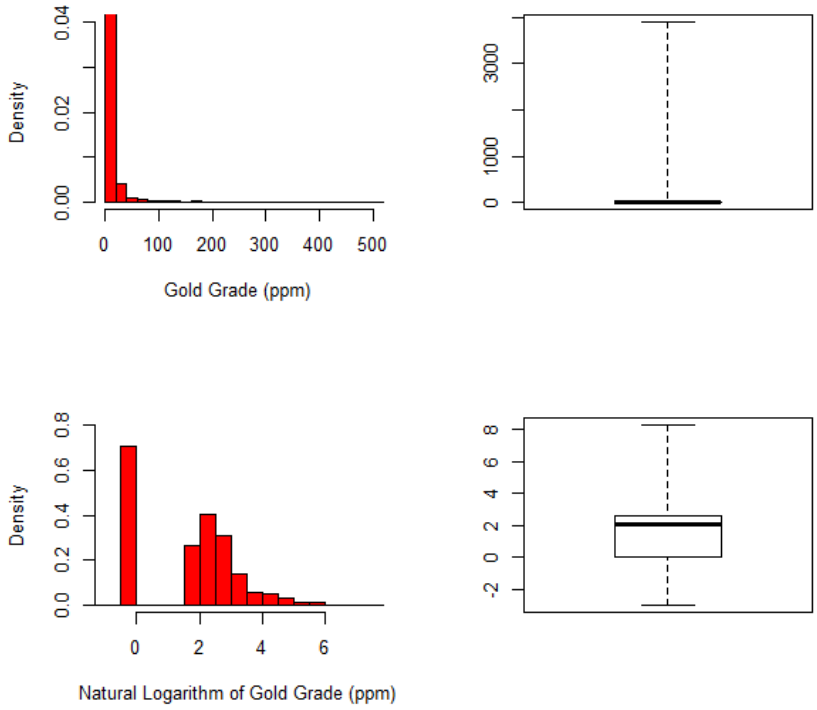


Fig. 2.4 Histogram and boxplot of grades (upper), histogram and boxplot of natural logarithm of grades (lower)

However, the copula analysis may pick out localised areas of high gold grades. The purple, red and blue lines on (Fig. (2.6b)) are the fitted Normal, kernel smoother and Gumbel distribution respectively to the residuals extracted from model. Three transformation of the residuals $\{r_i\}$ to $[0,1]$ were considered corresponding to: an assumed normal distribution $r_i = \Phi\left(\frac{r_i}{s_r}\right)$ where s_r is the standard deviation of the residuals; an assumed Gumbel distribution $r_i = e^{-e^{-\left(\frac{r_i-\hat{\xi}}{\hat{\theta}}\right)}}$ where $\hat{\theta}$ is the scale parameter and $\hat{\xi}$ is the location parameter; and the kernel margin approximated by the empirical distribution $r_{i:n} = \frac{i}{n+1}$ where $r_{i:n}$ is the ith smallest residual when the residuals are sorted in ascending order.

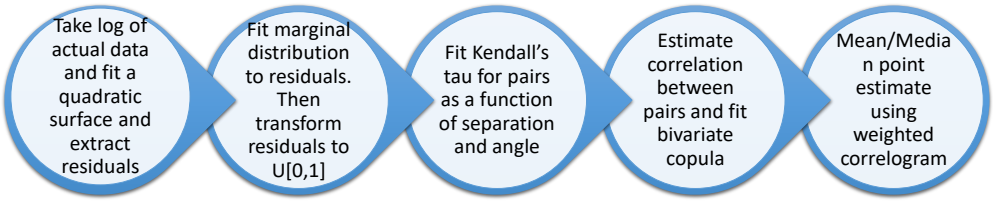


Fig. 2.5 Flow chart showing how spatial pair-copula are used to estimate mean and median at a point

Three hundred meter (0 - 300m) lag classes were found to give sufficient pairs in each class to fit the pair-copula model. Figure 2.7 and 2.8 shows the Kendall tau values against the mean of the distances between pairs for all pairs in the lag classes for the directions (135°) and (45°) directions respectively.

Table 2.2 Estimated coefficients of the fitted generalised least square model (range: 735 and nugget: 0.5)

Coefficient	Estimate	Estimated Standard Error	P-value
β_o	1.496	0.161	0.000
β_1	0.283	0.114	0.013
β_2	-0.453	0.118	0.000
β_3	0.025	0.105	0.811
β_4	-0.008	0.120	0.945
β_5	0.351	0.132	0.008

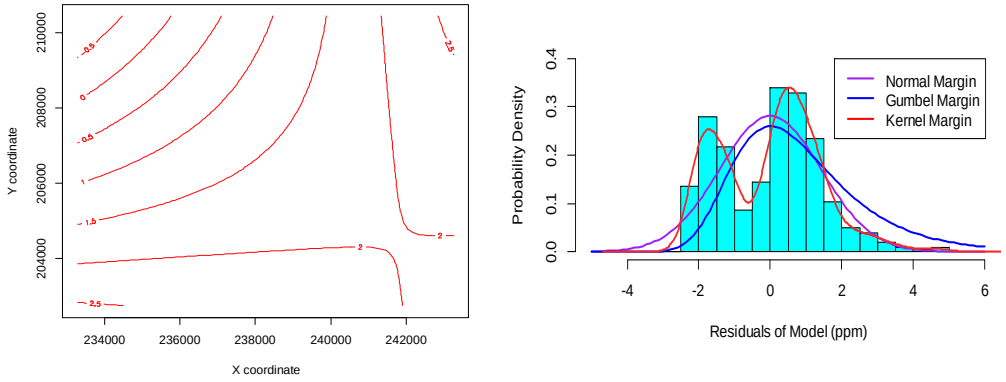


Fig. 2.6 a Contour plot of fitted quadratic surface (left) and **b** Histogram of residuals of model with fitted margins (right) [Normal: purple; Gumbel: blue; and Kernel: red]

From (Figs. 2.7 and 2.8), the estimated ranges of the correlograms for the direction 135° and 45° appears to be around 1,000m and (800 – 1,800m) respectively. The autocorrelation for the mean of all the distance classes was estimated using a polynomial fit with a cubic fit up to the range of the Kendall tau values for the direction 135° and 45° respectively. The weighted correlogram between direction 135° and 45° was estimated from the range parameters using Eq. (8). Figure 2.9, shows the Kendall tau values against the mean of distance classes for the weighted direction (between 135° and 45°).

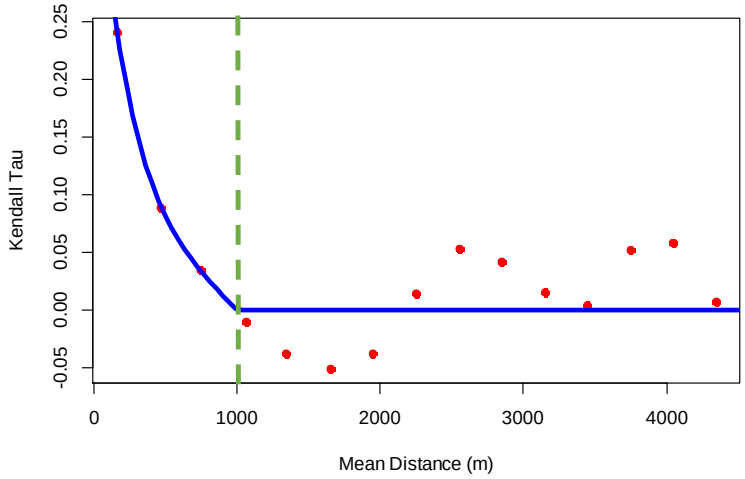


Fig. 2.7 Kendall tau values against the mean of the distance classes (●) for direction (135°)

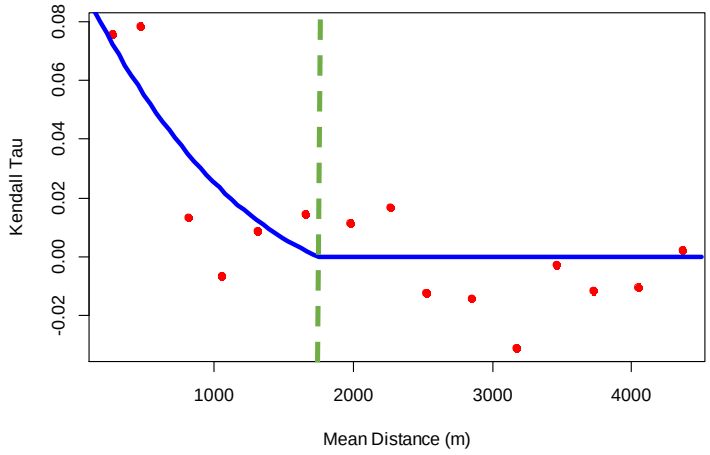


Fig. 2.8 Kendall tau values against the mean of the distance classes (●) for direction (45°)

The Kendall's tau (τ) of the copulas are given in terms of dependence parameters in Appendix B. The method of moments estimate of a dependence parameter can be obtained by evaluating the inverse function, which gives (θ) in terms of (τ), at the estimated value ($\hat{\tau}$) of (τ). However, maximum likelihood is generally considered a more efficient method of estimation and is implemented by the R function *spcopula* (Gräler 2014).

In all ten bivariate copulas, that is Gaussian, Student-*t*, Clayton, Frank, Gumbel, JoeBi, Tawn, and the survival versions of the Clayton, Gumbel and JoeBi copulas were competing at each distance class. The bivariate copulas with the highest log-likelihood value at each distance was chosen as the best fitting copula for that class.

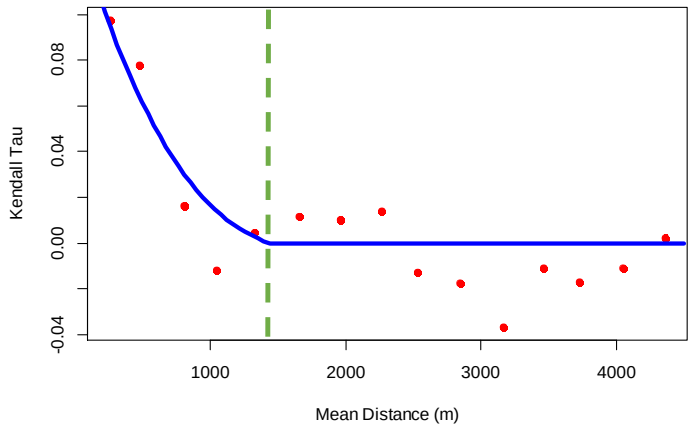


Fig. 2.9 Kendall tau values against the mean of the distance classes (•) for weighted direction (135°/45°)

Table 2.3 Best-fit copulas for each distance class for direction 135°

Class	Copula	Dependence Parameter
0 – 300	Survival Gumbel	1.318
300 – 600	Survival JoeBi	1.170
600 – 900	Survival JoeBi	1.054
900 - 1200	Independence	-
⋮	⋮	⋮
4200 - 4500	Independence	-

The results are given for the use of kernel transformation of the residuals as this was found to give the best results in cross-validation. Table 2.3, 2.4 and 2.5 shows the best fitting bivariate spatial copulas for the direction 135°, 45° and the weighted direction between 45°/135° respectively. Independence copula are assumed beyond the ranges of 900, 1,500 and 1,200 for direction 135°, 45° and weighted 45°/135° respectively.

When the number of pairs within a class are large enough then the kernel smoothing density estimate function *kdecopula* (Nagler 2017) in R was used to generate the contours. Figure 2.10 and 2.11 shows the empirical copula contours for selected distance classes for direction 135° and 45° respectively. If the datasets were isotropic the empirical copula contour plots for the orthogonal directions 135° and 45° would depict the same contours at all distance classes. However, comparing (Figs. 2.10 (a) (b) (c)) and (Figs. 2.11 (a) (b) (c)) it can be observed that at distance of (0 – 300m), (600 – 900m) and (900 – 1,200m) the spatial structure for direction 135° and 45° shows different empirical contour plots. The bivariate empirical copulas are necessarily symmetric. In the 135° direction which is along quartz veins there is a high correlation at relatively low values of grade, for the first three distance classes.

Table 2.4 Best fit copula for each distance class for direction 45°

Class	Copula	Dependence Parameter
0 – 300	JoeBi	1.137
300 – 600	Frank	0.516
600 – 900	Survival Clayton	0.072
900 – 1200	Survival JoeBi	1.040
1200 - 1500	Survival JoeBi	1.021
1500 - 1800	Independence	-
⋮	⋮	⋮
4200 - 4500	Independence	-

Table 2.5 Best fit copula for each distance class for weighted direction between 45° and 135°

Class	Copula	Dependence Parameter
0 – 300	Survival JoeBi	1.184
300 – 600	Frank	0.573
600 – 900	Survival Clayton	0.062
900 – 1200	Survival JoeBi	1.025
1200 - 1500	Independence	-
⋮	⋮	⋮
4200 - 4500	Independence	-

(a) **(b)** **(c)**

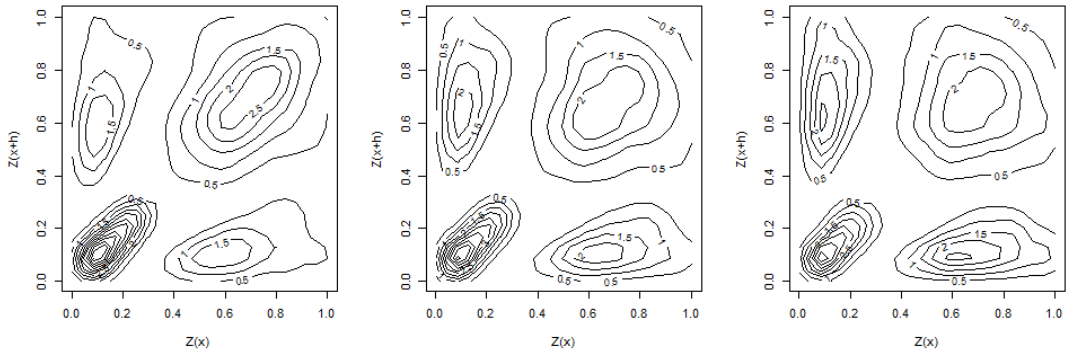


Fig. 2.10 Empirical copula contours of residuals for direction 135° **(a)** 0 - 300m, **(b)** 600 – 900m and **(c)** 900 – 1200m

There is same correlation at higher values, but the conditional distributions appear to be slightly bimodal. That is high values of grade tend to be associated with high values of grade or low values of grade rather than mid-range values of grade at neighbouring points. Similar remarks apply for 45° direction at the second and third distance class, although the bimodality at higher values is more pronounced.

(a) **(b)** **(c)**

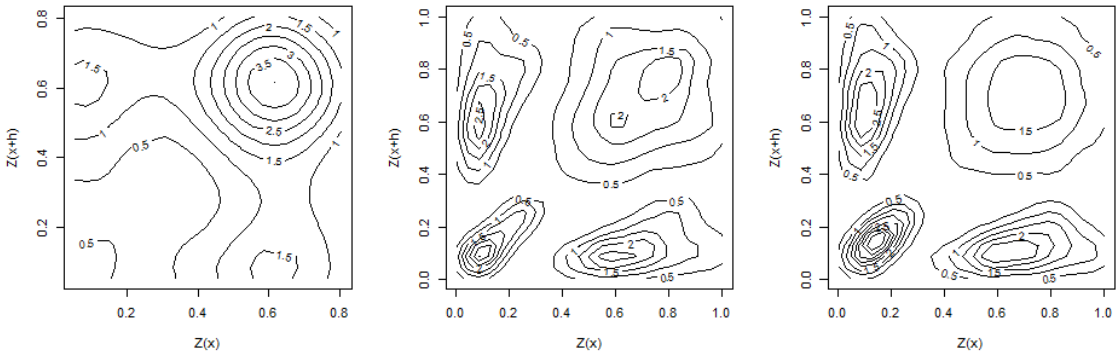


Fig. 2.11 Empirical copula contours of residuals for direction 45° **(a)** 0 - 300m, **(b)** 600 – 900m and **(c)** 900 – 1200m

The first class at 45° shows a higher correlation for high values of grade. However, this is based on a small number of pairs and given that this is an unusual pattern it may be best to attributed to random variation. The difference in the contour plots is further evidence for anisotropy.

To be able to validate the pair-copula model, cross-validation was carried out using log-normal kriging. Variogram analysis were carried out on the residual of the model using Geostatistics for Windows (*GeostatWIN*) software, the Omni directional variogram and contour plot of the variogram on a plane showed two main structures.

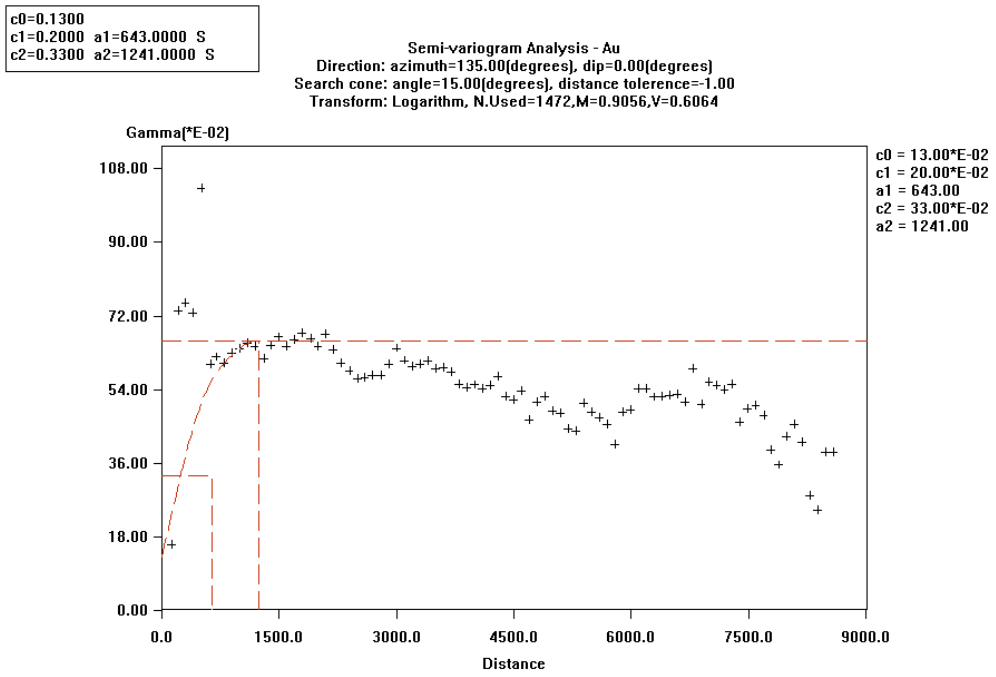


Fig. 2.12 Directional variogram (+) in direction NW-SE (Azimuth of 135°)

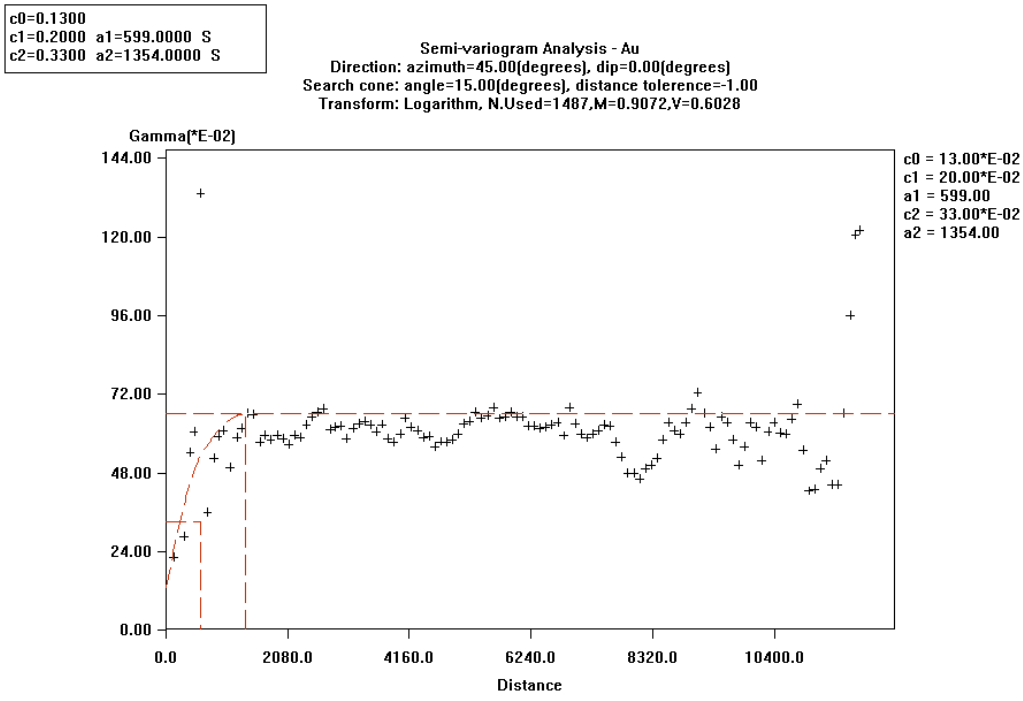


Fig. 2.13 Directional variogram (+) in direction NE-SW (Azimuth of 45°)

Directional variograms were explored in all directions, but the azimuth of 135° North-West (NW) – South-East (NE) showed good spatial structures and hence was used as the major direction, the minor direction was North-East (NE) – South-West (SW). This azimuth direction supports the Au grades data as most of the surface soils samples lie on that azimuth direction (see supplementary material). In all directional variograms two structures were modelled, (Figs. 2.12 and 2.13), shows the directional spherical variograms in the directions NW – SE and NE – SW. An anisotropy ratio of 0.92 calculated from the variogram parameters in (Figs. 2.12 and 2.13) was used in the kriging interpolation process.

Leave-one-out cross-validation was used with 10 nearby location samples in the spatial interpolation process. The performance of the copula-based models and log-normal kriging were evaluated using the MAE and the RMSE on the transformed data. The data were back transformed by adding the predicted residual value to the fitted values of the regression. The exponent of the results were taken after summing, because the surface of the regression were fitted to a log values.

Table 2.6 Summary of the cross-validation

Margin	**Approach**	**RMSE**	**MAE**
Normal	Pair-Copula-Mean	113.002	17.674
Normal	Pair-Copula-Median	113.047	17.436
Gumbel	Pair-Copula-Mean	114.751	18.032
Gumbel	Pair-Copula-Median	115.276	18.312
Kernel	Pair-Copula-Mean	113.025	17.650
Kernel	Pair-Copula-Median	113.027	17.371
Kriging	Log-Normal-Kriging	110.454	18.259

The mean and median estimators for the pair-copula model with normal and kernel margins performed better in terms of the MAE compared with the pair-copula model with Gumbel margins and log-normal kriging. Table 2.6 shows the cross-validation results, it can be observed that the pair-copula-median with kernel margins had the smallest MAE (17.371) followed by the pair-copula-median with normal margin.

2.4 Conclusion

The gold grades of 1,500 surface soil samples from a mine in Ghana have been analysed and show slight, but statistically significant, non-stationarity. The non-stationarity has been accounted for by fitting a quadratic regression surface, by generalized least squares with a spherical variogram to approximate the spatial correlation. The residuals from the regression were considered as a realisations of a stationary process. All possible pairs of residuals from this regression were grouped by their distance of separation, within two orthogonal angles (45 and 135 degrees). So the pair copula relates to a notional average between the two correlogram. The Kendall tau (τ) rank correlation was calculated from the pairs of grades in each distance class. There was evidence of spatial correlation which depends on the direction of the vector joining the points. The correlation had a longer range in the 45 degrees direction than in the 135 degrees direction. This anisotropy was also apparent in the contour plots of the empirical copulas. The spatial pair-copula was fitted by maximum likelihood and gives the full conditional distribution of the grade at any location.

In general, the k-dimension copula can be used to predict the grade at a point, conditional on known grades at the nearest $(k-1)$ neighbouring points. In this application there was no benefit from increasing k beyond 11. The spatial copula is the multivariate distribution of a grade at k points, and is proportional to the conditional distribution of grade at a single point when grades at the remaining $k-1$ points are known. The mean or median of the conditional

distribution can then be used as the estimate of grade. Moreover, the full conditional distribution allows for limits of prediction to be set. In this investigation, the errors at all points in the field were calculated as the difference between predicted values, the predictions being made from the nearest ten (10) neighbouring points, and the observed values. Then MAE and the RMSE were calculated. This procedure was repeated for lognormal kriging.

Kriging is optimal if a random field is Gaussian, and the logarithms of grade are typically approximated as Gaussian. However, in this application the logarithms of grade still had a positively skewed distribution, and this is not unusual. Extreme outlying values are sometimes capped before kriging but this will tend to bias results. Copulas are relatively insensitive to outlying values and their use is better justified than capping. The spatial pair-copula mean and median model with normal and kernel margins had slightly smaller MAE than log-normal kriging. However, the log-normal kriging model had slightly lower RMSE compared with pair-copula models. In the mining industry it is common to use MAE as a measure of performance because this is less sensitive to extreme outlying values. On the basis of this study the pair copula model appears to have a slight advantage in terms of MAE, a similar conclusion to that of Musafer et al. (2016). The main advantage of copulas is that they provide for realistic modelling of the uncertainty associated with predictions, rather than assuming normal distributions of logarithms of grade. This is important when companies incorporate uncertainty in their business decisions. Moreover, if there are sufficient data copulas be based on local marginal distributions, rather than assuming a common distribution for the entire field, while the correlation structure is estimated from the entire field. Spatial pair-copula models however have some limitations; high computational time are required to fit the full conditional distribution at nearby locations where the number of samples are large. Bardossy and Li (2008), proposed as approximation to estimating the number of nearby samples that are used to fit the full conditional distribution.

The main practical objective of this analysis was to model the anisotropic surface soils samples grades and determine regions of highest gold values within the field for future drilling campaign. From the spatial pair-copula predictions of surface soil samples grades, the mining company can make informed business decision to proceed with a drilling campaign. The advice is to concentrate on the north east corner of the surveyed field.

Acknowledgements

This research is supported by Australian Government Research Training Program Scholarship awarded to Mr. Emmanuel Addo Jr. The authors will like to thank the mining company for providing the surface soil sample datasets used in this case-study. The authors will like express their gratitude to the reviewers for their comments and suggestions, which have improved the practical application of this manuscript.

Appendix A

Constructing the Vine Copula

The aim of this appendix is to explain the concept of pair-copula models in the context of spatial distributions. The general principles can be demonstrated with a trivariate copula for modelling grade at three locations. There are three stages: defining the model; fitting the model; and making predictions from the model.

Defining the Model

A trivariate pdf for the grades at three locations $f(z_1, z_2, z_3)$ can be factorized as using the definition of conditional probability, and the corresponding multiplicative rule of probability

$$f(z_1, z_2, z_3) = f(z_3|z_1, z_2)f(z_2|z_1)f(z_1)\,.$$

From the definition of the copula

$$f(z_1, z_2) = c_{12}\big(F(z_1), F(z_2)\big)f(z_1)f(z_2)\,.$$

But also

$$f(z_1, z_2) = f(z_2|z_1)f(z_1),$$

So

$$f(z_2|z_1) = c_{12}\big(F(z_1), F(z_2)\big)f(z_2)\,.$$

The same argument can be applied with all statements conditional on z_1, to give

$$f(z_3|z_2, z_1) = c_{23|1}\big(F(z_2|z_1), F(z_3|z_1)\big)f(z_3|z_1),$$

$$= c_{23|1}\big(F(z_2|z_1), F(z_3|z_1)\big)c_{13}\big(F(z_1), F(z_3)\big)f(z_3)\,.$$

Therefore, combining the above results gives

$$f(z_1, z_2, z_3)$$
$$= c_{23|1}\big(F(z_2|z_1), F(z_3|z_1)\big)c_{12}\big(F(z_1), F(z_2)\big)c_{13}\big(F(z_1), F(z_3)\big)f(x_3)f(x_2)f(x_1)\,.$$

This is the justification for Eq. (4a) and the corresponding diagram in (Fig. 1(a))

In general the marginal densities $f(x_1), f(x_2)$ and $f(x_3)$ are different but in the spatial application there is only one marginal distribution of grade and $f(\,)$ is the common distribution. The copulas c_{12} and c_{13} are defined by their type and values of their parameters. The conditional copula $c_{23|1}$ is also defined by its type and the value(s) of its parameter(s). However the arguments of the conditional copula are also conditional on z_1, but they can be expressed in terms of unconditional copulas and this is key to the pair-copula construction. The key result is

$$F(z_2|z_1) = \frac{\partial C_{12}\big(F(z_1), F(z_2)\big)}{\partial F(z_1)},$$

which is proved by the following argument. From its definition

$$F(z_2|z_1) \approx P\left(Z_2 < z_2|\, z_1 - \frac{\partial z_1}{2} < Z_1 < z_1 + \frac{\partial z_1}{2}\right),$$

for small ∂z_1. By the definition of conditional probability

$$= \frac{P\left(Z_2 < z_2 \; and \; z_1 - \frac{\partial z_1}{2} < Z_1 < z_1 + \frac{\partial z_1}{2}\right)}{P\left(z_1 - \frac{\partial z_1}{2} < Z_1 < z_1 + \frac{\partial z_1}{2}\right)},$$

$$= \frac{F\left(z_1 + \frac{\partial z_1}{2}, z_2\right) - F\left(z_1 - \frac{\partial z_1}{2}, z_2\right)}{f(z_1)\partial z_1}.$$

Now let $\partial z_1 \to 0$ to obtain

$$F(z_2|z_1) = \frac{\partial F(z_1, z_2)}{\partial z_1} \cdot \frac{1}{f(z_1)}.$$

Now, by definition of the copula

$$F(z_1, z_2) = C\big(F(z_1), F(z_2)\big),$$

so

$$\frac{\partial F(z_1, z_2)}{\partial z_1} = \frac{\partial C\big(F(z_1), F(z_2)\big)}{\partial F(z_1)} \cdot \frac{\partial F(z_1)}{\partial (z_1)},$$

$$= \frac{\partial C\big(F(z_1), F(z_2)\big)}{\partial F(z_1)} f(z_1),$$

and substitution into the last expression for $F(z_2|z_1)$ gives the result. Writing $F(z_1) = u_1$, $F(z_2) = u_2$ results in

$$C(u_2|u_1) = \frac{\partial C_{12}(u_1, u_2)}{\partial u_1}.$$

$C(u_2|u_1)$ is often referred to as an "h" function. $h(u_2, u_1) = C(u_2|u_1)$

Joe (1996) proves this result in a general case of multiple conditioning variables. The conditional cumulative distribution functions required for generating the density of the full

copula are calculated using the partial derivatives of all the copulas involved. The set of indices of the conditional variables are denoted by v and the set of indices excluding j by $-j$

$$\mathrm{F}_{i|v}(z_i) = \frac{\partial C_{ij|v-j}\left(\mathrm{F}_{i|v-j}(z_i), \mathrm{F}_{j|v-j}(z_j)\right)}{\partial \mathrm{F}_{j|v-j}(z_j)}.$$

Fitting the Copula

The choice of a suitable marginal distribution of grade can be based on the histogram of the grades, or some transformation of grades. If there are many data an empirical kernel smoother may be preferable to some parametric distribution.

The two variable copulas that are considered for the pair-copula construction all have a single parameter that determines the strength of association. The parameter can be expressed as a function of Kendall's tau and the methods of moments estimate of the parameter is obtained by substituting the sample estimate of Kendall's tau. In the case of the unconditional copulas, all possible pairs of grades were divided into distance classes and Kendall's tau was estimated within distance class. These estimates can be interpolated to give an estimate of Kendall's tau for any distance separating the two points.

The value of Kendall's tau for the conditional copula can be obtained as follows. Within each distance class consider all possible pairs of pairs with a common location, that is of the form $\{(z_2,z_1), (z_3,z_1)\}$, and for each pair of pairs construct a pair (z_2,z_3) that is conditioned on z_1. The distances between z_2 and z_1, and between z_3 and z_1 are approximately equal as they are in the same distance class, but the distance between z_2 and z_3 is not so constrained. So the (z_2,z_3) pairs can be classed by the distance between z_2 and z_3. Then Kendall's tau can be calculated within these classes, and hence interpolated for any distances between z_2 and z_3 for a given category of z_1.

Predictions

The fitted model is defined in terms of: the fitted marginal distribution, the forms of the copulas and the parameters for each copula which are obtained from the estimated Kendall's tau, which in turn depends on the distances between the three points. The aim of fitting the copula is to predict grade at one of the points given numerical values of grade at the other two. The purpose of the prediction may be to: predict grade at an unknown point in the region; predict grade at a point slightly outside the region; or to predict grade at a point where the grade is known so as to calculate a prediction error and compare prediction strategies. The method is the same in all three cases. The fitted copula is the trivariate distribution $f(z_3|z_1, z_2)$ and the conditional distribution of z_3 is required, say, on z_1 and z_2, $f(z_1, z_2, z_3)$ and then the prediction is the mean, or possibly median of this distribution. The Metropolis-Hastings (M-H) algorithm (e.g., Chibb and Greenberg, 1995) provides a neat solution for finding the mean or median. The M-H algorithm is used to make random draws from $f(z_3|z_1, z_2)$ and so build up the conditional distribution as the distribution of these draws. The mean, median and prediction intervals can all be calculated from this conditional distribution. It is easiest to sample $F(z_3)$ from the pair-copula and then apply F^{-1} to obtain the corresponding conditional distribution of z_3. As for any conditional distribution $c\big(F(z_3)|F(z_1), F(z_2)\big)$ is proportional to $c\big(F(z_1), F(z_2), F(z_3)\big)$ with $F(z_2)$ and $F(z_3)$ replaced by their numerical values. Is referred to as $c\big(F(z_1), F(z_2), F(z_3)\big)$ with numeric values substituted for $F(z_2)$ and $F(z_3)$ as $g(w)$, where w= $F(z_3)$.A great advantage of the M-H algorithm is that the constant of proportionality (normalizing factor) is not needed. A suitable M-H algorithm implementation is

1. Select an initial value for w, $w_i (where\ i = 1)$
2. Randomly draw U_1 and U_2 from a uniform distribution on [0,1]

$$w^* = w_i + \left(0.5 + \frac{U_1}{10}\right),$$

$if\ g(w^*) > g(w_i)\ then\ w_i + 1 = w^*$,

$if\ \frac{g(w^*)}{g(w)} < U_2\ then\ w_{i+1} = w^*$,

$else\ w_{i+1} = w_i$.

3. Continue with Step 2 as many times as needed to build up the distribution (e.g., 10^5 or 10^6)

Appendix B

See Table 2.7.

Table 2.7 The density functions of all used copulas. Nelsen (2016) and Joe (1996)

Frank Copula	$-\frac{1}{\theta}\log\left[1+\frac{(\exp(-\theta u_1)-1)(\exp(-\theta u_2)-1)}{\exp(-\theta)-1}\right]$	$\theta \in [-1,1]$	$\tau(\theta)=1-\frac{4}{\theta}\left(1-\frac{1}{\theta}\int_0^{\theta}\frac{a}{e^a-1}da\right)$
Joe Copula	$1-\left[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right]$	$\theta \in [1,\infty]$	$\tau(\theta)=1+\frac{4}{\theta^2}\int_0^1 x\log(x)(1-x)^{2(1-\theta)/\theta}\,dx$
Clayton Copula	$\left[\max\{{u_1}^{-\theta}+{u_2}^{-\theta}-1;0\}\right]^{-1/\theta}$	$\theta \in [(-1,\infty)\backslash\{0\}]$	$\tau(\theta)=\frac{\theta}{\theta+2}$
Gumbel Copula	$\exp\left[-\left((-\log(u_1))^{\theta}\right)+\left(-\log(u_2)^{\theta}\right)\right]^{1/\theta}$	$\theta \in [1,\infty]$	$\tau(\theta)=1-\theta^{-1}$
Independence	u_1u_2	-	-
Gaussian Copula	$\Phi_2\{\Phi^{-1}(u_1),\Phi^{-1}(u_2);\theta\}$	$\theta=\rho$	$\tau(\theta)=\frac{2}{\pi arcsin(\theta)}$
Student Copula	$\left(t_v^{-1}(u_1),\ldots.,t_v^{-1}(u_d)\right)$	$\theta=\rho$	$\tau(\theta)=\frac{2}{\pi arcsin(\theta)}$
Survival Clayton Copula	$\left[\max\{{u_1}^{-\theta}+{u_2}^{-\theta}-1;0\}\right]^{-1/\theta}$	$\theta \in [(-1,\infty)\backslash\{0\}$	$\tau(\theta)=\frac{\theta}{\theta+2}$
Survival Gumbel Copula	$\exp\left[-\left((-\log(u_1))^{\theta}\right)+\left(-\log(u_2)^{\theta}\right)\right]^{1/\theta}$	$\theta \in [1,\infty]$	$\tau(\theta)=1-\theta^{-1}$
Survival Joe Copula	$1-\left[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right]$	$\theta \in [1,\infty]$	$\tau(\theta)=1+\frac{4}{\theta^2}\int_0^1 x\log(x)(1-x)^{2(1-\theta)/\theta}\,dx$
Tawn Copula	$\exp(\log u_1u_2)A\left(\frac{\log u_2}{\log(u_1,u_2)}\right)$	A ∈ [0,1]	$\tau(A)=\int_0^1\frac{t(1-t)}{A(t)}dA'(t)$

References

Aas, K, Czado, C, Frigessi, A & Bakken, H 2009, 'Pair-copula constructions of multiple dependence', *Insurance: Mathematics and economics*, vol. 44, no. 2, pp. 182-198.

AghaKouchak, A 2014, 'Entropy–copula in hydrology and climatology', *Journal of Hydrometeorology*, vol. 15, no. 6, pp. 2176-2189.

Bárdossy, A 2006, 'Copula-based geostatistical models for groundwater quality parameters', *Water Resources Research*, vol. 42, no. 11, pp. 1- 12.

Bárdossy, A & Kundzewicz, ZW 1990, 'Geostatistical methods for detection of outliers in groundwater quality spatial fields', *Journal of Hydrology*, vol. 115, no. 1, pp. 343-359.

Bárdossy, A & Li, J 2008, 'Geostatistical interpolation using copulas', *Water Resources Research*, vol. 44, no. 7.

Bedford, T & Cooke, RM 2002, 'Vines: A new graphical model for dependent random variables', *Annals of Statistics*, pp. 1031-1068.

Cherubini, U, Luciano, E & Vecchiato, W 2004*, Copula methods in finance, John Wiley & Sons.*

Chib, S & Greenberg, E 1995*, 'Understanding the Metropolis-Hastings algorithm', The american statistician, vol. 49, no. 4, pp. 327-335.*

Chollete, L, Heinen, A & Valdesogo, A 2009, 'Modeling international financial returns with a multivariate regime-switching copula', *Journal of financial econometrics*, p. nbp014.

Czado, C 2010, 'Pair-copula constructions of multivariate copulas', *Copula theory and its applications*, Springer, pp. 93-109.

De Michele, C & Salvadori, G 2003, 'A generalized Pareto intensity-duration model of storm rainfall exploiting 2-copulas', *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D2.

Embrechts, P, Lindskog, F & McNeil, A 2001, 'Modelling dependence with copulas', *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

Erhardt, TM, Czado, C & Schepsmeier, U 2015, 'R-vine models for spatial time series with an application to daily mean temperature', *Biometrics*, vol. 71, no. 2, pp. 323-332.

Favre, AC, El Adlouni, S, Perreault, L, Thiémonge, N & Bobée, B 2004, 'Multivariate hydrological frequency analysis using copulas', *Water Resources Research*, vol. 40, no. 1.

Genest, C & Favre, A-C 2007, 'Everything you always wanted to know about copula modeling but were afraid to ask', *Journal of hydrologic engineering*, vol. 12, no. 4, pp. 347-368.

Goovaerts, P 1998, 'Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties', *Biology and Fertility of soils*, vol. 27, no. 4, pp. 315-334.

Goovaerts, P, AvRuskin, G, Meliker, J, Slotnick, M, Jacquez, G & Nriagu, J 2005, 'Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan', *Water Resources Research*, vol. 41, no. 7.

Gräler B., & Appel, M (2015). "scopula". http://r-forge.r-project. org/projects/spcopula/.

Graler, B 2014, 'Modelling skewed spatial random fields through the spatial vine copula', Spatial Statistics, vol. 10, Nov, pp. 87-102.

Gräler, B 2014, 'Developing Spatio-temporal Copulas', PhD dissertation, Institute for Geoinformatics, University of M*ü*nster.

Gräler, B & Pebesma, E 2011, 'The pair-copula construction for spatial data: a new approach to model spatial dependency', *Procedia Environmental Sciences*, vol. 7, pp. 206-211.

Haff, IH, Aas, K & Frigessi, A 2010, 'On the simplified pair-copula construction—Simply useful or too simplistic?', *Journal of Multivariate Analysis*, vol. 101, no. 5, pp. 1296-1310.

Hu, L 2006, 'Dependence patterns across financial markets: a mixed copula approach', *Applied financial economics*, vol. 16, no. 10, pp. 717-729.

Joe, H 1996, 'Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters', *Lecture Notes-Monograph Series Kolkata: Institute of Mathematical Statistics*, pp. 120-141.

Journel, A & Alabert, F 1989, 'Focusing on spatial connectivity of extreme-valued attributes: Stochastic indicator models of reservoir heterogeneities', *AAPG Bull.;(United States)*, vol. 73, no. CONF-890404-.

Journel, A & Deutsch, C 1997, 'Rank order geostatistics: A proposal for a unique coding and common processing of diverse data', *Geostatistics Wollongong*, vol. 96, no. 1, pp. 174-187.

Journel, AG & Alabert, FG 1990, 'New method for reservoir mapping', *Journal of Petroleum technology*, vol. 42, no. 02, pp. 212-218.

Kazianka, H & Pilz, J 2010, 'Copula-based geostatistical modeling of continuous and discrete data including covariates', *Stochastic Environmental Research and Risk Assessment*, vol. 24, no. 5, pp. 661-673.

Kazianka, H & Pilz, J 2011, 'Bayesian spatial modeling and interpolation using copulas', *Computers & Geosciences*, vol. 37, no. 3, pp. 310-319.

Kurowicka, D & Cooke, RM 2006, *Uncertainty analysis with high dimensional dependence modelling*, John Wiley & Sons.

Marchant, B, Saby, N, Jolivet, C, Arrouays, D & Lark, R 2011, 'Spatial prediction of soil properties with copulas', *Geoderma*, vol. 162, no. 3, pp. 327-334.

Musafer, GN & Thompson, MH 2016, 'Non-linear optimal multivariate spatial design using spatial vine copulas', *Stochastic Environmental Research and Risk Assessment*, pp. 1-20.

Musafer, GN, Thompson, MH, Kozan, E & Wolff, RC 2013, 'Copula-Based Spatial Modelling of Geometallurgical Variables.pdf', *The Second Ausimm International Geometallurgy Conference / Brisbane, Qld, 30 September - 2 October 2013*.

Nagler, T 2017 'kdecopula: Kernel Smoothing for Bivariate Copula Densities', pp. 1 -21. R package version 0.9.0, URL:https://CRAN.R-project.org/package=kdecopula.

Nelsen, R 2006, 'An introduction to copulas', *Lecture Notes in Statistics. New York: Springer*.

Pinheiro J, Bates D, DebRoy S, Sarkar D and R Core Team (2016). _nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-128, <URL: http://CRAN.R-project.org/package=nlme>.

Rockafellar, RT & Uryasev, S 2002, 'Conditional value-at-risk for general loss distributions', Journal of banking & finance, vol. 26, no. 7, pp. 1443-1471.

Rodriguez, JC 2007, 'Measuring financial contagion: A copula approach', *Journal of empirical finance*, vol. 14, no. 3, pp. 401-423.

Rossi, RE, Mulla, DJ, Journel, AG & Franz, EH 1992, 'Geostatistical tools for modeling and interpreting ecological spatial dependence', *Ecological monographs*, vol. 62, no. 2, pp. 277-314.

Sklar, A 1959, 'Fonctions Deépartion à n Dimensions et Leurs Marges', *Publications de l'Institut de Statistique de l'Université de Paris. l'Université de Paris, Paris*.

Supplementary Material

The concentration of soil samples along lines in the azimuth direction allows for a detailed one-dimensional analysis. The longest unbroken record in the azimuth direction runs from an eastings of 233,514.5E to 238,951.7E and a northing of 202,914.9N to 205,221.2N, and has one hundred and nineteen soil samples points. Figure 2.S1 illustrates the specific sampling locations for the selected line.

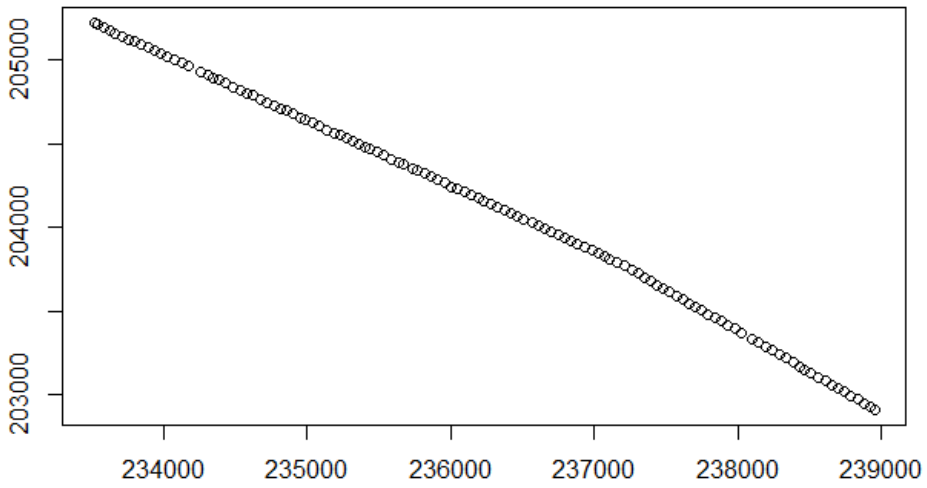


Fig. 2.S1 Specific sampling locations selected from main project area, X and Y axis represent eastings and northings of the project area respectively.

A pair-copula was fitted to triples (x_i, x_{i+1}, x_{i+2}) where $i = 1,2 \ldots . 117$. The results of the fitted copula are rotated Tawn copula 180° (relate parameter to formula in Appendix B) with dependence parameters (par = 1.96 and par1 = 0.46) for (x_i, x_{i+1}), Survival Joe copula with dependence parameter (par = 1.55) for (x_i, x_{i+2}) and finally rotated Tawn copula (180°) with dependence parameters (par = 2.01 and par1 = 0.48) for (x_{i+1}, x_{i+2}) Contour plots for the three fitted copulas are shown in fig. 2.S2.

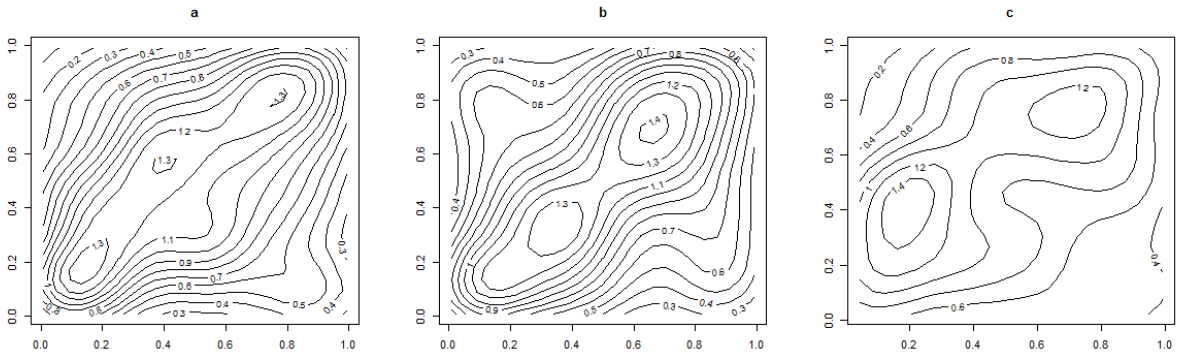


Fig. 2.S2 Empirical copula contours for (x_i, x_{i+1}), (x_i, x_{i+2}) and $(x_{i+1}, x_{i+2})|x_i$ as (**a**), (**b**) and (**c**) respectively

The contour plot for the fitted copula corresponding to (x_i, x_{i+1}) and (x_{i+1}, x_{i+2}) are similar. The contours for the copula corresponding to $(x_{i+1}, x_{i+2})|x_i$ are more spread out because the reduced correlations at lag 2.

An alternative model is to allow for a quadratic trend in the logarithms of grade. Figure 2.S3 (left) shows the scatter plot for all 119 grade samples, whilst fig. 2.S3 (right) illustrates the scatter plot for logarithms of all 119 grade samples.

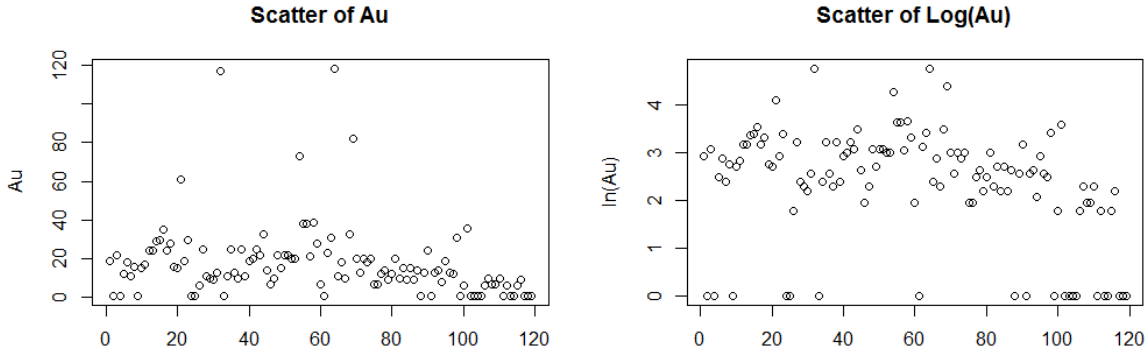


Fig. 2.S3 Scatter plot for all 119 grade samples (right), scatter plot for logarithm of all 119 grade samples (right)

A pair copula was fitted to triples (x_i, x_{i+1}, x_{i+2}) where $i = 1,2 \ldots . 117$ of the residuals from the fitted quadratic trend. The results of the fitted copula are Student-*t* copula with dependence

parameter (par = 0.15 with degree of freedom = 4) for (x_i, x_{i+1}), Frank copula with dependence parameter (par = 0.28) for (x_i, x_{i+2}) and finally Survival Joe copula with dependence parameter (par = 1.15) for (x_{i+1}, x_{i+2}). Contour plots for the three fitted copulas are shown in fig. 2.S4

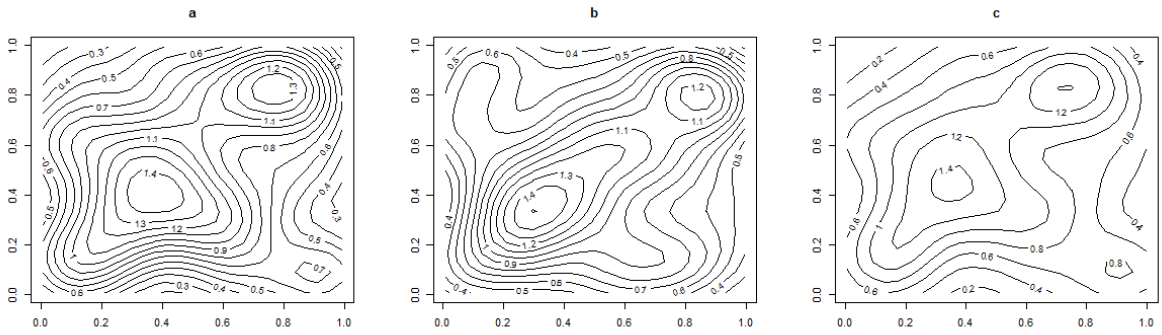


Fig. 2.S4 Empirical copula contours of residuals for the fitted trend for (x_i, x_{i+1}), (x_i, x_{i+2}) and $(x_{i+1}, x_{i+2})|x_i$ as (**a**), (**b**) and (**c**) respectively.

The contour plot for the fitted copula corresponding to (x_i, x_{i+1}) and (x_{i+1}, x_{i+2}) are similar. The contours for the copula corresponding to $(x_{i+1}, x_{i+2})|x_i$ are more spread out because the reduced correlations at lag 2. These conclusions are similar to that drawn from the analysis in fig. 2.S2.

Chapter 3

Estimation of direction of increase of gold mineralisation using pair-copulas (*Paper 2*)

Emmanuel Addo Jr, Emmanuel K. Chanda and Andrew V. Metcalfe

22nd International Congress on Modelling and Simulation Conference, Hobart-Tasmania, published - December 2017

Statement of Authorship

Title of Paper	**Estimation of direction of increase of gold mineralisation using pair-copulas**
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	E Addo Jr, EK Chanda, and AV Metcalfe. "Estimation of direction of increase of gold mineralisation using pair-copulas".In: MODSIM (2017), mssanz.org.au/modsim2017

Principal Author

Name of Principal Author (Candidate)	Emmanuel ADDO JUNIOR
Contribution to the Paper	Developed methodology, conducted programming and execution of methods. Wrote the manuscripts and acted as the corresponding author.
Overall percentage (%)	80%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/01/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Emmanuel KNOX CHANDA
Contribution to the Paper	Supervised the development of work and assisted in reviewing the manuscript
Signature	Date 10/1/2019

Name of Co-Author	Andrew VIGGO METCALFE
Contribution to the Paper	Supervised the development of work and assisted in reviewing the manuscript
Signature	Date 14/1/2019

Abstract

The case study is based on a gold deposit in western Ghana. One thousand five hundred surface soils samples are available on an area of 7810 hectares. The distribution of the grade appears to be a realization of a non-stationary anisotropic random process. The objectives of the analysis are to model the gold grade and to extrapolate into the near surrounding area to determine regions of highest gold value for a future drilling campaign. In the analysis we compare possible transformations of the data to reduce the influence of outliers, and in the case of copulas achieving a marginal uniform distribution. The anisotropy of the gold grades is described with empirical copula density plots for each distance class and for two orthogonal (135° and 45°) directions. The non-stationarity is modelled by regression methods including periodic variation if appropriate. The residuals from the regression are modelled with spatial pair copulas. An investigation of the possible benefits of increasing the number of nearby locations modeled in the spatial pair copula construction is utilized. Predictions are done for unknown sample locations to the North-East (NE) and South-West (SW) of the main field. The different approaches of increasing the nearby points are compared in terms of minimum, maximum and average predicted grade at all unknown sampling locations outside the main field.

KEYWORDS: Copula, Geostatistical modelling, Pair-copula, Kriging

3.1 Introduction

In many geostatistical applications the objective is to predict a spatial variable at points, for which there are no direct measurements, from a neighboring set of measurements. This requires modelling of spatial dependence between points. Regression methods can be used to model spatial trends, and the errors in the regression model can then be taken to represent a stationary field variable. The regression model can be fitted to the field data and the residuals are treated as data from a stationary field. Kriging is the standard method for estimation of grade at unknown points. Kriging depends on the variogram which represents expected values of the squared differences in grade at two points as a function of distance between them. Kriging is optimum if the distribution of the grade is multivariate Gaussian (MVG). However, gold grades are not always well modelled as MVG, even after transformation, as extreme outliers are common. Indicator kriging, which models grade as above or below some threshold, is a modification that is less sensitive to outliers (Journal and Alabert, 1990 and Goovaerts et al. 2005). The indicator approach is highly empirical and is not based on any stochastic model, so it requires large data sets for adequate precisions. In particular, indicators are fitted for each threshold separately which can lead to problems with monotonicity of the estimates.

Copulas are multivariate uniform distributions, provide a versatile model for the spatial distribution of grade variables, including multivariate Gaussian as special case, and offer an alternative to indicator kriging. Examples of spatial applications of copulas include: Bárdossy and Li (2008) used spatial copulas to model groundwater parameters, Gräler (2014) used pair-copula to model particulate matter concentrations (PM10); and Musafer et al. (2016) applied copulas to mining applications.

This paper describes the modelling of anisotropic surface soil sample gold grades using spatial pair-copulas in order to extrapolate into the near surrounding area to determine

regions of highest gold value for a future drilling campaign. The anisotropy of the gold grades is described with empirical copula contour plots for a direction of maximum range and the orthogonal direction. The non-stationarity is modelled by regression and the residuals from the regression are modelled with spatial pair-copulas. The spatial pair-copula model is used to predict gold grade at 20 exterior locations (NE corner and SW corner).

3.2 Methods

3.2.1 Theory of copulas

A copula is a multivariate uniform distribution with all margins uniform over the interval ($U[0,1]$). This is very versatile as any marginal probability distribution can be transformed to $U[0,1]$. Specifically, if Z is a random variable with cumulative distribution (cdf) $F(\,)$ then $F(Z) \sim U[0,1]$. Two examples of bivariate copula cdf are the Gaussian, which is equivalent to fitting a bivariate Gaussian distribution after transforming margins to normality, and Archimedean copulas which have cdf $C(u_1, u_2, .., u_p) = \psi\left(\psi^{-1}(u_1) + \psi^{-1}(u_2) + \cdots + \psi^{-1}(u_p)\right)$, where $\psi: [0,1] \rightarrow [0,\infty]$ is a continuous, strictly decreasing generator function with $\psi(1) = 0$. Sklar's Theorem (1959) states that any multivariate distribution can be expressed as a copula, and an expansion of a copula with any marginal distribution, which can all be different, is a valid multivariate distribution. A key result for pair-copula construction is that the probability density function (pdf) of a bivariate distribution $f(z_1, z_2)$ can be expressed as $f(z_1, z_2) = c(z_1, z_2) f(z_1) f(z_2)$ where $c(,)$ is the copula pdf. Most bivariate copula have a single parameter, which can be expressed as some function of Kendall's tau (Musafer et al., 2016), and which controls the dependence.

3.2.2 Pair copulas

In this application we compare the results obtained with the multivariate distribution of grade at four, six and eleven points. The canonical vine pair-copula construction is a factorization of the multivariate pdf into bivariate copulas (Aas et al. 2009). For three variables the factorization can be expressed as shown in Eq. (1)

$$f_{123}(z_1, z_2, z_3) = f_1(z_1).f_2(z_2).f_3(z_3).c_{12}\big(F_1(z_1), F_2(z_2)\big).$$

$$c_{13}\big(F_2(z_1), F_3(z_3)\big).c_{23|1}\left(F_{2|1}(z_2|z_1), F_{3|1}(z_3|z_1)\right) \quad (1)$$

For four variables the factorisation can be expressed as shown in Eq. (2)

$$f_{1234}(z_1, z_3, z_3, z_4) = f_1(z_1).f_2(z_2).f_3(z_3).f_4(z_4).c_{12}\big(F_1(z_1), F_2(z_2)\big)$$

$$.c_{13}\big(F_1(z_1), F_3(z_3)\big).c_{14}\big(F_1(z_1), F_4(z_4)\big).c_{23|1}\left(F_{2|1}(z_2|z_1), F_{3|1}(z_3|z_1)\right)$$

$$.c_{24|1}\left(F_{2|1}(z_2|z_1), F_{4|1}(z_4|z_1)\right).c_{34|12}\left(F_{3|12}(z_3|z_1, z_2), F_{4|12}(z_4|z_1, z_2)\right) \quad (2)$$

The pattern continues. The steps for fitting and estimation of the bivariate copulas follow.

Step 1: Empirical Bivariate Copula Contour or Densities for Orthogonal Directions

The initial step of copula based geostatistical modelling is to estimate the marginal univariate distribution function of the variable of interest, (Z). Under the stationarity assumptions the distribution for every location of the variable are identical. $F(z)$ can be calculated from all the observations $(z_1, \dots\dots, z_N)$ where N is the total number of samples in the domain. The estimated distribution function is used to transform all the observations to a unit interval $[0, 1]$.

Distances between every pair of observations $z_i(x_i, y_i)\ and\ z_j(x_j, y_j); i \neq j, \forall i, j = 1,2, \dots\dots, N$ are calculated as $h = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Each paired datum,

$\left\{F\left(z(x_i)\right), F\left(z(x_j)\right)\right\}$, is calculated and placed in its respective distance class $[(0, h_1), (h_1, h_2), \ldots\ldots, (h_{l-1}, h_l)]$, where h_l is define as the distance beyond which there is no significant dependence between the observations. The coordinates for each pair$\left\{F\left(z(x_i)\right), F\left(z(x_j)\right)\right\}$, that is in a two-dimensional (2D) case; pt1 $= (x_i, y_i)$ *and* pt2 $= (x_j, y_j)$ are rotated using the rotation matrix in Eq. (3). Given a defined tolerance angle (ε) and an angle(θ), all pairs of points that lies within an angle $\pm\varepsilon$ are selected as the new pairs of points for that directional angle(θ). This step is repeated for all pairs of data in all distance class. The mean distance between pairs of each distance class is computed and that represents the distance for any particular class interval. The principal axis is taken as the direction with the longest range.

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = \begin{pmatrix} cos\,\theta & sin\,\theta \\ -sin\,\theta & cos\,\theta \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \tag{3}$$

same rotation is used to transform (x_j, y_j) to (x_j', y_j'). Points are considered to lie in the direction class θ if

$$slope = \left|\frac{(y_j' - y_i')}{(x_j' - x_i')}\right| < \varepsilon \tag{4}$$

The weighted correlogram, Figure 2.1 is defined as

$$\rho_{(\theta)} = \rho_{(135^\circ)}\left(\frac{\theta}{90}\right) + \rho_{(45^\circ)}\left(\frac{90-\theta}{90}\right) \tag{5}$$

The number of pairs for each distance class is chosen to be large enough for the kernel smoothing density function (KernSmooth) in R to be used to generate contour plots.

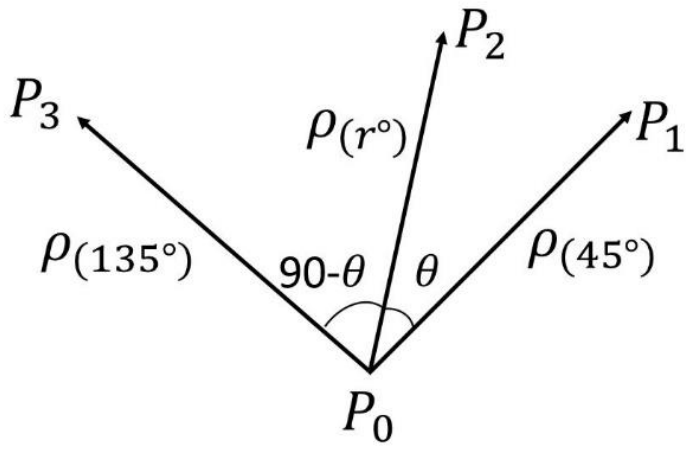


Fig. 3.1 Schematic diagram showing how a weighted correlogram is defined. P_0P_1 is the direction of maximum range

Step 2: Theoretical Bivariate Copula Densities and Spatial Copula Construction

For each distance class we have a number of grade pairs. The grades are transformed to uniform distribution, using the empirical cdf of all grades, and various types of copula are fitted by maximum likelihood. The copula with the highest value of the likelihood is chosen. Furthermore, Kendall's tau, which is a non-parametric measure of correlation based on ranks, is also calculated for each distance class. All selected spatial copulas for all distance class are used to obtain the distance-direction-dependent convex combinations of copula given in Eq. (6). The copula $C_{.,h}$ used in the convex combination in this instance depends on the distance and direction. This can be described by the distance-direction dependent copula $\mathrm{C}_h(\mathrm{u},\mathrm{v})$. This spatial copula must approach the Fréchet–Hoeffding bound Nelsen (1999) $M(u,v) = min(u,v)$ (which shows perfect positive dependence) when the distance approaches zero and the product copula $\prod(u,v) = uv$ (explains independence) when the distance approaches the range where the data are spatially correlated. Instead of restricting the spatial copula to a single

family, the convex combination of copulas for several distances-directions are used. That is $h_1, \ldots\ldots, h_l$, M denotes zero separation and Π denotes the maximum range h_l. Then the spatial copula is given with $\lambda_i := (h_i - h)/(h_i - h_{i-1})$ by

$$C_h(u,v) = \begin{cases} \lambda_1 M(u,v) + (1-\lambda_1)C_{1,h}(u,v), & 0 \le h \le h_1 \\ \lambda_i C_{i-1,h}(u,v) + (1-\lambda_i)C_{i,h}(u,v), & h_{i-1} \le h \le h_i \\ \lambda_l C_{l-1,h}(u,v) + (1-\lambda_l)C_{l,h}(u,v), & h_{l-1} \le h \le h_l \end{cases} \quad (6)$$

Step 3: Pair-Copula Construction and Spatial Interpolation

Finally, the random variable Z at unknown location x_0 also follows the distribution $F(\mathrm{z}|z_1, \ldots, z_k)$ which is conditioned on the known values of grade at the nearest neighbours $x_1, \ldots\ldots, x_k$. The full multivariate distribution of grade $z_1, \ldots, z_k$ is given by

$$F(x, z_1, \ldots, z_k) = P(Z < z, Z_1 < z_1 \ldots, Z_k < z_k) \quad (7)$$

where k is the total number of observations. The conditional distribution is then expressed in terms of the conditional pair-copula $C_{x,k}$

$$F(z|z_1, \ldots z_k) = c(u|u_1, \ldots u_k) f(z) \quad (8)$$

Point estimates of grade at unknown locations x_0 can then be obtained by calculating the mean or the median (Bardossy and Li 2008)

$$\hat{Z}_{mean}(x_0) = \int_0^1 F^{-1}(u) c(u|u_1, \ldots\ldots, u_n) du \quad (9)$$

$$\hat{Z}_{median}(x_0) = F_{\propto}^{-1}\left(u = C^{-1}(0.5|u_1, \ldots., u_n)\right) \quad (10)$$

The pair-copula modeling approach provides the full conditional distribution of the realisations at unknown locations, this full distribution can be used to generate prediction intervals.

3.3 Application

3.3.1 Overview of Project Area

Surface soil samples data from an operating mine in western Ghana have been used to demonstrate the copula based approach for geostatistical modelling. A total of 1500 surface soil samples (2-dimensional data) were taken over an area of 7810 hectares. The spatial correlation of the gold grade in surface soil samples is relatively low due to the mineralization style and the distances between the samples. In the local grid system of the mine, the project extends from an easting of 243258.6E to 233183.8E and a northing of 210425.9N to 202674.5N Figure 3.2 shows a 2D plot of the project area and locations of the surface soil samples. Also shown are the locations at which the grade is to be predicted. All the surface soil grades are in parts per million (ppm). All statistical analyses and the fitting of spatial pair-copula models were conducted using R (R Core Team 2016).

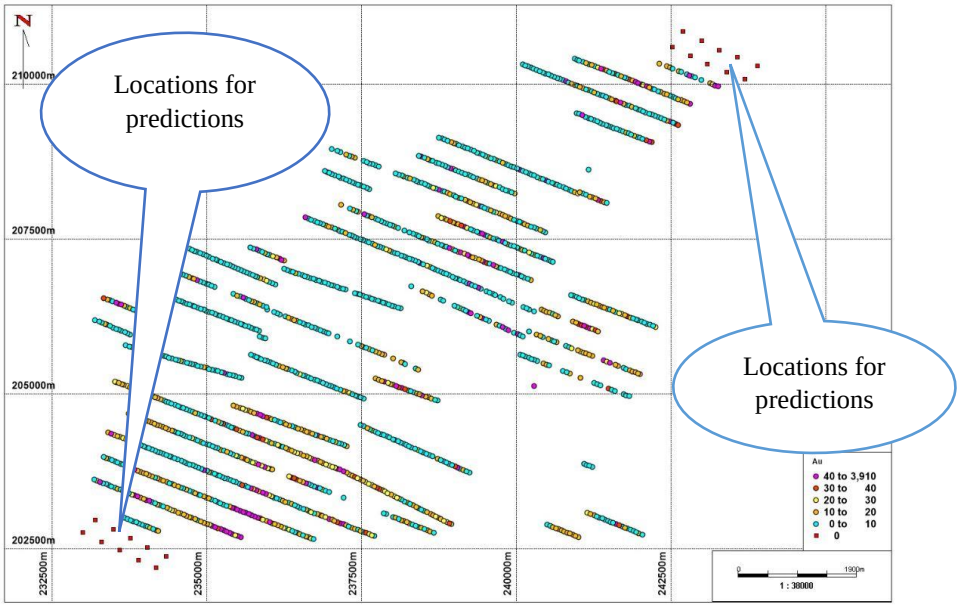


Fig. 3.2 Location of points with soil samples (●) and exterior points (■)

3.3.2 Summary Statistics and Fitting/Predicting from Quadratic Surface

Summary statistics of the surface soil sample grades are given in Table 3.1. The histogram of the grades in Fig.3.3 shows extremely high positive skewness. The non-stationarity of the surface soil grades was modelled by fitting a quadratic surface by regression.

Table 3.1 Summary statistics of the surface soil samples grades

Statistics value	
Number of Samples	1500
Minimum Value (ppm)	0.05
First Quartile (ppm)	1
Median (ppm)	8
Mean (ppm)	20.81
Standard Deviation (ppm)	114.66
Kurtosis	891.82
Skewness	27.19
Third Quartile (ppm)	14
Maximum Value (ppm)	3906
MAE (ppm)	23.91

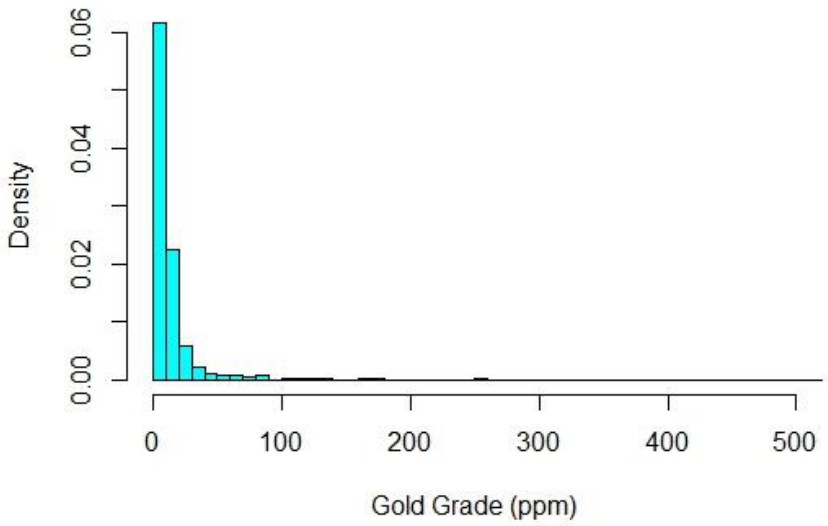


Fig. 3.3 Histogram of the surface soil samples grades

The natural logarithm (w) of the grade value was regressed on the x (X coordinate), y (Y coordinate), x^2, y^2 and the cross product xy. Equation 11 defines the relationship between the logarithm of grade and the coordinates. Where ε is the random error, which is expected to be spatially correlated with mean 0. The spatial correlation of the errors is estimated as a first step in fitting copulas or kriging (section 3.3). It does not affect the estimation of the coefficients but it does affect their standard errors. The estimated standard deviation of the errors is 1.421 on 1494 degrees of freedom, which is slightly smaller than the standard deviation of the logarithm of grade (1.480). The coefficient of multiple determination (R^2) is 0.082.

$$w = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy + \varepsilon \quad (11)$$

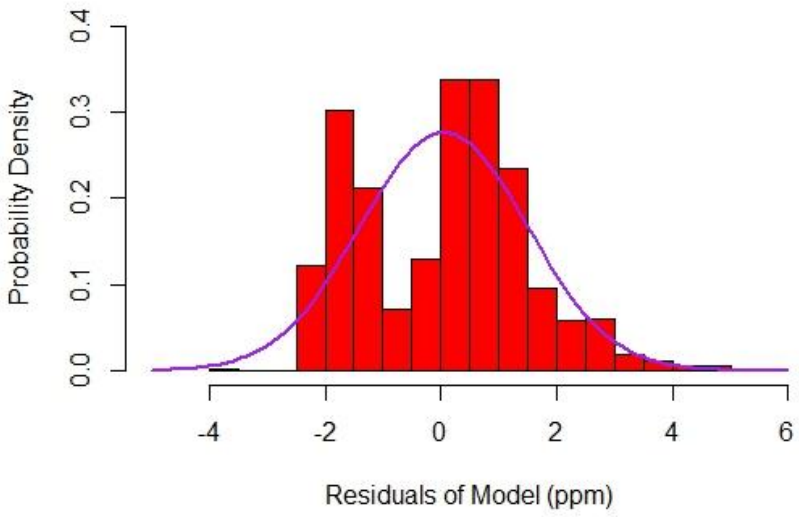


Fig. 3.4 Histogram of residuals from regression model and superimposed Gaussian pdf.

Whilst this is quite low it is highly statistically significant with a sample of this size, even after allowing for the spatial correlation of the errors. We assume the residuals (Fig. 3.4) of the regression are a realisation of a stationary spatial process. The fitted regression surface can be used to make predictions within or slightly beyond the study area. Figure 3.5 shows the contour plot of the fitted quadratic surface. In total, twenty (20) sampling locations outside the project

area (10 in the NE corner and 10 in the SW corner) were identified based on the fitted regression contour as shown in Fig. 3.5 to be the best areas for further exploration. Ten sampling locations to the NE and SW respectively were identified. The fitted regression was used to predict the grade for all exterior points.

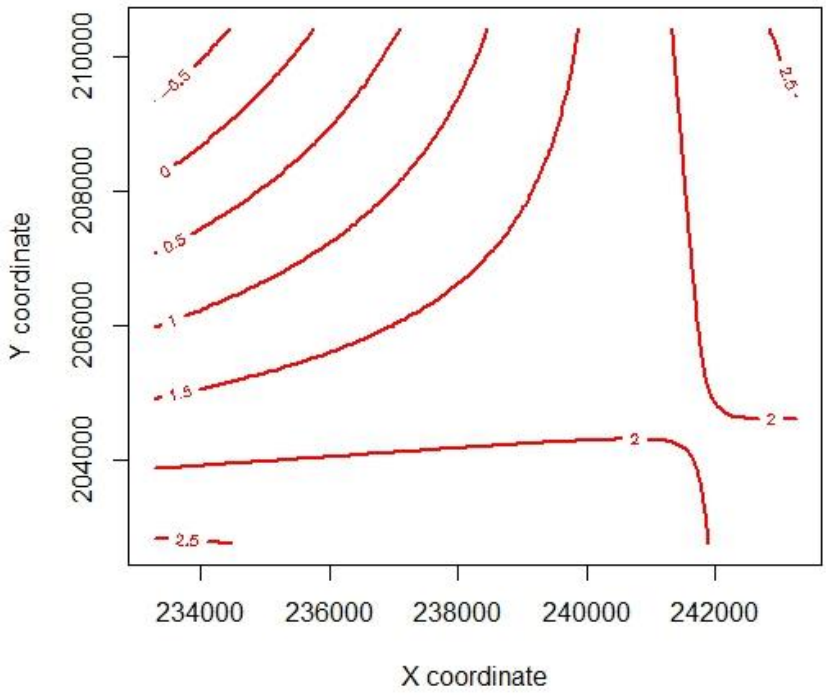


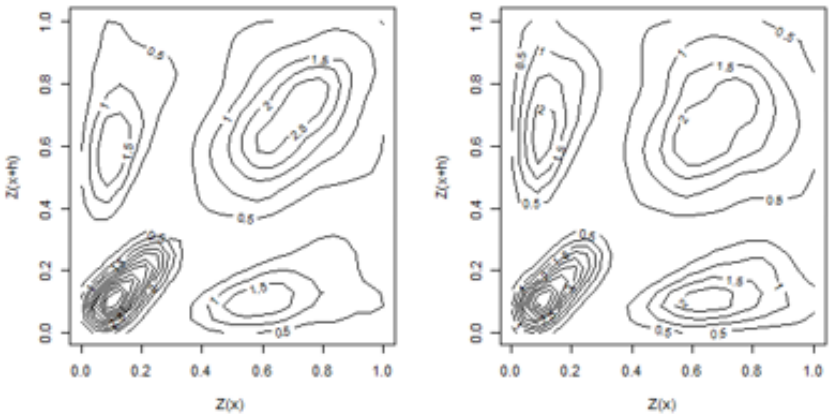
Fig. 3.5 Contour of the fitted regression surface

The expected value of logarithm of grade at a point with coordinates (x_p, y_p) is given by the equation $\hat{w} = \hat{\beta}_o + \hat{\beta}_1 x_p + \cdots + \hat{\beta}_5 x_p y_p$ where $\hat{\beta}_i$the estimated coefficients in equation (11). The means of the predicted logarithms of grade at the 10 SW exterior points and the 10 NE exterior points are both 2.61. However, these predictions ignore the information available from the residuals which are assumed to be correlated with grade at neighbouring points.

3.3.3 Constructing Empirical Copula Contours and Spatial Copula Construction

The residuals from the regression were considered as a realisation of a stationary process. A normal marginal distribution was fitted to the residuals as shown in Fig 3.4, and the residuals appear to be positively skewed. Therefore a rank transformation $r_{i:n} \rightarrow \frac{i}{n+1}$, where $r_{i:n}$ is the ith smallest residual was used to transformed the residuals to the uniform interval [0, 1]. This enabled the construction of empirical copula contours to explore the spatial dependence structure at all lag distances and directions. Using the kernel smoothing density package in R to generate the contours. Figure 3.6 shows the empirical copula contours for directions 135° and 45°, where 45° corresponds to the longest range. It can be observed from Fig. 3.6 that at distance of (0-300m) and (0-600m) the spatial structure for direction 135° and 45° shows different empirical contour plots. The difference in the contour plots provides evidence for anisotropy.

(a) 0-300m (b) 300-600m



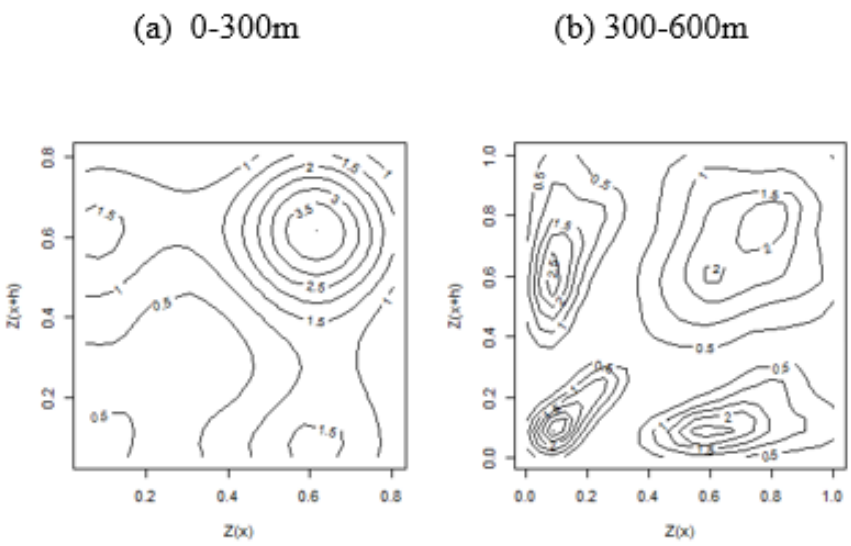


Fig. 3.6 Empirical copula contours of residuals for direction 135° (upper) and 45° (lower)

The Kendall's tau correlogram in Fig. 3.7a & 3.7b for the directions 135° and 45° were estimated with ranges around 1000 and 1100m respectively. Weighted correlogram between 135° and 45° was estimated using equation (5).

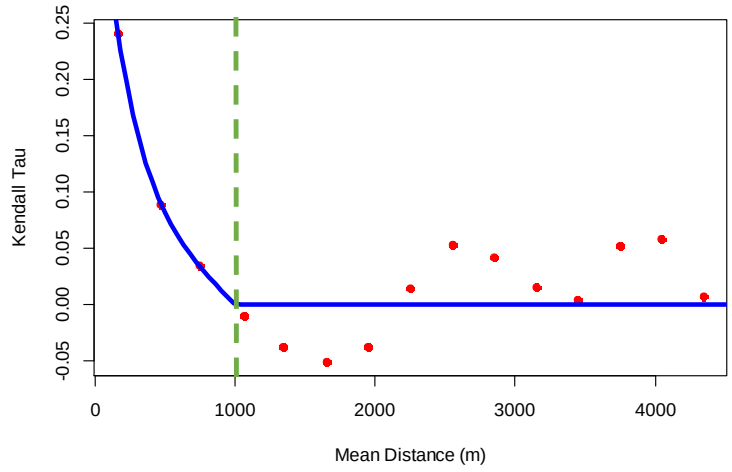


Fig 3.7a Kendall tau values against the mean of the distance classes for direction (135°)

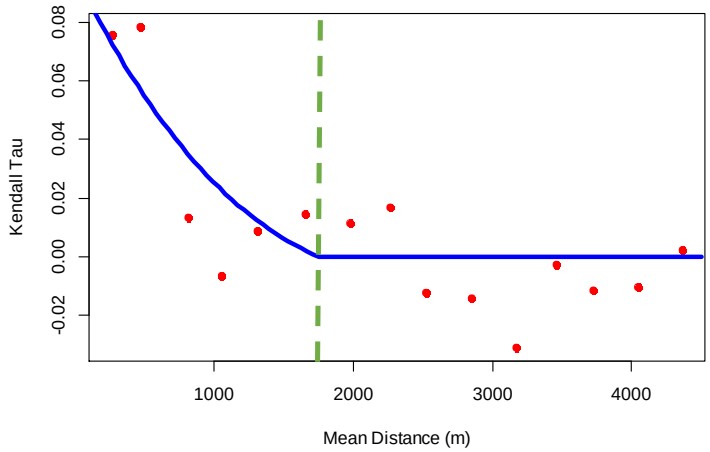


Fig. 3.7b Kendall tau values against the mean of the distance classes for direction (45°)

In all, ten bivariate copulas competed at each distance class, their dependence parameters were estimated by maximum likelihood. The bivariate copulas with the highest likelihood value at each distance was chosen as the best fitting copula for that class.

Table 3.2 Estimates of deviation of logarithm of grade from regression surface using copula (*Results using a randomly selected 95% of the data set are bold and bracketed in italic*)

Nos of Neighbouring Points	**NE Points (Mean Prediction)**			**SW Points (Mean Prediction)**		
	Min Grade	Max Grade	Mean Grade	Min Grade	Max Grade	Mean Grade
3	0.59 ***(0.54)***	2.29 ***(2.18)***	1.59 ***(1.51)***	0.67 ***(0.65)***	2.37 ***(2.39)***	1.79 ***(1.79)***
5	0.80 ***(0.74)***	2.33 ***(2.19)***	1.77 ***(1.68)***	0.77 ***(0.69)***	2.50 ***(2.57)***	1.85 ***(1.82)***
10	0.85 ***(0.83)***	2.48 ***(2.49)***	1.80 ***(1.74)***	1.79 ***(1.48)***	2.64 ***(2.77)***	2.23 ***(2.24)***

The performance of the copula is compared with kriging by calculating the mean absolute error (MAE) at interior points for 10 neighbouring points, as shown in Table 3.3.

Table 3.3 Comparison of pair copula and kriging

Model	MAE
Pair copula (Gaussian Margin)	17.7
Lognormal kriging	18.3

The estimate of expected grade at the exterior points is given by $exp(\hat{w}_p + \hat{r}_p + s^2/2)$ where s^2 is the estimated variance of the errors in equation (11). The adjustment of $exp(\hat{w}_p + \hat{r}_p)$ by the factor of $exp(s^2/2)$ is to allow for the difference between the median and mean of an assumed log-normal distribution. The results using 3, 5 and 10 points for the copula are given in Tables 3.3 and 3.4. An approximate standard error is taken as one quarter of the width of the 95% prediction interval for the mean grade based on the regression surface, which is $\pm$ 4%. This is an approximation because: it does not allow for the correlation between the 10 neighbouring points and so underestimates the error; but it does not allow for the reduction in uncertainty due to use of the copula. The mean estimator for the pair-copula model with normal margin was used. The data were back transformed by adding the predicted residual value to the fitted values of the regression and taking exponential.

Table 3.4 Estimate of the expected grade at the exterior points (*Results using a randomly selected 95% of the data set are bold and bracketed in italic*)

Nos of Neighbouring Points	**NE Points (Mean Prediction)**			**SW Points (Mean Prediction)**		
	Min Grade	Max Grade	Mean Grade	Min Grade	Max Grade	Mean Grade
3	59.86	409.82	231.29	65.16	457.13	263.62
	(55.26)	***(348.76)***	***(201.43)***	***(63.20)***	***(480.31)***	***(266.99)***
5	78.46	434.98	258.45	76.48	522.03	276.59
	(71.45)	***(351.25)***	***(220.17)***	***(71.11)***	***(577.14)***	***(275.15)***
10	71.49	463.56	270.17	208.12	581.39	358.64
	(73.39)	***(447.74)***	***(242.39)***	***(148.47)***	***(627.15)***	***(380.21)***

3.4 Discussion and Conclusions

Gold grades from a mine in Ghana have been analysed. The 1500 gold grades surface soil samples within the field were analysed and show slight, but statistically significant, non-stationarity. The non-stationarity was accounted for by fitting a quadratic regression surface. The residuals from the regression were considered as a realisations of a stationary process. All possible pairs of residuals from this regression were grouped by their distances of separation, and the angle between them. The Kendall tau τ rank correlation was calculated from the corresponding pairs of grades. There was evidence of spatial correlation which depends on the direction of the vector joining the points. The correlation had a longer range in the 45° direction than in the 135° direction. This anisotropy was apparent in the contour plots of the empirical copulas as shown in (Figure 3.6). Copulas were fitted by maximum likelihood and were used to predict the distribution of the grades of residuals outside the main surveyed field (Total of 20 sampling points, 10 apiece at NE and SW respectively). The minimum, maximum and mean predicted grades for all sampling points were estimated for (3, 5 and 10) neighbouring points. Table 3.2 shows the predicted grades based on the contouring of the grade deviation with the regression surface. The minimum, maximum and mean predicted grades for the NE and SW exterior points are highlighted in the table 3.4. The estimates change with the number of neighbouring points which suggest that at least 10 points should be used in this application, particularly in the SW corner. We recommend that the company should concentrate further exploration in the SW corner of the main field.

Acknowledgements

This research is supported by Australian Government Research Training Program Scholarship awarded to Mr. Emmanuel Addo Jr. The authors will like to thank the mining company for providing the surface soil sample datasets used in this case-study. The authors will like express their gratitude to the reviewers for their comments and suggestions, which have improved the practical application of this manuscript.

References

Aas, K, Czado, C, Frigessi, A & Bakken, H 2009, 'Pair-copula constructions of multiple dependence', Insurance: Mathematics and economics, vol. 44, no. 2, pp. 182-198.

Bárdossy, A & Li, J 2008, 'Geostatistical interpolation using copulas', Water Resources Research, vol. 44, no. 7.

Goovaerts, P, AvRuskin, G, Meliker, J, Slotnick, M, Jacquez, G & Nriagu, J 2005, 'Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan', Water Resources Research, vol. 41, no. 7.

Gräler, B 2014, 'Developing Spatio-temporal Copulas'.

Journel, AG & Alabert, FG 1990, 'New method for reservoir mapping', Journal of Petroleum technology, vol. 42, no. 02, pp. 212-218.

Musafer, GN & Thompson, MH 2016, 'Non-linear optimal multivariate spatial design using spatial vine copulas', Stochastic Environmental Research and Risk Assessment, pp. 1-20.

Nelsen, RB 1999, 'Introduction', An Introduction to Copulas, Springer New York, New York, NY, pp. 1-4.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/)

Sklar, A 1959, 'Fonctions Deépartion à n Dimensions et Leurs Marges', Publications de l'Institut de Statistique de l'Université de Paris. l'Université de Paris, Paris.

Chapter 4

A comparison of Gaussian, Student-*t* and Vine copulas for modelling geophysical measurements along a rock drill core (*Paper 3*)

Emmanuel Addo Jr, Andrew V. Metcalfe and Emmanuel K. Chanda

ANZIAM Journal, Volume 59, published - October 2018

Statement of Authorship

Title of Paper	**A comparison of Gaussian, Student-t and Vine copulas for modelling geophysical measurements along a rock drill core**
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	E Addo Jr, EK Chanda, and AV Metcalfe. "A comparison of Gaussian, Student-t and Vine copulas for modelling geophysical measurements along a rock drill core". In: ANZIAM JOURNAL (2018), pp. C216-C230. doi: dx.doi.org/10.21914/anziamj.v59i0.12646

Principal Author

Name of Principal Author (Candidate)	Emmanuel ADDO JUNIOR
Contribution to the Paper	Developed methodology, conducted programming and execution of methods. Wrote the manuscripts and acted as the corresponding author.
Overall percentage (%)	80%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/01/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Emmanuel KNOX CHANDA
Contribution to the Paper	Supervised the development of work and assisted in reviewing the manuscript
Signature	Date 10/1/2019

Name of Co-Author	Andrew VIGGO METCALFE
Contribution to the Paper	Supervised the development of work and assisted in reviewing the manuscript
Signature	Date 14/1/2019

Abstract

This case study is based on measurements made approximately at 20cm lengths along a down-the-hole diamond drill core from a pyrite mine in South Australia. The measurements are the P-wave velocity, magnetic susceptibility and impedance. The trivariate distribution is modelled using Gaussian, Student-*t* and vine copulas and the results are compared in terms of goodness of fit and differences in extreme values from distributions obtained by simulation from the copulas. The vine copula provides the best fit for the variables. Trivariate linear spatial Gaussian, Student-*t* and vine copulas are used to predict magnetic susceptibility one step below the depth of the drill core. The vine copula allows for more detailed modelling of the error structure, and so provides more accurate 90% prediction intervals. The 90% prediction interval for the vine copula is wider than that for the Student-*t* copula, and both are wider than the interval obtained with the Gaussian copula. In general, copulas provides a more realistic modelling of geological variables and hence allows for accurate assessment of risk and uncertainty.

KEYWORDS: Copula; Estimation; Mining

4.1 Introduction

Deep seated orebodies, low metal grades and fluctuating commodity prices have a high impact on the mining industry potentially reducing profit margins. It follows that accurate modelling of all geological variables is needed to reduce the risk associated with mineral prospects. It is claimed that, the precise modelling of geological variables is the most significant factor in the success of mining projects [8]. In many geostatistical applications, kriging is used for predicting geological variables at unknown locations, however this method is only optimum if the distribution of the variables are multivariate Gaussian MVG. In reality, most geological variables exhibit a skewed distribution, which makes kriging inaccurate for predicting at unknown locations. Copula models are ideal in dealing with highly skewed and tail dependent distribution, these models encompasses all multivariate distributions including the MVG [1, 6].

In this paper, the versatility and potential advantages of copula modelling of multivariate relationships in the context of a down-hole diamond drill core taken during prospecting has been demonstrated. The geological variables are impedance, magnetic susceptibility and P-wave velocity. These three geological variables are important path-finder variables when exploring for pyrite mineralisation.

The first objective is to describe the multivariate distribution of the three variables using copulas. The second objective is to compare Gaussian, Student-*t* and vine copula models for predicting magnetic susceptibility at further depths. The vine copula allows for detailed modelling of the error structure and therefore provided an accurate 90% prediction interval for magnetic susceptibility at further depth.

4.2 Methods

This section gives a brief overview of trivariate Gaussian and Student-*t* copulas [7], and a brief description of pair copula construction [5].

Suppose that Z is a continuous random variable with a cumulative distribution function (cdf) $F(z)$. Since Z is a random variable so too is $F(Z)$, and it follows from its definition that $F(Z)$ has a uniform distribution on the interval $[0, 1]$. This is known as the probability integral transform of Z. Copulas are multivariate uniform distributions. They encompass all multivariate distributions because the marginal cdfs are uniformly distributed. Moreover, copulas also yield multivariate distributions by expanding the uniform margins to any probability distributions which can all be different.

The Gaussian copula is equivalent to a multivariate Gaussian distribution and the Student-*t* copula is equivalent to the multivariate t distribution. They can both be applied to multivariate data with a large number of components. There is also a very wide range of different forms of bivariate copulas. However, these bivariate copulas do not generally extend beyond bivariate data without restrictions, such as equal correlations between components that limit their applicability. Vine copulas, also known as pair-copulas, provide a neat solution to this limitation.

The trivariate Gaussian copula belongs to the family of elliptical copulas and is equivalent to the standard MVG model. The cdf of the trivariate Gaussian copula is:

$$C(u_1, u_2, u_3;\ \Sigma) = \ \Phi_\Sigma[\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)], \qquad (1)$$

where Φ_Σ is the standardised MVG cdf and Φ^{-1} is the inverse standard Gaussian cdf. The trivariate Student-*t* copula is defined, with respect to its multivariate Student-*t* distribution, in a similar fashion.

It follows from the multiplicative rule of probability that any multivariate distribution can be factorised in several ways using conditional distributions. In particular, a copula can be factorised as a product of the marginal distributions and the bivariate conditional copulas. Such factorisations are called pair-copula models. The pair-copula decomposition is however not unique, for example, a five dimensional density can have about 240 different forms of construction. Each decomposition expresses the full copula density differently. [5, 2] used a graphical model called the regular vines to arrange the large number of pair-copula constructions. This regular vine copula are made up of special cases of D-vines and canonical vines [3]. Canonical vines are used when one can find a key variable that controls the relationships of the entire datasets. In this application, the canonical vine was selected. One of the variables used in this case study was found to control all the other variables. Moreover, that same variable is an important key path-finder for the exploration of pyrite mineralisation. Figure 4.1 shows the derivation of the canonical vine (C-vine) copula [5] for three variables, and has the density:

$$\mathrm{f}_{123}(z_1, z_2, z_3) = \mathrm{f}_1(z_1).\mathrm{f}_2(z_2).\mathrm{f}_3(z_3).\mathrm{c}_{12}\big(\mathrm{F}_1(z_1), \mathrm{F}_2(z_2)\big),$$
$$.\,\mathrm{c}_{13}\big(\mathrm{F}_1(z_1), \mathrm{F}_3(z_3)\big).\mathrm{c}_{23|1}\left(\mathrm{F}_{2|1}(z_2|z_1), \mathrm{F}_{3|1}(z_3|z_1)\right). \qquad (2)$$

4.3 Results

The Brukunga mine site is located in the Southern Mount Lofty Ranges in South Australia. The town is located 5 km north-east of Nairne and 40 km east of Adelaide. The geology of the project area is primarily the Cambrian calc-silstones lying within the north-south trending Kanmantoo Trough, the youngest sequence in the southern part of the Adelaide Geosyncline. Iron-sulphide mineralisation occurs as three steeply east dipping conformable lenses that are

separated by waste beds. Mineralization is pyrite and pyrrhotite with some minor sphalerite, chalcopyrite, galena and arsenopyrites.

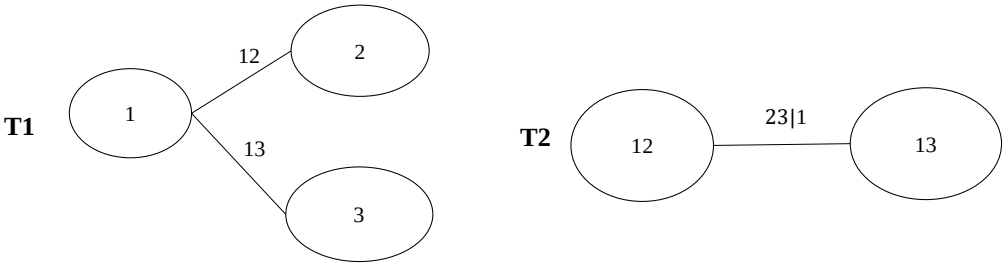


Fig. 4.1 Canonical vine for three variables

A down-the-hole diamond drill core drilled to a total depth of 324.02 m, with an average sampling interval of 20 cm was sampled concurrently for the measurements of magnetic susceptibility, P-wave velocity and impedance. These three variables are the main path-finder variables for the exploration of pyrite mineralisation. Summary statistics of all three geological variables are given in Table 4.1. Histograms of the geological variables is shown in Figure 4.2, negative skewness is evident in impedance and P-wave velocity whilst positive skewness is evident in magnetic susceptibility. The Pearson correlations between the three geological variables are also shown in Figure 4.2. All the correlations are positive and the highest (0.95) is that between the impedance and P-wave velocity.

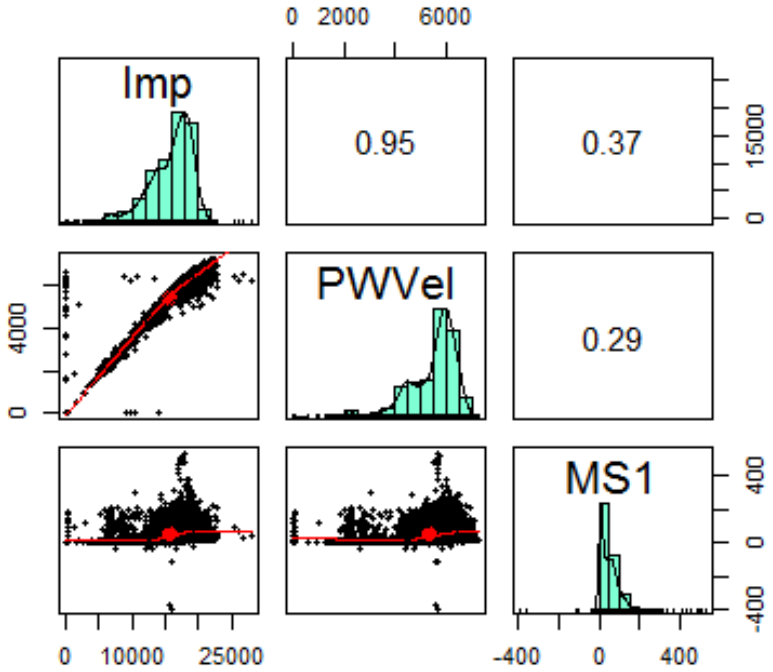


Fig. 4.2 Pearson correlation matrix between all three variables

Table 4.1 Summary statistics of trivariate dataset

Statistics	Impedance (ohm)	P-wave Velocity (km/s)	Magnetic Susceptibility (1)
Number of Samples	11079	11079	11079
Minimum Value	0	0	-389.80
First Quartile	13799	4804	8.89
Median	16625	5742	41.21
Mean	15709.72	5370.29	52.10
Standard Deviation	3540.16	1099.14	49.96
Kurtosis	2.14	3.10	7.69
Skewness	-1.25	-1.57	1.62
Third Quartile	18268	6116	77.02
Maximum Value	27975	7270	532.67

4.3.1 Fitting marginal distributions

The first step in copula modelling is to estimate the marginal distribution functions of the variables. Therefore, the cdfs can be calculated from all ($N = 11079$) observations. Without

any prior information from all three distributions, a non-parametric estimation was chosen for the marginal distributions [9]. A kernel margin approximated by the empirical distribution functions (ecdf) were fitted for all three variables. For example

$$F_{X,T}(x) = \frac{1}{T+1}\sum_{i=1}^{T} I(X_i \leq x) = \frac{\text{rank}(X_i)}{T+1} \quad (3)$$

where I( ) is 1 if its argument is true and 0 otherwise.

4.3.2 Fitting trivariate copulas

The ecdfs are used to transform the trivariate dataset to uniform scale. The trivariate Gaussian copula was fitted by maximum likelihood using the function *fitCopula* in the package Vine Copula [10] in R. Figure 4.3 (upper row) shows contour plots from the trivariate Gaussian copula. These plots represent the marginal bivariate copula density functions of the fitted copula for all the pairs from the triples. The trivariate Student-*t* copula was fitted to the datasets using the function *fitCopula* in the package Vine Copula [10] in R. The degrees of freedom for the fitted *t* distribution was *v* equal to 6. Figure 4.3 (middle row) shows the contour plots from the fitted Student-*t* copula. The C-vine structure was fitted by using the decomposition expressed in the Equation (2), the joint density function of the datasets was estimated. Unconditional survival Joe-Frank (BB8) and survival Joe-Gumbel (BB6) were fitted by maximum likelihood in the first tree using the functions *BiCopSelect* and *RVineMatrix* in the R package Vine Copula [10]. The density of the full pair-copula is a product of all the bivariate copula densities following the decomposition in Tree 1, a rotated 90°Joe-Frank (BB8) copula was fitted in the last tree. A table of all competing copulas for each tree fitting is given in Appendix A.

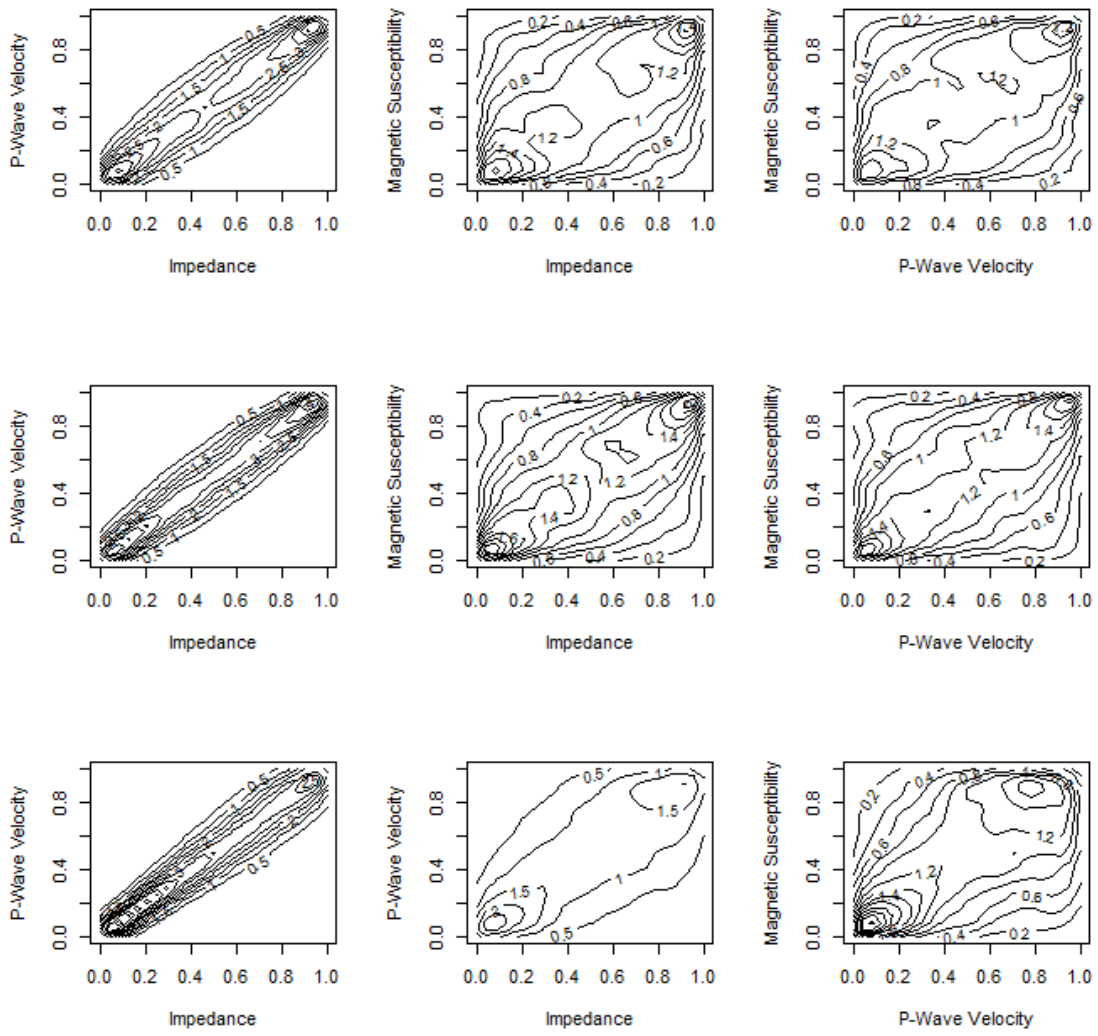


Fig. 4.3 Fitted theoretical contour plots for pairs of variables using Gaussian (upper row), Student-*t* (middle row) and Vine copulas (lower row)

The canonical vine structure explicitly includes the single variable marginal distribution (in this application, the impedance) as the root of the decomposition. Impedance was found to be the key variable that controls the relationship of the dataset based on the Pearson's correlations with all other variables. In addition, impedance is the main path-finder variable for the prospecting of pyrite orebodies. Figure 4.3 (lower row) shows the bivariate contour plots from the fitted C-vine structure.

4.3.3 Goodness of fit test

The cdfs of the three fitted copulas obtained with a Monte-Carlo procedure were plotted against the empirical copula $\hat{\mathrm{C}}_{\mathrm{N}}$ in Equation 4 and Figure 4.4.

$$\hat{\mathrm{C}}_{\mathrm{N}}(u,v,w) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{I}\left(\frac{\mathrm{D}_i}{\mathrm{N}+1} \leq u, \frac{\mathrm{P}_i}{\mathrm{N}+1} \leq v, \frac{\mathrm{M}_i}{\mathrm{N}+1} \leq w\right) \quad (4)$$

where D_i, P_i, M_i are the ranks of the impedance, P-wave velocity and magnetic susceptibility respectively. The plotted points corresponds to $u = v = w$ from 0.01 to 0.99 in steps of 0.01. The measure of fit is based on how close the points are to the diagonal line $y = x$. The vine copula provides a better fit compared with the Gaussian and Student-*t* copulas, because the points appear closer to the $y = x$ diagonal line. This was quantified by calculating the mean of the squared vertical distances between the empirical copula and the fitted copula just before (when the empirical copula is below the fitted copula), or just after (when the empirical copula is above the fitted copula), each of the 11079 data. This gave values of 0.00138, 0.00117 and 0.00023 for the Gaussian, Student-*t* and vine copulas respectively. The best fit, by this measure, is given by the vine copula. We formally compare the fit of the vine copula with the Kolmogorov-Smirnov test. The maximum deviation between the copula cdf and empirical density function is 0.0404. The probability of a value as large or large than this when sampling 11079 random deviates from the vine copula is found by simulation. The differences in the tail proportions of the three copulas and comparison of the field data was investigated by generating (105) random deviates from each copula. This was achieved with the R function *mvdc* in the package copula [4]. The proportion of triples with all three components below 0.05, and above 0.95, quantiles are shown in Table 4.2. The difference between the goodness of fit test for the Gaussian and Student-*t* copula is not apparent from Figure 4.4 (left) and 4.4 (centre). However, the proportion below and above 0.05 and 0.95 quantiles respectively shows that the

Student-*t* copula draws large number of trivariate points to its upper and lower tails which is a consequence of its tail dependence.

Table 4.2 Number (parts per million) of triples below and above 0.05 quantiles and 0.95 quantiles for all three variables

Data	Number of Triples	Number below 0.05	Number above 0.95
Field Observation	11079	3520	632
Gaussian Copula	10^5	8190	8140
Student-*t* Copula	10^5	12990	13010
Vine Copula	10^5	11480	1390

The vine copula allows for tail dependence when it is apparent but can also allow for lower levels of correlation in the tails than the Gaussian copula. In this application, the vine copula gives a considerably better fit in the upper tails, which are more important in mining applications. There is less variation in the fits for the lower tails, but it can be seen that the vine copula is not constrained to be symmetrical between the two tails. Comparing all three fitted theoretical copula to the empirical field observations it seems that the vine copula provides a better fit to the trivariate datasets.

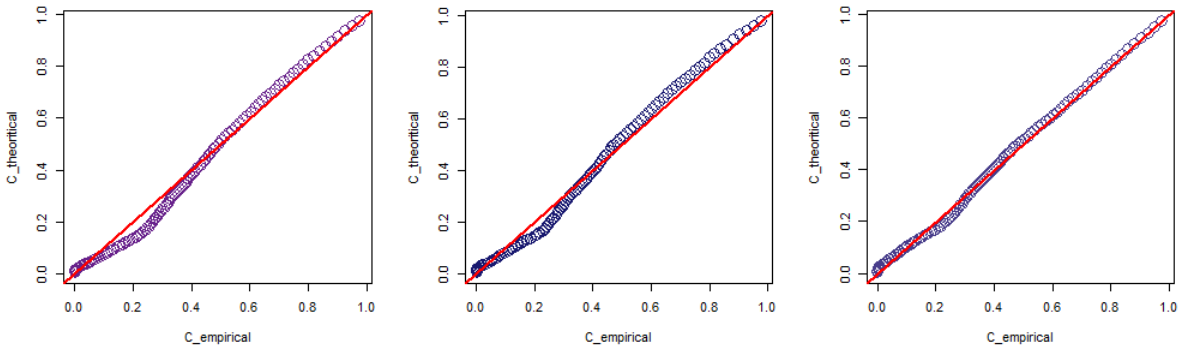


Fig. 4.4 Fitted Gaussian against empirical copulas (left), fitted Student-*t* against empirical copulas (centre) and fitted vine against empirical copulas (right).

4.3.4 Models for predicting at further depths

The trivariate Gaussian, trivariate Student-*t* and vine copulas were used to predict magnetic susceptibility which is the primary path-finder variable for pyrite mineralisation, one step below the depth of drill core. The principle is to fit a trivariate copula to contiguous triples in the linear spatial series using maximum likelihood. The trivariate Gaussian copula fitted to magnetic susceptibility had a correlation of 0.97 at a lag of 1 and 0.93 at a lag of 2. The fitted trivariate Student-*t* copula to the magnetic susceptibility had a correlation of 0.99 at a lag of 1, and 0.97 at a lag of 2, with 3 degrees of freedom. The correlation for the fitted magnetic susceptibility for the first tree (Tree 1) in the C-vine were 0.99 and 0.97, the last tree (Tree 2) had a correlation of 0.82.

The predictions one step below the drill core for all three copulas, after back transforming are given in Table 4.3 below. The vine copula had the lowest predicted value compared with the Student-*t* and Gaussian copulas but the differences are slight by comparison with the prediction interval. The 90% prediction intervals are narrowest for the Gaussian copula and widest for the vine copula. Both the Student-*t* copula and vine copula can allow for tail dependence if it is appropriate, and the vine copula allows for more flexibility in the modelling of the error structure than the Student-*t* copula. The wider prediction intervals are a consequence of the more detailed modelling and the 90% prediction interval from the vine copula should be the most accurate.

Table 4.3 Gaussian, Student-*t* and Vine copulas predictions and 90% predictions intervals

Model	Prediction	90% Prediction Interval
Gaussian Copula	5.76	(4.93, 6.95)
Student-*t* Copula	6.08	(5.32, 7.32)
Vine Copula	5.71	(4.26, 8.99)

4.4 Conclusions

Two applications of copulas have been considered, and for each application three copula types have been compared. The first application was the modelling of trivariate distribution of impedance, P-wave velocity and magnetic susceptibility using copulas. The magnetic susceptibility is highly associated with pyrite mineralization which is the primary economic variable. However, the impedance and P-wave velocity provide further confirmation about the response properties of the rocks, such as the grindability and fragmentation that affect the cost of processing. Estimating the relationship between these variables is important when deciding whether to exploit a prospect. The vine copula is more flexible than the Student-*t* copula or Gaussian copula and provided better fit in terms of closeness to the fitted empirical copula and the correspondence between its upper tails and those of the empirical copula.

The other application was for the prediction of magnetic susceptibility one step beyond the end of the drill core. Such predictions are valuable for establishing points for further exploration. The vine copula provided smaller prediction and far wider 90% prediction interval one step further depth. This finding is a consequence of the very large number of outlying values in the distribution of magnetic susceptibility (kurtosis of 7.69). In general, copulas provide a means for dealing with outliers as they only contribute to the fitting of the copula through their ranks. This is particularly valuable for geological data where outlying values are common and should not be dismissed as erroneous. In addition, copulas provides more accurate model, and also the uncertainty associated with predictions are more reliable.

Acknowledgements

This research is supported by Australian Government Research Training Program Scholarship awarded to Mr. Emmanuel Addo Jr. The authors are thankful to the mining company for providing the down-the-hole diamond drill core datasets used in this case-study. The authors

are also grateful to the reviewers for their comments and suggestions, which have improved the practical application of this manuscript.

Appendix A

See Table 4.4.

Table 4.4 Competing bivariate copulas for C-vine fitting

Copula Models			
Independence	Gaussian	Student-*t*	Clayton
Frank	Gumbel	Joe	BB1
BB7	BB8	BB6	Survival Clayton
Survival BB1	Survival Gumbel	Survival Joe	Survival BB6
Survival BB7	Rotated Clayton (90°)	Survival BB8	Rotated Gumbel (90°)
Rotated BB6 (90°)	Rotated Joe (90°)	Rotated BB1 (90°)	Rotated BB7 (90°)
Rotated Gumbel (270°)	Rotated BB8 (90°)	Rotated Clayton (270°)	Rotated Joe (270°)
Rotated BB7 (270°)	Rotated BB1 (270°)	Rotated BB6 (270°)	Rotated BB8 (270°)

References

[1] E Addo, EK Chanda, and AV Metcalfe. “Spatial Pair-Copula Model of Grade for an Anisotropic Gold Deposit”. In: *Mathematical Geosciences (2018)*, pp. 1–26. doi: 10.1007/s11004-018-9757-7 (cit. on p. C218).

[2] Tim Bedford and Roger M Cooke. “Vines: A new graphical model for dependent random variables”. In: *Annals of Statistics (2002)*, pp. 1031–1068. url: http://www.jstor.org/stable/1558694 (cit. on p. C219).

[3] Claudia Czado. “Pair-copula constructions of multivariate copulas”. In: *Copula theory and its applications. Springer, 2010*, pp. 93–109. doi: 10.1007/978-3-642-12465-5_4 (cit. on p. C219)

[4] Marius Hofert et al. “copula: Multivariate dependence with copulas”. *In: R package version 0.999-9*, URL http://CRAN. R-project. org/package= copula (2014). url: http://copula.r-forge.r-project.org/ (cit. on p. C225).

[5] A. Frigessi H. Bakken K. Aas C. Czado. “Pair-copula constructions of multiple dependence” *In: Mathematics and Economics 44.2 (2009)*. doi: https://doi.org/10.1016/j.insmatheco.2007.02.001 (cit. on pp. C218, C219, C220).

[6] G Nishani Musafer and M Helen Thompson. “Non-linear optimal multivariate spatial design using spatial vine copulas”. *In: Stochastic environmental research and risk assessment 31.2 (2017)*, pp. 551–570. doi: 10.1007/s00477-016-1307-6 (cit. on p. C218).

[7] R. B. Nelsen. “An Introduction to Copulas”. In: *Springer, New York, 2nd Edition (2006) (cit. on p. C218)*.

[8] Richard Peattie and Roussos Dimitrakopoulos. “Forecasting Recoverable Ore Reserves and Their Uncertainty at Morila Gold Deposit, Mali: An Efficient Simulation Approach and Future Grade Control Drilling”. In: *Mathematical Geosciences 45 (8 November 2013*), pp. 1005–1020. doi: 10.1007/s11004-013-9478-x (cit. on p. C218).

[9] Olivier Scaillet, Arthur Charpentier, and Jean-David Fermanian. “The estimation of copulas: Theory and practice”. In: (2007) (cit. on p. C222).

[10] Ulf Schepsmeier et al. VineCopula: Statistical Inference of Vine Copulas. R package version 2.0.5. 2016. url: https://github.com/tnagler/VineCopula (cit. on pp. C223, C224).

Chapter 5

Prediction of copper recovery from geometallurgical data using D-vine copulas (*Paper 4*)

Emmanuel Addo Jr, Andrew V. Metcalfe and Emmanuel K. Chanda

Journal of the Southern African Institute of Mining and Metallurgy, Accepted - December 2018

Statement of Authorship

Title of Paper	**Prediction of copper recovery from geometallurgical data using D-vine copulas**
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Emmanuel ADDO JUNIOR
Contribution to the Paper	Developed methodology, conducted programming and execution of methods. Wrote the manuscripts and acted as the corresponding author.
Overall percentage (%)	80%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/01/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Andrew METCALFE
Contribution to the Paper	Supervised the development of work and assisted in reviewing the manuscript
Signature	Date 14/1/2019

Name of Co-Author	Emmanuel Knox CHANDA
Contribution to the Paper	Supervised the development of work
Signature	Date 10/1/2019

Name of Co-Author	Exequiel Manuel SEPULVEDA ESCOBEDO		
Contribution to the Paper	Wrote and reviewed the data imputation algorithm in the manuscript		
Signature		Date	10/01/2019

Name of Co-Author	Amir ADELI		
Contribution to the Paper	Reviewed and supplied the datasets for the manuscript		
Signature		Date	8 Jan 2019

Name of Co-Author	Winfred ASSIBEY-BONSU		
Contribution to the Paper	Reviewed the manuscript		
Signature		Date	

Emmanuel Addo has emailed Winfred on 09/01/2019 and 10/01/2019. Winfred is not responding to email. He is currently on leave but has not given any indication of when he will be returning. He is holidaying with family in Ghana and we do not have other means for contacting him.

His contribution to the paper was minimal, so the current lack of signature is not a reflection of disagreement.

Please accept the thesis for review and we will provide his signature as soon as we receive it.

15/01/2019

michael Leonard

(Primary Supervisor)

Abstract

The accurate modelling of geometallurgical data can significantly improve decision-making and help optimize mining operations. This case study compares models for predicting copper recovery from three indirect test measurements that are typically available, to avoid the cost of direct measurement of recovery. Geometallurgical data from 930 drill core samples, with an average length of 19 m, from an orebody in South America have been analysed. The data includes copper recovery and the results of three other tests: Bond mill index test; resistance to abrasion and breakage index; and semi-autogenous grinding power index test. A genetic algorithm is used to impute missing data at some locations so as to make use of all 930 samples. The distribution of the variables is modelled with D-vine copula and predictions of copper recovery are compared with those from regressions fitted by ordinary least squares and generalized least squares. The D-vine copula model had the least mean absolute error.

KEYWORDS: copula, geometallurgy, modelling, regression, mining

5.1 Introduction

In this paper we compare the use of D-vine copula, generalized least squares (GLS), and ordinary least squares (OLS) for modelling geometallurgical data from an orebody in South America. The first objective is to construct models for predicting copper recovery (Rec) from the Bond mill index test (BWi); resistance to abrasion and breakage index (A*b); and semi-autogenous grinding (SAG) power index test (Spi). This involves fitting a D-vine copula and regression models fitted by OLS and GLS. The second objective is to investigate the performance of the fitted models for predicting Rec (Willmott and Matsuura 2005).

Traditional resource model approaches either ignore the mineral processing characteristics of extracted tonnages or treat processing as an independent component of a mining operation. The net present value (or any other objective) can be truly optimized only by considering the mining operation as an integrated system in which net value is defined as the end-product that the company sells. This approach requires the resource model to be extended to include all relevant rock properties and processing responses.

Comminution performances and mineral processing recovery factors have a substantial effect on production and the final value of the product. Hence their prediction in the early stages of a mining operation is crucial. The accurate and precise prediction of these variables is important for mine planning and project risk assessment. Commonly used tests for determining comminution performances are BWi, Spi, and A*b. Better understanding of the physical and chemical principles on which these performance indices are based has contributed to the acceptance and use of geometallurgy in resource modelling, referred to in a wider context as grade engineering.

Lishchuk et al. (2015) define geometallurgy as a multidisciplinary approach that integrates geology, mineralogy, mineral processing, and metallurgy to create spatially based models for production and operational decisions. The primary geological rock properties (e.g., grade,

alteration, texture, and grain size) are proxies for predicting metallurgical responses (e.g., type of processing, throughput, recovery, energy consumption, reagent usage, and grindability) (Coward et al. 2009; Dowd, Xu, and Coward, 2016). Incorporating these variables into the resource model in way that can be used effectively in mine planning poses a challenge to geostatisticians and resource modellers. In most projects, the lack of appropriate geometallurgical data collection and analysis leads to unreliable metallurgical response models.

The relatively large difference between the number of samples recorded in the geological database (logging, assays etc.) and the relatively few metallurgical test work samples further hinders the integration of metallurgical responses into the resource model using existing geostatistical methods (Hunt, Kojovic, and Berry, 2013). Also, there is often the problem of missing values of metallurgical variables, which may not be measured at all locations. Retaining only data where all variables are sampled could result in removing a large amount of data from the geometallurgical programme (Deutsch, 2013), which can lead to poor geostatistical modelling in areas where more data (of some variables) is actually sampled.

In addition, most geological and geometallurgical variables have complex multivariate relationships that are the result of a succession of several chaotic nonlinear natural processes which are often not well modelled by parametric multivariate probability distribution (Deutsch 2013). Moreover the non-additive and compositional nature of geological/geometallurgical variables makes their modelling more difficult (Walters and Kojovic 2006; Williams and Richardson, 2004). An alternative modelling strategy that can capture all these complex multivariate relationships is crucial for successful modelling of geometallurgical variables. Multivariate D-vine copulas are ideal for modelling complex multivariate relationships, skewed distributions, and tail-dependent distributions. Moreover, the D-vine copula models encompass all multivariate distribution, including the multivariate Gaussian distribution

(MVG).

This paper is comprised of three main sections. The 'Method' section describes the theory of copulas, pair copulas, and vine copulas (D-vine) construction models. The 'Application' section describes data imputation, modelling of copper recovery in terms of A*b, BWi, and Spi, and finally the prediction of copper recovery from A*b, BWi and Spi. The final section is Discussion and Conclusion.

5.2 Methods

This section gives an overview and summarizes the principles of copulas, pair copulas (D-vine) construction for four variables. Further details about the concept of copulas can be found in Joe (1996) and Nelsen (2006). In addition, more detailed explanation of the pair copula and vine copula models can be found in Aas *et al*. (2009), Bedford and Cooke (2002), and Kurowicka and Cooke (2006). Spatial applications of pair copulas can be found in Gräler and Pebesma (2011), Gräler (2014), Musafer *et al*. (2013), Musafer and Thompson (2016), and Addo, Chanda, and Metcalfe (2018).

5.2.1 Theory of Copulas

A copula is a multivariate uniform distribution. It follows that any multivariate distribution has a copula form because the marginal cumulative distribution functions (cdfs) can be transformed to uniform distributions. Conversely, the uniform margins of any copula can be transformed to any continuous probability distributions, which can differ for different margins. Therefore copulas provide a very flexible approach in modelling multivariate data. Consider a random variable $Z = (z_1, \dots, z_d)$ and define $u_i = F(z_i)$. We can define a copula by its cdf $C(u_1, u_2, \dots, u_d)$ and the corresponding probability density function (pdf) is

$$c(u_1, u_2, \dots, u_d) = \frac{\partial C(u_1, u_2, \dots, u_d)}{\partial u_1 \, \partial u_2 \dots \partial u_d}, \quad [1]$$

The copula pdf links the marginal pdfs to the multivariate pdf:

$$f(z_1, \dots, z_d) = c(u_1, \dots, u_d) f(z_1) \dots . f(z_d) \quad [2]$$

Generally, we often require multivariate distributions of more than two variables. The elliptical copulas (*i.e.*, Gaussian and Student-*t* copula) can easily be extended to more than two variables, but this is not generally the case for the Archimedean copulas (*i.e.*, Clayton, Frank, and Gumbel copula). A more flexible approach to modelling such multivariate distributions is the pair-copula D-vines as described by Aas *et al*. (2009), Bedford and Cooke (2002), and Kurowicka and Cooke (2006).

5.2.2 Pair Copula

Any multivariate distribution can be factorized in different ways using its conditional distributions. Specifically, a copula can be factorized as a product of the marginal distributions and the bivariate conditional copulas. We often term such factorization as pair-copula models. Joe (1996) presented a construction for a pair-copula model for a multivariate copula based on the distribution functions. After Joe's construction of a copula based on the distribution functions, Bedford and Cooke (2002) also presented a construction in terms of the densities. In their work, they organized the constructions in a graphical way involving a sequence of nested trees, which they refer to as 'regular vines'. They defined two popular classes of pair-copula construction (PCC) models, which they refer to as the D-vines and canonical (C) vines. Their work was further developed by Kurowicka and Cooke (2006). The derivation of a D-vine model, which is used in this application, is outlined below.

5.2.2.1 D-vines

Generally the pair copula can be seen as a multivariate copula that aims to approximate the

target copula, since not all copulas can be expressed as a vine copula (Haff, Aas, and Frigessi, 2010). This decomposition is, however, not unique; for example, a five-dimensional density can have about 240 different constructions. In the D-vines, the decomposition of the joint density consists of the pair-copula densities evaluated at conditional distributions functions and for specified indices and marginal densities (Bedford and Cooke 2002; Czado 2010 and Gräler 2014). Figure 5.1, which is reproduced from Aas *et al*. (2009), shows the graphical model used to demonstrate the D-vines for four variables. This consists of three trees: T_j $j = 1,2,3$. Tree T_j has $n + 1 - j$ nodes, where n is the number of variables. Using the decomposition shown in Figure 5.1 and Equation [3], the joint density function of four random variables can be expressed using the D-vines as

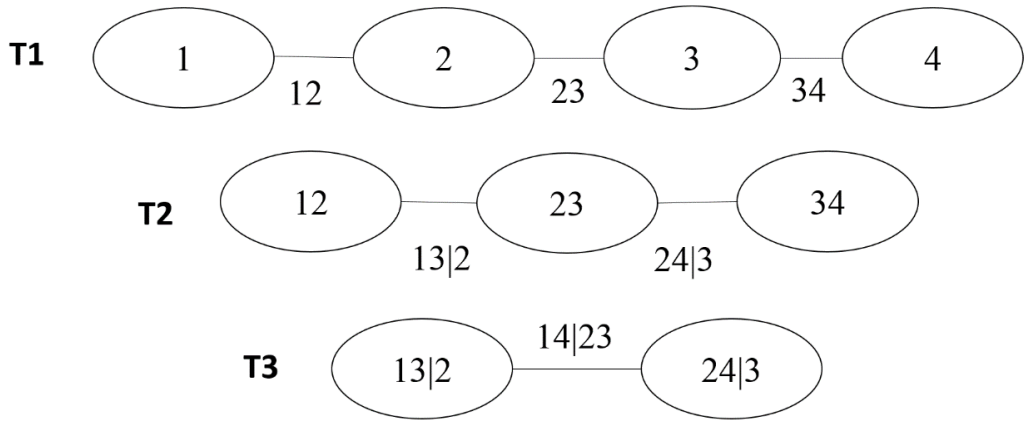


Fig. 5.1 D-vines for four variables.

$$f_{1234}(z_1, z_2, z_3, z_4) = f_1(z_1).f_2(z_2).f_3(z_3).f_4(z_4)$$
$$.c_{12}\big(F_1(z_1), F_2(z_2)\big).c_{23}\big(F_2(z_2), F_3(z_3)\big).c_{34}\big(F_3(z_3), F_4(z_4)\big)$$
$$.c_{13|2}\left(F_{1|2}(z_1|z_2), F_{3|2}(z_3|z_2)\right).c_{24|3}\left(F_{2|3}(z_2|z_3), F_{4|3}(z_4|z_3)\right)$$
$$.c_{14|23}\left(F_{1|23}(z_1|z_2, z_3), F_{4|23}(z_4|z_2, z_3)\right) \quad [3]$$

As shown in Equation [3], the D-vine distribution requires the computation of several conditional distribution functions and conditional bivariate copulas. From Joe (1996) and Aas, Frigessi, and Bakken (2009), the conditional distribution functions $F(z|v)$ for an m-dimensional vector $v = (v_1, \dots, v_m)$ can be obtained from the following recursive relationship

$$h(z|v) \coloneqq F(z|v) = \frac{\partial C_{zv_j|v-j}\left(F(z|v_{-j}), F(v_j|v_{-j})\right)}{\partial F(v_j|v_{-j})}, \quad [4]$$

where $v_j (j = 1, \dots, m)$ is an arbitrary component of v, and $v_{-j} = (v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_m)$ denotes the vector v excluding element v_j. The bivariate copula function is also specified by $C_{zv_j|v_j}$. Given $u_i (i = 1, \dots, n)$ to denote $F_i(z_i)$, we can derive the conditional distribution function $F(u_3|u_1, u_2)$ that is needed as an argument for $C_{14|23}$ in a four-dimensional D-vine copula density using Equation [4]. From Figure 5.1 $Tree$ 3 ($T3$) the argument $C_{14|23}$, namely $F_{1|23}(x_1|x_2, x_3)$, can be evaluated using the h function (Kraus and Czado 2016) associated with $C_{13,2}, C_{12}, and\ C_{23}$ from the first two trees $T1\ and\ T2$ as

$$\begin{aligned} F_{1|23}(z_1|z_2, z_3) &= h_{1|3;2}\left(F_{1|2}(z_1|z_2)F_{3|2}(x_3|x_2)\right) \\ &= h_{1|3;2}\left(h_{1|2}\left(F_1(z_1)|F_2(z_2)\right)\middle|h_{3|2}\left(F_3(z_3)|F_2(z_2)\right)\right). \end{aligned} \quad [5]$$

5.2.3 D-vine Copula-Based Conditional Forecasting Model

With the defined conditional distributions function in Equation [5], the inverse forms can also be defined, and can be used in forecasting. Using the bivariate case, the conditional distribution function of two random variables $z_1\ and\ z_2$ is $h(u_2|u_1)$. The main goal is to be able to obtain u_2 based on the information available at u_1. Given some fixed probabilities τ, we can derive u_2 from $C_{u_2|u_1}$using an explicit function $u_2 = u_2 = C^{-1}_{u_2|u_1}(\tau; u_1) = h^{-1}(\tau|u_1)$, where

$C_{u_2|u_1}^{-1}$ is the inverse of the copula function known as the τ quantile curve of the copula (Xu and Childs, 2013). The τth copula-based conditional quantile function of variable z_2 is

$$Q_z(\tau|z_1) = F^{-1}(u_2) = F^{-1}\left(C_{u_2|u_1}^{-1}(\tau; u_1)\right) = F^{-1}\left(h^{-1}(\tau|u_1)\right) \quad [6]$$

where F^{-1} is the inverse of u_2. For the four-dimensional case, the τth conditional quantile function of z_4, $Q_{x_4}(\tau|z_1, z_2, z_3)$, can be deduced by the recursive computation

$$Q_{z_4}(\tau|z_1, z_2, z_3) = F^{-1}(u_4)$$
$$= F^{-1}\left(h^{-1}\left\{h^{-1}\left[h^{-1}\left(\tau|h(h(u_3|u_1)|h(u_2|u_1))\right)|h(u_2|u_1)\right]u_1\right\}\right). \quad [7]$$

Hence we can forecast z_4 based on the variables z_1, z_2, and z_3. Moreover,$Q_{z_4}(\tau|z_1, z_2, z_3)$ is monotonically increasing in τ, so the crossing of quantile functions corresponding to different quantile levels is not possible. Bernard and Czado (2015) proved that linear regression quantile functions may cross if a non-Gaussian data is modelled.

In general, the multivariate D-vine copula model for the four-dimensional vine model can be implemented by following the steps below. Further details of the method can be found in Kraus and Czado (2016) and Liu *et al*. (2015)

1. Fit an appropriate marginal distribution to each of the variables, z_1, z_2, z_3, and z_4, where z_4 is the predicted variable and all the others are the explanatory variables.
2. Model the joint dependence structure of all the four variables using Equation [5] for the D-vine model.
3. Estimate all the appropriate bivariate copula for each pair copula using the R library $VineCopula$ (Schepsmeier *et al*., 2015).
4. Estimate the conditional distribution function of variable z_4 conditioned on the given

variables z_1, z_2, z_3 using Equation [5].

5. Finally, generate the predicted values of z_4 based on the given variable, z_1, z_2, z_3 using the copula-based quantile function as given in Equation [7].

5.2.4 Performance of Models

The mean absolute error (MAE) and root mean square error (RMSE) have been used to assess the prediction performance of the models.

$$MAE(A_i, \hat{A}_i) = \frac{1}{N}\sum_{i=1}^{n}|A_i - \hat{A}_i| \quad [8]$$

$$RMSE(A_i, \hat{A}_i) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(A_i - \hat{A}_i)^2} \quad [9]$$

where A_i is the observed recovery, and $\hat{A}_i$ is the predicted recovery made using a fitted model to all $N = 930$. So, the performance measures are calculated within the entire sample.

5.3 Application

Nine hundred and thirty (with some missing values) drill core samples with an average length of 19 m from a mine in South America were sampled for geometallurgical attributes of copper recovery (Rec), Bond mill index test (BWi), resistance to abrasion and breakage index (A*b), and semi-autogenous grinding power index test (Spi). Typical of most geometallurgical datasets, there are missing values that have not been sampled at some locations. There are 299 non-missing datasets that are sampled at all locations. To be able to use all 930 georeferenced drill core sample for the analysis, we employed a data imputation algorithm to predict missing values at some locations.

5.3.1 Data Imputation

The data-set has 930 georeferenced samples with four attributes of interest: Rec, BWi, Spi, and A*b. Table 5.1 shows the summary of descriptive statistics for all four attributes.

Table 5.1 Summary statistics of all four attributes.

Variable	Number of non-missing values	Number of missing values	Minimum	Maximum
Rec	560	370	36.20	99.30
BWi	840	90	9.12	15.58
Spi	539	391	10.69	98.60
A*b	300	630	32.38	175.66

Data imputation was formulated as an optimization problem seeking to preserve two main properties: the reproduction of the individual histograms and the bivariate correlation among the variables. Histograms of each of the variables were calculated using non-missing (informed) values, and correlations were calculated using samples where all variables have non-missing values. Table 5.2 indicates the bivariate correlations, and number of missing and non-missing values for the pair attributes. The diagonal shows the number of missing values, the upper triangle shows the Pearson correlation between two attributes, and the lower triangle shows the number of non-missing values for the pair attributes.

Table 5.2 Description of all four attributes and simple statistics.

	Rec	Bwi	Spi	A*b
Rec	**370**	0.08	0.11	-0.02
BWi	*470*	**90**	0.31	-0.28
Spi	*469*	*539*	**391**	-0.74
A*b	*300*	*300*	*299*	**630**

The datasets were decomposed in two sets: non-missing (informed) values and missing values. Hence, the multivariate datasets was define as: $X = \{X_1, \dots, X_D\}$, where D is the number of attributes. Each X_i is also define as follows: $X_i = V_i \cup M_i$, where V_i and M_i represent the informed and missing values respectively, and $V_i \cup M_i$ is used to indicate V_i if available, or

M_i otherwise.

The histogram function used for the imputation was denoted by $H(X)$, and used 21 regular bins or class intervals. The correlation is given by

$$CORR(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})/(n-1)}{S_x S_y} \quad [10]$$

where $\bar{x}, \bar{y}, S_x$, and S_y are the mean and standard deviation of x and y, and n is the sample size. The optimization problem for the data imputation was formulated as follows.

5.3.1.1 Decision Variables

As the main objective is to impute values to all missing values, the decision variables correspond to the set $\{M_1, \dots, M_D\}$. According to Table 5.1, there are 1481 missing values to impute.

5.3.1.2 Objective Function

Minimization of the mean quadratic error between the histogram of each variable with and without imputed data and the mean quadratic error between the correlations of each pair of variables with and without the imputed data

$$\boldsymbol{argmin} \sum_{d=1}^{D} \|H(V_d) - H(V_d \cup M_d)\| + \sum_{\substack{u=1,v=1 \\ u>v}}^{D} \|CORR(V_u,V_v) - CORR(V_u \cup M_u, V_v \cup M_v)\|. \quad [11]$$

5.3.1.3 Constraints

The only imposed constraint is the lower and upper bounds for each attribute, for which the minimum and maximum values are observed from the samples

$$min(V_d) \leq W_i^d \leq max(V_d), \forall i \in \{1, \dots, N^d\}, d \in \{1, \dots, D\}. \quad [12]$$

This formulation is nonlinear and may have no unique solution. Metaheuristics are optimization methods that can deal with these kinds of problems successfully. We therefore solved the optimization formulation by genetic algorithm (GA) metaheuristics due to its flexibility and good performance (Whitley, 1994). The GA is a stochastic method, hence different seeds of random number generator may generate different solutions. In this application, our experiments show that the imputed values change slightly in response to varying the seed, but the histograms and correlations are very stable. We use one representative set of imputed data found by one execution of GA. Table 5.3 shows the parameters used in the GA for data imputation.

Table 5.3 Parameters used by GA for data imputation.

Parameter	Value	Description
npop	1000	Number of individuals in the population
ngen	500	Number of generations (iterations)
Crossover operator	Uniform crossover	50% probability of getting the gene from parent 1 (and 50% from parent 2)
Mutation operator	Gaussian mutation	10% of genes at random, new value = current value + N(0,1)
Selection operator	Tournament selection	Tournament size of 10 individuals
cxpb	0.9	Probability of applying crossover
mutpb	0.4	Probability of applying mutation

The Pearson's correlation computed for imputed and non-missing samples shows that the correlations were perfectly reproduced. Figure 5.2 (upper and lower panel) shows the

histogram of the imputed data and with non-missing values respectively. Figure 5.3 also shows all four non-missing variables (*i.e.*, Rec, BWi, Spi, and A*b) in space. We discuss Figure 5.3 under the Discussion and Conclusion section.

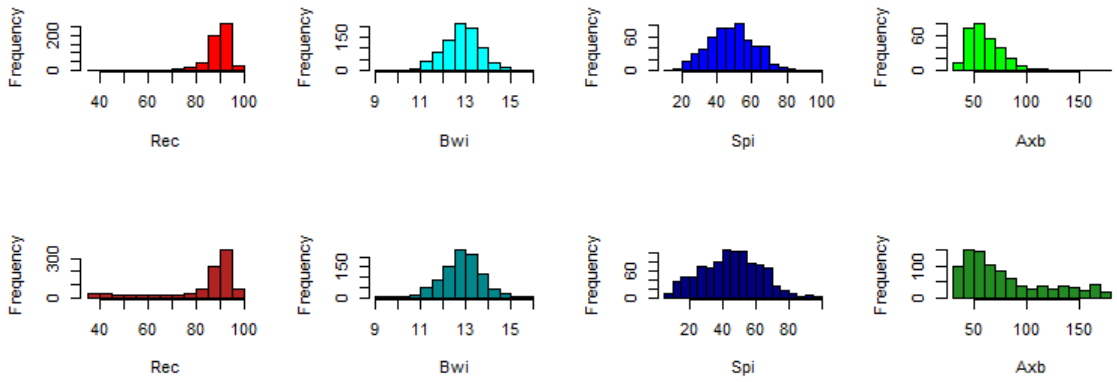


Fig. 5.2 Histogram of four variables with non-missing data (upper panel), histogram of four variables with imputed data (lower panel).

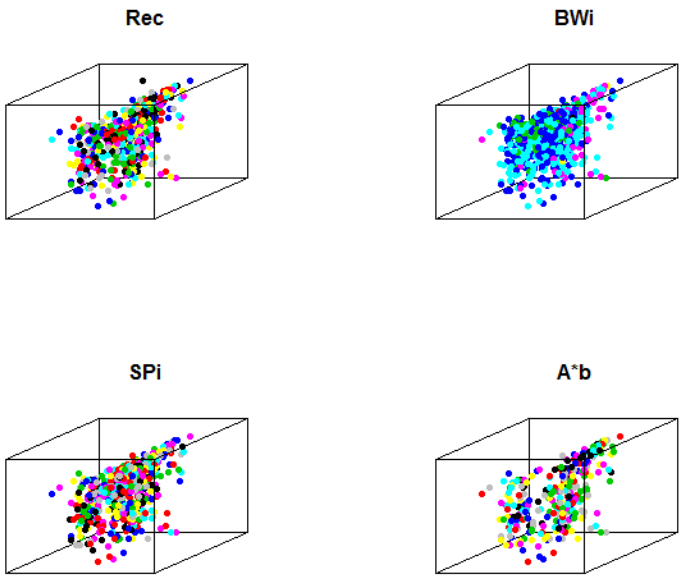


Fig. 5.3 The 3D spatial position of the samples showing non-missing values of Rec (top left), BWi (top right), SPi (bottom left), and A*b (bottom right)

5.3.2 Analysis

The imputed data, consisting of 930 drill core samples for four variables (Rec, BWi, Spi, and A*b), was used for the analysis, the explanatory variables being Spi, BWi, and A*b. The hypothesis of stationarity was tested by fitting a regression of recovery on the mean corrected eastings (x), mean corrected northings (y), mean corrected elevations (z) , x^2, y^2, and the cross-product xy in Equation [13]. This model, referred to as $Model1$, is

$$A = \beta_o + \beta_1 x + \beta_2 y + \beta_3 z + \beta_4 x^2 + \beta_5 y^2 + \beta_6 xy + \varepsilon \quad [13]$$

where ε is the random error, which is expected to be spatially correlated, with mean of 0 and standard deviation σ_o. $Model1$ was initially fitted by OLS. Then, a spherical variogram was fitted to the residuals and the variogram parameters were used for fitting with a GLS function $gls(\,)$ in the R library $nlme$ (Pinheiro and DebRoy, 2016). The fitted spherical variogram and model parameters are shown in Figure 5.4. The estimated coefficients for the GLS fit to $Model1$ are shown in Table 5.4.

Table 5.4 Estimated coefficients of the fitted GLS model (range, 230 and nugget, 0.5).

Coefficient	Estimate	Estimated standard error	P-value
β_o	133.05554	25.54233	0.00
β_1	0.00247	0.00413	0.55
β_2	-0.00062	0.00158	0.69
β_3	-0..02129	0.00578	0.00
β_4	0.00002	0.9e-5	0.03
β_5	0.3e-5	0.1e-5	0.02
β_6	0.6e-5	0.4e-5	0.19

The standard deviation of the residuals is 14.26 on 923 degrees of freedom, which is smaller than the standard deviation of the recovery data (15.02). While this reduction in standard deviation is small, the sample size is relatively large and two of the coefficients in the fitted quadratic surface are highly significant statistically. We then assume the residuals from the

GLS regression are a realization of a stationary spatial process.

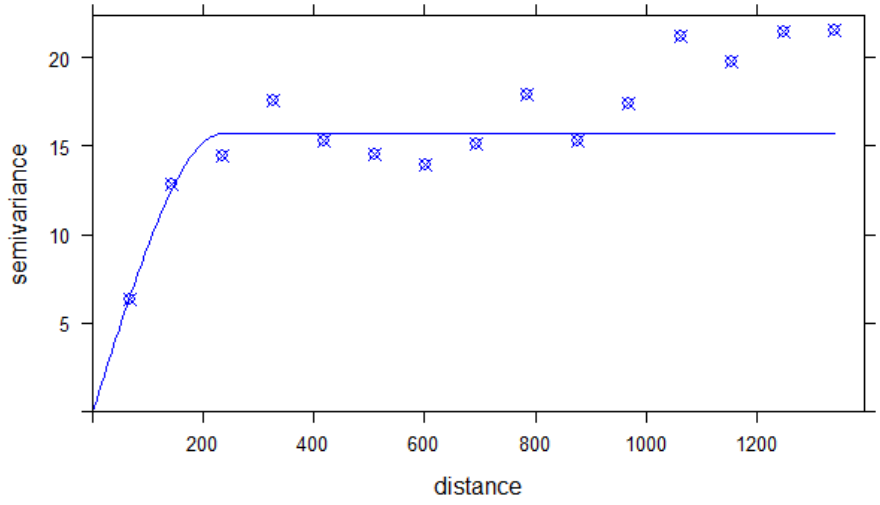


Fig. 5.4 Fitted spherical variogram with range 230, sill 15.7, and nugget 0.3

The residuals from the model *res* (Rec) were taken as the response variable for recovery (Rec) and were used together with the BWi, Spi, and A*b here (referred to as B, C, and D respectively) to fit the D-vine copula. Variables B, C, and D were mean-corrected, to avoid the excessively large values of quadratic terms and ill-conditioned matrices that would result if the original data was used. Quadratic terms and interactions between the explanatory variables included were BD, CD, B^2, C^2, D^2. So a 10-dimensional D-vine based model was constructed, which is made up of$res(Rec), B, C, D, BC, BD, CD, B^2, C^2, and\ D^2$. In the following, the $res(Rec)$ will be referred to as A.

We fitted an appropriate kernel marginal distribution to each of the mean-corrected variables. After obtaining well-fitted marginal distributions, a 10-dimensional D-vine copula was used to join the margins and model the joint dependence structure. To be able to establish the 10-dimensional D-vine copula, we fitted the best fitting bivariate copula for each pair copula using the R library $CDVine$ developed by Schepsmeier and Brechman (2015). The fitting was done

by equating the Kendall's tau to the value of Kendall's tau implied by the dependence parameter ($\theta_1, \theta_2, and\ \rho$), which is referred to as the 'method of moments'. A limitation of the method of moments is that it does not lead easily to a criterion for choosing between the copula forms. Therefore, the method of maximum likelihood was used for choosing between the copula forms, the form with the highest likelihood being chosen. The rotated version of the bivariate copulas (*i.e.*, BB6, BB7, and BB8) with angles 90°, 180°, and 270° was selected by maximum likelihood. In addition, Student-*t*, normal, and Frank copulas were also selected using maximum likelihood for some trees. Figure 5.5 illustrates the fitted bivariate copulas and their fitted dependence parameters ($\theta_1, \theta_2, \rho, and\ df$) for the 10-dimensional D-vine model. The final forecasting performance of the 10-dimensional D-vine copula model was calculated using a 10-D version of Equation [7]. The predicted values was back-transformed to the original unit (recovery per cent, *Rec*) by adding the predicted values from the 10D model to the predicted values from $Model1$.

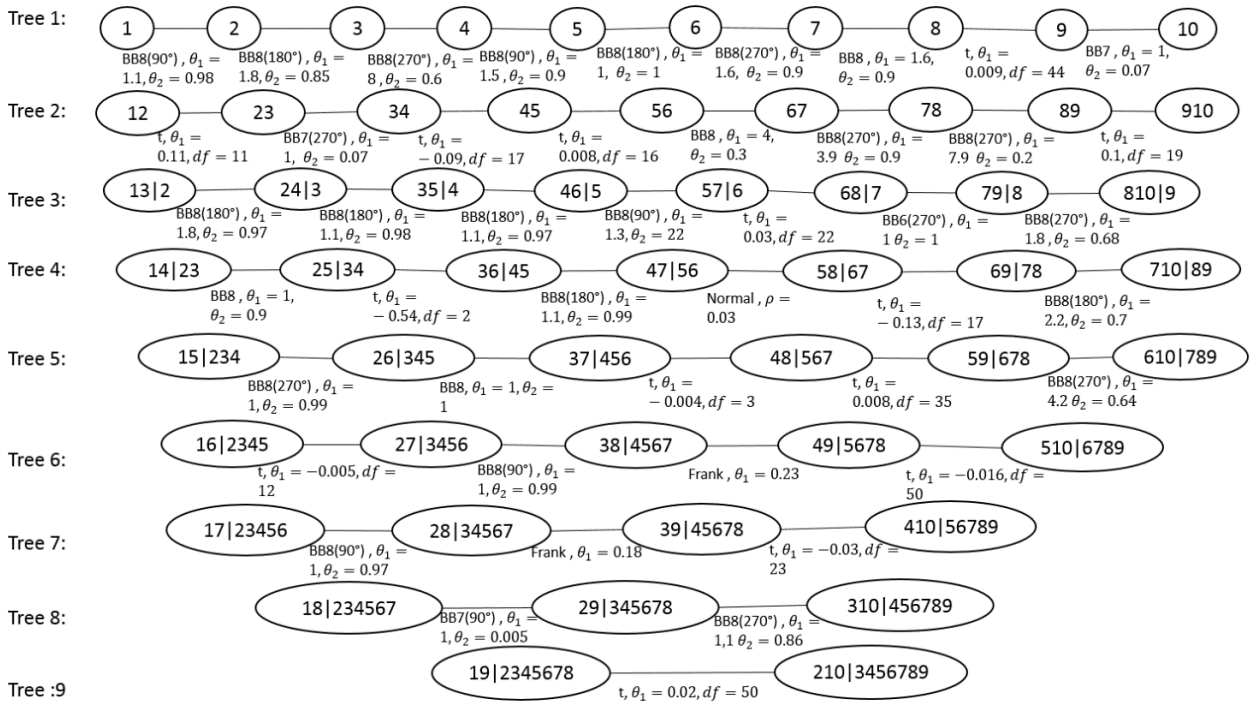


Fig. 5.5 Structure of the 10-dimension D-vine model, where 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 are $A, B, C, D, BC, BD, CD, B^2,\ C^2,\ D^2$respectively.

We compared the predicted recovery from the 10-dimensional D-vine copula with an OLS regression and a GLS regression model. The residual from model referred to in this application as A, which is the response variable, was regressed on the explanatory variables $B, C, D, BC, BD, CD, B^2, C^2, and\ D^2$. Equation [14] shows the OLS regression model fitted, and Table 5.5 shows the estimated coefficients for the OLS regression.

$$A^* = \beta_o + \beta_1 B + \beta_2 C + \beta_3 D + \beta_4 BC + \beta_5 BD + \beta_6 CD + \beta_7 B^2 + \beta_8 C^2 + \beta_9 D^2 + \varepsilon \quad [14]$$

Table 5.5 Estimated coefficients of the fitted OLS regression.

Coefficient	Estimate	Estimated standard error	P-value
β_o	1.11996	0.81077	0.17
β_1	-0.22724	0.58045	0.69
β_2	0.22607	0.04255	1.35e-07
β_3	0.10797	0.02407	8.19e-06
β_4	-0.02021	0.04535	0.66
β_5	-0.03634	0.02158	0.09
β_6	0.00116	0.00162	0.47
β_7	0.63571	0.43801	0.15
β_8	-0.00179	0.00233	0.44
β_9	-0.00089	0.00049	0.07

The linear regression model was used to predict recovery and the predicted values were back-transformed to the original units (recovery per cent, *Rec*) by adding the predicted values from $Model1$.

The residual from the model referred as A was regressed on the explanatory variables $B, C, D, BC, BD, CD, B^2, C^2$, and D^2 using the GLS model with spherical variogram parameters of range: 230, nugget: 0.5, and sill: 15.7. The GLS model is given in Equation [14], and the estimated model coefficients are shown in Table 5.6.

Table 5.6 Estimated coefficients of the fitted GLS model (range, 230 and nugget, 0.3).

Coefficient	Estimate	Estimated standard error	P-value
β_o	0.78122	0.86786	0.37
β_1	-0.22520	0.58289	0.69
β_2	0.22476	0.00430	0.00
β_3	0.10409	0.02456	0.00
β_4	-0.01938	0.04549	0.67
β_5	-0.03494	0.02161	0.11
β_6	0.00122	0.00163	0.46
β_7	0.65855	0.43858	0.13
β_8	-0.00157	0.00236	0.51
β_9	-0.00078	0.00050	0.12

The GLS model was used to make predictions of recovery at each point, and the predicted value was back-transformed to the original unit (recovery per cent *Rec*) by adding the predicted values from $Model1$. A summary of the cross-validation results using all three models is presented in Table 5.7. There is little difference between the GLS and OLS fits. The D-vine performs better according to the MAE but not according to RMSE. We discuss this in the next section. Figure 5.6 illustrates a scatter plot of the observed versus predicted recoveries from the GLS, D-vine, and OLS models. Further comparisons were made by making out-of-sample predictions. Proportions of 10% and 30% were removed at random locations from the 930 complete geometallurgical data. The models (OLS, GLS and D-vine) were fitted to the remaining 90% and 70% of the data.

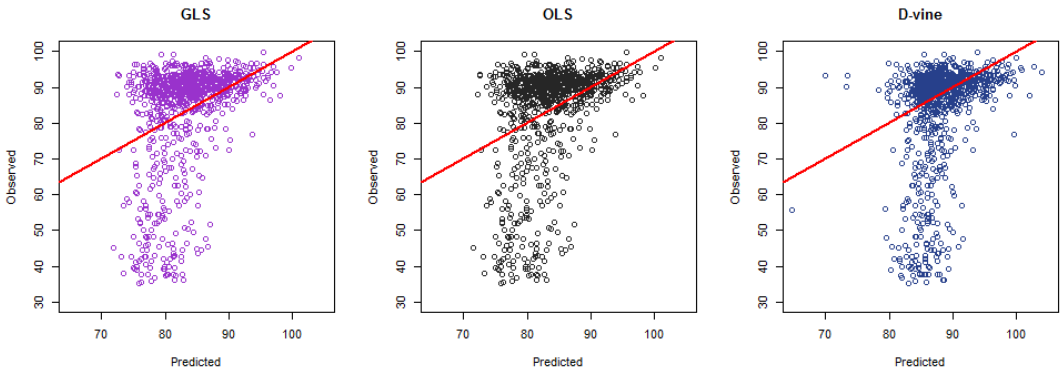


Fig. 5.6 Predicted vs observed recoveries for GLS, OLS, and D-vine models.

Table 5.7 Summary of the cross-validation.

Model	MAE	RMSE
D-vine copula regression	9.12	14.92
GLS regression	10.09	13.91
OLS regression	10.04	13.91

Out-of-sample predictions were generated and compared with the known values of the data removed. Figure 5.7(left) and Figure 5.7(right) shows a box plot of the removed CuRec (A), and the out-of-sample predictions using OLS (B), GLS (C) and D-vine (D), for 90% and 70% data used for predictions respectively. Summary statistics are given in Table 5.7.

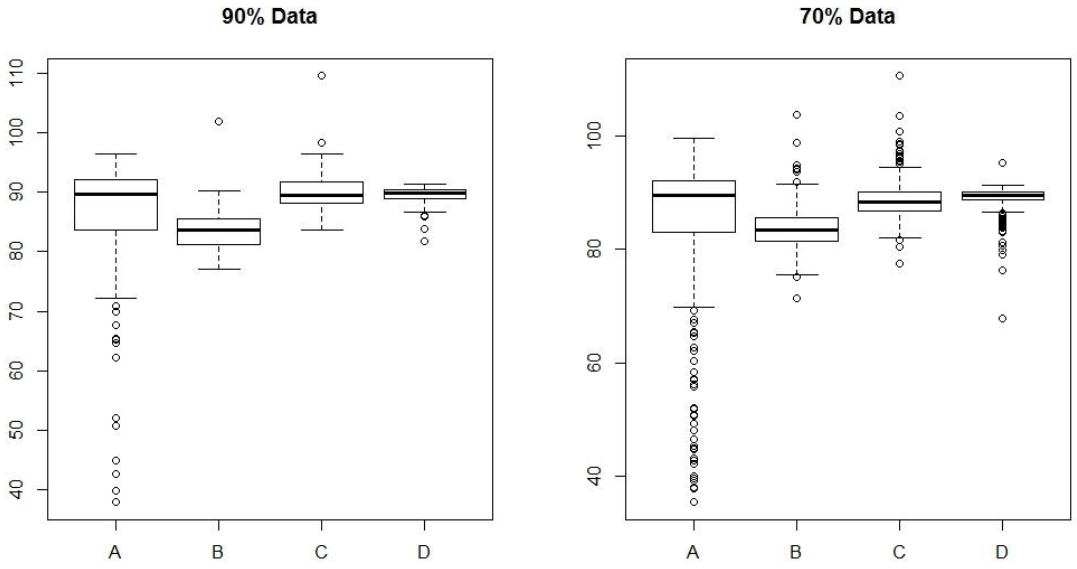


Fig. 5.7 Box plot of removed CuRec (A), out-of-sample predictions of OLS (B), GLS (C) and D-vine (D) models for 90% (left) and 70% (right) of data.

The D-vine performs best in terms of MAE, and GLS regression is an improvement on OLS regression, in terms of MAE for both 10% and 30% data removed. However, OLS regression is slightly better than both the D-vine copula and GLS regressions in terms of RMSE.

Table 5.8 Summary of the out-of- sample predictions

Model	MAE (90%)	RMSE (90%)	MAE (70%)	RMSE (70%)
D-vine copula regression	7.86	13.96	9.33	16.25
GLS regression	8.46	14.29	10.07	16.47
OLS regression	9.49	13.11	11.07	15.20

5.4 Discussion and Conclusion

Data on four variables (Rec, BWi, Spi, and A*b) from 930 drill core samples at know locations was available. Two hundred and ninety-nine drill cores had a complete record, but there were some missing items from the other 631 cores. In order to use all 930 drill core samples, a genetic algorithm (GA) was used to impute missing items at these locations. The objective function formulated in this case study was designed to reproduce precisely the individual histograms and the linear correlations between pairs of variables. This criterion is, however, subjective and in cases when missing data comes from preferential sampling, the histogram of the imputed data may differ from the actual underlying distribution. For example, it is common to perform metallurgical test work only in ore zones with a high grade profile, hence recovery for low-grade zones will not be well represented in the distribution. The objective function in the data imputation method should be adjusted according to the knowledge of the datasets. The recovery (Rec), the response in this application, shows a slight but statistically significant, non-stationarity. The non-stationarity of Rec has been accounted for by fitting a quadratic trend regression surface by the GLS model, with a spherical variogram to approximate the spatial correlation. The residuals from the GLS model were considered as a realization of a stationary process.

The residuals from the model and the mean corrected values for BWi, Spi, and A*b, together with two variable interactions and squares for the three variables, were used to fit a 10-dimensional D-vine copula model. The fitted 10-dimensional D-vine copula model was used to predict recovery, and MAE and RMSE were calculated. These predictions were compared

with predictions from regressions fitted by OLS and GLS. The D-vine copula model had the smallest MAE. However, the regression models had lower RMSE. A comparison of the scatter plots suggests that the D-vine gives more accurate and precise predictions for high levels of copper recovery. However, the D-vine appears to overestimate the copper recovery at low levels rather more than the regression models. Out-of-sample predictions using the three models were compared as a further check on the D-vine copula regression model. A proportion of the data (i.e., 10% and 30%) were removed at random locations from the complete geometallurgical datasets. The models were fitted to the remaining 90% and 70% of the data, and out-of-sample predictions were estimated and compared with the known values of 10% and 30% data removed. Results from the analysis shows that the D-vine model had the least MAE for both 90% and 70% data, although OLS regression was slightly better on RMSE. An explanation for the finding that the D-vine copula is better on MAE yet slightly worse on RMSE is that the D-vine copula is less affected by outliers. The outliers will make a major contribution to the RMSE, and regression fits the coefficients by minimizing the RMSE. This has the effect that outlying observations are highly influential in the fitting process, drawing the fitted surface towards them and so reducing the RMSE. For this reason the MAE is considered more useful in the mining industry, where outlying values are common and the implicit assumption of a Gaussian distribution, under which GLS would be optimum, is not realistic. The D-vine copula is preferable to capping, which introduces a downward bias. Moreover, the D-vine will generally produce more accurate prediction intervals than a regression model, because it allows for a general form of the distribution of the errors. Generally, geometallurgical tests are expensive and a modelling approach that can provide accurate and precise predictions of some variables from others will save money.

Acknowledgements

This research is supported by Australian Government Research Training Program Scholarship awarded to Mr. Emmanuel Addo Jr. The authors will like to thank the mining company for providing the geometallurgical datasets used in this case-study. The authors will like express their gratitude to the reviewers for their comments and suggestions, which have improved the practical application of this manuscript.

References

Aas, K., Czado, C., Frigessi., A., and Bakken., H. (2009). Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics, 44 (2), 182-198.

Addo, E, Chanda, E & Metcalfe, A 'Spatial Pair-Copula Model of Grade for an Anisotropic Gold Deposit', Mathematical Geosciences, pp. 1-26.

Bedford, T. and Cooke, R.M. (2002). Vines: A new graphical model for dependent random variables. Annals of Statistics, 30 (4), 1031-1068.

Bernard, C. and Czado, C. (2015). Conditional quantiles and tail dependence. Journal of Multivariate Analysis, 138, 104-126.

Coward, S., Vann, J., Dunham, S., and Stewart, M. (2009). 'The Primary-Response framework for geometallurgical variables. Proceedings of the 7th international Mining Geology Conference, Perth, WA, 17-19 August 2009. Australasian Institute of Mining and Metallurgy, Melbourne. pp. 109-113.

Deutsch, C.V. (2013). Geostatistical modelling of geometallurgical variables - problems and solutions. Proceedings of the Second AusIMM International Geometallurgy Conference, Brisbane, QLD, 30 September - 2 October 2013. Australasian Institute of Mining and Metallurgy, Melbourne. pp. 7-16.

Dowd, P. Xu, C., and Coward, S. (2016). Strategic mine planning and design: some challenges and strategies for addressing them. Mining Technology, 125 (1), 22-34.

Erhardt, T.M., Czado, C., and Schepsmeier, U. (2015). R-vine models for spatial time series with an application to daily mean temperature. Biometrics, 71 (2), 323-332.

Fenske, N., Kneib, T., and Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Journal of the American Statistical Association, 106 (494), 494-510.

Gräler, B. (2014). Developing spatio-temporal copulas. PhD dissertation, Institute for Geoinformatics, University of Münster.

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.

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.

Hunt, J., Kojovic, T., and Berry, R. (2013). Estimating comminution indices from ore mineralogy, chemistry and drill core logging. Proceedings of the Second AUSIMM

International Geometallurgy Conference, Brisbane, QLD, 30 September - 2 October 2013. Australasian Institute of Mining and Metallurgy, Melbourne. pp. 173-176.

Joe, H. (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. Lecture Notes - Monograph Series. Institute of Mathematical Statistics, Hayward, CA. pp. 120-141. https://projecteuclid.org/euclid.lnms/1215452614

Kraus, D. and Czado, C. (2016). D-vine copula based quantile regression. Computational Statistics and Data Analysis. https://arxiv.org/abs/1510.04161v4

Kurowicka, D. and Cooke, R.M. (2006). Uncertainty Analysis with High Dimensional Dependence Modelling. Wiley.

Lishchuk, V., Koch, P-H., Lund, C., and Lamberg, P. (2015). The geometallurgical framework. Malmberget and Mikheevskoye case studies. Minerals and Metallurgical Engineering Division, Luleå University of Technology.

Liu, Z., Zhou, P., Chen, X., and Guan, Y. (2015). A multivariate conditional model for streamflow prediction and spatial precipitation refinement. Journal of Geophysical Research: Atmospheres, 120 (19). https://doi.org/10.1002/2015JD023787

Musafer, G.N. and Thompson, M.H. (2016). Non-linear optimal multivariate spatial design using spatial vine copulas. Stochastic Environmental Research and Risk Assessment, 31 (2), 551–570.

Musafer, G.N., Thompson, M.H., Kozan, E., and Wolff, R.C. (2013). Copula-based spatial modelling of geometallurgical variables. Proceedings of the Second AusIMM International Geometallurgy Conference, Brisbane, QLD, 30 September - 2 October 2013. Australasian Institute of Mining and Metallurgy, Melbourne. pp 239-246.

Nelsen, R. (2006). An introduction to copulas. Lecture Notes in Statistics. Springer, New York.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team. (2016). _nlme: Linear and nonlinear mixed effects models_. R package version 3.1-128. http://CRAN.R-project.org/package=nlme

Schepsmeier, U., Stöber, J., Brechmann, E.C., Graeler, B., Nagler, T., and Erhardt, T. (2016). VineCopula: Statistical inference of vine copulas. R package version 2.0.5. https://github.com/tnagler/VineCopula

Walters, S. and Kojovic, T. (2006). Geometallurgical mapping and mine modelling (GEMIII) - the way of the future. Proceedings of SAG 2006, Vancouver, Canada, 23-27 September 2006. Vol. 4. pp. 411-425.

Whitley, D. (1994). A genetic algorithm tutorial. Statistics and Computing, 4 (2), 65–85.

Williams, S.R. and Richardson, J. (2004). Geometallurgical mapping: A new approach that reduces technical risk. Proceedings of the 36th Annual Meeting of the Canadian Mineral Processors. Canadian Institute of Mining, Metallurgy and Petroleum, Montreal. pp. 241-268.

Willmott, C.J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30 (1), 79-82.

Xu, Q. and Childs, T. (2013). Evaluating forecast performances of the quantile autoregression models in the present global crisis in international equity markets. Applied Financial Economics, 23 (2), 105-117.

Chapter 6

Conclusions and Future Work

Resource and reserve estimation is significant because large capital investment decisions are often made based on these models. Major decisions like advancing or retreating with further exploration of orebodies/deposits are based on these models. The process of resource estimation is most often very challenging because geological and geometallurgical variables exhibit very complex features, which makes capturing all of them in a model very cumbersome and complex. The main problem is that the current geostatistical modelling tools fail to model and preserve the univariate/multivariate non-linearity and spatial non-linearity present in these geological and geometallurgical variables. Moreover, due to the massively multivariate nature of these geological and geometallurgical variables, current modelling tools are either very cumbersome or inefficient to model accurately these datasets. Finally, most geometallurgical variables are usually under sampled, this is as a result of the legacy nature or the very expensive sampling cost. These main practical issues present to the geostatistician to either retain data where all variables are sampled or remove large amounts of data from the modelling process. Retaining samples at only locations where all variables are sampled or removing data from the modelling process may result in the introduction of bias or loss of information and finally a poor geostatistical model.

Advances in pair-copula theory and efficient data imputation algorithms means there is considerable potential to address these problems and improve methods of resource estimation. The primary aim of this thesis was to develop alternate univariate/multivariate modelling tools and a geostatistical data imputation algorithm to model geological and geometallurgical variables. Specifically three main objectives are listed below:

a) Univariate model: Extend on the existing copula-based models (Gräler and Pebesma 2011; Musafer and Thompson 2013) and develop a novel copula based anisotropy

model to estimate geological variables that exhibit skewness and non-linear spatial dependence.

b) Multivariate model: To apply the copula-based spatial model in a multivariate setting to model and predict geometallurgical variables whilst capturing non-linear spatial dependence and complex multivariate relationships. In addition, develop a novel predicting framework using copula based multivariate model.

c) Data Imputation: To apply a novel geostatistical data imputation algorithm to missing geometallurgical datasets, this new algorithm seeks to preserve the individual histograms and bivariate correlation among the geometallurgical variables.

In this thesis, copula geostatistical models have been employed to model geological and geometallurgical variables. The main practical advantages of copula based models over traditional geostatistical modelling tools are that they represent the non-linear spatial dependence structures and other complexities that come with geological and geometallurgical variables. Furthermore, copula based models are able to model more accurately the error structure associated with predicting geological and geometallurgical variables. Having a reliable or accurate error structure model may play a vital role in quantifying the risk associated with estimating orebodies/deposits.

6.1 Conclusions

This thesis achieves the proposed objectives outlined above by the following outcomes:

Paper 1: Spatial Pair-Copula Model of Grade for an Anisotropic Gold Deposit

- Proposed an alternative copula based geostatistical model to predict geological variables, this model was able to preserve and capture the univariate non-linear spatial relationship present in the data.

- The copula based geostatistical model has only been applied in a mining context to model isotropic orebodies. This is the first ever application of a copula based geostatistical model to anisotropic orebodies. This paper elaborates on the step by step guidelines in using the copula based approach to model an anisotropic orebody in a practical application.
- The copula based geostatistical model gives 5% better predictive performance on mean absolute error compared with a traditional geostatistical model (lognormal kriging) when non-linear spatial dependence and high outlying values are present in the geological variables.

Paper 2: Estimation of Direction of Increase of Gold Mineralisation using Spatial Pair-Copulas

- The copula based geostatistical model developed for anisotropic orebodies in paper 1 was used to make predictions of unknown sampling points outside the main geological field. This enhanced model was able to preserve the non-linear spatial relationship of the geological variables and therefore results in accurate predictions and prediction intervals of the unknown points.
- The traditional geostatistical method of making predictions at unknown locations is to use the variogram model to estimate the spatial relationship of the variables and then krige. However, the variogram fails to model the co-movement of the extreme outlying values because it is simply a measure of linear dependence over the distribution of the variables. In circumstances where the geological variables show high outlying values and non-linear spatial relationship, the variogram model will underestimate the predicted variable at unknown locations. Alternatively, the copula based geostatistical model is able to capture and preserve the non-linear spatial relationship and is not affected by high outlying values. In this case study, the copula based models gave 5%

more predictions at unknown sample locations, therefore the prediction intervals from the copula based model are more reliable compared with traditional log-normal kriging.

- Drilling is a crucial and expensive part of the mining chain and hence models used to make predictions outside main geological fields needs to be accurate and reliable. These models serves as a template for designing drill holes and therefore needs to be able to model all the complex features that geological variables exhibit.

Paper 3: A comparison of Gaussian, Student-*t* and Vine Copulas for Modelling Geophysical Measurements along a Rock Drill Core

- Existing traditional multivariate geostatistical models for spatial variables must include co–regionalisation models. However, these multivariate models ignore non-linear dependence between all the variables and therefore fail to reproduce the within-variable spatial dependence across all locations. Alternatively, copula based geostatistical models are able to model the multivariate non-linear dependence between all variables and in addition are easy to fit and model.
- A new predicting framework using a copula based multivariate model was developed and presented. The new method was able to preserve the multivariate non-linear dependence between all the variables and the within-variable spatial dependence at all locations. The 90% confidence intervals generated from the vine copula are 10% more on lower and upper limits, which is a result of the detailed modelling of the error structure using vine copulas.
- This new predicting framework and model are easy to use and not time consuming, because the run time is much faster and the mathematics is more tractable. However, traditional geostatistical modelling of multiple spatial variables is complex and time consuming due to the requirements of several cross-variograms, with an increasing number of variables.

Paper 4: Prediction of Copper Recovery from Geometallurgical Data using D-Vine Copulas

- Proposed a new data imputation algorithm to impute missing data at some locations where geometallurgical variables are absent. This data imputation algorithm preserves vital statistics (histograms and bivariate correlation) of the geometallurgical data sampled at all locations.
- This paper also proposed a D-vine copula model to predict geometallurgical variables at unsampled locations. This model is able to capture non-linear relationships usually found among predictors and dependent variables. The D-vine copula model resulted in a 10% less on the mean absolute error for predicted values compared with the other traditional models.

6.2 Future Work

There is substantial improvement achieved through using copulas to model geological and geometallurgical variables. However, there are nonetheless some challenges and limitations in their application which provides motivation for ongoing research:

- A two dimensional (2-D) anisotropic orebody soil sample dataset was presented in paper 1 and 2 of this thesis. However, in most mining applications, resource and reserve classifications are conducted on a three-dimensional (3-D) drillhole database. A 3-D model could be achieved if slight addition is made to the current model and algorithm to capture the third (z) dimension.
- Choosing the optimum number of nearby sample locations for predictions at unknown location can be computationally challenging and time consuming when using a copula based geostatistical model. Therefore, new methods are needed which can efficiently sample and select nearby samples for predictions. Numerical integration methods that can improved the selection of nearby samples should be considered in future research.

- The proposed copula based geostatistical model presented in paper 3 only preserved and captured the non-linearity present in multivariate spatial geological variables. However, some geological and geometallurgical datasets could exhibit both non-linearity and heteroscedasticity. Therefore, work remains to extend the framework to incorporate both the non-linearity and heteroscedasticity. This would require an update of the current model presented in this thesis.
- Geometallurgical response variables presented in paper 4 in this thesis are assumed to scale up linearly from the laboratory scale to the plant scale. This is a very strong assumption, and is not applicable in most real mining cases. Therefore, field-based studies would be required to test the validity of this assumption. A challenge in performing such a study would be poor metallurgical data which exist in most mining databases.
- The proposed Genetic Algorithm (GA) presented in paper 4 for data imputation yielded good results in the test case presented. However, GAs do not guarantee near optimal solutions. Furthermore, GAs are a stochastic method and hence different runs may lead to different results. Therefore, robustness of solutions when using different models needs further exploration.
- Finally the Genetic Algorithm (GA) presented in paper 4 of this thesis was designed to reproduce just two main statistics (i.e., the histogram and bivariate correlations). This can be improved to capture most of the general statistics of the geological and geometallurgical datasets. This would require an update of the current algorithm developed in this thesis.

The research conducted in this thesis has identified further potential research opportunities that could be addressed in future research work. In paper 1 and 2 (Chapter 2 and 3), the soil sample datasets enabled a simple and yet a comprehensive approach when using the spatial pair-copula

to explain anisotropic dependence structures. The anisotropy was explained with empirical copula density plots for each distance and direction. However, in mining applications, resource and reserve estimations are conducted on a three-dimensional (3-D) drillhole database. Therefore, future work should aim at using the pair-copulas to explore anisotropic 3-D drillhole geological data. Furthermore, choosing the optimal number of nearby sample locations for the case study presented in paper 1 and 2 (Chapter 2 and 3) was computationally challenging and time consuming. An efficient search algorithm, example as direct search simulated annealing can be integrated into the current proposed algorithm to reduce the current computational time. This should be the focus for future research work.

In paper 3 (Chapter 4), the proposed copula based geostatistical model presented preserves the non-linearity of the multivariate geological and geometallurgical datasets. However, in most practical mining applications, multivariate spatial geological and geometallurgical datasets could exhibit both non-linearity and heteroscedasticity. Therefore, future work should concentrate on using the copula based geostatistical model to capture and preserve the non-linearity and heteroscedasticity present in geological and geometallurgical variables.

In paper 4 (Chapter 5), the GA algorithm presented was conditioned to reproduce only two main statistics (i.e., histograms and bivariate correlations), this could be improved to capture vital statistics of geometallurgical variables. This area should be further researched in future work.

This thesis presented method to model univariate/multivariate geological and geometallurgical variables. The improvement provided by these methods means that geological and geometallurgical variables could be modelled more accurately and precisely. Furthermore, a much more realistic prediction intervals could be generated from these models which can assist in quantifying the risk associated in modelling geological and geometallurgical variables.

While there is not yet a full 3-D demonstration, it is nonetheless recommended that the methods could be used for 2-D applications. A number of avenues for further work have been outlined, which are recommended as a worthwhile research activity given the potential for copula-based geostatistical methods.

Appendix C

Model Code for Paper 1 and 2

```
require (copula);

require (spcopula);

library (spcopula);

require (VineCopula);

bins = calcBins(data1, var = "LNAu", nbins =3, boundaries = seq(from=300, to=4500, by=300), cor.method = "kendall");

calcKTauPol = fitCorFun(bins, degree=3, cutoff= 600,cor.method = "kendall", weighted = F);

curve(calcKTauPol,0, 4500, col="blue",add=TRUE,lwd=2.5);

families = list(normalCopula(0), tCopula(0),claytonCopula(1), frankCopula (1), gumbelCopula (1.5), joeBiCopula (1.5),tawnCopula (1), surClaytonCopula (1.5),surGumbelCopula (1.5),surJoeBiCopula (1.5));

loglikTau= loglikByCopulasLags(bins, data1, families, calcKTauPol, lagSub = 1:length(bins$meanDists));

bestFitTau = apply(apply(loglikTau$loglik, 1, rank, na.last=T), 2, function(x) which(x==10));

colnames(loglikTau$loglik)[bestFitTau];

spCop = spCopula(append(families[bestFitTau[1:2]], indepCopula()), distances= bins$meanDists[1:3], spDepFun= calcKTauPol, unit="m");

vineDim = 11L;

data1Neigh = getNeighbours(dataLocs = data1, var = "LNAu", size = vineDim, min.dist =0.01);

data1SpVine = spcopula:::fitSpVine(spVineCopula(spCop,vineCopula(as.integer(vineDim-1))), list(data1Neigh,data1));

data1SpVine = data1SpVine@copula;

predMedian = NULL

predMean = NULL

for (loc in 1:nrow(data1Neigh@data)){ ## loc = 1500

cat("Location: ", loc,"\n")

condSecVine = condSpVine(condVar = as.numeric(data1Neigh@data[loc,-1]),

dists = list(data1Neigh@distances[loc,,drop=F]),data1SpVine)
```

```
predMedian = c(predMedian, qMar(optimise(function(x)
abs(integrate(condSecVine,0,x)$value-0.5),c(0,1))$minimum))

 condExp = function(x){

   condSecVine(pMar(x))*dMar(x)*x

   }

   predMean = c(predMean, integrate(condExp,0,3000, subdivisions = 1e6)$value)

 }
```

Model Code for Paper 3

```
library (copula);
library (VineCopula);
#### Fitting of Gaussian Copula ####
u = cbind (Emac[,1],Emac[,2],Emac[,3]);
u = as.matrix (u);
fit.tau_n = fitCopula (normalCopula(dim=3, dispstr="un"),u,method = "ml");
rho = coef (fit.tau_n);
print (rho);
summary (fit.tau_n);
#### Fitting of t Copula ####
u1 = cbind(Emac[,1],Emac[,2],Emac[,3]);
u1 = as.matrix(u1);
fit.tau_t = fitCopula(tCopula(dim=3, dispstr="un", df=5),u1, method ='ml');
rho = coef(fit.tau_t);
print(rho);
summary(fit.tau_t);
#### Fitting of Vine Copula ####
Matrix = c(2,3,1,0,1,3,0,0,3);
Matrix = matrix(Matrix,3,3);
family = c(0,40,18,0,0,20,0,0,0);
family = matrix(family,3,3);
par1 = c(0.000,-5.341441,3.938817,0.0000,0.0000,4.034865,0,0,0);
par1 = matrix(par1,3,3);
par2 = c(0.0000,-0.4775409,2.2860431,0.00000,0.00000,0.7134058,0,0,0);
par2 = matrix(par2,3,3);
RVM = RVineMatrix(Matrix = Matrix,family = family,par = par1,par2 = par2,  names=
c("V1","V2","V3"));
cop    = normalCopula(param = c(0.9378,0.4857,0.3932), dispstr = "un", dim = 3)
x = mvdc(copula = cop, margins = c("unif","unif","unif"), paramMargins = list(list(min = 0,
max = 1), list(min = 0, max = 1, list(min = 0, max = 1)));
```

```
cop1 = tCopula(c(0.959,0.534,0.456), dispstr = "un", dim = 3, df= 6)

x1 = mvdc(copula = cop1, margins = c("unif","unif","unif"), paramMargins = list(list(min =
0, max = 1), list(min = 0, max = 1),list(min = 0, max = 1)));

set.seed(123);

cop2 = r270BB8Copula(param = c(-5.34,-0.31));

marg = list(list(min = 0, max = 1),list(min = 0,max = 1));

x2 = mvdc(copula = cop2, margins = c("unif","unif"),paramMargins = marg);

### Analysis of the tails for Gaussian fitted copula ####

set.seed(100)

Gsim = rMvdc(100000, x);

fhat0 = kde(x=Imp,h=hpi(Imp), gridtype = "linear", xmin = 5e-04, xmax = 27975.29);

fhat1 = kde(x=Vel,h=hpi(Vel), gridtype = "linear",xmin = 5e-04, xmax = 7269.714);

fhat2 = kde(x=Ms,h=hpi(Ms), gridtype = "linear",xmin = -389.8024, xmax = 532.6693);

Gsim[,1] = qkde(Gsim[,1], fhat = fhat0);

Gsim[,2] = qkde(Gsim[,2], fhat = fhat1);

Gsim[,3] = qkde(Gsim[,3], fhat = fhat2);

mean1 = mean(Gsim[,1]);;

mean2 = mean(Gsim[,2]);

mean3 = mean(Gsim[,3]);;

sigma1 = sd(Gsim[,1]);

sigma2 = sd(Gsim[,2]);

sigma3 = sd(Gsim[,3]);

sim_1 = outer(Gsim[,1],mean1+2*sigma1, '>');

sim_2 = outer(Gsim[,2],mean2+2*sigma2, '>');;

sim_3 = outer(Gsim[,3],mean3+2*sigma3, '>')

sim_11 = outer(Gsim[,1],mean1-2*sigma1, '<');

sim_12 = outer(Gsim[,2],mean2-2*sigma2, '<');

sim_13 = outer(Gsim[,3],mean3-2*sigma3, '<');

sim_product = sim_1*sim_2*sim_3

sim_product1 = sim_11*sim_12*sim_13
```

```
a = colSums(sim_product);
b = colSums(sim_product1);;
((a+b)/100000)*1000000;
### Analysis of the tails for t fitted copula ####
set.seed(100);
tsim = rMvdc(100000, x1);
tsim[,1] = qkde(tsim[,1], fhat = fhat0);
tsim[,2] = qkde(tsim[,2], fhat = fhat1);
tsim[,3] = qkde(tsim[,3], fhat = fhat2);
nmean1 = mean(tsim[,1]);
nmean2 = mean(tsim[,2]);
nmean3 = mean(tsim[,3]);
nsigma1 = sd(tsim[,1]);
nsigma2 = sd(tsim[,2]);
nsigma3 = sd(tsim[,3]);
sim_1 = outer(tsim[,1],nmean1+2*nsigma1, '>');
nsim_2 = outer(tsim[,2],nmean2+2*nsigma2, '>');
nsim_3 = outer(tsim[,3],nmean3+2*nsigma3, '>');
nsim_11 = outer(tsim[,1],nmean1-2*nsigma1, '<');
nsim_12 = outer(tsim[,2],nmean2-2*nsigma2, '<');
nsim_13 = outer(tsim[,3],nmean3-2*nsigma3, '<');
nsim_product = nsim_1*nsim_2*nsim_3;
nsim_product1 = nsim_11*nsim_12*nsim_13;
na = colSums(nsim_product);
na1 = colSums(nsim_product1);
((na+na1)/100000)*1000000;
```

Model Code for Paper 4

```
library(VineCopula);
library(copula);
library(vinereg);
library(nlme);
library(gstat);
library(sp);
### Fit Variogram for all four variables ###
vgm = variogram(A0~x+y+z, data = Data1)
fit.vgm = fit.variogram(vgm, model = vgm(0.5,"Sph",230, 0.27));
plot(vgm,fit.vgm,  col="blue", pch =13, cex= 1 );
### Linear model Model0  evidence against stationarity ###
model0 = lm(A~x+y+z+x2+y2+xy, Data);
summary(model0);
A1 = model0$residuals
names(Data)
coordinates(Data) = ~x+y+z;
### Model 1 is the gls model with the same terms as Model0 ###
model1 = gls(A~x+y+z+x2+y2+xy, Data);
summary(model1);
### Update new model with the variogram structure from model3 ###
model2 = update(model1, corr = corSpher(c(229.5,0.27),form =~x+y+z, nugget =TRUE));
summary(model2);
### Extract the residual from model2 and use that as the new Recovery ###
A11 = model2$residuals;
### Model 00 is the LINEAR REGRESSION ###
newxyz  = Data[c(1:3)];
y = cbind(b,c,d,bc,bd,cd,b2,c2,d2)
newData = as.data.frame(y);
model00 = lm(A11~b+c+d+bc+bd+cd+b2+c2+d2, newData) ## you need get the new Data for this regression
```

```
summary(model00)
lm_pred_BCD = predict(model00, newdata = newData); ## This is predicting Rec from B,C and D
summary(lm_pred_BCD);
pred_xyz = predict(model2, newdata = newxyz); ## Thisis predicting x, y and z from model2
pred_xyz;
### Final prediction from the linear regression model ###
lm_pred_Rec = pred_xyz + lm_pred_BCD;
summary(lm_pred_Rec);
### Model 3 is the LEAST SQUARE REGRESSION MODEL  ###
model3 = gls(A11~b+c+d+bc+bd+cd+b2+c2+d2, data = newData);
### Update new model with the variogram structure from model3 ###
model4 = update(model3, corr = corSpher(c(229.5,0.27),form =~x+y+z, nugget =TRUE));
summary(model4);
gls_pred_BCD = predict(model4, newdata = newData);
gls_pred_BCD;
pred_xyz = predict(model2, newdata = newxyz); ## Thisis predicting x, y and z from model2
pred_xyz;
### Final prediction from the least square regression model ###
gls_pred_Rec = pred_xyz + gls_pred_BCD;
summary(gls_pred_Rec);
### Fit vine D-copula to the new data and make predictions ###]
fit_vine_par = vinereg(
  A11~b+c+d+bc+bd+cd+b2+c2,
  data = newData,
  selcrit = "loglik",
  family_set = "parametric"
)
fit_vine_par$order
summary(fit_vine_par$vine)
contour(fit_vine_par$vine)
```

```
### Predict from the fitted D vine copula ###
newData1 = fit_vine_par$model_frame;
newData1 = newData1[c(2:9)];
pred_DVine = predict(fit_vine_par, newdata = newData1, alpha = 0.5);
Dvine_pred_Rec = pred_DVine$`0.5` + pred_xyz;
summary(Dvine_pred_Rec)
```