On the spatial modelling of mixed and constrained geospatial data

by

Hassan Talebi

(MESc)

A Thesis with Publications presented to Edith Cowan University in fulfilment of the requirement for the degree of
Doctor of Philosophy

School of Science
Edith Cowan University
Joondalup, WA 6027, Australia

2018

Abstract

Spatial uncertainty modelling and prediction of a set of regionalized dependent variables from various sample spaces (e.g. continuous and categorical) is a common challenge for geoscience modellers and many geoscience applications such as evaluation of mineral resources, characterization of oil reservoirs or hydrology of groundwater. To consider the complex statistical and spatial relationships, categorical data such as rock types, soil types, alteration units, and continental crustal blocks should be modelled jointly with other continuous attributes (e.g. porosity, permeability, seismic velocity, mineral and geochemical compositions or pollutant concentration). These multivariate geospatial data normally have complex statistical and spatial relationships which should be honoured in the predicted models.

Continuous variables in the form of percentages, proportions, frequencies, and concentrations are compositional which means they are non-negative values representing some parts of a whole. Such data carry just relative information and the constant sum constraint forces at least one covariance to be negative and induces spurious statistical and spatial correlations. As a result, classical (geo)statistical techniques should not be implemented on the original compositional data. Several geostatistical techniques have been developed recently for the spatial modelling of compositional data. However, few of these consider the joint statistical and/or spatial relationships of regionalized compositional data with the other dependent categorical information.

This PhD thesis explores and introduces approaches to spatial modelling of regionalized compositional and categorical data. The first proposed approach is in the multiple-point geostatistics framework, where the direct sampling algorithm is developed for joint simulation of compositional and categorical data. The second proposed method is based on two-point geostatistics and is useful for the situation where a large and representative training image is not available or difficult to build. Approaches to geostatistical simulation of regionalized compositions consisting of several populations are explored and investigated. The multi-population characteristic is usually related to a dependent categorical variable (e.g. rock type, soil type, and land use). Finally, a hybrid predictive model based on the advanced

geostatistical simulation techniques for compositional data and machine learning is introduced. Such a hybrid model has the ability to rank and select features internally, which is useful for geoscience process discovery analysis.

The proposed techniques were evaluated via several case studies and results supported their usefulness and applicability.

Keywords: compositional data, two-point geostatistics, multiple-point geostatistics, machine leaning, spatial predictive models.

Declaration

I certify that this thesis does not, to the best of my knowledge and belief:

i. Incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education;

ii. Contain any material previously published or written by another person except where due reference is made in the text;

iii. Contain any defamatory material.

I also grant permission for the Library at Edith Cowan University to make duplicate copies of my thesis as required.

Signed: Hassan Taleb [redacted]

Data: 15/10/2018

Acknowledgements

I wish to express my sincere gratitude to those people without whom it would not be possible for me to undertake this program during my years at Edith Cowan University.

I sincerely thank my Principal Supervisor, Associate Professor Ute Mueller, and Associate Supervisor, Dr. Johnny Lo, for their invaluable time, continuous moral support, insightful guidance and constructive criticism since the start of the program. I would like to acknowledge Dr Raimon Tolosana-Delgado and Professor Karl Gerald van den Boogaart for their contribution to the development of proposed algorithms and implementations.

I am thankful to the technical and administrative staff of the School of Science at ECU for their assistance with regard to the project. The PhD research project was made possible by an Edith Cowan University International Postgraduate Research Scholarship (ECU-IPRS).

The anonymous reviewers of different internationally recognized journals and the thesis are also thanked for their precious time and scholarly comments which were helpful in further developing the manuscripts.

List of Journal Publications Arising from this Candidature

Published Book Chapter

- Talebi H, Lo J, and Mueller U (2017). A hybrid model for joint simulation of high-dimensional continuous and categorical variables. In: J.J. Gómez-Hernández, J. Rodrigo-Ilarri, M.E. Rodrigo-Clavero, E. Cassiraga and J.A. Vargas-Guzmán (Editors), *Geostatistics Valencia 2016*. Springer International Publishing, Cham, pp. 415-430.

Accepted Journal Paper

- Talebi H, Mueller U, Tolosana-Delgado R, van den Boogaart K G (2018). Geostatistical simulation of geochemical compositions in the presence of multiple geological units - Application to mineral resource evaluation. *Mathematical Geosciences*, DOI: 10.1007/s11004-018-9763-9.

Journal Papers Under Review

- Talebi H, Mueller U, Tolosana-Delgado R, (2018). Joint simulation of compositional and categorical data via direct sampling technique - Application to improve mineral resource confidence. *Computer & Geosciences*, Under Review.
- Talebi H, Mueller U, Tolosana-Delgado R, Grunsky, E C, McKinley J M, Caritat P de (2018). Surficial and deep earth material prediction from geochemical compositions, a spatial predictive model, *Natural Resources Research*, Under Review.

Table of Contents

List of Figures

List of Tables

Chapter 1

General Introduction[1]

1.1 Research background

In many geoscience applications such as evaluation of mineral resources, characterization of oil reservoirs, hydrology of groundwater, and contaminated site characterization and remediation, spatial uncertainty modelling and prediction of regionalized variables from various sample spaces (e.g. continuous and categorical) is required. Some of these variables are discrete or qualitative such as rock types, soil types, land uses, alteration or mineralization and some of them are continuous or quantitative such as mineral grade, porosity, permeability, water or oil saturation, and pollutant concentration. Several geostatistical models have been developed for the spatial modelling of categorical or continuous variables (Chilès and Delfiner 2012; Deutsch and Journel 1998; Goovaerts 1997; Wackernagel 2003), but for the joint modelling of such data little has been done because of the difficulties of integrated multivariate modelling of data of different characteristics. As the spatial distributions of these multivariate data are often interdependent, separate modelling of them is insufficient (Emery and Silva 2009; Maleki and Emery 2015; Talebi et al. 2017; van den Boogaart et al. 2018). Multivariate continuous data in the form of percentages, proportions, frequencies, and concentrations are common in geosciences (e.g. geochemical or mineralogical data, proportions of rock types in a mining block, and proportions of soil types or land uses in a study area). Such data are compositional in their nature which means they are non-negative and bounded, representing some parts of a whole (Aitchison 1982; Aitchison 1986). Compositional data carry just relative information and the constant sum constraint forces at least one covariance to be negative inducing spurious statistical and spatial correlations (Aitchison 1986; Pawlowsky-Glahn et al. 2015; Pawlowsky-Glahn and Buccianti 2011; Tolosana-Delgado 2006; van den Boogaart and Tolosana-Delgado 2013). As a result, classical (geo)statistical techniques should not be implemented on the original compositional data (Pawlowsky-Glahn and Egozcue 2016; Pawlowsky-Glahn and Olea 2004; Tolosana-Delgado 2006). The spatial analysis of

[1] This thesis is presented and organised as “Thesis with publication” format.

compositional data is an open area of research (Buccianti and Grunsky 2014; McKinley et al. 2016; Mueller et al. 2014; Mueller et al. 2017; Pawlowsky-Glahn and Egozcue 2016; Tolosana-Delgado et al. 2015a; Tolosana-Delgado et al. 2016; Tolosana-Delgado et al. 2015b; Tolosana-Delgado and van den Boogaart 2014; van den Boogaart et al. 2017; van den Boogaart and Tolosana-Delgado 2013; van den Boogaart et al. 2018). Demands for spatial modelling of such constrained multivariate data with different categorical variables simultaneously add more complexity (Talebi et al. 2017; van den Boogaart et al. 2018). To jointly model such mixed and constrained data, and to reproduce complex relationships between them, existing geostatistical techniques need to be modified and adapted.

1.2 Literature review

1.2.1 Spatial modelling of compositional data

A random vector with non-negative components representing parts of a whole which carries relative information (ratios between components carry information and not the absolute values) is a composition (Aitchison 1982, 1986). Statistical analysis of compositional data and the log-ratio approaches were first introduced by Aitchison (1982, 1986). Many of the regionalized variables predicted via geostatistical approaches are compositional such as ore grades, mineral and geochemical data, contaminants, porosity, saturation, and many other petro-physical variables. Spurious spatial correlations between such regionalized compositional variables, were first recognised by Pawlowsky-Glahn (1984). Spurious correlation is generated when compositional data are treated as real data, with the usual Euclidean geometry (Pawlowsky-Glahn and Egozcue 2016). Indeed, compositions are equivalence classes, so a closed composition is just a representation (Pawlowsky-Glahn et al., 2015). The result of compositional analyses under the assumption of equivalence classes are valid for any other representations and are fully addressed via the implementation of log-ratio transformations. The first attempt to construct spatial models of regionalised compositions was the implementation of the additive log-ratio (alr) transformation and cokriging the log-ratios (Aitchison, 1982; Aitchison, 1986; Pawlowsky-Glahn and Olea, 2004). However, this approach has some limitations. For instance,

computation of variances and covariances using the alr coordinates may be problematic (See Pawlowsky-Glahn and Olea, (2004) for more information). Orthogonal projection of compositional data into real (Cartesian coordinates) space leads to easy use of many (geo)statistical algorithms (Mateu-Figueras et al., 2011). Nowadays analysis of compositional data is commonly summarised by working on coordinates (projection to orthogonal coordinates known as isometric log-ratio transformation (Egozcue et al. 2003)), where compositional data are projected to real space (unbounded and unconstrained) and multivariate geostatistical algorithms can be implemented for spatial modelling purposes, followed by back-transformation to compositional space.

In the case of geostatistical simulation, having a multivariate Gaussian distribution is a primary assumption for experimental data (log-ratios in the case of compositional data). Several methods have been proposed to address this assumption. The results of geostatistical simulation achieved by simple methods of transformation to normal space, like the normal score transformation (Deutsch and Journel 1998), or in the case of high dimensional data with complicated relationships, advanced transformation methods like the stepwise conditional transformation (Leuangthong and Deutsch 2003) and projection pursuit method (Barnett et al. 2014), are not independent of the choice of log-ratio transformation. In the geostatistical treatment of compositional data, it is desirable to have invariant results in each step (Tolosana-Delgado 2006). To transform the log-ratios into a multivariate standard normal distribution, van den Boogaart et al. (2017) proposed a method based on a continuous affine-equivariant multivariate kernel density deformation (flow anamorphosis) which is quite useful for joint geostatistical simulation of compositional data. Several applications have shown that the transformed data via flow anamorphosis are not only multivariate normal but often exhibit absence of spatial cross-correlation which make the geostatistical simulation of such orthogonal factors, more straightforward (Mueller et al. 2017; van den Boogaart et al. 2017). Flow anamorphosis is also capable of reproducing complex patterns in input data including presence of outliers, presence of several populations, nonlinearity, and heteroscedasticity.

Although many studies have been conducted on the spatial modelling of regionalized compositional data (Buccianti and Grunsky 2014; Grunsky et al. 2017;

Grunsky et al. 2014; McKinley et al. 2018; McKinley et al. 2016; Mueller et al. 2014; Pawlowsky-Glahn and Egozcue 2016; Pawlowsky-Glahn and Olea 2004; Tolosana-Delgado and McKinley 2016; Tolosana-Delgado et al. 2016; Tolosana-Delgado et al. 2015b; Tolosana-Delgado and van den Boogaart 2013; Tolosana Delgado 2006; van den Boogaart and Tolosana-Delgado 2013), few of these studies considered the spatial relationships between regionalized compositions and the other dependent categorical information (Talebi et al. 2017; van den Boogaart et al. 2018). The dependent categorical data such as rock type, soil type, mineralization type, and crustal blocks are related (statistically and spatially) to the compositional data. The multi-population characteristic of the input data is generally related to a dependent categorical variable. Most of the time, the input data are separated into purer subpopulations and geostatistical analyses are implemented on these subsets independently (this process is commonly known as domaining). Another approach is to apply nonstationary geostatistical algorithms. However, multivariate geostatistical simulation via flow anamorphosis introduces new ways for spatial modelling of complex compositional data. For instance, the need for domaining prior to geostatistical modelling (to fulfil stationarity assumptions) due to multi-population characteristic of input data may become unnecessary in some applications. More studies are needed to assess the potential of geostatistical simulation of compositional data via orthogonal projection (isometric log-ratio transformation) and flow anamorphosis. The complex relationships between compositional and categorical data should be honoured in the estimated or simulated models. More studies are needed to assess the effect of one or more dependent (statistically and spatially) categorical variable on spatial modelling of compositional data.

1.2.2 Two-point geostatistical modelling of mixed data

Two-point geostatistical algorithms are based on the moments up to second order (variogram, covariance and variance). The spatial autocorrelation (spatial variation for a single variable) and cross-correlation (spatial variation between different variables) are considered via calculating experimental (cross)variograms and fitting models to them (Chilès and Delfiner 2012; Isaaks and Srivastava 1989). As an early

solution to joint spatial modelling of multivariate data with different characteristics, modellers suggested the use of a deterministic model based on one categorical variable and prediction of continuous data within each category separately (Dowd 1986; Duke and Hanna 2001; Rossi and Deutsch 2014; Sinclair 1998; Sinclair and Blackwell 2002). Although this model is simple to apply, it does not consider the uncertainty in the layout of the categories (e.g. geological domains). In this approach geologists have to delineate the exact shape of each layout based on experimental data and their interpretation of earth science processes. Unfortunately, very few of such processes are understood well enough to allow modellers to use deterministic models (Isaaks and Srivastava 1989). As the experimental data become sparse and geology becomes more complex the likelihood of misclassification in the spatial model of categorical variable increases dramatically.

A solution to this shortcoming is to use probabilistic models to simulate the categorical data distributions in space and predict the continuous data in each simulated category independently (Alabert and Massonnat 1990; Boucher and Dimitrakopoulos 2012; Dubrule 1993; Jones et al. 2013; Roldão et al. 2012; Talebi et al. 2016). This method is known as a cascade or hierarchical approach. In this approach geostatistical simulations for categorical data are used to improve the domain definition and quantify the uncertainty in the position of their boundaries by generating multiple realizations. Many simulation models are available for simulating categorical data including sequential indicator (Deutsch 2006; Journel and Alabert 1990; Journel and Gomez-Hernandez 1993), Boolean (Lantuéjoul 2002), truncated Gaussian (Galli et al. 1994; Matheron et al. 1987), plurigaussian (Armstrong et al. 2011), and multipoint simulation (Mariethoz and Caers 2015; Strebelle 2002) and therefore a method suited to the specific data can be selected at this step. Although the cascade approach is simple and has powerful tools for measuring uncertainty in categorical and continuous data, it has some substantial drawbacks. The method does not consider the spatial relationship of continuous and categorical data and also does not take into account the spatial dependence of continuous data across domain boundaries and in turn generates abrupt transitions when crossing boundaries which is not always the case in practice (Kim et al. 2005; Larrondo et al. 2004; Ortiz and Emery 2006; Talebi et al. 2015; Tolosana-Delgado et al. 2014; Wilde and Deutsch 2012).

To account for the continuity of the continuous data across domain boundaries and the uncertainty in the spatial extent of these domains, a probabilistic approach can be applied based on geostatistical simulation of the categorical data and on the calculation of their probabilities of occurrence over the area of interest. These probabilities are then used for weighting the predictions of continuous data to derive predictions associated with each domain (Emery and González 2007a; Emery and González 2007b; Talebi et al. 2015). This approach is appropriate for reproducing gradual transitions in the realization of continuous variables across boundaries (soft boundaries). However, it provides just one scenario of variation for the continuous data in the area of interest which is not useful for uncertainty modelling and risk assessment purposes. On the other hand, the final integrated map may be over-smoothed due to the averaging nature of this algorithm.

To take into account the spatial correlations of continuous and categorical data and spatial correlations of continuous data across boundaries, and as well as considering the uncertainty of categorical and continuous data distributions simultaneously, one approach is to co-simulate these two kinds of variables. Bahar and Kelkar (2000) proposed a co-simulation approach in which one categorical variable is generated by truncating one Gaussian random field and one continuous variable by transforming an independent Gaussian random field. For reproducing the spatial dependencies of two variables they proposed a transformation function for the second Gaussian random field conditionally on the simulated domain. An alternative to this approach is to use a truncated Gaussian random field for categorical data and a correlated Gaussian random field for continuous data (Dowd 1994; Dowd 1997; Freulon et al. 1990). However, the two aforementioned models may have some shortcomings when multivariate data with complex spatial relationships are considered. These methods use several simplifications including using a restrictive coregionalization model for two Gaussian random fields, transforming the categorical variables into continuous Gaussian data without considering the effects of conditional continuous variables, assuming spatially ordered sequences of categories and using one model of anisotropy for them (therefore this model is not practical for modelling complex relationships of geological domains). A more general approach is to use an extension of multivariate Gaussian and plurigaussian models simultaneously (Cáceres and Emery 2010;

Emery and Silva 2009; Maleki and Emery 2015). In this model continuous data are associated with a multivariate Gaussian random field and categorical data with the truncation of one or more Gaussian random fields. Further, it is assumed that all Gaussian random fields are spatially cross-correlated so it is possible to reproduce the dependencies between the categorical and continuous data. This method offers several advantages such as accounting for the uncertainty in the spatial layout of the boundaries between different categories, the ability to reproduce soft boundaries and considering the spatial dependencies between categorical and continuous data. The model also has the ability to incorporate non-stationarity in the categorical data (Maleki and Emery 2017) and can be generalized to the joint simulation of several continuous and categorical variables by adding more Gaussian random fields. Although the method is very flexible and has several advantages over earlier models, there are still some shortcomings. This approach follows a co-simulation based on defining a linear model of coregionalization (LMC) to jointly simulate multivariate data. Simplicity of modelling and verification of the admissibility make the LMC a popular means for defining the spatial relationships of multivariate data (Goulard and Voltz 1992). However, defining symmetrical cross-covariances and using the same structure in the cross-covariance and related variables are shortcomings which decrease the flexibility of the method since in geoscience applications variables are cross-correlated with different support and different spatial behaviour.

To address the problems of LMC, Marcotte (2012) offered a generalized of LMC (GLMC) in which the observed variables are considered as linear combinations of few primary independent variables and some other variables which are deterministic functions of primary variables. A more flexible technique would be, in the multivariate case, the non-LMC approach. Through the use of a non-LMC approach any number of variables, with any number of components for each structure can be considered. Furthermore each component can be isotropic or anisotropic (Marcotte 2015).

High-dimensional data are very common in geosciences and as the number of variables and simulation domain increase, co-simulation approaches based on an LMC or non-LMC will need considerable computer processing to solve large systems of equations per simulated node. An alternative is to decompose the

variables under study into factors which are uncorrelated spatially. Such orthogonal factors can then be simulated independently. Statistical and spatial relationships between variables can be reimposed on the simulated model afterward. This approach for joint simulation offers better accuracy and computational efficiency as the number of attributes being simulated increases. Principal component analysis (PCA) (Davis 1987; Wackernagel 2003), minimum/maximum autocorrelation factors (MAF) (Bandarian et al. 2008; Desbarats and Dimitrakopoulos 2000; Rondon 2012; Vargas-Guzmán and Dimitrakopoulos 2003), and U-WEDGE (Mueller and Ferreira 2012) are some examples of decorrelation methods. As these decorrelation methods have not been developed enough for reproduction of complex relationships such as non-linearity, constraints, or heteroscedasticity, using a chained transformation might produce more satisfactory results (Barnett and Deutsch 2012; Barnett et al. 2014; Mueller et al. 2014). However, a sensitivity analysis must be done to find the optimum order of transformations in a chain. Furthermore, the aforementioned spatial decorrelation techniques were developed for joint simulation of multivariate continuous variables and none of them considered the effects of other dependent regionalized categorical variables. Spatial prediction and uncertainty modelling of a mixture of regionalized continuous and categorical variables is common in many geoscience applications. New spatial decorrelation techniques have to be developed with the ability to jointly simulate many dependent (statistically and spatially) continuous and categorical variables. Such techniques should be able to address the compositional nature of some continuous variables.

1.2.3 Multiple-point geostatistical modelling of mixed data

Two-point geostatistical techniques are constrained by using 2-point statistics only and are inefficient in reproducing complex spatial structures and patterns (Guardiano and Srivastava 1993; Mariethoz and Caers 2015; Strebelle 2000; Strebelle 2002). Such complex spatial patterns might not be properly modelled using traditional two point spatial statistics such as variograms (Journel and Zhang 2006). Multiple-point geostatistical simulation (MPS) techniques capture spatial patterns from so-called training images or training data. Using higher order

statistics makes the MPS algorithms capable of reproducing complex spatial patterns. However, large and representative training images or training data with desirable resolution are needed to model the spatial uncertainties properly. Many MPS algorithms have been developed in the recent years, however few of them are capable of running co-simulation of mixed data (Mariethoz et al. 2010; Peredo and Ortiz 2011; van den Boogaart et al. 2018).

A spatial predictive model was developed by van den Boogaart et al. (2018) which combines a multipoint geostatistical algorithm with a new form of distributional regression to estimate conditional distributions. The algorithm is capable of jointly simulating dependent spatial variables from various sample spaces (e.g. compositional, distributional, geometrical, and categorical). However, computational effort is substantial. The algorithm needs further development to simulate large mineral deposits or petroleum reservoirs. MPS algorithms for joint simulation of compositional and categorical data need to be developed or adapted which are easy to implement and fast enough to simulate many dependent variables on large simulation grids. Among the MPS techniques, the Direct Sampling (DS) technique (Mariethoz et al. 2010) is well suited to the co-simulation of mixed data since an explicit estimation of a model of co-dependence is not required, multivariate spatial patterns of different sizes are captured without the need to define a search template of specific size and geometry, and spatial patterns of different scales are captured without the need for a multigrid search strategy. However, DS is a distance based technique and requires measuring the distance between the spatial data events, which is problematic in the case of compositional data. Distances should not be measured from the original compositional data (data in form of proportions, percentages, probabilities, frequencies, and concentrations). The lack of sub-compositional coherence of Euclidean distances (Pawlowsky-Glahn et al. 2015) and the fact that these distances are massively dominated by the major components of the system (while the component of interest might be one of the small components) are some of the reasons why DS should not be implemented on the original compositional data, but on suitably transformed data. Other metrics for measuring the distance between spatial compositional patterns should be developed and implemented (such as Aitchison distance) or compositional data should be transformed to real space via an isometric transformation prior to

simulation via DS. New metrics should be defined to assess the performances of DS to simulate the compositional random function and spatial compositional patterns.

1.2.4 Application of machine learning algorithm for compositional data modelling

Over the past few years, many studies have involved the use of machine learning algorithms (MLAs) to explore the compositional patterns as footprints for geoscience process discovery analysis (Caritat et al. 2017; Carranza 2017; Grunsky et al. 2017; Grunsky et al. 2014; McKinley et al. 2018; Tolosana-Delgado et al. 2015a; Tolosana-Delgado and van den Boogaart 2014). Few of these studies have addressed the spatial correlations between geospatial data and the associated spatial uncertainty. Most of the machine learning algorithms are non-spatial techniques, which means they do not consider the multivariate spatial relationships between variables. As a result, the probability maps generated via MLAs cannot be accepted as the model of spatial uncertainty. In geostatistics, spatial relationships are taken into account via means such as second order ((cross)variograms) and/or higher order statistics (training images). In many applications of MLAs for spatial data, uncertainty associated with the input spatial data is ignored. However, this uncertainty can be incorporated into the machine learning algorithms by combining these non-spatial learners with geostatistical simulation. Each realization of random function can be used as an input (new observation) to a trained classifier. Ensemble classifiers which combine many simple learners (e.g. built from bootstrap samples) are preferable due to their stability, better predictive performance, ability to measure the performance and to select the most significant features internally (Breiman 1996). The estimated probabilities of different classes (e.g. rock or soil type as a categorical response variable) for all geostatistical realizations should be combined afterward. Such combination integrates elements of statistical and spatial uncertainties. However, care should be taken when combining these estimated probabilities to avoid any systematic bias. The new combined spatial uncertainty model can be used further to predict most probable classes. The proposed algorithm should address the compositional nature of data. Due to the high-dimensional

characteristics of compositional features (log-contrasts), developing a compositionally compliant feature selection will be useful for geoscience process discoveries.

1.3 Research objectives

The main objective of this research is to develop approaches for the joint modelling of regionalized compositional and categorical data. This study aims to address the following objectives:

1. To adapt the implementation of the direct sampling technique for the joint simulation of compositional and categorical data, and to introduce new metrics to evaluate the simulated compositional random function.

2. To develop a spatial decorrelation technique for joint two-point geostatistical simulation of high-dimensional continuous and categorical data. The compositional nature of some multivariate continuous variables will be addressed properly within the proposed algorithm.

3. To assess the capability of geostatistical simulation of complex regionalized compositional data via orthogonal projection (isometric log-ratio transformation) and flow anamorphosis. Effects of a dependent regionalized categorical variable on the predicted compositions will be assessed.

4. To adapt the implementation of machine learning algorithms (non-spatial ensemble classifiers in this study) to address the spatial uncertainty of input data. This will be achieved by combining the non-spatial classifiers (e.g. random forest) with geostatistical simulation. The estimated probabilities for several realizations of random function will be combined to integrate elements of statistical and spatial uncertainties. The new model of spatial uncertainty will be used further for prediction of various classes (e.g. rock or soil type as a categorical response variable). A coherent compositional feature selection will be introduced. The compositional nature of data will be addressed properly within all steps of proposed technique.

5. To assess the capability and performance of the developed techniques via implementing on real geoscientific case studies.

For situations where large and representative training images are available, multiple-point geostatistical methods are preferable. Due to the complexity of multivariate mixed and constrained geospatial data, the implementation of direct sampling technique is adapted for joint simulation of compositional and categorical data. The applicability and usefulness of the proposed algorithm is tested on one synthetic and one real case study.

The second objective of this PhD research is to develop a spatial decorrelation technique for joint simulation of high-dimensional continuous and categorical data. This method is appropriate for modelling projects where large and representative training images with proper resolution are not available. Along with generating predictions, the spatial uncertainty of regionalized continuous and categorical variables will be evaluated. The compositional nature of some multivariate continuous variables will be considered. The new method will be tested on a real mining case study.

Advanced geostatistical simulation of compositional data via orthogonal projection (isometric log-ratio transformation) and flow anamorphosis will be investigated. Ability of such algorithm to reproduce complex patterns such as presence of outliers, multi-population characteristic, and nonlinearity will be assessed. Multi-population characteristic and/or non-stationarity phenomenon might be related to a dependent categorical variable (e.g. geological units). Implementing such advanced geostatistical simulation technique may make the need for domaining and/or handling of non-stationarity unnecessary in some applications and situations. Effects of a dependent regionalized categorical variable on the whole process of spatial simulation of compositions will be investigated. The new method will be tested on a real mining case study.

Finally, to utilize the capability of machine learning algorithms to explore complex multivariate patterns and to select and rank features in a spatial framework, a hybrid spatial predictive model is developed based on the combined use of advanced geostatistical simulation techniques and machine learning algorithms. The spatial uncertainty of input compositional data is fully addressed. The new combined

spatial uncertainty model is used further for class prediction. A fully compositional feature selection is introduced. The developed hybrid model is used for surficial and deep earth material prediction through two real case studies.

1.4 Thesis structure

This thesis is presented and organised as "Thesis with publication" format[1]; and is structured in chapters as follows:

Chapter 1 presents the background of this PhD research and literature overview. The objectives of the research and the structures of the thesis are discussed in separate subsections.

Chapter 2 presents the developed method for joint simulation of compositional and categorical data via the direct sampling technique. The potential of the developed algorithm to improve mineral resource confidence is explored via one synthetic and one real mining case study.

Chapter 3 introduces a hybrid model for joint simulation of high-dimensional continuous and categorical variables in two-point geostatistical framework. The model is tested on a real mining case study.

Chapter 4 explores various approaches to geostatistical simulation of regionalized compositions consisting of several populations. Applications of such techniques to mineral resource evaluation are investigated.

Chapter 5 introduces a new workflow for implementation of a spatial predictive model (a hybrid of geostatistical simulation and machine learning). The potential of the new model is investigated through its application to surficial and deep earth material prediction from geochemical compositions.

[1] "Thesis with publication" format is an acceptable format of thesis for postgraduate research at ECU policy. The current thesis has been written based on the guideline provided at http://www.ecu.edu.au/GPPS/policies_db/policies_view.php?rec_id=0000000434. In this format, the submitted thesis can consist of publications that have already been published, are in the process of being published, or a combination of these.

Chapter 6 presents the general discussions on the developed techniques. Pros and cons of the developed techniques and the area of their application are discussed in this chapter.

Chapter 7 covers the overall conclusions of this PhD thesis and further recommendations.

1.5 Chapter references

Aitchison J (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 44, 139-177.

Aitchison J (1986). *The statistical analysis of compositional data*. London, UK, Chapman & Hall Ltd.

Alabert FG, Massonnat GJ (1990). Heterogeneity in a complex turbiditic reservoir: Stochastic modelling of facies and petrophysical variability. *SPE Annual Technical Conference and Exhibition*, https://doi.org/10.2118/20604-MS.

Armstrong M et al. (2011). *Plurigaussian Simulations in Geosciences*. Berlin Heidelberg, Berlin, Heidelberg, Springer, https://doi:10.1007/978-3-642-19607-2_3.

Bahar A, Kelkar M (2000). Journey from well logs/cores to integrated geological and petrophysical properties simulation: A methodology and application. *Society of Petroleum Engineers*, https://doi.org/10.2118/66284-PA.

Bandarian EM, Bloom LM, Mueller U (2008). Direct minimum/maximum autocorrelation factors within the framework of a two structure linear model of coregionalisation. *Computers&Geosciences*, 34, 190-200, https://doi.org/10.1016/j.cageo.2007.03.015.

Barnett RM, Deutsch CV (2012). Practical implementation of non-linear transforms for modeling geometallurgical variables. In: Abrahamsen P, Hauge R, Kolbjørnsen O (eds), *Geostatistics Oslo 2012*, vol 17. Quantitative Geology and Geostatistics. Netherlands, Springer, pp 409-422, https://doi:10.1007/978-94-007-4153-9_33.

Barnett RM, Manchuk JG, Deutsch CV (2014). Projection pursuit multivariate transform. *Mathematical Geosciences*, 46, 337-359, https://doi:10.1007/s11004-013-9497-7.

Boucher A, Dimitrakopoulos R (2012). Multivariate block-support simulation of the Yandi iron ore deposit, Western Australia. *Mathematical Geosciences*, 44, 449-468, https://doi:10.1007/s11004-012-9402-9.

Breiman L (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Buccianti A, Grunsky E (2014). Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes? *Journal of Geochemical Exploration*, 141, 1-5 https://doi.org/10.1016/j.gexplo.2014.03.022.

Cáceres A, Emery X (2010). Conditional co-simulation of copper grades and lithofacies in the Rio Blanco-Los Bronces copper deposit. In: Castro R, Emery X, Kuyvenhoven R (eds), *Proceedings of the IV international*

conference on mining innovation MININ 2010, Santiago, Chile, Gecamin Ltd, pp 311–320.

Caritat P de, Main PT, Grunsky EC, Mann AW (2017). Recognition of geochemical footprints of mineral systems in the regolith at regional to continental scales. *Australian Journal of Earth Sciences*, 64, 1033-1043, https://doi:10.1080/08120099.2017.1259184.

Carranza EJM (2017). Natural Resources Research publications on geochemical anomaly and mineral potential mapping, and introduction to the special issue of papers in these fields. *Natural Resources Research*, 26, 379-410, https://doi:10.1007/s11053-017-9348-1.

Chilès, J. P., Delfiner, P. (2012). *Geostatistics: modeling spatial uncertainty*. New York: Wiley.

Davis M (1987). Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology*, 19, 91-98, https://doi:10.1007/BF00898189.

Desbarats AJ, Dimitrakopoulos R (2000). Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. *Mathematical Geology*, 32, 919-942, https://doi:10.1023/A:1007570402430.

Deutsch CV (2006). A sequential indicator simulation program for categorical variables with point and block data: BlockSIS. *Computers & Geosciences*, 32, 1669-1681, http://dx.doi.org/10.1016/j.cageo.2006.03.005.

Deutsch CV, Journel AG (1998). *GSLIB: Geostatistical software library and user's guide*. New York, Oxford University Press.

Dowd PA (1986). Geometrical and geological controls in geostatical estimation and orebody modelling. *Proceedings of the 19th APCOM Conference: Society of Mining Engineers*, Inc., Littletown Colorado, p 81–89.

Dowd PA (1994). Geological controls in the geostatistical simulation of hydrocarbon reservoirs. *Arabian Journal for Science and Engineering*, 19, 237-247.

Dowd PA (1997). Structural controls in the geostatistical simulation of mineral deposits. Paper presented at the Baafi EY, Schofield NA (Eds.), *Geostatistics Wollongong*, Kluwer Academic, Dordrecht.

Dubrule O (1993). Introducing more geology in stochastic reservoir modelling. In: Soares A (ed), *Geostatistics Tróia*, vol 5. Quantitative Geology and Geostatistics. Springer Netherlands, pp 351-369.

Duke JH, Hanna PJ (2001). Geological interpretation for resource modelling and estimation. Monograph Series, *Australasian Institute of Mining and Metallurgy*, 23, 147-156.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279-300.

Emery X, González KE (2007a). Incorporating the uncertainty in geological boundaries into mineral resources evaluation. *J Geol Soc India*, 69(1), 29–38.

Emery X, González KE (2007b). Probabilistic modelling of lithological domains and it application to resource evaluation. *Journal of the Southern African Institute of Mining and Metallurgy*, 107, 803-809.

Emery X, Silva DA (2009). Conditional co-simulation of continuous and categorical variables for geostatistical applications. *Computers & Geosciences*, 35, 1234-1246.

Freulon XD, Fouquet C, Rivoirard J (1990). Simulation of the geometry and grades of a uranium deposit using a geological variable. In: *International Symposium on Applications of Computers and Operations Research in the Mineral Industry*, Technische Universität Berlin, Berlin, pp 649–659.

Galli A, Beucher H, Le Loc'h G, Doligez B, Group H (1994). The pros and cons of the truncated gaussian method. In: *Geostatistical Simulations*, vol 7. Quantitative Geology and Geostatistics. Netherlands, Springer, pp 217-233.

Goovaerts P (1997). *Geostatistics for Natural Resources Evaluation*. Applied Geostatistics. Oxford University Press.

Goulard M, Voltz M (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24, 269-286.

Grunsky EC, de Caritat P, Mueller UA (2017). Using surface regolith geochemistry to map the major crustal blocks of the Australian continent. *Gondwana Research*, 46, 227-239.

Grunsky EC, Mueller UA, Corrigan D (2014). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping. *Journal of Geochemical Exploration*, 141, 15-41.

Guardiano FB, Srivastava RM (1993). Multivariate Geostatistics: beyond bivariate moments. In: Soares A (ed), *Geostatistics Tróia* '92: Volume 1. Netherlands, Springer, Dordrecht, pp 133-144.

Isaaks EH, Srivastava RM (1989). *An introduction to applied geostatistics*. Oxford University Press.

Jones P, Douglas I, Jewbali A (2013). Modeling combined geological and grade uncertainty: Application of multiple-point simulation at the Apensu gold deposit, Ghana. *Mathematical Geosciences*, 45, 949-965.

Journel AG, Zhang T (2006). The necessity of a multiple-point prior model. *Mathematical Geology*, 38, 591-610.

Journel AG, Alabert FG (1990). New method for reservoir mapping. SPE-18324-PA, 42, 2012-2218.

Kim HM, Mallick BK, Holmes CC (2005). Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association*, 100, 653-668.

Lantuéjoul C (2002). *Geostatistical simulation.* Springer-Verlag Berlin Heidelberg.

Larrondo P, Leuangthong O, Deutsch CV (2004). Grade estimation in multiple rock types using a linear model of coregionalization for soft boundaries. Paper presented at the Proceedings of the 1st international conference on mining innovation. Gecamin Ltd, Santiago, Chile.

Leuangthong O, Deutsch CV (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology*, 35, 155-173.

Maleki M, Emery X (2015). Joint simulation of grade and rock type in a stratabound copper deposit. *Mathematical Geosciences*, 47, 471-495.

Maleki M, Emery X (2017). Joint simulation of stationary grade and non-stationary rock type for quantifying geological uncertainty in a copper deposit. *Computers & Geosciences*, 109, 258-267.

Marcotte D (2012). Revisiting the linear model of coregionalization. In: *Geostatistics Oslo, Quantitative Geology and Geostatistics.* Springer Netherlands, pp 67-78.

Marcotte D (2015). TASC3D: A program to test the admissibility in 3D of non-linear models of coregionalization. *Computers & Geosciences*, 83, 168-175.

Mariethoz G, Caers J (2015). *Multiple-Point geostatistics: Stochastic modeling with training images.* John Wiley & Sons, Ltd.

Mariethoz G, Renard P, Straubhaar J (2010). The direct sampling method to perform multiple-point geostatistical simulations. Water Resources Research, https://doi:10.1029/2008WR007621.

Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., (2011). The principle of working on coordinates. In: Pawlowsky-Glahn V, and Buccianti A (eds.), *Compositional Data Analysis*. https://doi.org/10.1002/9781119976462.ch3.

Matheron G, Beucher H, de Fouquet C, Galli A, Guerillot D, Ravenne C (1987). Conditional simulation of the geometry of Fluvio-Deltaic reservoirs. *Society of Petroleum Engineers*, https://doi.org/10.2118/16753-MS.

McKinley JM, Grunsky E, Mueller U (2018). Environmental monitoring and peat assessment using multivariate analysis of regional-scale geochemical data. *Mathematical Geosciences*, 50, 235-246.

McKinley JM et al. (2016). The single component geochemical map: Fact or fiction? *Journal of Geochemical Exploration*, 162, 16-28.

Mueller U, Ferreira J (2012). The U-WEDGE transformation method for multivariate geostatistical simulation. *Mathematical Geosciences*, 44, 427-448.

Mueller U, Tolosana-Delgado R, van den Boogaart KG (2014). Approaches to the simulation of compositional data – a nickel-laterite comparative case study. Paper presented at the Orebody Modelling and Strategic Mine Planning Symposium, Melbourne.

Mueller U, van den Boogaart KG, Tolosana-Delgado R (2017). A truly multivariate normal score transform based on lagrangian flow. In: Gómez-Hernández JJ, Rodrigo-Ilarri J, Rodrigo-Clavero ME, Cassiraga E, Vargas-Guzmán JA (eds.), *Geostatistics Valencia 2016*. Springer International Publishing, Cham, pp 107-118.

Ortiz JM, Emery X (2006). Geostatistical estimation of mineral resources with soft geological boundaries: a comparative study. *The South African Institute of Mining and Metallurgy*, 106, 577–584.

Pawlowsky-Glahn V, Egozcue JJ (2016). Spatial analysis of compositional data: A historical review. *Journal of Geochemical Exploration*, 164, 28-32.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modelling and analysis of compositional data*. John Wiley & Sons, Ltd.

Pawlowsky-Glahn V, Olea RA (2004). Geostatistical analysis of compositional data. Oxford University Press.

Pawlowsky-Glahn V, Buccianti A (2011). *Compositional data analysis: Theory and applications.* John Wiley & Sons, Ltd.

Pawlowsky-Glahn V (1984). On spurious spatial covariance between variables of constant sum. *Sci de la Terre Inf Geol*, 21, 107-113.

Pawlowsky-Glahn V, Burger H (1992). Spatial structure analysis of regionalized compositions. *Mathematical Geology*, 24, 675-691.

Peredo O, Ortiz JM (2011). Parallel implementation of simulated annealing to reproduce multiple-point statistics. *Computers & Geosciences*, 37, 1110-1121.

Roldão D, Ribeiro D, Cunha E, Noronha R, Madsen A, Masetti L (2012). Combined use of lithological and grade simulations for risk analysis in iron ore, Brazil. In: Abrahamsen P, Hauge R, Kolbjørnsen O (eds.), *Geostatistics Oslo 2012*, Quantitative Geology and Geostatistics. Springer Netherlands, pp 423-434.

Rondon O (2012). Teaching aid: minimum/maximum autocorrelation factors for joint simulation of attributes. *Mathematical Geosciences*, 44, 469-504.

Rossi ME, Deutsch CV (2014). *Mineral resource estimation*. Springer, Dordrecht.

Sinclair AJ (1998). Geological controls in resource/reserve estimation. *Exploration and Mining Geology*, 7, 29-44.

Sinclair AJ, Blackwell GH (2002). *Applied mineral inventory estimation*. Cambridge University Press.

Strebelle S (2000). Sequential simulation drawing structures from training images. Ph.D. Thesis, Stanford University.

Strebelle S (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 34, 1-21.

Talebi H, Asghari O, Emery X (2015). Stochastic rock type modeling in a porphyry copper deposit and its application to copper grade evaluation. *Journal of Geochemical Exploration*, 157, 162-168.

Talebi H, Hosseinzadeh Sabeti E, Azadi M, Emery X (2016). Risk quantification with combined use of lithological and grade simulations: Application to a porphyry copper deposit. *Ore Geology Reviews*, 75, 42-51.

Talebi H, Lo J, Mueller U (2017). A hybrid model for joint simulation of high-dimensional continuous and categorical variables. In: Gómez-Hernández JJ, Rodrigo-Ilarri J, Rodrigo-Clavero ME, Cassiraga E, Vargas-Guzmán JA (eds.), *Geostatistics Valencia 2016*. Springer International Publishing, Cham, pp 415-430.

Tolosana-Delgado R (2006). Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring. PhD thesis, University of Girona, Girona (Spain).

Tolosana-Delgado R, McKinley J (2016). Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). *Applied Geochemistry*, 75, 263-276.

Tolosana-Delgado R, McKinley J, van den Boogaart KG (2015). Geostatistical fisher discriminant analysis. Paper presented at the 17th annual conference of the International Association for Mathematical Geosciences, Freiberg (Saxony) Germany.

Tolosana-Delgado R, Mueller U, van den Boogaart KG (2016). Compositionally compliant contact analysis. In: Raju NJ (ed.), Geostatistical and Geospatial Approaches for the Characterization of Natural Resources in the Environment: Challenges, Processes and Strategies. Springer International Publishing, Cham, pp 11-14.

Tolosana-Delgado R, Mueller U, van den Boogaart KG, Ward C, Gutzmer J (2015). Improving processing by adaption to conditional geostatistical simulation of block compositions. *Journal of the Southern African Institute of Mining and Metallurgy*, 115, 13-26.

Tolosana-Delgado R, van den Boogaart KG (2014). Towards compositional geochemical potential mapping. J*ournal of Geochemical Exploration*, 141, 42-51.

Tolosana-Delgado R, van den Boogaart KG (2013). Joint consistent mapping of high-dimensional geochemical surveys. *Mathematical Geosciences*, 45, 983-1004.

van den Boogaart KG, Mueller U, Tolosana-Delgado R (2017). An affine equivariant multivariate normal score transform for compositional data. *Mathematical Geosciences*, 49, 231-251.

van den Boogaart KG, Tolosana-Delgado R (2013). *Analyzing compositional data with R.* Springer-Verlag Berlin Heidelberg.

van den Boogaart KG., Tolosana-Delgado R, Lehmann M, Mueller U (2018). On the joint multi point simulation of discrete and continuous geometallurgical parameters. In: Dimitrakopoulos R. (eds) Advances in Applied Strategic Mine Planning. Springer, Cham

Vargas-Guzmán JA, Dimitrakopoulos R (2003). Computational properties of min/max autocorrelation factors. *Computers & Geosciences*, 29, 715-723.

Wackernagel H (2003). *Multivariate geostatistics.* Springer-Verlag Berlin Heidelberg.

Wilde BJ, Deutsch CV (2012). Kriging and simulation in presence of stationary domains: Developments in boundary modeling. In: Abrahamsen P, Hauge R, Kolbjørnsen O (eds.), *Geostatistics Oslo 2012*, Quantitative Geology and Geostatistics. Springer Netherlands, pp 289-300.

Chapters 2, 3 and 4 are not included in this version of the thesis.

Chapter 3 has been published as:
Talebi H, Lo J, and Mueller U (2017). A hybrid model for joint simulation of high-dimensional continuous and categorical variables. In: J.J. Gómez-Hernández, J. Rodrigo-Ilarri, M.E. Rodrigo-Clavero, E. Cassiraga and J.A. Vargas-Guzmán (Editors), *Geostatistics Valencia 2016*. Springer International Publishing, Cham, pp. 415-430. https://doi.org/10.1007/978-3-319-46819-8_28

Chapter 4 has been published as:
Talebi H, Mueller U, Tolosana-Delgado R, van den Boogaart K G (2018). Geostatistical simulation of geochemical compositions in the presence of multiple geological units - Application to mineral resource evaluation. *Mathematical Geosciences*, 51: 129. https://doi.org/10.1007/s11004-018-9763-9

Chapter 5 has been published as:
Talebi, H., Mueller, U., Tolosana-Delgado, R., Grunsky, E. C., McKinley, J. M., & de Caritat, P. (2019). Surficial and deep earth material prediction from geochemical compositions. *Natural Resources Research*, *28*(3), 869-891. https://doi.org/10.1007/s11053-018-9423-2

Chapter 5

Surficial and deep earth material prediction from geochemical compositions - a spatial predictive model[1]

Abstract

Prediction of true classes of surficial and deep earth materials using multivariate geospatial data is a common challenge for geoscience modellers. Most geological processes leave a footprint that can be explored by geochemical data analysis. These footprints are normally complex statistical and spatial patterns buried deep in the high-dimensional compositional space. This paper proposes a spatial predictive model for classification of surficial and deep earth materials derived from the geochemical composition of surface regolith. The model is based on a combination of geostatistical simulation and machine learning approaches. A random forest predictive model is trained and features are ranked based on their contribution to the predictive model. To generate potential and uncertainty maps, compositional data are simulated at unsampled locations via a chain of transformations (isometric log-ratio transformation followed by the flow anamorphosis) and geostatistical simulation. The simulated results are subsequently back-transformed to the original compositional space. The trained predictive model is used to estimate the probability of classes for simulated compositions. The proposed approach is illustrated through two case studies. In the first case study the major crustal blocks of the Australian continent are predicted from the surface regolith geochemistry of the National Geochemical Survey of Australia project. The aim of the second case study is to discover the superficial deposits (peat) from the regional-scale soil geochemical data of the Tellus project. The accuracy of the results in these two case studies confirms the usefulness of the proposed method for geological class prediction and geological process discovery.

---

[1] This chapter has been submitted for publication as a full research paper in:
Talebi, H., Mueller, U., Tolosana-Delgado, R., Grunsky, E.C., McKinley, J.M., Caritat, P.D., 2018. Surficial and Deep Earth Material Prediction from Geochemical Compositions - a Spatial Predictive Model, Natural Resources Research, (In review).
Whilst efforts were made to retain original content of the article, minor changes such as number formats, font size and style were implemented in order to maintain consistency in the formatting style of the thesis.

Keywords: Compositional data, Log-ratio, flow anamorphosis, geostatistical simulation, machine learning

5.1 Introduction

Surficial and deep earth materials are normally representatives of several classes with different characteristics. Tectonic, lithological and alteration units, soil types, vegetation classes, plant species, and land uses are examples of such classes. Spatial maps of these classes and their associated uncertainties are vital components in current strategies for managing projects such as mineral exploration, animal and human health, environmental and ecological planning, efficient management of water resources, geo-hazard risk assessment, agriculture and sustainable food production. Class prediction and spatial uncertainty modelling using multivariate geospatial data are a common challenge for geoscience modellers. Mechanisms behind geological systems can be explained partly by geochemical data and methods (Buccianti and Grunsky 2014; Caritat et al. 2017; Grunsky et al. 2017; Grunsky et al. 2014; Harris and Grunsky 2015; McKinley 2015; McKinley et al. 2018; Tolosana-Delgado and McKinley 2016; Tolosana-Delgado and van den Boogaart 2014). Spatial or spatiotemporal geoscientific entities such as climate zones, ecosystems, landforms, and surface and subsurface geology are related to geochemistry derived from surface and near-surface materials (Drew et al. 2010; Grunsky et al. 2017; Grunsky et al. 2013; McKinley 2015; McKinley et al. 2018). Over the last decade, geochemical surveys at different scales (e.g. regional, national, transnational, and continent scales) have become widely available. These geochemical surveys normally constitute “big data” of high-dimensionality making the statistical and spatial analyses challenging (Grunsky 2010; Grunsky et al. 2014; Tolosana-Delgado and McKinley 2016). Most geological processes leave some sort of footprint that can be explored by advanced geochemical data analysis. These footprints are complex multivariate statistical and/or spatial patterns hidden deep in the geochemical compositional space. Advanced statistical and/or spatial compositional data analysis should be implemented to explore these patterns. Geochemical data are inherently compositional in nature, presenting several challenges for spatial predictive models (Pawlowsky-Glahn and Egozcue 2016;

Pawlowsky-Glahn and Olea 2004; Tolosana-Delgado 2006; Tolosana-Delgado and van den Boogaart 2013; van den Boogaart and Tolosana-Delgado 2013). Compositional data are multivariate, non-negative values that represent the abundance of some parts of a whole. In such data, the constant sum constraint forces at least one covariance to be negative and induces spurious statistical and spatial correlations and patterns. Furthermore, these data carry just relative information (Aitchison 1986) and interpretations are necessarily multivariate, dependent on all components. To transform compositional data into an unbounded space and to increase mathematical tractability, different log-ratio transformations (Aitchison 1986; Pawlowsky-Glahn and Olea 2004; Tolosana-Delgado 2006) can be applied prior to using standard (geo)statistical techniques. A geochemical survey normally produces thousands of samples and dozens of variables (log-ratios) and as such are practically impossible to effectively visualise and interpret without the assistance of computers and statistical tools. In addition, the underlying geological processes most of the time are obscure and difficult to understand. In such situations machine learning algorithms (MLAs) have been shown to perform well in the prediction of classes from spatially dispersed data and discovering the underlying geological processes (Harris and Grunsky 2015; Kanevski et al. 2009). However, MLAs are typically not spatially predictive algorithms, which means that they do not consider the multivariate spatial relationships between features. As a result, the probability maps generated via MLAs cannot be accepted as the model of spatial uncertainty. In a geostatistical treatment spatial relationships are taken into account via means such as second order ((cross)variograms) and/or higher order statistics (training images). To address this limitation of MLAs an alternative solution is proposed in this study based on the combined use of advanced multivariate geostatistical simulation and MLAs.

The proposed spatial compositional predictive model is twofold: spatial simulation of geochemical compositions at unsampled locations and class prediction for each simulated map via a trained random forest (RF) algorithm (Breiman 2001). Other spatial (Tolosana-Delgado et al. 2015) or non-spatial (Kuhn and Johnson 2013) predictive models can also be implemented, but RF is utilized in this study for its ease of implementation, robustness against over-fitting, ability to handle many types of predictors (sparse, skewed, continuous, categorical, etc.) without the need

to pre-process them, ability to handle missing data and to select the most relevant features (Kuhn and Johnson 2013). Once the spatial compositional vectors have been simulated in the study area, MLAs (RF in this study) can be implemented to predict the probability of occurrence of classes conditional to each realization of the compositional random function. To simulate the compositional random function at unsampled locations, the input geochemical compositions are transformed to real space via an isometric log-ratio (ilr, Egozcue et al., 2003) transformation. To avoid violating the assumption of multivariate multigaussianity of geostatistical simulation techniques (Chilès and Delfiner 2012), log-ratios are transformed to multivariate normal space via a flow anamorphosis (FA) algorithm (Mueller et al. 2017; van den Boogaart et al. 2017). FA is applied in this study because of its ability to reproduce complex patterns (e.g. presence of outliers, presence of several populations, nonlinearity, and heteroscedasticity) in the input data, its invariance property under the choice of log-ratio transformation, and its property of generating spatially orthogonal factors that makes geostatistical simulation straightforward. The turning bands (TB) algorithm (Emery 2008; Emery and Lantuéjoul 2006) is used to simulate the orthogonal factors at unsampled locations. Finally the simulated results are back-transformed to the original space to provide several simulated spatial maps of geochemical compositions. Based on the true classes for the input set, a random forest algorithm is trained using the generated features. The ability of RF to rank the features based on their contribution to the predictive model aids the discovery of underlying geological processes. Finally the trained RF is used to predict the probabilities of classes at unsampled locations using the simulated compositions. Minimum, expected, and maximum probability scenarios are defined for each class from simulated probabilities.

The objectives of this research is to introduce a new method to account for spatial uncertainty on classifiers based on a combination of geostatistical simulation and machine learning classification algorithms. The most probable geological classes are predicted out of geochemical survey data using the new model of spatial uncertainty. Finally, a compositional feature selection is introduced and implemented for geological process discovery studies.

The proposed approach is illustrated through two case studies. In the first case study surface regolith geochemistry data are used to predict the major crustal blocks of

the Australian continent. Discovering superficial peat deposits in Northern Ireland from regional-scale soil geochemical data is the aim of the second case study.

The organization of this paper is as follows: section 5.2.1 discusses the analysis of compositional data. Flow anamorphosis as a powerful technique for transforming input data to multivariate normal space is discussed in section 5.2.2. Section 5.2.3 presents the random forest predictive model and the recursive feature elimination with resampling technique. Steps of the proposed method for modelling spatial uncertainty are presented in section 5.2.4. Sections 5.3 and 5.4 present the implementation of the method and results and discussion for the two case studies. Finally, some conclusions and final thoughts are presented in section 5.5.

5.2 Methodology

5.2.1 Compositional data analysis

Compositions are multivariate data which components represent the relative contribution of some parts forming a whole. Typically, these non-negative components are measured on the same scale (proportions, percentages, ppm or ppb) and are constrained by a constant sum property. Regionalized compositions are consequently defined as follows:

$$\vec{Z}(u) = [z_1(u), z_2(u), \cdots, z_D(u)]; \begin{cases} z_i(u) \geq 0;\ i = 1,2,\cdots,D, u \in \boldsymbol{A} \\ \sum_{i=1}^{D} z_i(u) = m \end{cases}, \quad (5.1)$$

where $z_i(u)$ represents the ith component measured at location u within the study area $\boldsymbol{A}$ and m is the constant sum. Geochemical data are a typical example of compositional data. It is often the case that the data analysed do not add to the constant m, in which case an additional variable can be introduced, often called *filler* or *rest*, to ensure that the constant sum constraint is satisfied. Compositional data carry by definition relative information (Aitchison 1986) and the constant sum constraint is known to induce the problems of spurious statistical and spatial

correlations (Aitchison 1982; Pawlowsky-Glahn and Olea 2004). Constraints of positivity and constant sum and the spurious correlations can be appropriately addressed by implementing log-ratio transformations, for instance, making (geo)statistical treatment more amenable (Aitchison 1986; Pawlowsky-Glahn and Egozcue 2016; Pawlowsky-Glahn et al. 2015; van den Boogaart and Tolosana-Delgado 2013). Several families of log-ratio transformations exist in the literature. The pairwise log-ratio (pwlr), additive log-ratio (alr) and centred log-ratio (clr) transformations were introduced by Aitchison (1986) and the isometric log-ratio (ilr) transformation was proposed by Egozcue et al. (2003). The pairwise log-ratios are readily interpretable and are defined as follows:

$$\begin{aligned}\mathrm{pwlr}\left(\vec{Z}(u)\right) &= \begin{bmatrix} 0 & \ln\left(\frac{z_1(u)}{z_2(u)}\right) & \cdots & \ln\left(\frac{z_1(u)}{z_D(u)}\right) \\ \ln\left(\frac{z_2(u)}{z_1(u)}\right) & 0 & \dots & \ln\left(\frac{z_2(u)}{z_D(u)}\right) \\ \vdots & \vdots & & \\ \ln\left(\frac{z_D(u)}{z_1(u)}\right) & \ln\left(\frac{z_D(u)}{z_2(u)}\right) & \ddots & 0 \end{bmatrix} \\ &= \left[\xi_{ij}(u)\right].\end{aligned} \tag{5.2}$$

The additive log-ratios are one of the columns of the pwlr (ratios for which denominator is fixed to one of the components), typically the last one:

$$alr\left(\vec{Z}(u)\right) = \left[\ln\left(\frac{z_1(u)}{z_D(u)}\right), \ln\left(\frac{z_2(u)}{z_D(u)}\right), \dots, \ln\left(\frac{z_{D-1}(u)}{z_D(u)}\right)\right] = [\xi_{iD}(u)]. \tag{5.3}$$

The centred log-ratios present the logarithms of ratios of each component to the geometric mean of all components. They are obtained via the following formula:

$$\mathrm{clr}\left(\vec{Z}(u)\right) = \ln\left(\frac{\vec{Z}(u)}{\sqrt[D]{\prod_{i=1}^{D} z_i(u)}}\right). \tag{5.4}$$

Finally, the isometric log-ratio transformation is defined as follows:

$$\mathrm{ilr}\left(\vec{Z}(u)\right) = V \cdot \mathrm{clr}\left(\vec{Z}(u)\right), \tag{5.5}$$

where V is a $(D-1) \times D$ matrix whose columns are pairwise orthogonal vectors, each sums to zero. Each matrix V satisfying these conditions gives rise to an ilr transformation.

All the aforementioned log-ratio transformations are *log-contrasts*, that is: linear combinations of the components in log-scale with coefficients summing to zero:

$$\xi\,(u) = \sum_{i=1}^{D} \alpha_i \ln\left(z_i(u)\right);\ \sum_{i=1}^{D} \alpha_i = 0\ ;\ \alpha_i \in \mathbb{R} \tag{5.6}$$

Complex log-contrasts can be defined to discover the hidden underlying geological processes and classes. Many log-contrasts can be defined and the most appropriate ones depend on the aim of the analysis undertaken (McKinley et al. 2016; Pawlowsky-Glahn and Buccianti 2011).

5.2.2 Flow anamorphosis

As shown in the preceding section, compositional data do not have a unique, canonical representation and several log-ratio transformations are available. Invariance of the simulated results under the choice of log-ratio transform is thus highly desirable. This property is known as affine equivariance. Log-ratios are not commonly multivariate normal, so they have to be combined with a normal score transform prior to using geostatistical simulation techniques in order to not violate the assumption of multigaussianity of most of these simulation algorithms (Chilès and Delfiner 2012; Mueller et al. 2014). Conventional normal score transformations based on quantile matching are neither affine equivariant nor do provide multivariate normal transformed scores. The flow anamorphosis is a multivariate form of gaussian anamorphosis which is capable of transforming original multivariate data to multivariate normal space and at the same time is invariant under the choice of log-ratio transform (Mueller et al. 2017; van den Boogaart et al. 2017). The method continuously deforms a kernel density estimate of the given multivariate density of the observations into a standard multivariate normal

distribution. The transformation is dependent on the selection of the two parameters, σ_0 and σ_1 (initial and final spreads of the smoothing kernels of the kernel density estimates). Deformation of the underlying space is controlled by σ_0. Smaller values of σ_0 lead to a stronger deformation. The choice of a suitable value for σ_0 depends on the number of variables, sample size and complexity of the input data (Mueller et al. 2017). On the other hand, σ_1 controls the ranges of the transformed distributions. In our experience the FA-transformed data are not only multivariate normal but often also exhibit a lack of spatial cross-correlation, which makes the geostatistical simulation of such orthogonal factors straightforward (Mueller et al. 2017; van den Boogaart et al. 2017). The simulated results are subsequently back-transformed to the original space via FA^{-1}.

5.2.3 Random forest algorithm and feature selection

Tree-based classification models consist of several nested conditions on the predictors that partition the observations into purer subpopulations. Within these partitions, a model is used to predict the class of future observations. Tree-based models are very popular due to their ease of interpretation and implementation, their ability to handle many types of predictors (sparse, skewed, continuous, categorical, etc.) without the need to pre-process them, allow missing data and conduct feature selection (Kuhn and Johnson 2013). However, single decision trees are prone to instability, which means that slight changes in the input observations can drastically change the structure of the tree and, hence, the subsequent interpretations and predictions. Ensemble methods that combine many simple predictive models (e.g. built from bootstrap samples) into one predictive model have been developed to address this instability and have much better predictive performance (Breiman 1996). The other advantage of the ensemble models is that the predictive performance can be estimated internally, which correlates well with either cross-validation estimates or test set estimates. The left out observations from each bootstrap sample (called “out-of-bag”) are used to assess the predictive performance of each model in the ensemble. The average of the out-of-bag performance metrics can then be used to measure the overall predictive

performance of the entire ensemble. Algorithm 5.1 shows the processes of a general random forest algorithm (Breiman 2001), a well-known ensemble predictive model.

Algorithm 5.1 General RF algorithm

---

1. Select the number of trees in the forest $(\boldsymbol{t})$
2. **for** $\boldsymbol{i} = \mathbf{1}: \boldsymbol{t}$
3. Generate a bootstrap sample of the original observations
4. Train a decision tree on this sample
5. **for** each split in the tree
6. Randomly select a subset $(\boldsymbol{s} \ll \boldsymbol{R})$ of the predictors $(\boldsymbol{\xi_{r}}, \boldsymbol{r} = \mathbf{1}: \boldsymbol{R})$
7. Select the best predictor out of this subset and partition the observations
8. **end**
9. Build the ultimate tree without pruning
10. **end**

For each new observation each of the t trees in the forest is used to predict its class and the resulting t predictions are combined to give the forest prediction. The number of trees in the forest (t) and the number of randomly selected predictors for each split (s) are the most important parameters in the RF algorithm, which need to be tuned. It has been shown that the selection of a large t will not adversely affect the RF model and does not lead to over-fitting (Breiman 2001), however it increases the computational burden. Several experiments have shown that the random forest tuning parameter does not have a drastic effect on its accuracy (Kuhn and Johnson 2013). Several approaches have been proposed to quantify the importance of predictors in the RF model such as measuring the improvement of node purities for each predictor at each occurrence of that predictor across the whole forest and aggregating them to determine the overall importance. However, these approaches for measuring the importance of predictors are adversely affected by the correlations between predictors (Strobl et al. 2007).

Due to the high-dimensional characteristic of the *log-contrasts* (ξ) calculated from geochemical compositions, determining which subset of them should be included in a predictive model is a critical question. While decision trees are not affected by redundant predictors due to the built-in feature selection, RF shows a moderate degradation in its accuracy due to random selection of predictors for splitting (Kuhn and Johnson 2013). Given the potential negative impact of redundant information (collinearity within log-contrasts), there is a need to find a smaller subset of them by maximizing the predictive performance of the RF algorithm. Feature selection is primarily implemented for removing non-informative or redundant predictors from the model. Multiple predictive models (built from subsets s_i of significant predictors) are evaluated to find the optimal combination of predictors that maximizes model performance. A recursive feature elimination with resampling technique (Guyon et al. 2002; Kuhn and Johnson 2013) is used in this study to select the most informative subset of *log-contrasts* for the classification purpose. The final predictive model with the highest accuracy is built from this subset of significant predictors (Algorithm 5.2).

Algorithm 5.**2** Recursive feature elimination with resampling

1. **for** each iteration of resampling
2. Divide the input observations into training and test subsets via resampling
3. Build a predictive model on the training set using all the $\boldsymbol{R}$ predictors
4. Measure the model accuracy
5. Measure the rank of predictors
6. **for** each subset size $\boldsymbol{s_i}, \boldsymbol{i} = \mathbf{1}: \boldsymbol{S}$
7. Keep the $\boldsymbol{s_i}$ most important predictors
8. Build a predictive model on the training set using $\boldsymbol{s_i}$ predictors
9. Measure model performance on the test subset
10. **end**
11. **end**
12. Calculate the performance profile over the $\boldsymbol{s_i}$ using the test subsets
13. Determine the appropriate number of predictors
14. Determine the final ranks of predictors
15. Fit the final model based on the optimal $\boldsymbol{s_i}$ predictors using all the input observations

5.2.4 Spatial modelling of geological classes

To spatially predict geological classes from geochemical composition, the first step is to simulate the compositional random function at unsampled locations. Algorithm 5.3 shows the procedure of geostatistical simulation of regionalized compositions. In line 1 of this algorithm, any log-ratio transformation can be implemented as long as the selected anamorphosis is affine equivariant. An ilr transformation (Eq. 5.5) is used in this study for this purpose. After transforming the log-ratios to multivariate normal space via the FA algorithm, the spatially orthogonal multivariate normal scores are simulated at unsampled locations independently. In this study a turning bands algorithm will be used for this purpose (Emery et al. 2016; Emery and Lantuéjoul 2006). After generating L realizations of the

compositional random function, the expected spatial map of regionalized compositions is defined as follows:

$$\vec{Z}^*(u) = C\left[\left(\prod_{l=1}^{L} z_1^l\,(u)\right)^{1/L}, \left(\prod_{l=1}^{L} z_2^l\,(u)\right)^{1/L}, \cdots, \left(\prod_{l=1}^{L} z_D^l\,(u)\right)^{1/L}\right], \tag{5.7}$$

where C is the closure operator defined as:

$$C\left(\vec{Z}(u)\right) = \left[\frac{\left(\prod_{l=1}^{L} z_1^l\,(u)\right)^{1/L}}{\sum_{d=1}^{D}\left(\prod_{l=1}^{L} z_d^l\,(u)\right)^{1/L}}, \frac{\left(\prod_{l=1}^{L} z_2^l\,(u)\right)^{1/L}}{\sum_{d=1}^{D}\left(\prod_{l=1}^{L} z_d^l\,(u)\right)^{1/L}}, \cdots, \frac{\left(\prod_{l=1}^{L} z_D^l\,(u)\right)^{1/L}}{\sum_{d=1}^{D}\left(\prod_{l=1}^{L} z_d^l\,(u)\right)^{1/L}}\right]. \tag{5.8}$$

The conditional total compositional variation of the simulated composition at location u is given by:

$$\text{totvar}_{composition}\left(\vec{Z}(u)\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\text{var}\left(\ln\frac{z_i(u)}{z_j(u)}\right). \tag{5.9}$$

The map of the total compositional variations for the simulated compositions can be considered as a means to assess spatial uncertainty of the geochemical compositions. High values of this metric show the most uncertain areas (and vice versa) with respect to the simulated geochemical compositions.

Algorithm 5.**3** Geostatistical simulation of geochemical compositions

1. Transform the set of $\boldsymbol{D}$ closed components into a set of $\boldsymbol{D}-\mathbf{1}$ unbounded log-ratios, by means of a log-ratio transformation
2. Transform the log-ratios to multivariate normal space via an affine equivariant anamorphosis
3. Simulate the multivariate normal scores at unsampled locations via any geostatistical simulation technique
4. Transform the simulated results back to the original (compositional) space

The second step is to build a predictive model based on the input labelled observations (input geochemical compositions). For such a predictive model, the features consist of *log-contrasts* (ξ). To extract relevant compositional information, a combination of the knowledge-driven *log-contrasts* (based on a geochemical understanding of the processes under consideration) and established mathematical representations (e.g. pwlr and clr) can be used as the input features (McKinley et al. 2016). These features together with the associated classes (e.g. rock types, soil types, mineralized material, etc.) are used to train the RF predictive model (Algorithm 5.1). The significant log-contrasts are recognised and ordered based on their contributions to the predictive model via Algorithm 5.2. The selected *log-contrasts* (out of many) and their ranks are very useful for geological process discovery and interpretation. The same selected *log-contrasts* are calculated from the simulated compositions at unsampled locations. The trained RF is used to predict classes at these locations. For each location u and for each realization l of the compositional random function, RF generates a discrete prediction (geological classes $I^l(u)=k; k=1,\dots,K\ and\ l=1,\dots,L$) and a vector of probabilities $\vec{p}^l(u)=\left[p_1^l(u),p_2^l(u),\cdots,p_K^l(u)\right]$. However the local uncertainty of the discrete predictions is underestimated and should not be used for spatial classification purposes. As an example consider the information in the Table 5-1, where there are three geological classes ($k=1,2,3$) and at location u a compositional random function has been simulated 5 times ($l=1,\dots,5$). Running a predictive model on

these realizations (uncertain inputs) will generate different sets of probabilities. Although the probability of other classes occurring is non-zero for each realisation, the final decision for location u would be class 3 with zero uncertainty, which is not true. This example shows that the spatial uncertainty of geological classes generated by a predictive model might be misleading.

Table 5-1 Prediction with uncertain inputs

Realization number ($\boldsymbol{l}$)	$\boldsymbol{p_1(u)}$	$\boldsymbol{p_2(u)}$	$\boldsymbol{p_3(u)}$	Most probable class ($\boldsymbol{k}$)
1	0.10	0.20	0.70	3
2	0.15	0.25	0.60	3
3	0.05	0.30	0.65	3
4	0.10	0.25	0.65	3
5	0.15	0.30	0.55	3
Final decision for location $\boldsymbol{u}$				= 3

As a result, discrete predictions of RF for each realization of geochemical compositions should be ignored and predicted probabilities ($\vec{p}^{\,l}(u) = [p_1^l(u), p_2^l(u), \cdots, p_K^l(u)]$) should be treated as follows: For a location u the probability of occurrence of a specific class k varies from $min(p_k^l(u))$ to $max(p_k^l(u))$ while the vector of expected probabilities is defined as closure of the vector of geometric means of the probabilities for each class:

$$\vec{q}(u) = C\left[\left(\prod_{l=1}^{L} p_1^l\,(u)\right)^{1/L}, \left(\prod_{l=1}^{L} p_2^l\,(u)\right)^{1/L}, \cdots, \left(\prod_{l=1}^{L} p_K^l\,(u)\right)^{1/L}\right]. \qquad (5.10)$$

The expected spatial probability model $\vec{q}(u)$ combines the statistical uncertainty (e.g. bootstrapping in the RF model) and the spatial uncertainty (L realizations of the geostatistical model). For the example in Table 5-1 probability of class 1 varies

from $\min_{l=1,..,3}(p_1^l(u)) = 0.05$ to $\max_{l=1,...,3}(p_1^l(u)) = 0.15$ while the expected probability is 0.104 ($\vec{q}(u) = [0.104, 0.260, 0.636]$). The most probable class for location u should be defined from $\vec{q}(u)$ which is class 3 in this example. Finally, the conditional total variation of geological classes for a location u is given by:

$$\text{totvar}_{probability}(u) = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{K}\text{var}\left(\ln\frac{p_i(u)}{p_j(u)}\right) \tag{5.11}$$

High values of this metric show the most uncertain areas (and vice versa) with respect to the predicted geological classes.

5.3 Major crustal blocks prediction using surface regolith geochemistry

5.3.1 Dataset

In this first case study multi-element near-surface geochemical compositions from the National Geochemical Survey of Australia (NGSA) are used to predict the exposed to deeply buried major crustal blocks (MCBs) of the Australian continent. The NGSA is a uniform and internally consistent geochemical database, covering approximately 81% of the continent of Australia (Caritat and Cooper 2011; Caritat and Cooper 2016). The NGSA dataset consists of 4 subsets based on the sampling depth and grain size. In this study the focus is on the "total" analysis of the fine-grained fraction (<75 μm) of the top outlet sediment samples (0–10 cm depth) (for further detail please see Grunsky et al. (2017)). Figure 5-1a shows the map of the major MCBs over Australia, while the distribution of surface lithology and the geological regions of Australia are shown in Figure 5-1b. The NGSA sample site locations are shown as black dots on these maps. The MCBs, derived from the major boundaries in the Australian crust as interpreted from geophysical and geological data by Korsch and Doublier (2015, 2016), reflect distinct tectonic domains comprised of early Archean to recent Cenozoic igneous, metamorphic and sedimentary rock assemblages. The MCBs were numbered in order of decreasing size. Of the 30 MCBs derived from the crustal boundaries, 22 are used in the present analysis as explained in Grunsky et al. (2017). In the present contribution we introduce and implement a new method for modelling spatial uncertainty of

Australian MCBs based on surface regolith geochemistry and for predicting MCBs in areas lacking/between geochemical samples. The most important log-contrasts for distinguishing crustal blocks are introduced and mapped for further geological discovery analysis.

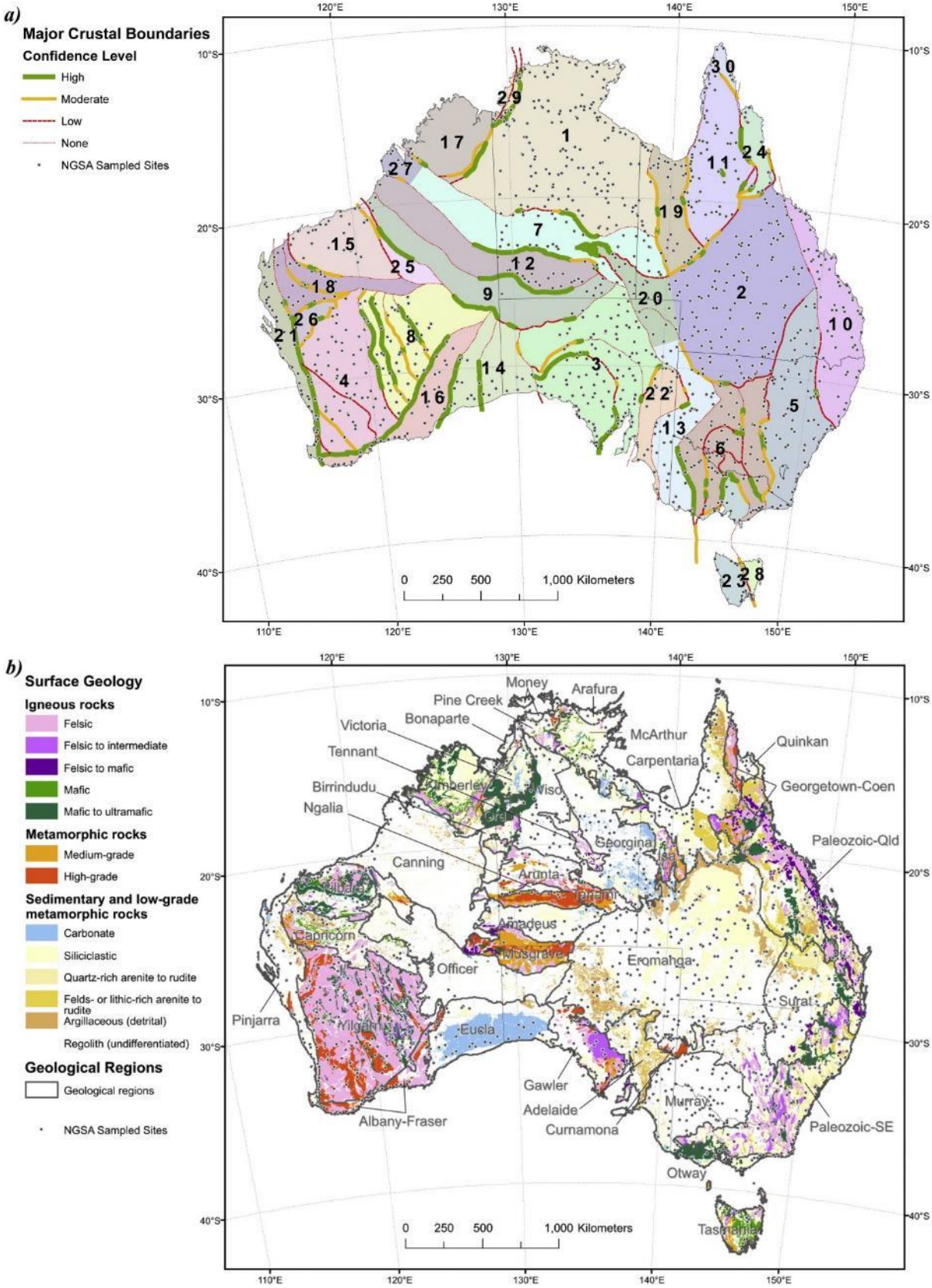

Figure 5-1 (a) Major crustal blocks of Australia (coloured and numbered). The line styles of the MCB boundaries reflect the confidence level in their position/existence (solid thick: high; solid thin: moderate; dashed: low; dot-dashed: none). (b) Surface geology and the geological regions of

Australia. The NGSA sample site locations are shown as black dots on both maps. Sources: Blake and Kilgour (1998), Caritat and Cooper (2011), Korsch and Doublier (2016), Nakamura and Milligan (2015), Raymond (2012). Modified after Grunsky et al. (2017)

5.3.2 Results and discussion

Input data (1067 compositional samples with 52 variables, 50 elements (Al, As, Au, Ba, Be, Bi, Ca, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, F, FeT, Ga, Gd, Ge, Hf, Ho, K, La, Lu, Mg, Mn, Na, Nb, Nd, Ni, P, Pb, Pr, Rb, Sc, Se, Si, Sm, Sn, Sr, Tb, Th, Ti, U, V, Y, Yb, Zn, Zr) plus LOI and *filler*) were transformed to real space via an ilr transformation (Eq. 5.5). As the ilr-transformed data were not multivariate normal, a transformation to normal space was needed prior to geostatistical simulation. The ilr-transformed scores were transformed to multivariate normal space via flow anamorphosis. Due to the complexity of the data and the number of variables, multivariate normality was not achieved by a single FA. Two successive FA with the same parameters ($\sigma_0 = 0.1$ and $\sigma_1 = 1.1$) were required to achieve multivariate normality. Spatial structural analysis (variography) showed further that the multivariate normal scores are spatially orthogonal, with Tercan's (Tercan 1999) $\bar{\tau}$ and $\bar{\kappa}$ equal to 0.0954 and 0.9073, respectively, so they could be simulated independently. The scores were simulated independently on a regular grid (25 km × 25 km) via a turning bands algorithm and back-transformed to compositions afterward. In total, 100 realizations of geochemical compositions were generated at unsampled locations. To illustrate the simulated model, the spatial distributions of three major elements (out of 52 jointly simulated variables), Ca, total Fe and Mg, are depicted in Figure 5-2. The expected maps were calculated via equation 5.7. Figure 5-3 shows the map of the conditional total compositional variations for the simulated compositions. This map can be considered as a means of assessing spatial uncertainty of the geochemical compositions. Close to sample locations where direct information is available variation is low, while in areas where no sample was taken, variation is high. Some MCBs generally show higher uncertainty than others, for instance MCB 6 shows less uncertainty than MCB 1 or southern parts of MCB 4 show higher uncertainty than its northern parts.

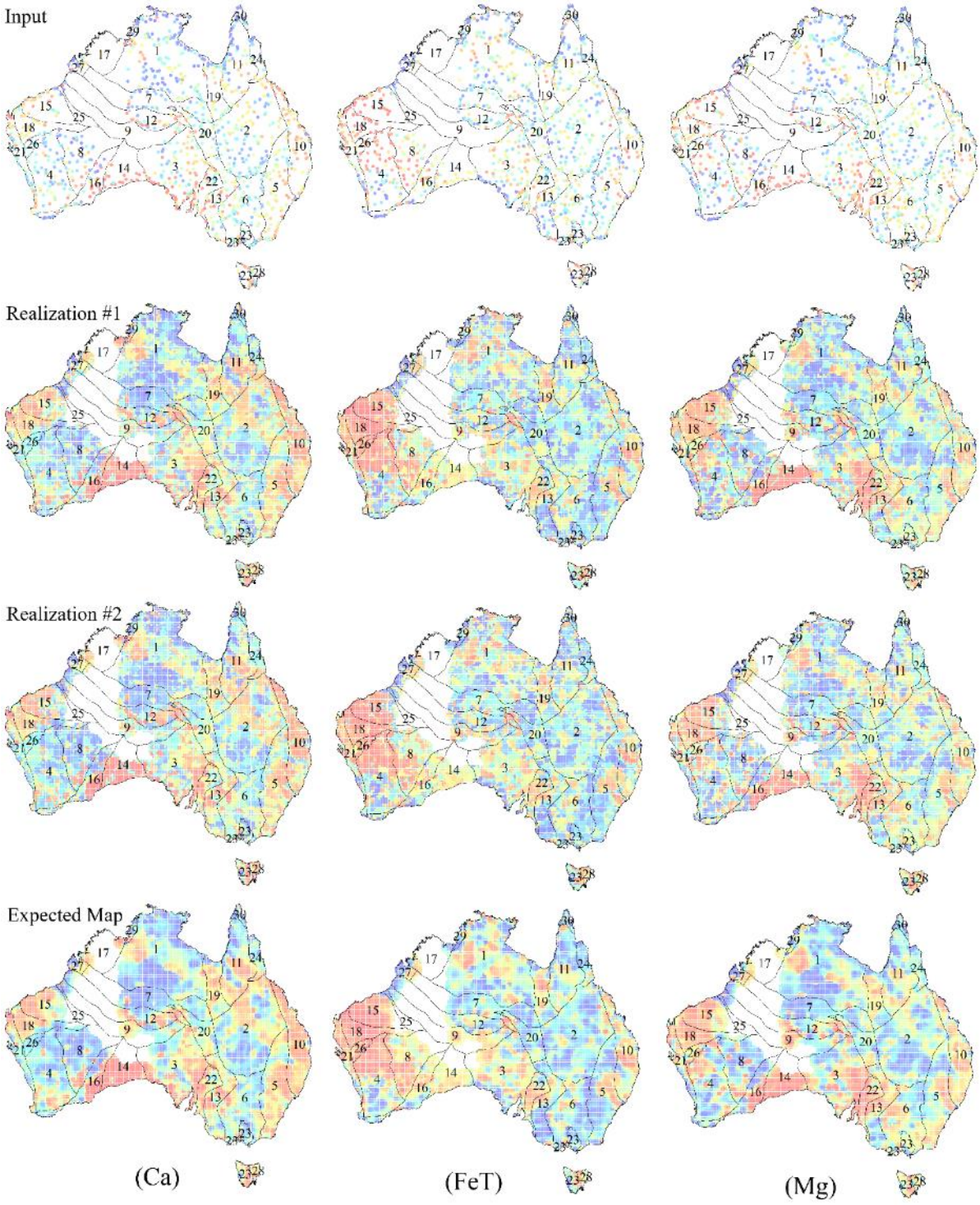

Figure 5-2 Input geochemical compositions, two realizations of the geostatistical simulation procedure and expected map for three major components Ca, total Fe and Mg (warm colours are associated with high values)

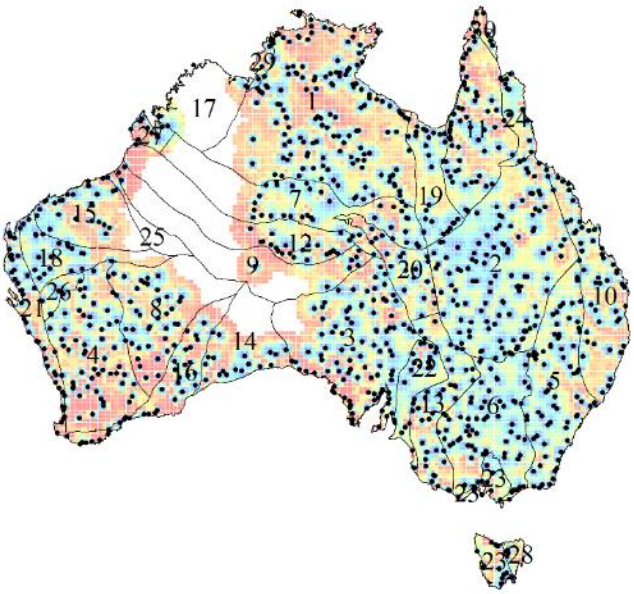

Figure 5-3 Conditional total compositional variation, a means to assess the spatial uncertainty of the geochemical compositions (warm colours are associated with high uncertainty and black dots are the location of samples)

The RF predictive model was trained based on the input labelled log-ratios. In this case only pairwise (1326 log-ratios) and centred log-ratios (52 log-ratios) were used as predictors and MCBs as the categorical response variable. Out of 30 MCBs, 8 were not considered due to an insufficient number of sample sites in each of these MCBs (Grunsky et al. 2017). Algorithm 5.2 was used to select the most informative subset of log-ratios for the classification purpose. The final predictive RF with the highest accuracy was associated with a subset of only 220 log-ratios (Figure 5-4). Figure 5-5 shows the top 30 (out of 220 selected log-ratios) most informative log-ratios for classification of MCBs. To determine the most significant log-ratios for discriminating a crustal block of interest from the remaining blocks, a binary response variable can be defined (e.g. 1 is the block of interest and 0 is all other blocks) and Algorithm 5.2 can be run again.

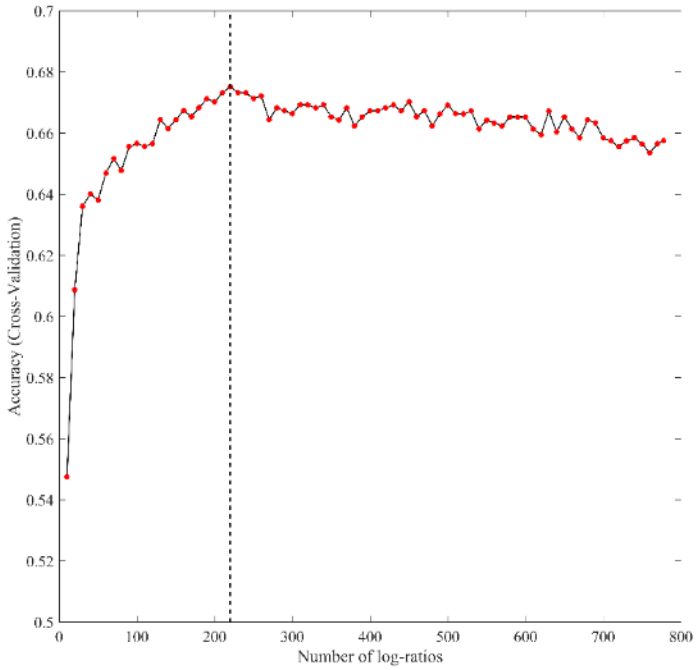

Figure 5-4 Recursive feature elimination with resampling to identify the most important subset of log-ratios

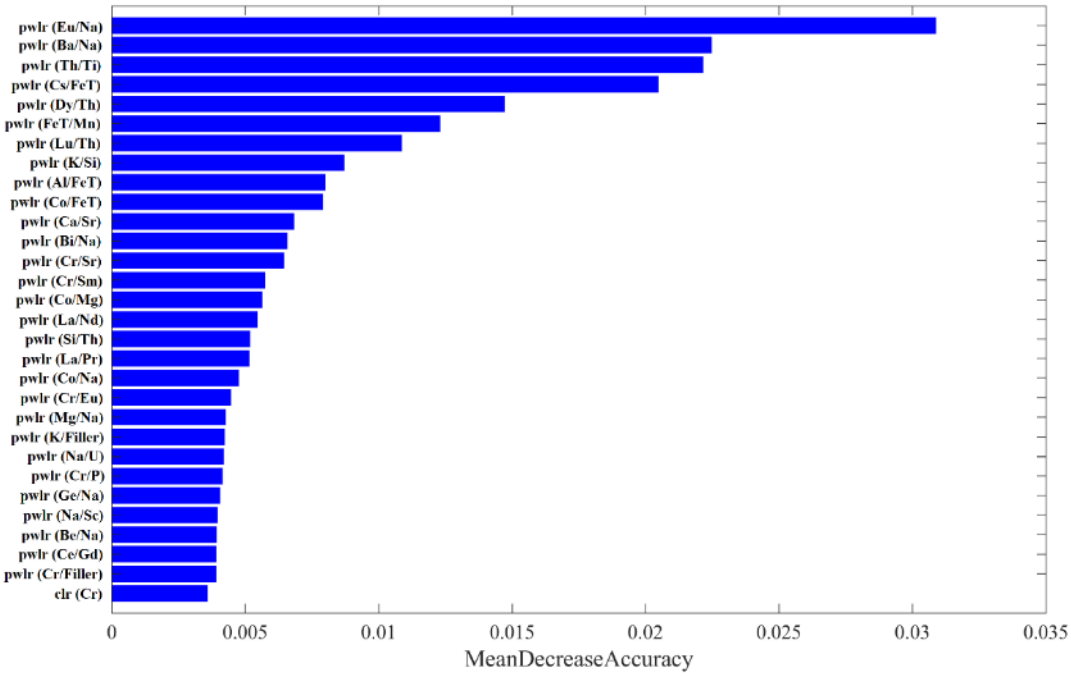

Figure 5-5 The top 30 most informative log-ratios for classification of all MCBs (the significance of selected log-ratios is decreasing from the top to bottom of the chart)

Table 5-2 shows the top 5 most important log-ratios (from left to right) for each MCB of interest. For example, for MCB01 and MCB02, pwlr(Eu/Na) and pwlr(Th/Ti) are the most significant predictors respectively. The simulated model for these two log-ratios are depicted in Figure 5-6. High values (warm colours) of pwlr(Eu/Na) and low values (cool colours) of pwlr(Th/Ti) are associated with MCB01 and MCB02 respectively.

Table 5-2 The top 5 most important log-ratios (from left to right) associated with each MCB

MCBs	**Top 5 most important log-ratios (from left to right)**				
MCB01	pwlr(Eu/Na)	pwlr(Ba/Na)	pwlr(Bi/Na)	pwlr(Co/Na)	pwlr(Mg/Na)
MCB02	pwlr(Th/Ti)	pwlr(Ca/Sr)	pwlr(K/Si)	pwlr(K/Filler)	pwlr(Eu/Na)
MCB03	pwlr(Co/Mg)	pwlr(Cs/FeT)	pwlr(FeT/Mn)	pwlr(Co/FeT)	pwlr(K/Si)
MCB04	pwlr(Dy/Th)	pwlr(Lu/Th)	pwlr(La/Nd)	pwlr(La/Pr)	pwlr(Ce/Nd)
MCB05	pwlr(Cs/FeT)	pwlr(FeT/Mn)	pwlr(Th/Ti)	pwlr(Co/FeT)	pwlr(Eu/Na)
MCB06	pwlr(Cs/FeT)	pwlr(Th/Ti)	pwlr(Al/FeT)	pwlr(Eu/Na)	pwlr(Dy/Th)
MCB07	pwlr.Dy/Th)	pwlr(Co/FeT)	pwlr(FeT/Mn)	pwlr(Th/Ti)	pwlr(Nb/Th)
MCB08	pwlr(Cr/Sr)	pwlr(Cr/Sm)	pwlr(Cr/Eu)	pwlr(Cr/P)	pwlr(Th/Ti)
MCB10	pwlr(FeT/Mn)	pwlr(Na/Zr)	pwlr(Th/Ti)	pwlr(Na/U)	pwlr(Eu/Na)
MCB11	pwlr(Cs/FeT)	pwlr(Th/Ti)	pwlr(FeT/Mn)	pwlr(Al/FeT)	pwlr(Co/FeT)
MCB12	pwlr(Cr/K)	pwlr(Co/Mg)	pwlr(Co/FeT)	pwlr(Th/Ti)	pwlr(Cr/Rb)
MCB13	pwlr(FeT/Mn)	pwlr(Eu/Na)	pwlr(Dy/Th)	pwlr(Ba/Na)	pwlr(Al/FeT)
MCB14	pwlr(Co/Mg)	pwlr(Cs/FeT)	pwlr(Th/Ti)	pwlr(Cr/Sm)	pwlr(Co/FeT)
MCB15	pwlr(Cu/LOI)	pwlr(Cr/Sm)	pwlr(Cs/FeT)	pwlr(Cr/Eu)	pwlr(Cr/Sr)
MCB16	pwlr(Cr/Sm)	pwlr(Cr/Eu)	pwlr(Cs/FeT)	pwlr(Dy/Th)	pwlr(FeT/Mn)
MCB18	pwlr(Cs/FeT)	pwlr(Cu/LOI)	pwlr(Co/FeT)	pwlr(Cr/Sr)	pwlr(Al/FeT)
MCB19	pwlr(Th/Ti)	pwlr(K/Si)	pwlr(Nb/Yb)	pwlr(Cs/FeT)	pwlr(Si/Th)
MCB20	pwlr(Th/Ti)	pwlr(Cs/FeT)	pwlr(FeT/Mn)	pwlr(Nb/Yb)	pwlr(K/Rb)
MCB21	pwlr(Ce/Gd)	pwlr(Dy/Th)	pwlr(Cs/FeT)	pwlr(Th/Ti)	pwlr(Gd/La)
MCB22	pwlr(Th/Ti)	pwlr(Cs/FeT)	pwlr(FeT/Mn)	pwlr(Co/Mg)	pwlr(Eu/Na)
MCB23	pwlr(Th/Ti)	pwlr(FeT/Mn)	pwlr(Cs/FeT)	pwlr(Co/FeT)	pwlr(Eu/Na)
MCB24	pwlr(Cs/FeT)	pwlr(Th/Ti)	pwlr(Co/FeT)	pwlr(Al/Cs)	pwlr(Cs/Rb)

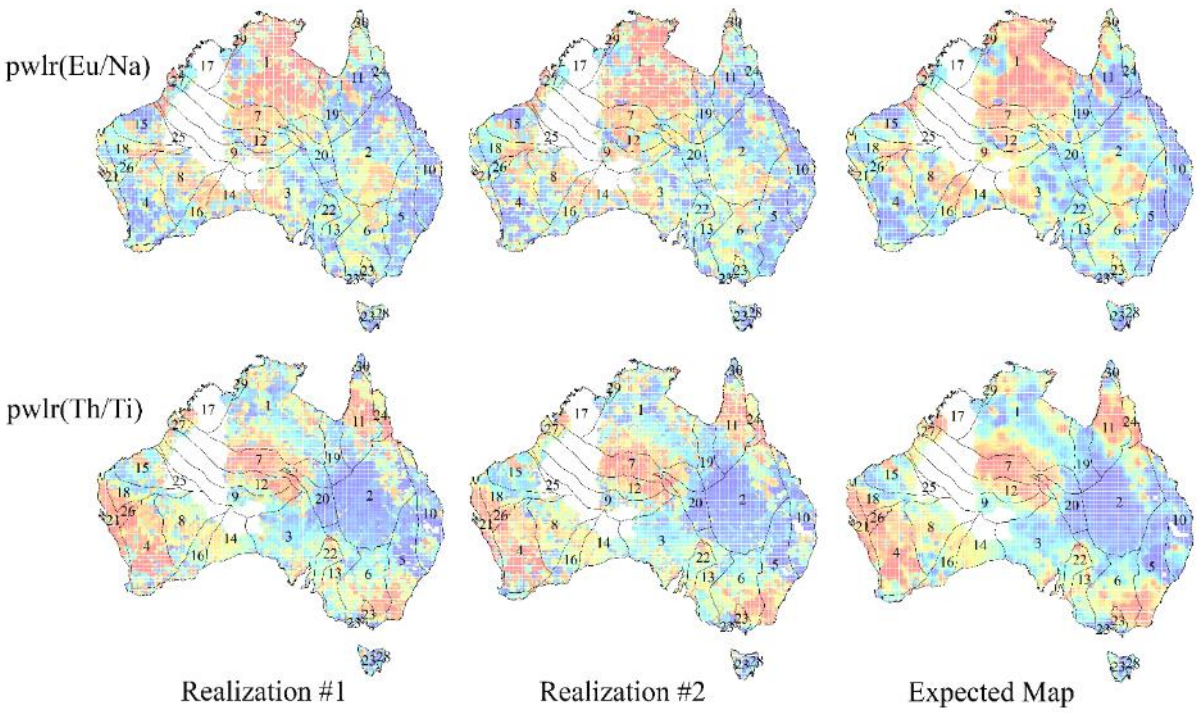

Figure 5-6 Simulated models (two randomly selected realizations) and expected maps for the most significant log-ratios associated with MCB 1 and 2 (warm colours are associated with high values)

The trained RF were used to estimate the probability of occurrence of MCBs at unsampled locations using pwlr and clr of simulated compositions as input predictors. For each location u of the study area and each MCB k, 100 probabilities were simulated. Maps of minimum, expected (Eq. 5.10) and maximum estimated probabilities are shown in Figure 5-7 for MCBs 1 to 4. Figure 5-8 shows conditional total variation of simulated MCBs calculated via Equation 5.11. Areas close to geochemical samples show lower uncertainty. MCBs 1, 2 and 10 show higher uncertainty than the other MCBs while MCBs 3, 6, 13 and 22 show low uncertainty. Finally, Figure 5-9 shows the most probable MCBs calculated via the proposed method. The predicted crustal blocks are broadly consistent with the known MCBs (continuous black lines in Figure 5-9). Discrepancies may be due to uncertain initial definition of crustal boundaries (e.g. due to ambiguity of geophysical data) or from surficial processes (e.g. chemical weathering and/or physical transport effects) that mask/shift the crustal block geochemical signature (see discussion in Grunsky et al. (2017)). In conclusion, the architecture of the MCBs of Australia can be predicted accurately from geochemical composition of the Australian surface regolith. These

results can be used further for managing projects such as mineral exploration, environmental and ecological planning, and efficient usage of water resources.

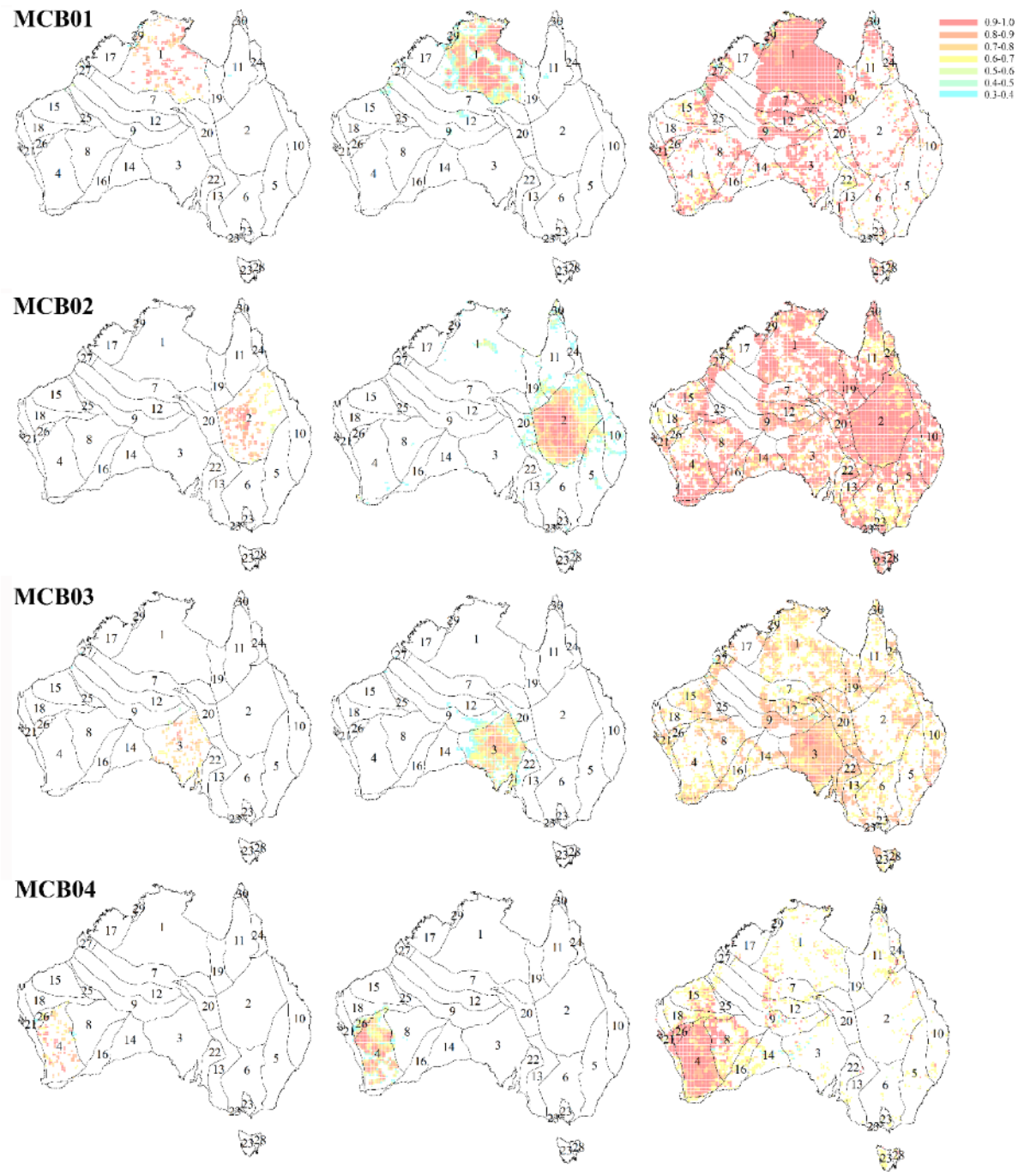

Figure 5-7 Maps of minimum (first column), expected (middle column) and maximum (last column) probability of occurrence for MCB 1 to 4

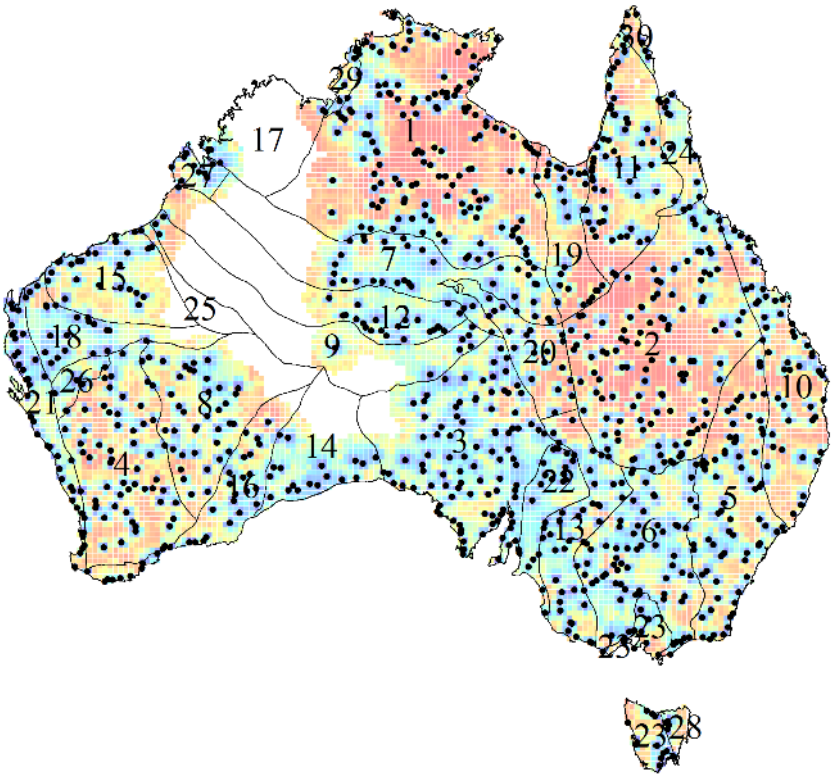

Figure 5-8 Conditional total variation of all simulated MCBs (warm colours show high values)

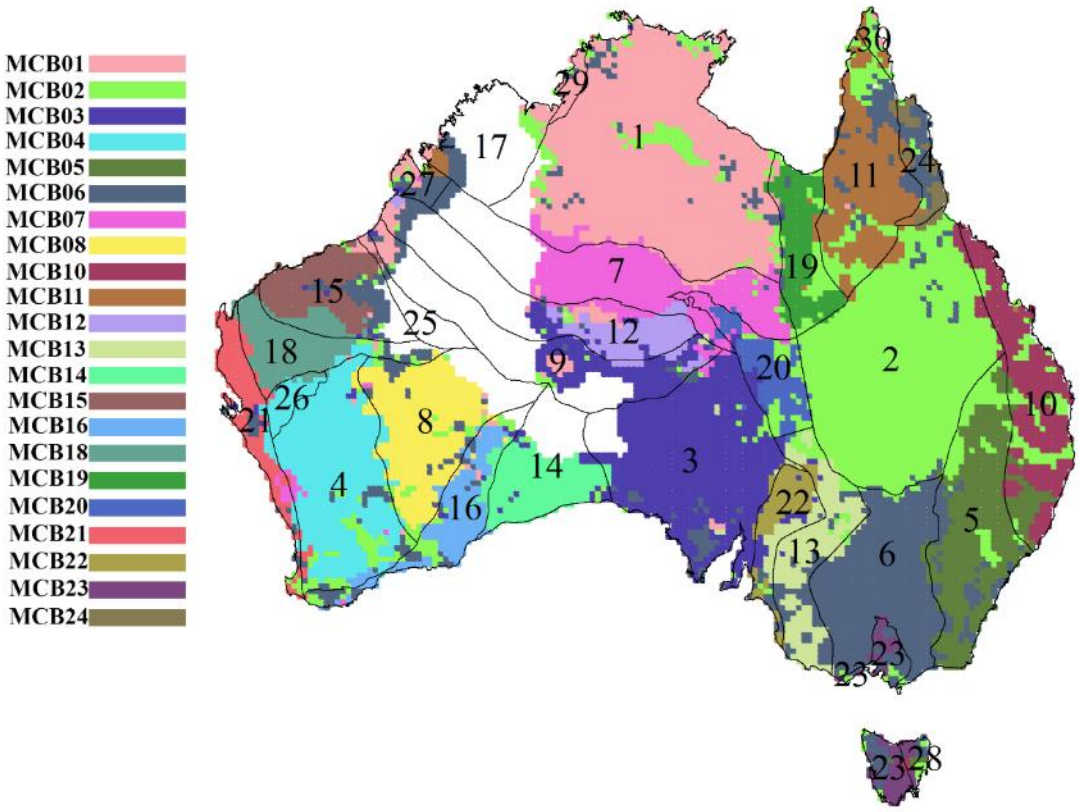

Figure 5-9 Map of most probable MCBs

5.4 Post-glacial deposits exploration for environmental monitoring

In this study, regional-scale soil geochemical dataset (obtained as part of the Tellus Project generated by the Geological Survey of Northern Ireland) is analysed to explore the relationship between soil geochemistry and post-glacial deposits (e.g. surficial peat deposits) for environmental monitoring of this fragile ecosystem. Superficial deposits (e.g. glacial till, post-glacial alluvium, and peat) in this area have been created due to the advance of ice sheets and their melt-waters over the last 100,000 years (Figure 5-10). Accurate mapping of peat-covered areas has become important because of the relatively high carbon density of peat and organic-rich soils.

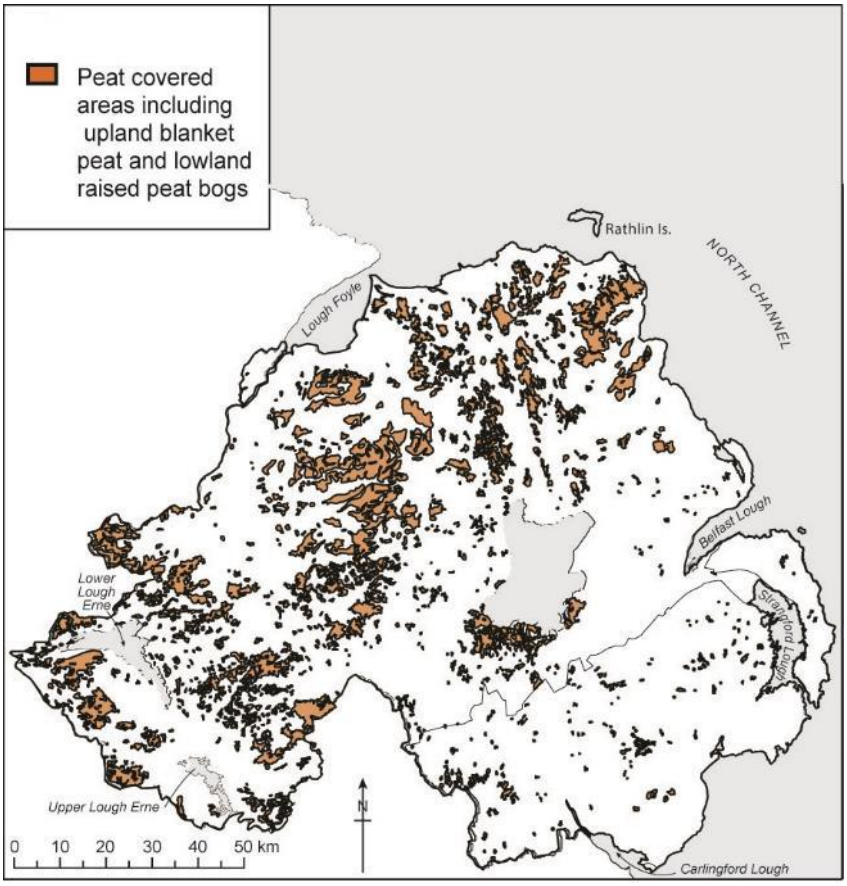

Figure 5-10 Post-glacial peat-covered areas; adapted from McKinley et al. (2018)

5.4.1 Dataset

The Northern Ireland Tellus Survey (GSNI 2007; Young and Donald 2013) consists of 6862 rural soil samples (X-ray fluorescence (XRF) analyses). Geochemical samples presented in this study were collected at 20-cm depth, with average spatial coverage of one sample site every 2 km^2. Each soil sample site was assigned to the post-glacial peat covered map (Figure 5-10), resulting in spatial data for one binary response variable (presence or absence of peat) and 50 continuous geochemical variables (Ag, Al_2O_3, As, Ba, Bi, Br, CaO, Cd, Ce, Cl, Co, Cr, Cs, Cu, Fe_2O_3, Ga, Ge, Hf, I, K_2O, La, MgO, MnO, Mo, Na_2O, Nb, Nd, Ni, P_2O_5, Pb, Rb, SO_3, Sb, Sc, Se, SiO_2, Sm, Sn, Sr, Th, TiO_2, Tl, U, V, W, Y, Yb, Zn, Zr, and *filler* which includes Loss on Ignition (LOI)). More information on Tellus Survey field methods and analytical methodology are available in Smyth (2007) and Young and Donald (2013).

5.4.2 Results and discussion

Input data were transformed to real space via ilr transformation (Eq. 5.5) and subsequently to multivariate normal space via flow anamorphosis. Two successive FA with the same parameters ($\sigma_0 = 0.1$ and $\sigma_1 = 1.1$) were required to achieve multivariate normality. The multivariate normal scores were simulated 100 times on a regular grid (1 km × 1 km) independently via the turning bands algorithm and back-transformed to compositions subsequently. Figure 5-11 shows the map of the conditional total compositional variations (spatial uncertainty of the geochemical compositions) calculated via Equation 5.9. Outlines of the peat covered areas are shown by black polygons. According to this map geochemical compositions show higher variation close to peat deposits. This may represent random disturbances of the geochemical signal at very small spatial scale due to peat cover.

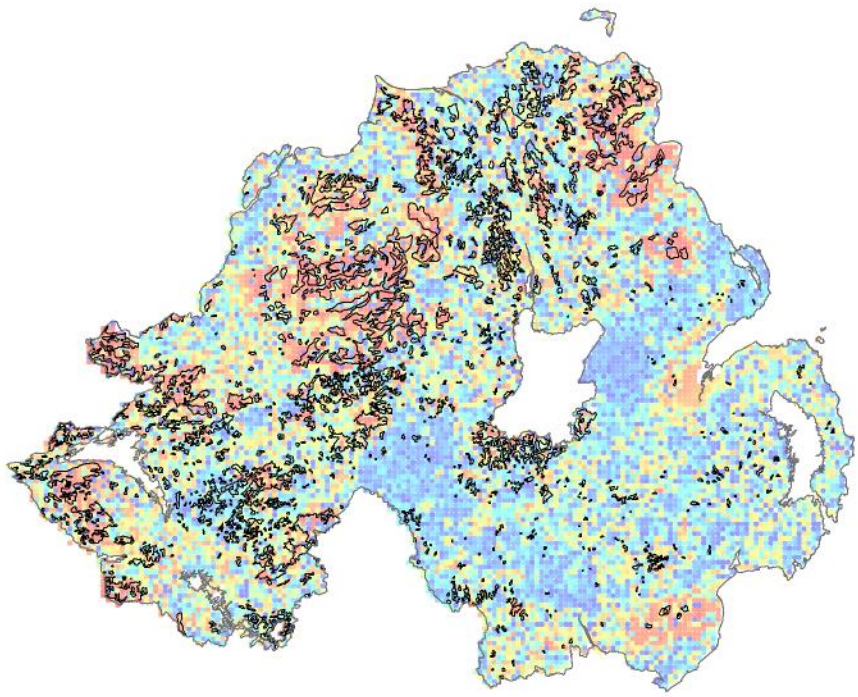

Figure 5-11 Conditional total compositional variation (warm colours are associated with high values and black polygons are peat covered areas)

The pairwise log-ratios (1225 log-ratios) and centred log-ratios (50 log-ratios) were used as predictors and peat/non-peat as the binary response variable to train a RF

predictive model. The most informative subset of log-ratios for discrimination of peat covered areas was selected using Algorithm 5.2. The final predictive RF with the highest accuracy was associated with a subset of only 150 log-ratios (Figure 5-12). Figure 5-13 shows the top 30 most significant log-ratios for discrimination of peat-covered areas. Figure 5-14 shows the spatial distribution (two randomly selected realizations and the expected map) of the most informative log-ratio, pwlr (Y/*filler*), where a coincidence between low values (cool colours) of this log-ratio and peat covered areas is clear. The most informative log-ratios, e.g pwlr (Y/*filler*), include the presence of LOI in the *filler* variable. This supports the previously known association between peat cover and LOI.

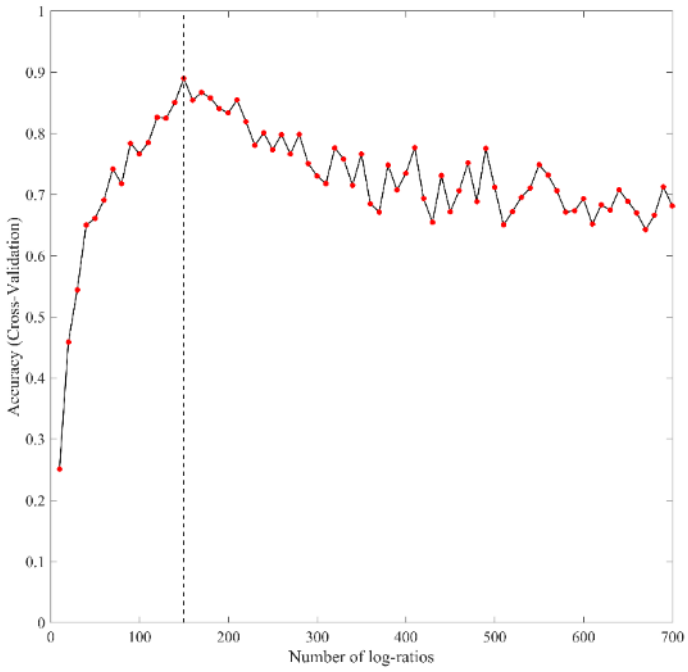

Figure 5-12 Recursive feature elimination with resampling to identify the most important subset of log-ratios (Northern Ireland Tellus Survey data)

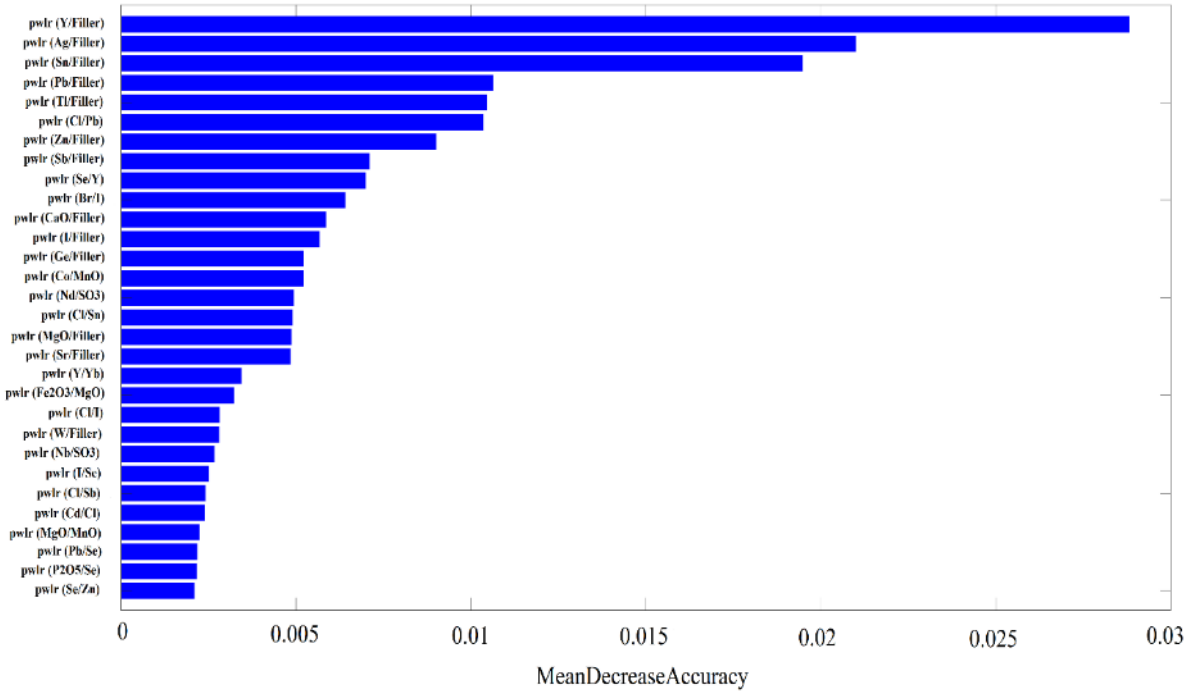

Figure 5-13 The top 30 most informative log-ratios for discrimination of peat covered areas (the significance of selected log-ratios is decreasing from the top to bottom of the chart)

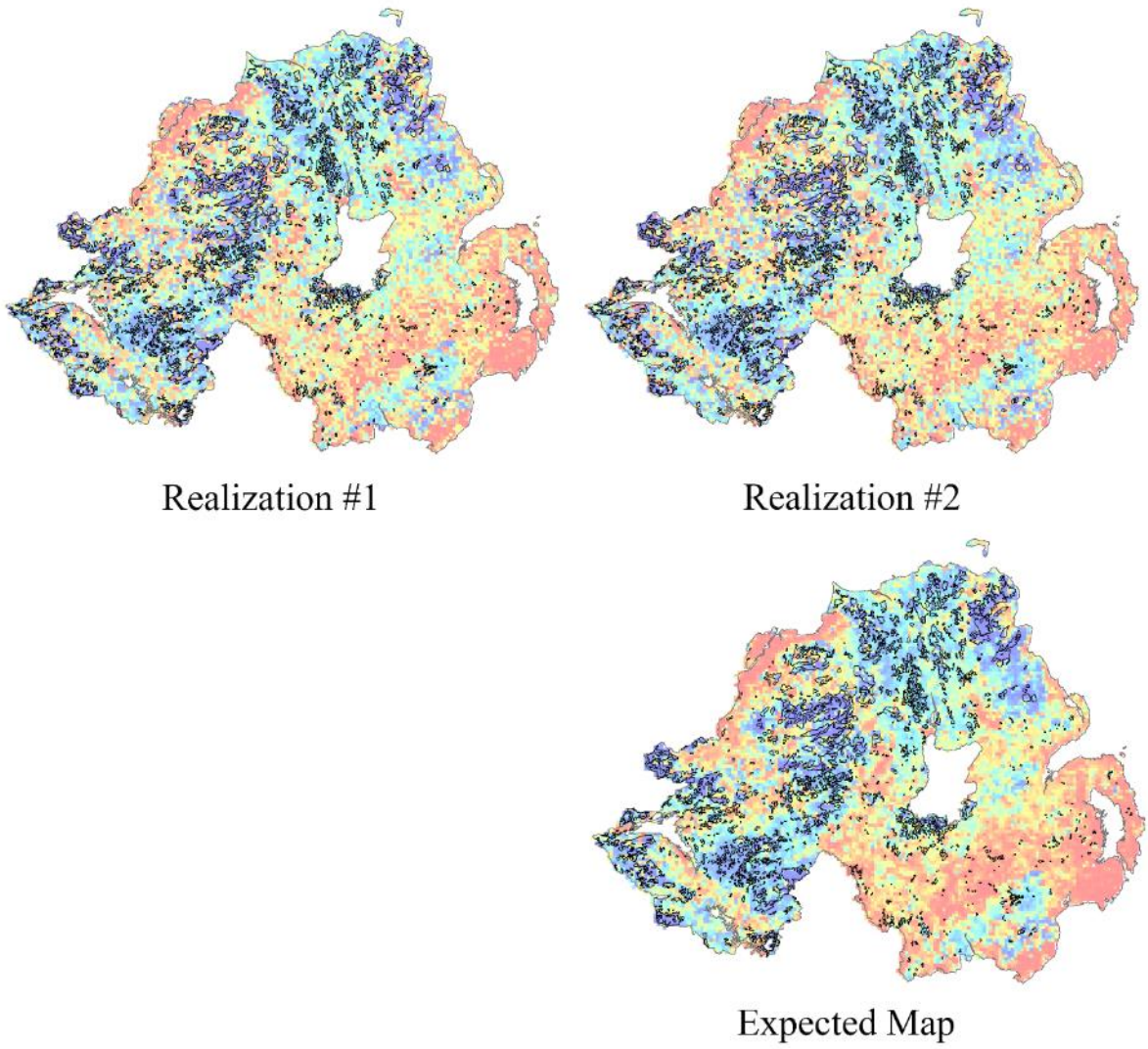

Figure 5-14 Simulated model (two randomly selected realizations) and expected map of the most significant log-ratio (pwlr (Y/*filler*)) for discrimination of peat covered areas (warm colours are associated with high values and black polygons are peat covered areas)

Finally the trained RF was used to predict the probability of occurrence of peat covered areas at unsampled locations. Maps of minimum, expected (Eq. 5.10) and maximum estimated probabilities of peat covered areas are shown in Figure 5-15 which demonstrate good consistency with the reported peat areas (Figure 5-10). Figure 5-16 shows conditional total variation of predicted peat covered areas calculated via Equation 5.11. Areas close to peat deposits show higher uncertainty. Figure 5-17 shows the most probable peat covered areas calculated via the proposed method. Although Figure 5-15 and Figure 5-17 show good match with the reported peat covered areas, inconsistencies may be due to uncertain initial definition of peat

covered areas (Figure 5-10) and/or degradation of peat-covered areas since the creation of the superficial deposit classification that mask the peat geochemical signature. Peat covered areas include upland blanket bog which is more extensive and spatially coherent, and lowland 'raised bogs' which are smaller more fragile ecosystems. Using the proposed spatial predictive model, the locations of the main upland blanket peat covered areas have been predicted accurately from geochemical composition of the Northern Ireland Tellus Survey. The association of LOI with peat covered areas helps to explain the most informative log-ratios, e.g. pwlr (Y/*filler*). However the approach has also identified the presence of potentially important marker elements (Y, Ag and Sn) which may have accumulated in peat which acts as a sink for toxic elements. The results can be used further for managing projects such as environmental and ecological planning. As the underlying geology and spatial distribution of soil types across Northern Ireland are similar to the UK (Jordan et al. 2001) and Northern Europe in general, the proposed techniques in this study can be applied on those areas.

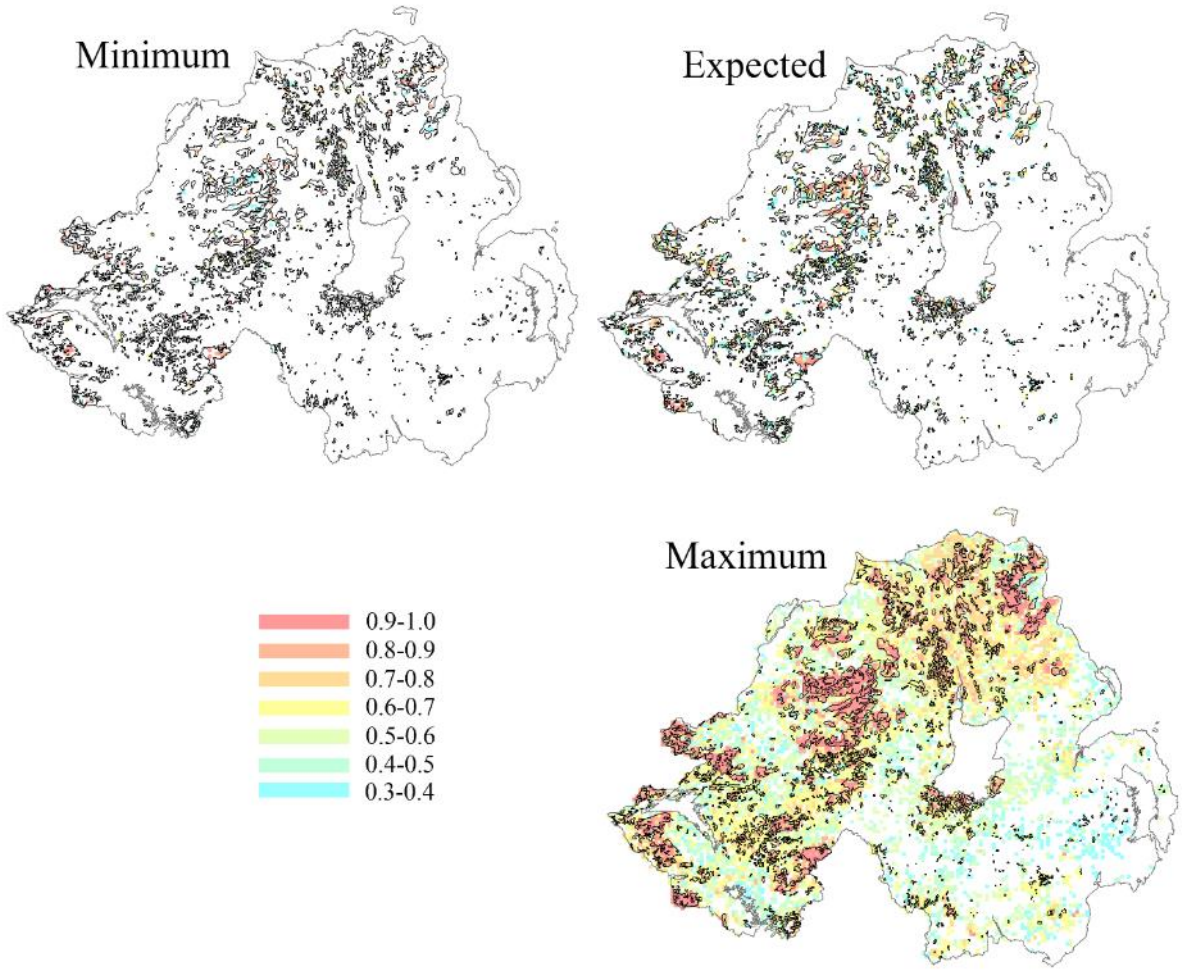

Figure 5-15 Maps of minimum, expected and maximum probability of occurrence for peat covered areas

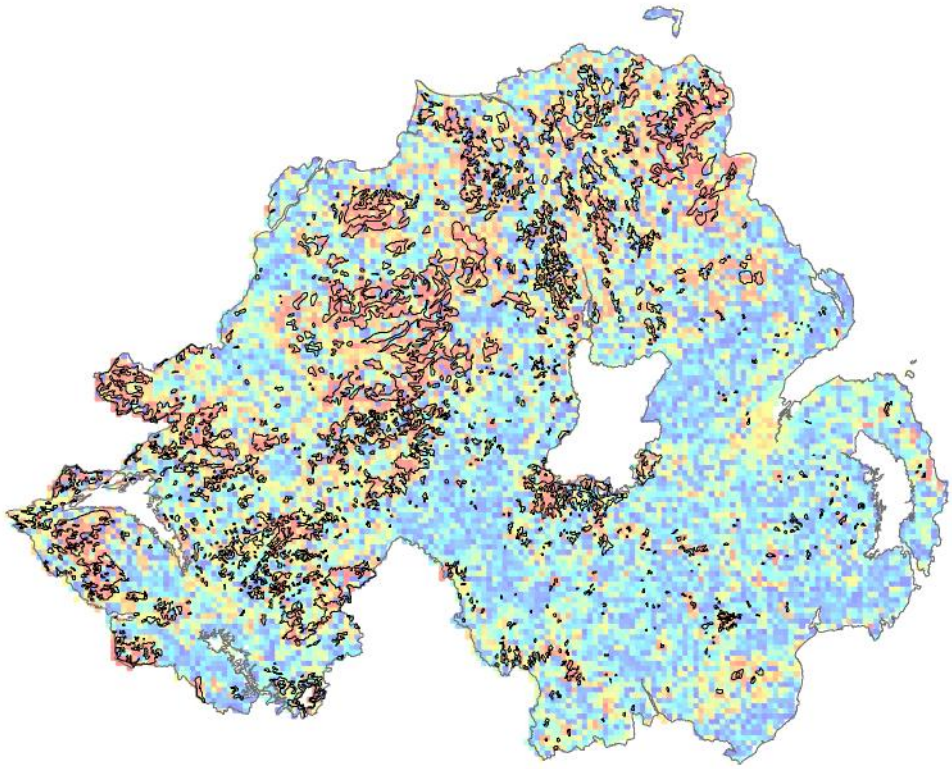

Figure 5-16 Conditional total variation of simulated peat covered areas (warm colours are associated with high values and black polygons are peat covered areas)

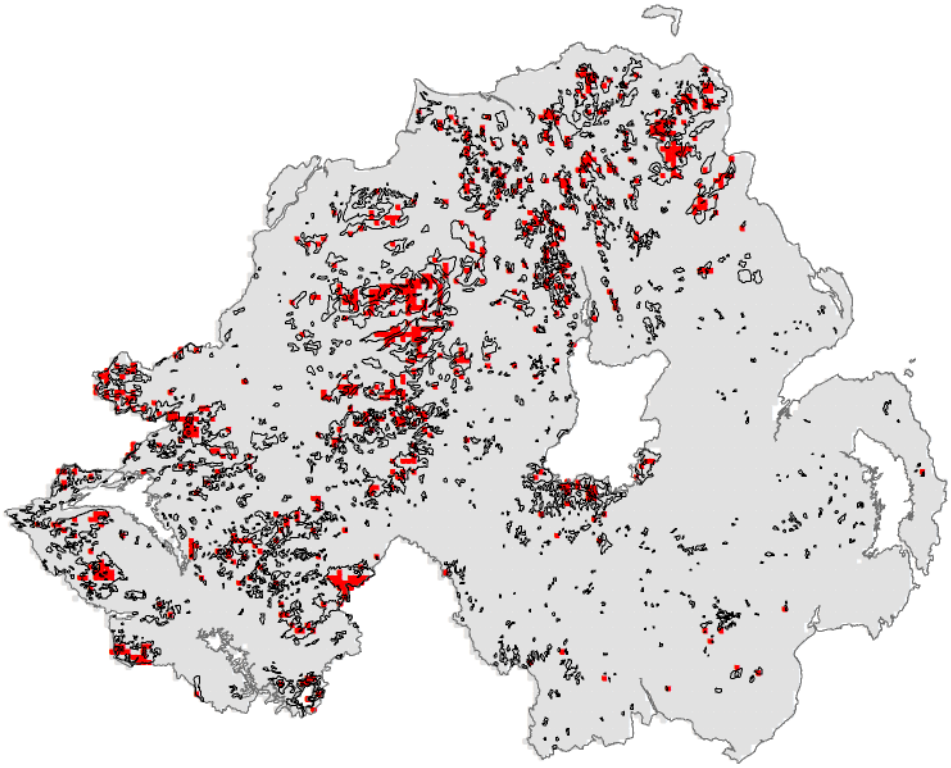

Figure 5-17 Map of most probable peat covered areas (shown by red colour)

5.5 Conclusions

This study introduces a novel approach for the spatial modelling of uncertainty and prediction of geological classes using geochemical compositions. The approach is based on the combined use of advanced geostatistical simulation for compositional data (geostatistical simulation using isometric log-ratio transformation and flow anamorphosis) and a random forests predictive model. Due to the high-dimensional characteristics of log-ratios, recursive feature elimination with resampling technique were used to select the most significant log-ratios for the classification purpose. Such a feature selection technique is known to lead to a more stable and accurate predictive model and can be used further as an exploratory data analysis tool for geological process discoveries. The proposed approach was applied on two case studies. In the first case study the major crustal blocks of the Australian continent were predicted from the surface regolith geochemical compositions while in the second case study the spatial distribution of superficial deposits (peat) were

predicted from regional-scale soil geochemical data of Northern Ireland (Tellus Project). The accuracy of the results in these two case studies confirmed the usefulness and applicability of the proposed method.

5.6 Acknowledgements

The first three authors acknowledge financial support through DAAD-UA grant CodaBlockCoEstimation. The National Geochemical Survey of Australia project was part of the Australian Government's Onshore Energy Security Program 2006–2011, from which funding support is gratefully acknowledged. The NGSA was led and managed by Geoscience Australia and carried out in collaboration with the geological surveys of every State and the Northern Territory under National Geoscience Agreements. The Geological Survey of Northern Ireland (GSNI) is thanked for the use of the Tellus dataset. The Tellus Project was carried out by GSNI and funded by The Department for Enterprise, Trade and Investment (DETINI) and The Rural Development Programme through the Northern Ireland Programme for Building Sustainable Prosperity.

5.7 References

Aitchison J (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society*, 44, 139-177.

Aitchison J (1986). *The statistical analysis of compositional data*. London, UK: Chapman & Hall Ltd.

Blake D, Kilgour B (1998). Geological regions of Australia 1:5,000,000 Scale [Dataset]. Canberra: Geoscience Australia, Available at: http://www.ga.gov.au/metadatagateway/metadata/record/gcat_a05f7892-b237-7506-e044-00144fdd4fa6/Geological+Regions+of+Australia%2C+1%3A5+000+000+scale.

Breiman L (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Breiman L (2001). Random forests. *Machine Learning*, 45, 5-32.

Buccianti A, Grunsky E C (2014). Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes? *Journal of Geochemical Exploration*, 141, 1-5.

Caritat P de, Cooper M (2011). National geochemical survey of Australia: The geochemical atlas of Australia. Geoscience Australia, Record 2011/20, Available at: http://www.ga.gov.au/metadata-gateway/metadata/record/gcat_71973.

Caritat P de, Cooper M (2016). A continental-scale geochemical atlas for resource exploration and environmental management: The national geochemical survey of Australia. *Geochemistry: Exploration, Environment, Analysis*, 16, 3-13.

Caritat P de, Main P T, Grunsky E C, Mann A W (2017). Recognition of geochemical footprints of mineral systems in the regolith at regional to continental scales. *Australian Journal of Earth Sciences*, 64, 1033-1043.

Chilès J P, Delfiner P (2012). *Geostatistics: modeling spatial uncertainty*. New York: Wiley.

Drew L J, Grunsky E C, Sutphin D M, Woodruff L G (2010). Multivariate analysis of the geochemistry and mineralogy of soils along two continental-scale transects in North America. *Science of The Total Environment*, 409, 218-227.

Egozcue J J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279-300.

Emery X (2008). A turning bands program for conditional co-simulation of cross-correlated Gaussian random fields. *Computers & Geosciences*, 34, 1850-1862.

Emery X, Arroyo D, Porcu E (2016). An improved spectral turning-bands algorithm for simulating stationary vector Gaussian random fields. *Stochastic Environmental Research and Risk Assessment*, 30, 1863-1873.

Emery X, Lantuéjoul C (2006). TBSIM: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method. *Computers & Geosciences*, 32, 1615-1628.

Grunsky E C (2010). The interpretation of geochemical survey data. *Geochemistry: Exploration, Environment, Analysis*, 10, 27-74.

Grunsky E C, Caritat P de, Mueller U (2017). Using surface regolith geochemistry to map the major crustal blocks of the Australian continent. *Gondwana Research*, 46, 227-239.

Grunsky E C, Drew L J, Woodruff L G, Friske P W B, Sutphin D M (2013). Statistical variability of the geochemistry and mineralogy of soils in the Maritime Provinces of Canada and part of the Northeast United States. *Geochemistry: Exploration, Environment, Analysis*, 13, 249-266.

Grunsky E C, Mueller U, Corrigan D (2014). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping. *Journal of Geochemical Exploration*, 141, 15-41.

Geological Survey Northern Ireland (GSNI) (2007). Tellus project overview, https://www.bgs.ac.uk/gsni/Tellus/index.html.

Guyon I, Weston J, Barnhill S, Vapnik V (2002). Gene selection for cancer classification using Support Vector Machines. *Machine Learning*, 46, 389-422.

Harris J R, Grunsky E C (2015). Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Computers & Geosciences*, 80, 9-25.

Jordan C, Higgins A, Hamill K, Cruickshank J (2001). The soil geochemical atlas of Northern Ireland. Department of Agriculture and Rural Development, NI.

Kanevski M, Pozdnoukhov A, Timonin V (2009). *Machine learning for spatial environmental data: theory, applications and software*. BocaRaton, USA: CRC Press.

Korsch R J, Doublier M P (2015). Major crustal boundaries of Australia, Scale 1:2 500 000. second ed. Canberra, Geoscience Australia, http://www.ga.gov.au/metadata-gateway/metadata/record/83223.

Korsch R J, Doublier M P (2016). Major crustal boundaries of Australia, and their significance in mineral systems targeting. *Ore Geology Reviews*, 76, 211-228.

Kuhn M, Johnson K (2013). Applied predictive modeling. New York: Springer.

McKinley J M (2015). Using compositional geochemical ground survey data as predictors for geogenic radon potential. Paper presented at the International Workshop on the European Atlas of Natural Radiation, Verbania, Italy.

McKinley J M, Grunsky E C, Mueller U (2018). Environmental monitoring and peat assessment using multivariate analysis of regional-scale geochemical data. *Mathematical Geosciences*, 50, 235-246.

McKinley J M, Hron K, Grunsky E C, Reimann C, Caritat P de, Filzmoser P et al. (2016). The single component geochemical map: Fact or fiction? *Journal of Geochemical Exploration*, 162, 16-28.

Mueller U, Tolosana-Delgado R, van den Boogaart K G (2014). Approaches to the simulation of compositional data – a nickel-laterite comparative case study. Paper presented at the Orebody Modelling and Strategic Mine Planning Symposium 2014, Melbourne.

Mueller U, van den Boogaart K G, Tolosana-Delgado R (2017). A truly multivariate normal score transform based on lagrangian flow. In J. J. Gómez-Hernández, J. Rodrigo-Ilarri, M. E. Rodrigo-Clavero, E. Cassiraga, J. A. Vargas-Guzmán (eds.), *Geostatistics Valencia 2016* (pp. 107-118). Springer

Nakamura A, Milligan P R (2015). Total magnetic intensity (TMI) colour composite image. Canberra: Geoscience Australia, http://www.ga.gov.au/metadata-gateway/metadata/record/82799/.

Pawlowsky-Glahn V, Egozcue J J (2016). Spatial analysis of compositional data: A historical review. *Journal of Geochemical Exploration*, 164, 28-32.

Pawlowsky-Glahn V, Egozcue J J, Tolosana-Delgado R (2015). *Modelling and analysis of compositional data*. Chichester (UK): Wiley.

Pawlowsky-Glahn V, Olea R A (2004). *Geostatistical analysis of compositional data*. Oxford University Press.

Pawlowsky-Glahn V, Buccianti A (2011). *Compositional data analysis: Theory and applications*. Chichester (UK): Wiley.

Raymond O L (2012). Surface geology of Australia, Data package, [Dataset]. Canberra, Geoscience Australia, https://www.ga.gov.au/products/servlet/controller?event=GEOCAT_DETAILS&catno=74855.

Smyth D (2007). Methods used in the Tellus geochemical mapping of Northern Ireland. British geological survey, open report or/07/022.

Strobl C, Boulesteix A L, Zeileis A, Hothorn T (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 25

Tercan A E (1999). Importance of orthogonalization algorithm in modeling conditional distributions by orthogonal transformed indicator methods. *Mathematical Geology*, 31, 155-173.

Tolosana-Delgado R (2006). Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring. PhD thesis, University of Girona, Spain.

Tolosana-Delgado R, McKinley J M (2016). Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). *Applied Geochemistry*, 75, 263-276.

Tolosana-Delgado R, McKinley J M, van den Boogaart K G (2015). Geostatistical fisher discriminant analysis. Paper presented at the 17th annual conference of the International Association for Mathematical Geosciences, Freiberg (Saxony) Germany.

Tolosana-Delgado R, van den Boogaart K G (2014). Towards compositional geochemical potential mapping. *Journal of Geochemical Exploration*, 141, 42-51.

Tolosana-Delgado R, van den Boogaart K G (2013). Joint consistent mapping of high-dimensional geochemical surveys. *Mathematical Geosciences*, 45, 983-1004.

van den Boogaart K G, Mueller U, Tolosana-Delgado R (2017). An affine equivariant multivariate normal score transform for compositional data. *Mathematical Geosciences*, 49, 231-251.

van den Boogaart K G, Tolosana-Delgado R (2013). *Analyzing Compositional Data with R*. Heidelberg (Germany): Springer.

Young M, Donald A (2013). A guide to the Tellus data. Belfast: Geological Survey of Northern Ireland.

Chapter 6

General discussion

Regionalized compositional data (in form of percentages, probabilities, proportions, frequencies, and concentrations) are common in geosciences. Geochemical and mineralogical data, proportions of material occupied the porous media in an aquifer or oil reservoir, proportions of rock types, soil types and land uses in the study area are examples of such compositional information. Most of the time regionalized compositional data are statistically and spatially related to one or more dependent categorical data such as rock types, soil types, alteration units, and continental crustal blocks. Complex statistical and spatial relationships between these mixed data should be honoured in the simulated and/or estimated models. Developing joint predictive models for such geospatial mixed data is necessary due to their applicability for geoscience modelling projects. This PhD thesis explored and introduced several approaches to spatial modelling of regionalized compositional and categorical data for different situations and applications. To this end, multiple-point geostatistical techniques have priority due to their capability for reproducing complex spatial patterns. However, to implement MPS techniques, large and dense compositional and categorical training images or training data are needed. For situations where such training information is not available and/or complex spatial patterns are not present in the study area (or such patterns are not our interest), two-point geostatistical algorithms can be implemented. Finally, several advantages of machine learning algorithms such as recognition of complex statistical patterns, internal feature selection and cross-validation can be used for the joint modelling of compositional and categorical data. However, care should be taken while implementing such techniques on geospatial data as they are non-spatial algorithms.

Algorithms were developed for all the aforementioned situations. The following subsections discuss the pros and cons of the developed algorithms, the area of their application, and proper ways of implementation.

6.1 Multiple-point framework

Among many MPS algorithms, the direct sampling technique was selected to be developed for the joint simulation of compositional and categorical data. DS is capable of running co-simulation of mixed data, capturing multivariate spatial patterns of different sizes without the need to define a search template with fixed size and geometry, and capturing spatial patterns of different scales without the need for a multigrid search strategy (Mariethoz and Caers 2015; Mariethoz and Renard 2010; Mariethoz et al. 2010). In the case of compositional data, the dissimilarity between spatial compositional patterns cannot be measured in the standard Euclidean metric, instead a compositional distance (known as Aitchison distance) (Aitchison 1986; Pawlowsky-Glahn et al. 2015) needs to be used. Another way is to transform the compositional data to real space via an isometric log-ratio transformation and to measure distances via commonly used distances for real data such as Euclidean distance. After scanning the training image and finding a close pattern, the whole compositional vector can be pasted in the associated node of the simulation grid. Simulating compositional vectors as a whole increases the speed of the DS algorithm. However, this approach is recommended only when a large compositional training image or a large set of compositional training data are available. The large size of the compositional training image guarantees reasonable total compositional variation in the simulated model. Whenever such a large compositional training image is not available, selecting a fully random path for simulation and simulating isometric log-ratios randomly at each node of the simulation grid generates reasonable compositional variation. Simulation based on a fully random path and isometric log-ratio transformation leads to generating compositions not present in the input data. The isometric log-ratio transformation also reduces the dimension of the compositional vector by one (making the simulation faster) while preserving the distances between compositions. Two case studies in chapter 2 showed that the sub-compositional patterns (shown in ternary diagrams) can be reproduced properly with this technique. To evaluate the realizations of compositional random function, new metrics (e.g., global Aitchison distance between the simulated results and validation set) were introduced in chapter 2 as in the case of compositional data, standard descriptive statistics are not informative. To increase the accuracy of the predictions for compositional and

categorical data via the proposed workflow, parameters of DS should be tuned properly and/or a sensitivity analysis for the appropriate size of the training data should be conducted.

6.2 Two-point framework

In chapter 3 a spatial decorrelation technique for joint two-point geostatistical simulation of high-dimensional continuous and categorical data was presented based on the plurigaussian model and min/max autocorrelation factors. Each categorical variable can be presented via one or more underlying Gaussian variables. As a result the proposed method is capable of simulating several categorical variables by defining several plurigaussian models (Armstrong et al. 2011). On the other hand, in the case of compositional data, they should be opened up to real space via one of the several available log-ratio transformations (Aitchison 1986; Egozcue et al. 2003). Any log-ratio transformation can be used as long as the transformation to normal space is based on a multivariate affine equivariant anamorphosis (van den Boogaart et al. 2017). In chapter 4 it has been shown that the classical transformation to normal space (Gaussian anamorphosis) is not capable of reproducing complex statistical patterns inside data. Geostatistical modelling via flow anamorphosis is capable of reproducing complex patterns in data including: outliers, multiple populations, nonlinearity, and heteroscedasticity. The invariance property of the flow anamorphosis gives modellers the freedom to select an appropriate log-ratio transformation (among many available log-ratio transformations). The transformed scores via this anamorphosis are multivariate normal, statistically independent, and spatially orthogonal. The orthogonality is particularly important in the case of high-dimensional data as the geostatistical modelling of such independent factors is straightforward. In chapter 4, it has been shown that in situations where continuous data show spatial correlation across the boundary between different categories (soft transitions) and consist of different statistical populations, geostatistical simulation via flow anamorphosis without domaining outperforms other approaches for spatial modelling of compositional data such as domaining and independent simulation or probabilistic weighted approach. The proposed method was implemented on a nickel-cobalt laterite

deposit and results were satisfactory based on several criterial (e.g. reproduction of probability distribution functions, sub-compositional patterns (checked via ternary diagrams), variograms, and grade-tonnage curves).

6.3 Machine learning – Spatial predictive implementation

Ensemble predictive models (such as the Random Forest algorithm) are very popular due to their ease of implementation, their ability to handle many types of predictors (sparse, skewed, continuous, categorical, etc.) without the need to pre-process them, allowing missing data, conducting feature selection and cross-validation internally, and stability of the predicted results (Kuhn and Johnson 2013). A major limitation of machine learning algorithms is that they generally do not consider the spatial relationships between observations and variables. As a result the uncertainty maps generated via MLAs cannot be considered a trustworthy spatial uncertainty model. This kind of limitation was addressed in chapter 5, where a hybrid model was developed based on the combined use of advanced geostatistical simulation (implementing a non-linear Gaussian anamorphosis) and random forest algorithm. In a first step, the random forest classifier was trained based on the available input data. To acknowledge the spatial uncertainty of compositional data, different realisations of the given compositional random function were used as input to the trained ensemble classifier. For each realisation of compositional random function, the probabilities of different classes (e.g., crustal blocks, deposit/non-deposit areas, rock type) were estimated. Some ideas were borrowed from compositional data analysis to merge these probabilities in order to combine elements of statistical (bootstrapping via RF) and spatial (turning bands algorithm) uncertainties. The most probable spatial map of categories was defined using the final spatial uncertainty model. A compositionally compliant feature selection was introduced to address the high-dimensionality characteristics of compositional features (log-contrast). The two case studies in chapter 5, proved the usefulness of the proposed algorithm for geological class prediction, spatial uncertainty modelling, and recognising significant features for geoscience processes discovery analysis. In the first case study, the spatial distribution of major crustal block of Australian continent was predicted accurately. In the second case

study, the spatial distribution of peat covered areas were predicted with high accuracy. The spatial maps of the most significants log-ratios (associated with each geological class) and the associated spatial uncertainties were generated for each case.

6.4 Chapter references

Aitchison J (1986). The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman & Hall Ltd., London.

Armstrong M et al. (2011). *Plurigaussian simulations in geosciences.* Springer Berlin Heidelberg, Berlin, Heidelberg.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279-300.

Kuhn M, Johnson K (2013). *Applied predictive modeling*. Springer-Verlag New York.

Mariethoz G, Caers J (2015). *Multiple-point geostatistics: Stochastic modeling with training images.* John Wiley & Sons, Ltd.

Mariethoz G, Renard P (2010). Reconstruction of incomplete data sets or images using direct sampling. *Mathematical Geosciences*, 42, 245-268.

Mariethoz G, Renard P, Straubhaar J (2010). The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, 46, W11536.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modelling and analysis of compositional data*. John Wiley & Sons, Ltd.

van den Boogaart KG, Mueller U, Tolosana-Delgado R (2017). An affine equivariant multivariate normal score transform for compositional data. *Mathematical Geosciences*, 49, 231-251.

Chapter 7

Overall conclusions and future recommendations

This PhD research pursued the development of approaches to spatial uncertainty modelling and prediction of a set of regionalized dependent compositional and categorical variables. The proposed approaches have many geoscience applications including in the evaluation of mineral resources, characterization of oil reservoirs or hydrology of groundwater, and contaminated site characterization and remediation. Through the development of the proposed techniques, the compositional nature of continuous data was addressed and fully incorporated in the joint modelling approach. Two main streams were followed for the spatial uncertainty modelling and prediction: two-point and multiple-point geostatistics.

For the geoscience modelling projects where a large and representative training images (or training data) for compositional and categorical information are available, the proposed approach (adapted implementation and evaluation of Direct Sampling algorithm) for the multiple-point stream is recommended due to its capability of correctly reproducing statistical and spatial, compositional and categorical, dependent patterns. The direct sampling algorithm is developed and presented in chapter 2 to this end. However, the proposed approach for simulating compositional data via DS should be examined further in terms of sub-compositional coherence and total compositional variation. In this study, parameters of DS algorithm were selected by the user. Numerical optimisation techniques can be developed and implemented to find the optimum parameters.

For the situation where the first stream is not applicable (e.g. lack of a representative training image), a hybrid model was developed and presented in chapter 3, based on plurigaussian models and min/max autocorrelation factors. This spatial decorrelation technique for two-point geostatistical simulation is capable of modelling several compositional and categorical variables. In this technique for each categorical variable, a separate plurigaussian model is defined. A Gibbs sampler algorithm was used to simulate the underlying Gaussian variables associated with each plurigaussian model. In the proposed algorithm the Gibbs

sampler was conditional only to categorical information. A Gibbs sampler algorithm conditional to both, categorical and compositional (in term on log-ratios) information, may produce more accurate predictions.

Regionalized compositions often consist of several populations and each population shows different statistical and spatial characteristics. The multi-population characteristics are usually related to a dependent categorical variable (e.g. rock types, soil types, and land uses). Several geostatistical simulation approaches were implemented for spatial modelling of regionalized compositional data with multi-population characteristic in chapter 4. The results proved that the flow anamorphosis is a vital element for geostatistical modelling of regionalized compositional data. Several applications were shown that the transformed data via flow anamorphosis are not only multivariate normal but also exhibit absence of spatial cross-correlation which make the geostatistical simulation of such orthogonal factors, more straightforward. Flow anamorphosis is capable of reproducing complex patterns in input data including presence of outliers, presence of several populations, nonlinearity, and heteroscedasticity. To pursue the capability and usefulness of the geostatistical simulation using flow anamorphosis for resources modelling, it is recommended to implement this technique to other multi-element deposits where several variables with complex statistical and spatial relationships need to be spatially simulated.

Finally, to explore complex compositional patterns and to select and rank significant features (log-contrasts) in a spatial framework, a hybrid spatial predictive model is developed based on the combined use of advanced geostatistical simulation and machine learning algorithms (Random forest in this case). The spatial uncertainty of compositional data was fully incorporated into an ensemble classifier. The estimated probabilities of geological classes associated with each realization of compositional random function were integrated to combine elements of statistical and spatial uncertainties. The new model of spatial uncertainty was used further to predict the most probable geological classes. Due to the high-dimensionality characteristic of log-contrasts, a compositionally compliant feature selection was introduced which is useful for geoscience process discovery analysis. The developed hybrid model was capable to predict surficial and deep earth classes of materials using soil geochemical compositional information with high accuracy.

The proposed method was implemented for two real case studies in chapter 5 and the final results indicated that the generated spatial uncertainty model is consistent with the geological understanding of the phenomenon of interest. The predicted map of geological classes via the proposed hybrid model can be improved further by the proportion correction technique, introduced in chapter 4. This spatial correction technique is especially useful in the situations where one or more classes have low proportions and are dominated by other classes with high proportions.

Appendices

Appendix A Permission of copyrighted material

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jul 18, 2018

This Agreement between Mr. Hassan Talebi ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4391740517445
License date	Jul 18, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	A Hybrid Model for Joint Simulation of High-Dimensional Continuous and Categorical Variables
Licensed Content Author	Hassan Talebi, Johnny Lo, Ute Mueller
Licensed Content Date	Jan 1, 2017
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	>50,000
Author of this Springer Nature content	yes
Title	n/a
Instructor name	n/a
Institution name	Edith Cowan University
Expected presentation date	Aug 2018
Requestor Location	Mr. Hassan Talebi School of Science Edith Cowan University Joondalup, WA 6027 Australia Attn: Mr. Hassan Talebi
Billing Type	Invoice
Billing Address	Mr. Hassan Talebi 56 Regents Park Road Joondalup, Australia 6027 Attn: Mr. Hassan Talebi
Total	0.00 USD

Terms and Conditions

Springer Nature Terms and Conditions for RightsLink Permissions

Springer Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g.

Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj. [JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.0

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Oct 14, 2018

This Agreement between Mr. Hassan Talebi ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4447891037500
License date	Oct 14, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Mathematical Geosciences
Licensed Content Title	Geostatistical Simulation of Geochemical Compositions in the Presence of Multiple Geological Units: Application to Mineral Resource Evaluation
Licensed Content Author	Hassan Talebi, Ute Mueller, Raimon Tolosana-Delgado et al
Licensed Content Date	Jan 1, 2018
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	>50,000
Author of this Springer Nature content	yes
Title	Mr
Institution name	Edith Cowan University
Expected presentation date	Oct 2018
Requestor Location	Mr. Hassan Talebi 56 Regents Park Road Joondalup, WA 6027 Australia Attn: Mr. Hassan Talebi
Billing Type	Invoice
Billing Address	Mr. Hassan Talebi 56 Regents Park Road Joondalup, Australia 6027 Attn: Mr. Hassan Talebi
Total	0.00 AUD
Terms and Conditions	

Springer Nature Terms and Conditions for RightsLink Permissions

Springer Nature Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT]** (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj. [JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix B Statement of co-authors contribution

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Joint simulation of compositional and categorical data via direct sampling technique – Application to improve mineral resource confidence.

Journal: Computer & Geosciences.

Paper status: Under review

List of authors: Hassan Talebi, Ute Mueller, Raimon Tolosana-Delgado

PhD candidate: Hassan Talebi

Scientific contributions to the paper:

The PhD candidate contributed to the article with regards to development of the idea, computer programming, numerical analysis, and manuscript writing constituting 80% of the work.

Ute Mueller contributed to the development of the idea, interpretation of the results, and critical revision of the manuscript (10%).

Raimon Tolosana-Delgado contributed to the development of the idea and the algorithm and revision of the manuscript (10%).

Signature, PhD student

---[redacted]

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Associate Professor Ute Mueller: [redacted]---------- Date: 8/8/2018

Dr Raimon Tolosana-Delgado: [redacted]---------- Date: 10.08.2018

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Book chapter title: A hybrid model for joint simulation of high-dimensional continuous and categorical variables.

Book: Geostatistics Valencia 2016, Springer International Publishing, Cham, pp. 415-430. Editors: J.J. Gómez-Hernández, J. Rodrigo-Ilarri, M.E. Rodrigo-Clavero, E. Cassiraga and J.A. Vargas-Guzmán.

List of authors: Hassan Talebi , Johnny Lo, and Ute Mueller

PhD candidate: Hassan Talebi

Scientific contributions to the paper:

The PhD candidate contributed to the article with regards to development of the idea, numerical analysis, and manuscript writing constituting 70% of the work.

Ute Mueller contributed to the development of the idea, interpretation of the results, and critical revision of the manuscript (20%).

Johnny Lo contributed to the development of the metrics for evaluation of the proposed method and critical revision of the manuscript (10%).

Signature, PhD student

---[redacted]

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Associate Professor Ute Mueller: [redacted]---------- Date: 8/8/2018

Dr Johnny Lo: ----------[redacted]---------- Date: 21/8/2018

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Geostatistical simulation of geochemical compositions in the presence of multiple geological units - Application to mineral resource evaluation.

Journal: Mathematical Geosciences.

Paper status: Accepted for publication, 15/07/2018

List of authors: Hassan Talebi, Ute Mueller, Raimon Tolosana-Delgado, K. Gerald van den Boogaart

PhD candidate: Hassan Talebi

Scientific contributions to the paper:

The PhD candidate contributed to the article with regards to development of the idea, computer programming, numerical analysis, and drafting the manuscript constituting 75% of the work.

Ute Mueller contributed to the development of the idea, interpretation of the results, and critical revision of the manuscript (10%).

Raimon Tolosana-Delgado contributed to the development of the idea and the algorithm and critical revision of the manuscript (10%).

K. Gerald van den Boogaart contributed to the development of the idea and revision of the manuscript (5%).

Signature, PhD student

---[redacted]

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Associate Professor Ute Mueller: [redacted]--------- Date: 8/8/2018

Dr Raimon Tolosana-Delgado: -----[redacted]---------- Date: 10/08/2018

Professor K. Gerald van den Boogaart: -[redacted]---------- Date: 9/8/2018

Co-authorship statement for publications with the PhD

With reference to ECU thesis with publication policy, statements from co-author attesting to the PhD candidate's contribution to the joint publications must be included in the appendix.

Paper title: Surficial and deep earth material prediction from geochemical compositions - A spatial predictive model.

Journal: Natural Resources Research

Paper status: Submitted.

List of authors: Hassan Talebi, Ute Mueller, Raimon Tolosana-Delgado, Eric C Grunsky, Jennifer M. McKinley, Patrice de Caritat.

PhD candidate: Hassan Talebi

Scientific contributions to the paper: The PhD candidate contributed to the article with regards to development of the idea, numerical analysis, and manuscript writing constituting 70% of the work.

Ute Mueller contributed to the development of the idea, interpretation of the results, and critical revision of the manuscript (10%).

Raimon Tolosana-Delgado contributed to the development of the idea and the algorithm and critical revision of the manuscript (5%).

Eric C Grunsky contributed to the preparation of the data and interpretation of the results for the first case study and critical revision of the manuscript (5%).

Patrice de Caritat contributed to the preparation of the data and interpretation of the results for the first case study and critical revision of the manuscript (5%).

Jennifer M. McKinley contributed to the preparation of the data and interpretation of the results for the second case study and critical revision of the manuscript (5%).

Signature, PhD student [redacted]

I, as a co-author, endorse that this level of contribution by the candidate indicated above is appropriate.

Signatures, co-authors

Associate Professor Ute Mueller: [redacted] Date: 08/08/2018

Dr Raimon Tolosana-Delgado: [redacted] Date: 08/08/2018

Professor Eric C Grunsky: [redacted] Date: 27/08/2018

Dr Jennifer M. McKinley: [redacted] Date: 31/7/18.

Dr Patrice de Caritat: [redacted] Date: 27/8/2018.