



Prediction of copper recovery from geometallurgical data using D-vine copulas

E. Addo Jr¹, A.V. Metcalfe², E.K. Chanda¹, E. Sepulveda^{1,5},
W. Assibey-Bonsu³, and A. Adeli⁴

Affiliation:

¹School of Civil, Environmental and Mining Engineering, University of Adelaide, Australia

²School of Mathematical Sciences, University of Adelaide, Australia

³Group Geostatistician and Evaluator, Gold Fields Limited, South Africa

⁴Department of Mining Engineering, University of Chile, Chile

⁵School of Mining Engineering, University of Talca, Chile

Correspondence to:

E. Addo Jr

Email:

emmanuel.addojunior@adelaide.edu.au

Dates:

Received: 10 Sep. 2018

Revised: 1 Apr. 2019

Accepted: 1 Apr. 2019

Published: April 2019

How to cite:

Addo Jr, E., Metcalfe, A.V., Chanda, E.K., Sepulveda, E., Assibey-Bonsu, W., and Adeli, A. Prediction of copper recovery from geometallurgical data using D-vine copulas.

The Southern African Institute of Mining and Metallurgy

DOI ID:

<http://dx.doi.org/10.17159/2411-9717/319/2019>

ORCID ID:

<https://orcid.org/0000-0002-5046-8934>

This paper was first presented at the *Geometallurgy Conference 2018 'Back to the Future'*, 7–8 August 2018, Lagoon Beach Conference Venue, Cape Town, South Africa.

Synopsis

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.

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 (BW_i); 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 BW_i, 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 a 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,

Prediction of copper recovery from geometallurgical data using D-vine copulas

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. This is followed by the 'Discussion and Conclusion'.

Method

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).

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).

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 '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.

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 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 1 and Equation [3], the joint density function of four random variables can be expressed using the D-vines asw

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 $F(z|v)$

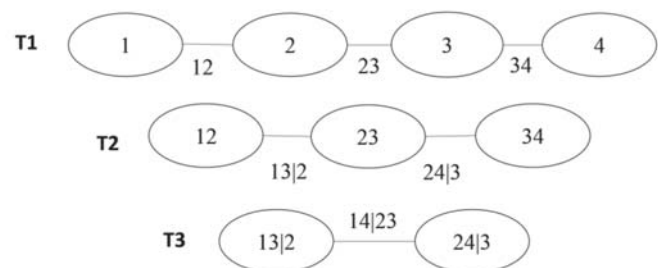


Figure 1 - D-vines for four variables

Prediction of copper recovery from geometallurgical data using D-vine copulas

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

for an m -dimensional vector $v=(v_1, \dots, v_m)$ can be obtained from the following recursive relationship:

$$\begin{aligned}
 h(z|v) := F(z|v) = \\
 \frac{\partial C_{zv_j|v-j}(F(z|v-j), F(v_j|v-j))}{\partial F(v_j|v-j)}
 \end{aligned} \quad [4]$$

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 1 *Tree 3 (T3)* the argument $C_{14|23}$, namely $F_{123}(x_1|x_2, x_3)$, 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 T_1 and T_2 as

$$\begin{aligned}
 F_{123}(x_1|x_2, x_3) = & h_{1|3;2}(F_{1|2}(x_1|x_2)F_{3|2}(x_3|x_2)) \\
 = & h_{1|3;2}(h_{1|2}(F_1(x_1)|F_2(x_2))) \\
 & h_{3|2}(F_3(x_3)|F_2(x_2))
 \end{aligned} \quad [5]$$

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

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

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

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

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].

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 obtained using a fitted model to all $N=930$. So, the performance measures are calculated within the entire sample.

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 (BW_i), resistance to abrasion and breakage index (A*_b), and semi-autogenous grinding power index test (Spi). Typical of most geometallurgical data-sets, there are missing values that have not been sampled at some locations. There are 299 non-missing data-sets that are sampled at all locations. To be able to use all 930 georeferenced drill core samples for the analysis, we employed a data imputation algorithm to predict missing values at some locations.

Data imputation

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

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

Prediction of copper recovery from geometallurgical data using D-vine copulas

Table I
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

using non-missing (informed) values, and correlations were calculated using samples where all variables have non-missing values. Table II 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.

The data-sets were decomposed in two sets: non-missing (informed) values and missing values. Hence, the multivariate data-set was defined as: $X = \{X_1, \dots, X_D\}$, where D is the number of attributes. Each X_i is also defined 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.

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 I, there are 1481 missing values to impute.

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.

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

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.

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

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 III shows the parameters used in the GA for data imputation.

Table II
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

Table III
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
cspb	0.9	Probability of applying crossover
mutpb	0.4	Probability of applying mutation

Prediction of copper recovery from geometallurgical data using D-vine copulas

The Pearson's correlation computed for imputed and non-missing samples shows that the correlations were perfectly reproduced. Figure 2 (upper and lower panel) shows the histogram of the imputed data and with non-missing values respectively. Figure 3 also shows all four non-missing variables (*i.e.*, Rec, BWi, Spi, and A*b) in space. We discuss Figure 3 under the 'Discussion and Conclusion' section.

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

$$A = \beta_0 + \beta_1x + \beta_2y + \beta_3z + \beta_4x^2 + \beta_5y^2 + \beta_6xy + \varepsilon \quad [13]$$

where ε is the random error, which is expected to be spatially correlated, with mean of zero and standard deviation σ_ε . *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 4. The estimated coefficients for the GLS fit to *Model1* are shown in Table IV.

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.

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 (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*.

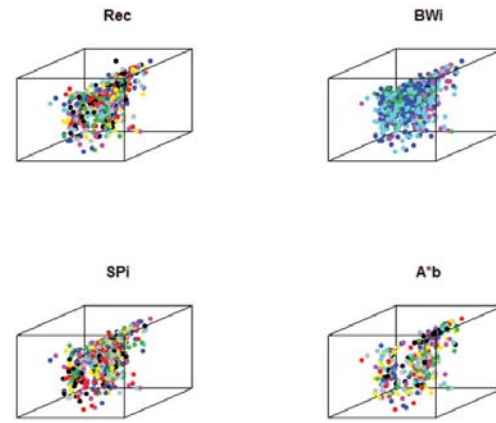


Figure 3—The 3D spatial position of the samples showing non-missing values of Rec (top left), BWi (top right), Spi (bottom left), and A*b (bottom right)

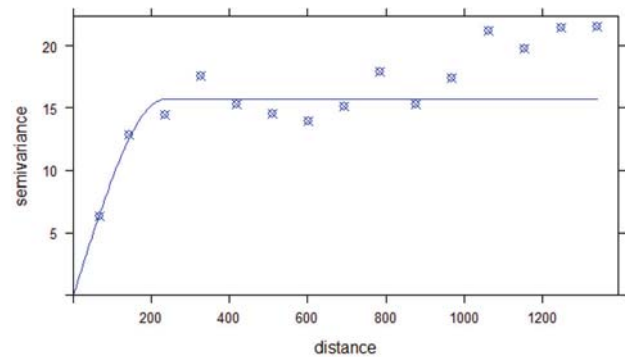


Figure 4—Fitted spherical variogram with range 230, sill 15.7, and nugget 0.3.

Table IV

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

Coefficient	Estimate	Estimated standard error	P-value
β_0	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

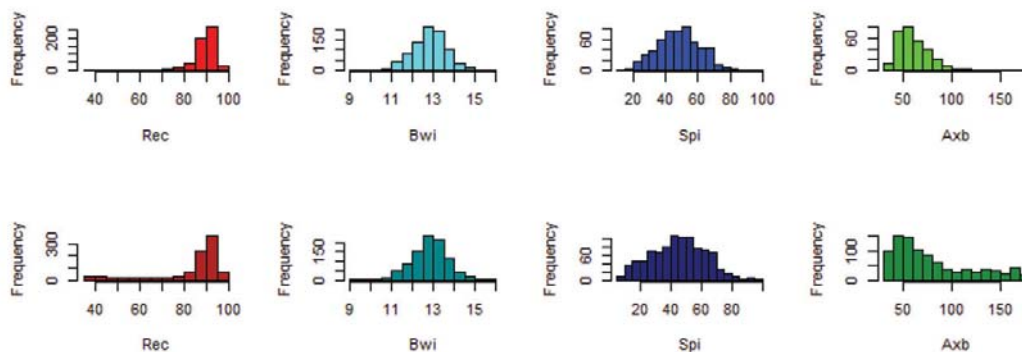


Figure 2—Histogram of four variables with non-missing data (upper panel), histogram of four variables with imputed data (lower panel)

Prediction of copper recovery from geometallurgical data using D-vine copulas

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 (θ_1, θ_2 , and ρ), 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 $90^\circ, 180^\circ$, 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 illustrates the fitted bivariate copulas and their fitted dependence parameters (θ_1, θ_2, ρ , 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 10D version of Equation [7]. The predicted values were 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*.

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, B², C², and D²*. Equation [14] shows the OLS regression model fitted, and Table V shows the estimated coefficients for the OLS regression.

$$A^* = \beta_0 + \beta_1 B + \beta_2 C + \beta_3 D + \beta_4 BC + \beta_5 BD + \beta_6 CD + \beta_7 B^2 + \beta_8 C^2 + \beta_9 D^2 + \varepsilon \quad [14]$$

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, B², C², and D²* 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 VI.

Table V
Estimated coefficients of the fitted OLS regression

Coefficient	Estimate	Estimated standard error	P-value
β_0	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

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

Coefficient	Estimate	Estimated standard error	P-value
β_0	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	0.001673	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

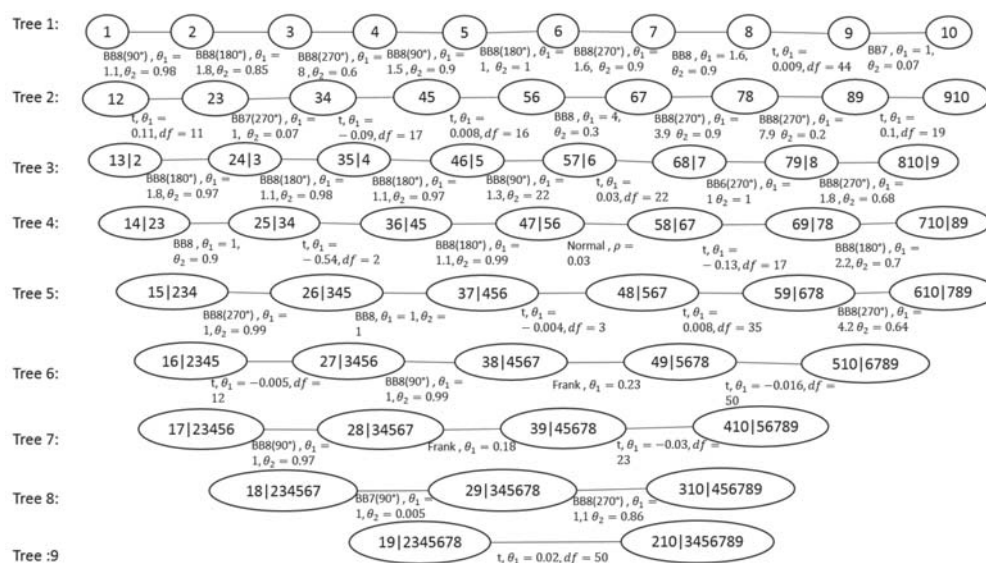


Figure 5—Structure of the 10-dimension D-vine model, where 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 are A, B, C, D, BC, BD, CD, B², C², and D² respectively

Prediction of copper recovery from geometallurgical data using D-vine copulas

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 VII. 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 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.

Out-of-sample predictions were generated and compared with the known values of the data removed. Figure 7 shows box plots 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. Summary statistics are given in Table VIII.

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.

Discussion and conclusion

Data on four variables (*Rec*, *BWi*, *Spi*, and *A*b*) from 930 drill

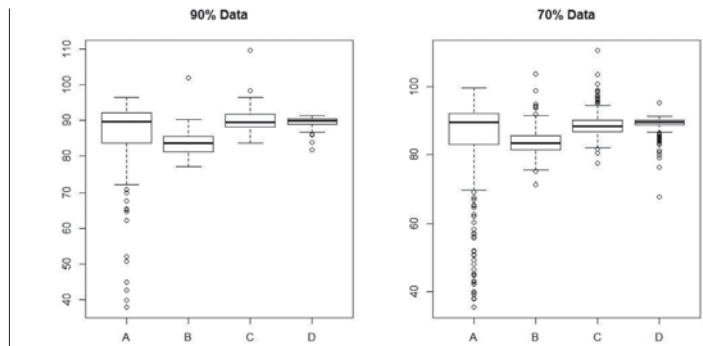


Figure 7—Box plots of removed CuRec (A), out-of-sample predictions of OLS (B), GLS (C), and D-vine (D) models for 90% (left) and 70% (right) of data

core samples at known 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 data-sets. The recovery (*Rec*), the response in this application, shows a slight, but statistically significant, nonstationarity. The nonstationarity 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.

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

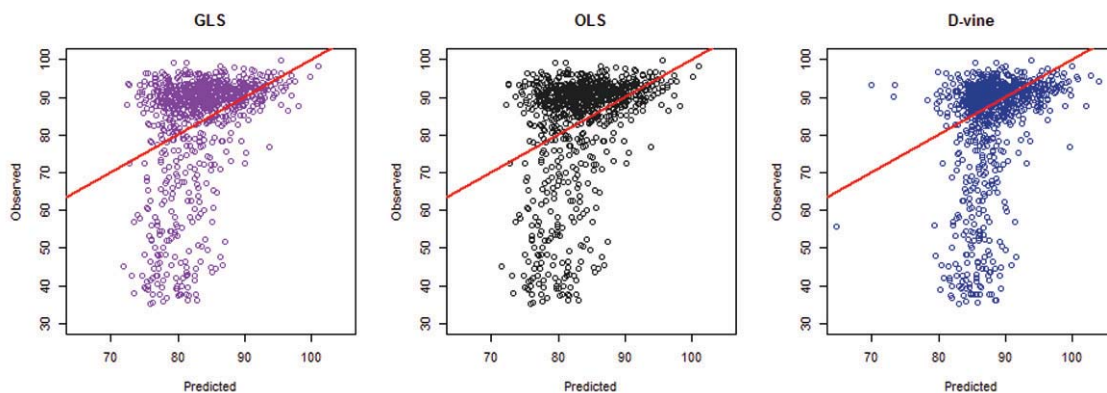


Figure 6—Predicted vs observed recoveries for GLS, OLS, and D-vine models

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

Prediction of copper recovery from geometallurgical data using D-vine copulas

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%) was removed at random locations from the complete geometallurgical data-sets. 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 an Australian Government Research Training Program Scholarship awarded to Emmanuel Addo Jr. The authors thank the mining company for providing the geometallurgical data-sets used in this case study. We express our 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*, vol. 44, no. 2. pp. 182–198.
- ADDO, E., CHANDA, E., and METCALFE, A.V. Spatial pair-copula model of grade for an anisotropic gold deposit. *Mathematical Geosciences*, July. pp. 1–26. doi: 10.1007/s11004-018-9757-7
- BEDFORD, T. and COOKE, R.M. 2002. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, vol. 30, no. 4. pp. 1031–1068.
- BERNARD, C. and CZADO, C. 2015. Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, vol. 138. pp. 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*, vol. 125, no. 1. pp. 22–34.
- ERHARDT, T.M., CZADO, C., and SCHEPFSMEIER, U. 2011. R-vine models for spatial time series with an application to daily mean temperature. *Biometrics*, vol. 71, no. 2. pp. 323–332.
- FENSKÉ, 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*, vol. 106, no. 494. pp. 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*, vol. 7. pp. 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*, vol. 101, no. 5. pp. 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*, vol. 120, no. 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*, vol. 31, no. 2. pp. 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>
- SCHEPFSMEIER, U., STÖBER, J., BRECHMANN, E.C., GRAELER, B., NÄGLER, 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*, vol. 4, no. 2. pp. 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*, vol. 30, no. 1. pp. 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*, vol. 23, no. 2. pp. 105–117. ◆