Set- and Measure-indexed Random Fields and the De Wijs Process

- With Application on Soil Data From Barro Colorado Island in Panama -

Master Thesis

Ib Thorsgaard Jensen

Aalborg University
Department of Mathematics

AALBORG UNIVERSITY

STUDENT REPORT

Title:
Set- and Measure-indexed Random Fields and the De Wijs Process

Theme:
Master Thesis

Project Period:
Spring and Autmn Semester 2019

Participant:
Ib Thorsgaard Jensen

Supervisor:
Rasmus Plenge Waagepetersen

Copies: 1

Page Numbers: 74

Date of Completion:
January 6, 2020

Abstract:

This thesis has the De Wijs process as its primary focus. Theory on generalised functions as well as stochastic processes indexed on test functions, Borel sets, and measures is described. Estimation of the parameters in a De Wijs Plus White Noise (WWN) Process using restricted maximum likelihood is detailed. The covariance matrix of relevant contrasts is constructed, and a data-analysis is carried out. Parameter estimation is performed for the WWN Process on a soil dataset from Barro Colorado Island containing 19 different minerals, and the variogram is introduced and used as a summary statistic. Then kriging for set-indexed random fields is outlined and a 10-fold cross-validation is carried out on the WWN Process fitted to the data. Here it is found that for all of the minerals in the dataset the WWN Process performs better than predicting using the sample mean. Furthermore, on all but one mineral, the WWN process also performs better than an intrinsic random field of order 0 with polynomial generalised covariance. When restricting only to the data on regularly spaced grid, there are a few minerals where the power or exponential model performs better than the WWN process. This suggests that Peter McCullagh's notion of a *loi du terroir* does not extend to forest soil data.

Contents

Preface

The present report is a master thesis, which serves as a graduation project of the master program in mathematics at Aalborg University (AAU). Basic theory from probability theory, statistics and random fields as well as basic statistical programming in R is assumed knowledge. The proofs written in the report are independent work unless a source is explicitly cited.

In Chapter 1 the concept of generalised functions is introduced. In Section 2.1-2.2 stochastic processes with arbitrary index-sets are introduced and Kolmogorov's Theorem is proven. Then, in Section 2.3 and 3.1, stochastic processes indexed by test functions, Borel sets, and measures are introduced, in Section 3.2 the model known as the De Wijs process is introduced, and in Section 3.3 it is described how to estimate the relevant parameters of De Wijs Plus White Noise (WWN) process using restricted maximum likelihood estimation. Then, in Section 4.1, the Barro Colorado Island Dataset is introduced, and in the rest of Chapter 4 and in Section 5.1-5.2 it is analysed using a WWN process, and in Section 5.3 these results are compared to two other models, namely the power model and the exponential model. Finally, in Chapter 6 the results are discussed and the conclusions from the previous chapters are summarised.

Aalborg University, January 6, 2020

Ib Thorsgaard Jensen
<itje14@student.aau.dk>

Chapter 0

Introduction

Traditionally, geostatistical data are modelled using *random fields*, which are collections of random variables indexed on a Euclidean space (typically $\mathbb{R}^2$). However, there is a wide variety of possible index-set that may be utilised, and the Euclidean spaces may not always be the most suitable. In this project we consider an index set of test functions (functions in $\mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R})$ with bounded support), an index set of Borel sets and an index-set of signed measures. The relationships between these index-sets is discussed in Section 3.1. In Chapter 1 we introduce the notion of generalised functions, which serves as a preliminary to generalised stochastic processes introduced in Section 2.3. Generalised stochastic processes are collections of random variables indexed by test functions, which has e.g been used for analysing stochastic differential equations and stochastic integral equations (see Schäffler [2018, Section 4.4] and Dawson [1970, Section 5]).

In Section 3.1 we introduce set-indexed random fields and generalised random fields. The former is a collection of random variables indexed by Borel sets, and the latter is a collection of random variables indexed by signed measures that are zero on $\mathbb{R}^d$. The idea behind set-indexed random fields is that, in reality we cannot sample information on an infinitesimally sized area, so we drop this abstraction, and instead of indexing by points in a Euclidean space, we index on subsets of a Euclidean space representing the whole area where the data are collected. This is indispensable for some datasets, such as crop yield data, where sampling on an infinitesimal point is nonsensical. Generalised random fields are introduced to enable the analysis of contrasts of set-indexed random fields.

In Section 3.2 we introduce the De Wijs process which is a specific model of generalised random field. The De Wijs process has been studied extensively by Peter McCullagh and David Clifford in the application to crop yield data (see Clifford and McCullagh [2006] and Clifford et al. [2006]). Peter McCullagh goes so far as to call the De Wijs Plus White Noise (WWN) process a *loi du terroir* [McCullagh, 2003], or *law of the soil*, since it was found that although the WWN process is the simplest model in its class, it fits as well as the full model on all crops in all climates (see McCullagh [2003] and Clifford and McCullagh [2006, Section 9]). The De Wijs process has also been studied in more theoretical settings. For example, the De Wijs process has been found to be connected to a so-called Recurrent Brownian Motion (see Mondal [2015]), and it has been used to study the relation between geostatistics and Markov Random Fields using First Order Intrinsic Autoregressions (see Besag and Mondal [2005]).

In Section 3.3 we introduce restricted maximum likelihood estimation in order to estimate the parameters of the WWN Process. This is then applied in Chapter 4 and 5, where we use the

WWN process for analysing a soil sample dataset from a rain-forest on Barro Colorado Island in Panama (see Harms et al. [2004]). Here we introduce the variogram and use it as a summary statistic in Section 5.1 and in Section 5.2 we introduce kriging for a set-indexed random field, and use it to evaluate the WWN Process using a 10-fold cross-validation. In Section 5.3 we briefly analyse the data with two other geostatistical models, the power model and the exponential model, in order to compare them to the WWN Process. Finally, in Chapter 6, we compare the results found using the WWN process to the results found by using a so-called *intrinsic random field of order 0* with a polynomial generalised covariance function, as described in Jensen and Fitzhugh [2018].

Chapter 1

Generalised Functions

In this chapter we introduce so-called *Generalised Functions*. Before these can be defined we must first define the so-called *test functions*. The chapter is based on Schäffler [2018, Chapter 1].

1.1 Test Functions

In the following, let $\mathcal{C}_d^\infty$ denote all infinitely differentiable continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. We define the support of such a function, f, by

$$\operatorname{supp}(f) = \operatorname{cl}\Big(\{x \in \mathbb{R}^d : f(x) \neq 0\}\Big),$$

where $\operatorname{cl}(\cdot)$ denotes the closure of a set.

We may now define test functions in the following way.

Definition 1.1 (Test Functions). *A function $\varphi \in \mathcal{C}_d^\infty$ is called a test function, if $\operatorname{supp}(\varphi)$ is bounded. The set of test functions from $\mathbb{R}^d$ to $\mathbb{R}$ is denoted $\mathcal{T}_d$.*

Let $\varphi \in \mathcal{T}_d$, and consider the support when scaling. For $a \in \mathbb{R}$ we get

$$\operatorname{supp}(a\varphi) = \begin{cases} \operatorname{supp}(\varphi), & a \neq 0 \\ \emptyset, & a = 0. \end{cases}$$

Thus $a\varphi \in \mathcal{T}_d$. Now let $\varphi_1, \varphi_2 \in \mathcal{T}_d$, and consider the support of $\varphi_1 + \varphi_2$. We have that

$$\operatorname{supp}(\varphi_1 + \varphi_2) \subseteq \operatorname{supp}(\varphi_1) \cup \operatorname{supp}(\varphi_2),$$

so since both $\operatorname{supp}(\varphi_1)$ and $\operatorname{supp}(\varphi_2)$ are bounded, it follows that $\operatorname{supp}(\varphi_1 + \varphi_2)$ is also bounded, and thus $\varphi_1 + \varphi_2$ is a test function. This, along with other properties which hold trivially, shows that $\mathcal{T}_d$ is a vector space.

We now introduce an important definition regarding the meaning of convergence of a sequence of test functions.

Definition 1.2 (Convergence of Test Functions). *Let $\{\varphi_i\}_{i\in\mathbb{N}}$ be a sequence of test functions. The sequence is said to be convergent to a function $\varphi \in \mathcal{T}_d$, if*

- *The set $\bigcup_{i=1}^{\infty} \text{Supp}(\varphi_i)$ is bounded.*
- *The sequence $\{\varphi_i - \varphi\}$ converges uniformly to the zero function in $\mathbb{R}^d$.*
- *The sequence $\partial^\alpha(\varphi_i - \varphi)$ converges uniformly to the zero function in $\mathbb{R}^d$ for any multi-index, α.*

It turns out that it is possible to approximate any continuous function, with compact support, with a sequence of test functions. To prove this, we first introduce the function $\psi : \mathbb{R}^d \to \mathbb{R}$ defined by,

$$\psi(x) = \begin{cases} 0, & \text{when } 1 \leq \|x\|^2, \\ \exp(-\frac{1}{1-\|x\|^2}), & \text{when } 1 > \|x\|^2, \end{cases}$$

where $\|\cdot\|$ is the Euclidean norm. Note that ψ is symmetric, has bounded support and is infinitely differentiable. Let $I_\psi = \int_{\mathbb{R}^d} \psi(x)\mathrm{d}x < \infty$. Furthermore, define

$$\psi_1(x) = \frac{\psi(x)}{I_\psi} \text{ for } x \in \mathbb{R}^d.$$

We use these functions to define an important function for approximation of continuous functions using test functions,

$$\psi_R(x) = \frac{\psi_1(x/R)}{R^d} \tag{1.1}$$

with $R > 0$. The support of ψ_R is the d-dimensional ball $\bar{B}(0, R)$, thus it is also a test function. Considering the integral of $\psi_R(x)$, using integration by substitution and setting $y = \frac{x}{R}$, we have

$$\int_{\mathbb{R}^d} \psi_R(x)\mathrm{d}x = \int_{\mathbb{R}^d} \frac{\psi_1(\frac{x}{R})}{R^d}\mathrm{d}x = \int_{\mathbb{R}^d} \frac{\psi_1(y)}{R^d} \det\left(J(y)\right) \mathrm{d}y = 1, \tag{1.2}$$

since $J(y)$ is the Jacobian matrix of the function $x = yR$, which is just an identity matrix scaled by R.

We can now formulate a theorem about the approximation of continuous functions, with compact support, using test functions.

Theorem 1.3 (The approximation theorem). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function with compact support. For each $\varepsilon > 0$, there exists a function $\varphi \in \mathcal{T}_d$, such that,*

$$|f(x) - \varphi(x)| < \varepsilon, \text{ for all } x \in \mathbb{R}^d.$$

Specifically, when defining for any $R > 0$ that

$$\varphi_R(x) = \int_{\mathbb{R}^d} f(u)\psi_R(u - x)\mathrm{d}u,$$

it holds that $\varphi_R \to f$ uniformly for $R \to 0$.

Proof. This proof originates in [Schäffler, 2018, proof of Theorem 1.2].

Since f is continuous with compact support, it is also uniformly continuous, meaning that for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$|f(x) - f(y)| < \varepsilon, \text{ for all } x, y \in \mathbb{R}^d \text{ with } \|x - y\| < \delta.$$

Additionally we have that

$$\begin{aligned}\varphi_R(x) &= \int_{\mathbb{R}^d} f(u)\psi_R(u - x)\mathrm{d}u = \int_{\bar{B}(x,R)} f(u)\psi_R(u - x)\mathrm{d}u \\ &= f(y) \int_{\bar{B}(x,R)} \psi_R(u - x)\mathrm{d}u = f(y) \text{ for some } y \in \bar{B}(x, R).\end{aligned}$$

where the second equality is due to the fact that the support of $\psi_R(u - x)$ is $\bar{B}(x, R)$, and the third equality is from the mean value theorem of integrals (see e.g Amann and Escher [2008, Ch. 6.4]). The last equality is due to (1.2). With $R < \delta$, we have

$$|f(x) - \varphi_R(x)| = |f(x) - f(y)| < \varepsilon, \text{ because } \|x - y\| \leq R < \delta.$$

Thus as $R \to 0$ the function $\varphi_R \to f$ uniformly, proving the result. □

For an example of this, consider the function

$$f(x) = \begin{cases} 1 - \frac{1}{2}|x|, & \text{when } |x| < 2, \\ 0, & \text{when } |x| \geq 2. \end{cases} \tag{1.3}$$

Since $f(x)$ is continuous with compact support, we may approximate it using φ_R as defined in Theorem 1.3. A plot of $f(x)$ can be seen in Figure 1.1. In the plots in Figure 1.2 and Figure 1.3, we can see that φ_R approximates $f(x)$ better for the lower value of R. Note that the integral that defines φ_R has been approximated numerically, resulting in the rather jagged shape in Figure 1.2. For large values of R the function φ_R is a smoothed version of $f(x)$.

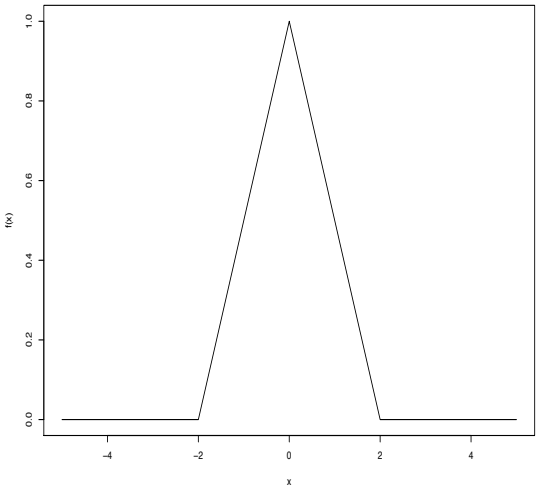

Figure 1.1: Plot of the function $f(x)$ as defined in (1.3)

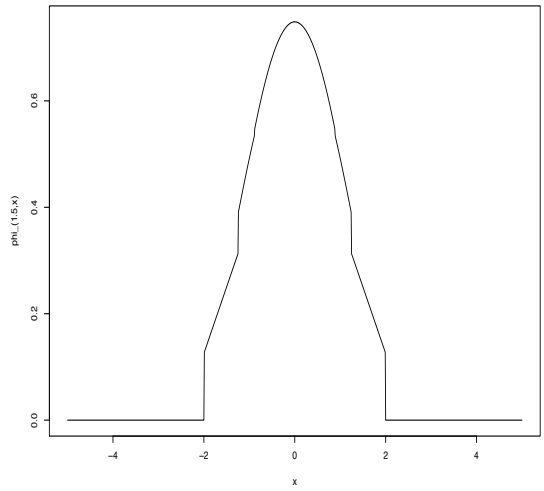

Figure 1.2: Plot of the function $\varphi_{1.5}(x)$.

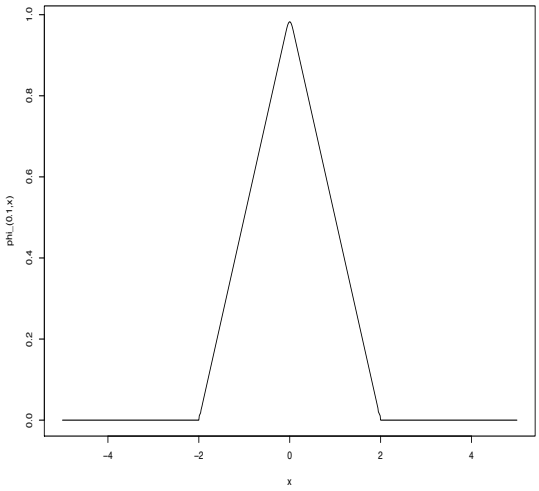

Figure 1.3: Plot of the function $\varphi_{0.1}(x)$.

Test functions are of central importance in the next section, where so-called functionals are introduced.

1.2 Functionals

In this section we introduce the notion of functionals. We define the term in a very general setting. Before defining functionals, we note that $\mathcal{A}$ being a vector space over a field $\mathcal{B}$, means that, among other things, $\mathcal{A}$ is closed under multiplication with elements of $\mathcal{B}$.

Definition 1.4 (Functional). *For a given vector space $\mathcal{A}$ over a field, $\mathcal{B}$, a mapping*

$$F : \mathcal{A} \to \mathcal{B},$$

is called a functional.

Furthermore, a functional F is said to be linear if for all $a \in \mathcal{B}$, and $x, y \in \mathcal{A}$,

$$F(ax) = aF(x)$$
$$F(x+y) = F(x) + F(y).$$

With functionals defined in a general setting, we turn our attention to a specific functional. For this purpose, let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally integrable function, meaning that $\int_A |f(x)|\mathrm{d}x < \infty$ for any bounded set A. For each such f, we may define the mapping $F_f : \mathcal{T}_d \to \mathbb{R}$, by

$$F_f(\varphi) = \int_{\mathbb{R}^d} f(x)\varphi(x)\mathrm{d}x. \tag{1.4}$$

Since $\mathcal{T}_d$ is a vector space over $\mathbb{R}$, as shown in Section 1.1, the mapping F_f defines a functional. Furthermore, F_f is a linear functional, since for $a \in \mathbb{R}$ and $\varphi_1, \varphi_2 \in \mathcal{T}_d$ we have

$$F_f(a\varphi) = \int_{\mathbb{R}^d} af(x)\varphi(x)\mathrm{d}x = aF_f(\varphi)$$
$$F_f(\varphi_1 + \varphi_2) = \int_{\mathbb{R}^d} f(x)\left(\varphi_1(x) + \varphi_2(x)\right)\mathrm{d}x = F_f(\varphi_1) + F_f(\varphi_2).$$

In the following, we describe how a functional defined as in (1.4) can be used to give meaning to the derivative of a function that are not differentiable. Suppose $f : \mathbb{R} \to \mathbb{R}$ is locally integrable and differentiable. Suppose furthermore that the derivative f' is locally integrable, then for any $\varphi \in \mathcal{T}_1$ we get, using integration by parts, that

$$\int_{\mathbb{R}} f'(x)\varphi(x)\mathrm{d}x = [f(x)\varphi(x)]_{-\infty}^{\infty} - \int_{\mathbb{R}} f(x)\varphi'(x)\mathrm{d}x. \tag{1.5}$$

Since $\varphi \in \mathcal{T}_1$, the set $\mathrm{supp}(\varphi)$ is bounded, and thus the first term in (1.5) is zero, and

$$\int_{\mathbb{R}} f'(x)\varphi(x)\mathrm{d}x = -\int_{\mathbb{R}} f(x)\varphi'(x)\mathrm{d}x, \tag{1.6}$$

meaning $F_{f'}(\varphi) = -F_f(\varphi')$. In the above, the left hand side is defined for differentiable functions f, but the right hand side is defined for any locally integrable functions f.

Now let $\varphi(u) = \psi_{R,x}(u)$, where $\psi_{R,x}(u)$ is defined as

$$\psi_{R,x}(u) = \psi_R(u - x), \tag{1.7}$$

where ψ_R is defined as in (1.1). For any differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ we have

$$F_g(\psi_{R,x}) = \int_{\mathbb{R}^d} g(u)\psi_{R,x}(u)\mathrm{d}u = \int_{\bar{B}(x,R)} g(u)\psi_{R,x}(u)\mathrm{d}u$$
$$= g(y)\int_{\bar{B}(x,R)} \psi_{R,x}(u)\mathrm{d}u = g(y)$$

for some $y \in \bar{B}(x, R)$, where the second to last equality is a result of the mean value theorem for integrals. Taking the limit as $R \to 0$, we have that,

$$\lim_{R\to 0} F_g(\psi_{R,x}) = g(x), \text{ for } x \in \mathbb{R}^d, \tag{1.8}$$

which can be used to give meaning to the derivative of non-differentiable locally integrable functions in the following way. Once again let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable, locally integrable function. Then by (1.6) and (1.8) we get that

$$\lim_{R\to 0} -F_f(\psi'_{R,x}) = \lim_{R\to 0} F_{f'}(\psi_{R,x}) = f'(x), \tag{1.9}$$

where we see that $\lim_{R\to 0} -F_f(\psi'_{R,x})$ is well defined as long as f is locally integrable, even if it is not differentiable.

We now introduce the notion of continuity of functionals from the set of test functions to the real numbers.

Definition 1.5 (Continuity of Functionals). *A functional $F : \mathcal{T}_d \to \mathbb{R}$ is said to be continuous if*

$$\lim_{i\to\infty} F(\varphi_i) = F(\varphi),$$

for all sequences $\{\varphi_i\}$ which converge to the function $\varphi \in \mathcal{T}_d$ (see Definition 1.2).

Using Definition 1.5 it can be shown that functionals on the form in (1.4) are continuous.

Proposition 1.6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally integrable function. A linear functional $F_f : \mathcal{T}_d \to \mathbb{R}$ on the form,*

$$F_f(\varphi) = \int_{\mathbb{R}^d} f(x)\varphi(x)\mathrm{d}x,$$

is continuous.

Proof. This proof follows the arguments presented in [Schäffler, 2018, p. 13-14].

The zero function, $\varphi(x) = 0$ for all $x \in \mathbb{R}^d$ is a test function, as

$$\mathrm{supp}(\varphi) = \emptyset,$$

which is closed and bounded. Thus the functional F is continuous if,

$$\lim_{i\to\infty} F(\varphi_i) = 0,$$

for all sequences of test functions which converge to the zero function (see Definition 1.2).

Now let $\{\varphi_i\}_{i\in\mathbb{N}}$ be sequence of test functions that converges to the zero function. Then for some $\varepsilon > 0$

$$\mathrm{supp}(\varphi_i) \subseteq B(0,\varepsilon), \text{ for all } i \in N.$$

Using this fact, we have that

$$\begin{aligned}
\lim_{i\to\infty} |F_f(\varphi_i)| &= \lim_{i\to\infty} \left| \int_{\mathbb{R}^d} f(x)\varphi_i(x)\mathrm{d}x \right| = \lim_{i\to\infty} \left| \int_{B(0,\varepsilon)} f(x)\varphi_i(x)\mathrm{d}x \right| \\
&\leq \lim_{i\to\infty} \int_{B(0,\varepsilon)} |f(x)||\varphi_i(x)|dx \leq \lim_{i\to\infty} \left(\sup_{x\in B(0,\varepsilon)} |\varphi_i(x)| \int_{B(0,\varepsilon)} |f(x)|\mathrm{d}x \right), \\
&= \left(\int_{B(0,\varepsilon)} |f(x)|\mathrm{d}x \right) \lim_{i\to\infty} \left(\sup_{x\in B(0,\varepsilon)} |\varphi_i(x)| \right) = 0,
\end{aligned}$$

where the last inequality results from test functions being bounded, and the last equality from $|\varphi_i| \to 0$ as $i \to \infty$, proving the result. □

Next we see how functionals lead to the definition of so-called generalised functions.

1.3 Generalised Functions

As discussed in the previous section we can define a continuous and linear functional for a locally integrable function f, from the space of test functions to the real numbers. This particular functional is a so-called generalised function, which we now formally define on the space of test functions.

Definition 1.7 (Generalised Functions from $\mathcal{T}_d$). *A linear and continuous functional $F : \mathcal{T}_d \to \mathbb{R}$ is called a generalised function. Furthermore, if there exists a locally integrable function, f, such that $F = F_f$, then F is called a regular generalised function.*

The set of generalised function from $\mathcal{T}_d$ is denoted $\mathcal{G}_d$. Defining the scalar multiple, and addition of generalised function as

$$aF(\varphi) = F(a\varphi), \text{ for } a \in \mathbb{R}, \varphi \in \mathcal{T}_d, F \in \mathcal{G}_d$$
$$(F + G)(\varphi) = F(\varphi) + G(\varphi), \text{ for } \varphi \in \mathcal{T}_d, F, G \in \mathcal{G}_d$$

the space $\mathcal{G}_d$ is a vector space over $\mathbb{R}$, as all the other requirements are trivially satisfied. The derivative of a generalised function may be defined analogously with (1.6) as follows.

Definition 1.8 (Derivative of Generalised Functions). *Let $F : \mathcal{T}_1 \to \mathbb{R}$ be a generalised function. Then F' is said to be a the derivative of F if*

$$F'(\varphi) = -F(\varphi').$$

We note that the above definition implies that regular generalised functions satisfy $(F_f)'(\varphi) = F_{f'}(\varphi)$ when f is differentiable.

The so-called Dirac distribution[1], δ_{x_0}, is an example of a generalised function. The Dirac distribution is defined as

$$\delta_{x_0}(\varphi) = \varphi(x_0), \text{ for } \varphi \in \mathcal{T}_d \text{ and } x_0 \in \mathbb{R}^d.$$

This is an example of a generalised function, which is not a regular generalised function. We formulate this minor result as a proposition.

Proposition 1.9. *The Dirac distribution, $\delta_{x_0}(\varphi) = \varphi(x_0)$, is not a regular generalised function.*

Proof. This proof follows the arguments presented in [Schäffler, 2018, p. 16-17].

We prove this by contradiction, so suppose that there exists a locally integrable function, f, such that

$$\int_{\mathbb{R}^d} f(x)\varphi(x)\mathrm{d}x = \varphi(x_0), \text{ for all } \varphi \in \mathcal{T}_d, \tag{1.10}$$

[1] Some authors prefer to refer to generalised functions as 'distributions', however there is no relation to cumulative distribution functions.

that is, we assume that the Dirac distribution is a regular generalised function. Since f is locally integrable, then for all $x \in \mathbb{R}^d$ and all $\varepsilon > 0$

$$\int_{\bar{B}(x,\varepsilon)} |f(u)| \mathrm{d}u = d_x < \infty.$$

Thus, for all $x \in \mathbb{R}^d$, there exists an ε_x sufficiently small that $d_x < 1$. Consider now the function ψ_{ε,x_0} defined as in (1.7), which is a test function for all $\varepsilon > 0$. Thus by the assumption in (1.10), we must have,

$$\int_{\mathbb{R}^d} f(x)\psi_{\varepsilon,x_0}(x)\mathrm{d}x = \psi_{\varepsilon,x_0}(x_0).$$

However,

$$\begin{aligned}
\int_{\mathbb{R}^d} f(x)\psi_{\varepsilon_{x_0},x_0}(x)\mathrm{d}x \leq \left|\int_{\mathbb{R}^d} f(x)\psi_{\varepsilon_{x_0},x_0}(x)\mathrm{d}x\right| &\leq \int_{\mathbb{R}^d} |f(x)\psi_{\varepsilon_{x_0},x_0}(x)|\mathrm{d}x \\
&\leq \sup_{\bar{B}(x_0,\varepsilon_{x_0})}|\psi_{\varepsilon_{x_0},x_0}(x)| \int_{\bar{B}(x_0,\varepsilon_{x_0})} |f(x)|\mathrm{d}x \\
&= \psi_{\varepsilon_{x_0},x_0}(x_0)d_{x_0} < \psi_{\varepsilon_{x_0},x_0}(x_0),
\end{aligned}$$

where third inequality comes from the fact that $|f(x)\psi_{\varepsilon,x_0}(x)| \leq |f(x)|\sup_{\bar{B}(x_0,\varepsilon)}|\psi_{\varepsilon,x_0}(x)|$ for all $x \in \mathbb{R}^d$, and the final inequality comes from the fact that $d_{x_0} < 1$ by the choice of ε_{x_0}. This proves the result. □

Chapter 2

Stochastic Processes

2.1 Preliminaries

This section is based on Schäffler [2018, Section 2.1] and Billingsley [1995, Section 36].

In this section we define a stochastic process in a very general setting.

Definition 2.1 (Stochastic Processes). *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $(\Gamma, \mathcal{G})$ be a measurable space and let I be some non-empty set. Then*

$$Z : \Omega \times I \to \Gamma$$

is called a stochastic process with index-set I, if $Z(\cdot, i)$ is an $\mathcal{F} - \mathcal{G}$ measurable function for any fixed $i \in I$.

For a fixed $\omega \in \Omega$ we call the mapping $i \mapsto Z(\omega, i)$ the *path* of Z, and if this function is continuous for all $\omega \in \Omega$ we say that Z has continuous paths (or almost surely continuous paths if it only holds for almost all $\omega \in \Omega$). We usually suppress the dependence on ω and denote $Z(\omega, i)$ simply as $Z(i)$. Furthermore, in this report we only consider the case where $(\Gamma, \mathcal{G}) = (\mathbb{R}, \mathbb{B})$, where $\mathbb{B}$ is the Borel σ-algebra.

Note that the notion of stochastic processes in Definition 2.1 is a very general one. Often the term stochastic process is used to refer only to the case where the index-set is a subset of $\mathbb{R}$. A stochastic process where the index-set is some $D \subseteq \mathbb{R}^d$ is called a random field. Thus a stochastic process is usually thought of as a special case of a random field, where $d = 1$, but with Definition 2.1 a random field is a special case of a stochastic process.

In order to discuss the distributional behaviour of stochastic processes, the notion of finite dimensional distributions is defined. Note that we denote the d-dimensional Borel σ-algebra as $\mathbb{B}_d$.

Definition 2.2 (Finite-dimensional Distribution). *Let Z be a stochastic process with index-set I and let $i_1, \ldots, i_n \in I$. Then the measure, $\mu_{i_1,\ldots,i_n}$, satisfying*

$$\mu_{i_1,\ldots,i_n}(A) = P\left(\Big(Z(i_1), \ldots, Z(i_n)\Big) \in A\right),$$

where $A \in \mathbb{B}_n$, is said to be a finite-dimensional distribution of Z.

From the above definition, two consistency properties arise. First suppose we have a Borel set on the form $A = A_1 \times \cdots \times A_n$ and a permutation π. Then we have the symmetry property, that is $\mu_{i_1,\ldots,i_n}$ is invariant under permutation, since

$$\begin{aligned}\mu_{i_1,\ldots,i_n}(A_1 \times \cdots \times A_n) &= P\left(Z(i_1) \in A_1, \ldots, Z(i_n) \in A_n\right) \\ &= P\left(Z(i_{\pi(1)}) \in A_{\pi(1)}, \ldots, Z(i_{\pi(n)}) \in A_{\pi(n)}\right) \\ &= \mu_{i_{\pi(1)},\ldots,i_{\pi(n)}}(A_{\pi(1)} \times \cdots \times A_{\pi(n)}). \end{aligned} \tag{2.1}$$

Secondly we have the compatibility property. That is, since, by definition, $Z(\omega, i_n) \in \mathbb{R}$ for all $\omega \in \Omega$, we have

$$\begin{aligned}\mu_{i_1,\ldots,i_n}(A_1 \times \cdots \times A_{n-1} \times \mathbb{R}) &= P\left(Z(i_1) \in A_1, \ldots, Z(i_{n-1}) \in A_{n-1}, Z(i_n) \in \mathbb{R}\right) \\ &= P\left(Z(i_1) \in A_1, \ldots, Z(i_{n-1}) \in A_{n-1}\right) \\ &= \mu_{i_1,\ldots,i_{n-1}}(A_1 \times \cdots \times A_{n-1}). \end{aligned} \tag{2.2}$$

As we see above all finite dimensional distributions of a stochastic Process, Z, satisfy (2.1) and (2.2). The question is, if the converse is also the case. That is, for a class of probability measures $\{\mu_{i_1,\ldots,i_n}\}_{i_1,\ldots,i_n \in I}$, is there a stochastic process, Z, with index-set I over a probability space $(\Omega, \mathcal{F}, P)$, such that

$$\mu_{i_1,\ldots,i_n}(A) = P\left((Z(i_1), \ldots, Z(i_n)) \in A\right)$$

for any $\{i_1, \ldots, i_n\} \subseteq I$ and any $A \in \mathcal{F}$? It turns out that the answer is yes. This is a result known as *Kolmogorov's Theorem*, but before we can give the proof, a considerable amount of preparation is necessary. Firstly we note that, in fact, the symmetry and compatibility conditions can be consolidated into one requirement, as we see in the following.

Let π be a permutation and define a mapping $\xi_\pi : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\xi_\pi(x_1, \ldots, x_n) = \left(x_{\pi^{-1}(1)}, \ldots, x_{\pi^{-1}(n)}\right).$$

Note that since the inverse of a permutation is also a permutation, ξ_π is simply a function that permutes the coordinates. Since $\xi_\pi^{-1}(A_1 \times \cdots \times A_n) = A_{\pi(1)} \times \cdots \times A_{\pi(n)}$, it follows that (2.1) can be expresses as $\mu_{i_1,\ldots,i_n}(A) = \mu_{i_{\pi(1)},\ldots,i_{\pi(n)}}\left(\xi_\pi^{-1}(A)\right)$. Now define $\rho_p : \mathbb{R}^n \to \mathbb{R}^{n-p}$ as projection, that is

$$\rho_p(x_1, \ldots, x_n) = (x_1, \ldots, x_{n-p}).$$

Since it follows that $\rho_p^{-1}(A_1 \times \cdots \times A_{n-p}) = A_1 \times \cdots \times A_{n-p} \times \mathbb{R}^p$, we get that for $A = A_1 \times \cdots \times A_{n-1}$, (2.2) can be expressed as $\mu_{i_1,\ldots,i_{n-1}}(A) = \mu_{i_1,\ldots,i_n}\left(\rho_1^{-1}(A)\right)$.

Now suppose we have $i_1, \ldots, i_n \in I$ and $t_1, \ldots, t_m \in I$ such that $m \geq n$ and $\{i_1, \ldots, i_n\} \subset \{t_1, \ldots, t_m\}$. Then $(i_1, \ldots, i_n)$ must be the first n components of some permutation of $(t_1, \ldots, t_m)$, and thus there exists a permutation, π, such that $t_{\pi^{-1}(1)} = i_1, \ldots, t_{\pi^{-1}(n)} = i_n$. Now define $\psi_{\pi,p} : \mathbb{R}^m \to \mathbb{R}^{m-p}$ such that

$$\psi_{\pi,p}(x_1, \ldots, x_m) = \rho_p\left(\xi_\pi(x_1, \ldots, x_m)\right) = \left(x_{\pi^{-1}(1)}, \ldots, x_{\pi^{-1}(m-p)}\right)$$

Thus we have that $\psi_{\pi,p}^{-1}(A) = \xi_\pi^{-1}\left(\rho_p^{-1}(A)\right)$. Thus the condition

$$\mu_{i_1,\ldots,i_n}(A) = \mu_{t_1,\ldots,t_m}\left(\psi_{\pi,m-n}^{-1}(A)\right), \tag{2.3}$$

implies both (2.1) and (2.2).

2.2 Kolmogorov's Theorem

This section is based on Billingsley [1995, Section].

In this section we seek to prove Kolmogorov's Theorem. In Billingsley [1995] two proofs are given. The one we present here is a proof by construction, so before we can give it, a probability space must be constructed. To this end we introduce the concept of product spaces.

Definition 2.3 (Product Space). *Let I be a non-empty set. Then the set of all functions on the form $f : I \to \mathbb{R}$ is called the product space of I and is denoted $\mathbb{R}^I$.*

We note that if $I = \{1, \dots, n\}$, then there is a one-to-one correspondence between $\mathbb{R}^I$ and $\mathbb{R}^n$. As such all Euclidean spaces are, in some sense, a special case of this general setting, and thus when I is infinite, $\mathbb{R}^I$ can be seen as an infinite-dimensional Euclidean space.

Now let $\zeta : \mathbb{R}^I \times I \to \mathbb{R}$ be defined as

$$\zeta(x, i) = x(i) \tag{2.4}$$

and denote $\zeta_i(\cdot) := \zeta(\cdot, i)$. Such functions are referred to as *coordinate functions* and are important in the proof of Kolmogorov's Theorem. Let $\mathbb{B}_I$ be the σ-algebra generated by all coordinate functions, that is $\mathbb{B}_I$ is the smallest σ-algebra such that all coordinate functions are measurable. In other words, $\mathbb{B}_I$ is generated by the sets on the form

$$\zeta_i^{-1}(A) = \{x \in \mathbb{R}^I : \zeta(x, i) \in A\} = \{x \in \mathbb{R}^I : x(i) \in A\}, \tag{2.5}$$

where $i \in I$ and $A \in \mathbb{B}$.

The sets included in $\mathbb{B}_I$ can be quite irregular, and in practice it is not always necessary to consider all of $\mathbb{B}_I$. In fact, throughout most of this section, we only need to examine so-called *finite dimensional sets*.

Definition 2.4 (Finite-dimensional Sets). *Let I be a non-empty set, and let A be a set on the form*

$$A = \left\{x \in \mathbb{R}^I : (\zeta(x, i_1), \dots, \zeta(x, i_n)) \in B\right\} = \left\{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_n)) \in B\right\}. \tag{2.6}$$

where $i_1, \dots, i_n \in I$ are distinct and $B \in \mathbb{B}_n$. Then A is called a finite-dimensional set.

The class of all finite-dimensional sets is denoted $\mathbb{B}_I^0$. Clearly a set on the form in (2.5) is a finite-dimensional set, so it would be tempting to conclude that $\mathbb{B}_I$ is a subset of $\mathbb{B}_I^0$. However this is not, in general, the case, since $\mathbb{B}_I$ is *generated* by sets on the form in (2.5), whereas $\mathbb{B}_I^0$ merely contain the finite-dimensional sets. In fact, when I is infinite, $\mathbb{B}_I^0$ is not a σ-algebra. To see this, let $\{i_k\}_{k\in\mathbb{N}}$ be a sequence of elements in I without repetitions and let $\{C_k\}_{k\in\mathbb{N}}$ be a sequence of non-empty one-dimensional Borel sets, where $C_k \neq \mathbb{R}$ for all $k \in \mathbb{N}$ and $\bigcap_{k=1}^{\infty} C_k = \emptyset$. Then let

$$A_k = \left\{x \in \mathbb{R}^I : x(i_k) \in C_k\right\}.$$

Now note that for any $n \in \mathbb{N}$ we have

$$\bigcap_{k=1}^{n} A_k = \left\{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_n)) \in C_1 \times \cdots \times C_n\right\}. \tag{2.7}$$

Thus we see that $x \in \bigcap_{k=1}^{n} A_k$ implies that $x(i_k) \in C_k$ for $k = 1, \ldots, n$. Thus $x \in \bigcap_{k=1}^{\infty} A_k$ must imply that $x(i_k) \in C_k$ for all $k \in \mathbb{N}$. Now suppose there exists a Borel set, D, and $j_1, \ldots, j_n \in I$ such that

$$\bigcap_{k=1}^{\infty} A_k = \left\{ x \in \mathbb{R}^I : (x(j_1), \ldots, x(j_n)) \in D \right\} =: B. \tag{2.8}$$

But then there must exist some $m \in \mathbb{N}$ such that $i_m \notin \{j_1, \ldots, j_n\}$ (in fact there must be an infinite number of such m). Now note, that the only requirement for a function, $x : I \to \mathbb{R}$, to be in B concerns is its behaviour on $j_1, \ldots, j_n$ and there are no requirements about which values x may attain on $I \backslash \{j_1, \ldots, j_n\}$. Thus for any non-empty one-dimensional Borel set, E, there exists an $x \in B$ such that $x(i) \in E$ for some $i \in I$ where $i \notin \{j_1, \ldots, j_n\}$. Then, specifically, there exists an $x \in B$ such that $x(i_m) \in \mathbb{R} \backslash C_m$. That is $x(i_m) \notin C_m$, which means that

$$x \in B, \text{ but } x \notin \bigcap_{k=1}^{\infty} A_k,$$

which is a contradiction, and thus $\mathbb{B}_I^0$ is not closed under countable intersections, and thus it is not a σ-algebra.

While $\mathbb{B}_I$ is not, in general, a subset of $\mathbb{B}_I^0$, it turns out that the reverse is the case.

Lemma 2.5. *Let I be a non-empty set. Then $\mathbb{B}_I^0 \subseteq \mathbb{B}_I$.*

Proof. Let

$$A = \left\{ x \in \mathbb{R}^I : (x(i_1), \ldots, x(i_n)) \in C \right\},$$

where $C \in \mathbb{B}_n$ and $i_1, \ldots, i_n \in I$. To show the result, we must show that all finite-dimensional sets can be generated by sets on the form seen in (2.5). We already saw that this holds when $C = C_1 \times \cdots \times C_n$, where $C_k \in \mathbb{B}$ for $k = 1, \ldots, n$ (see (2.7)). Then it must also hold when C_k, $k = 1, \ldots, n$ are all intervals. Now note that all n-dimensional Borel sets can be constructed from n-dimensional intervals using countable unions, countable intersections or complements. So now let $\{A_k\}_{k \in \mathbb{N}}$ be a sequence of finite-dimensional sets such that

$$A_k = \left\{ x \in \mathbb{R}^I : (x(i_1), \ldots, x(i_n)) \in C_k \right\},$$

where $\{C_k\}_{k \in \mathbb{N}}$ is a sequence of Borel sets. Then

$$\bigcap_{k=1}^{\infty} A_k = \left\{ x \in \mathbb{R}^I : (x(i_1), \ldots, x(i_n)) \in \bigcap_{k=1}^{\infty} C_k \right\}, \tag{2.9}$$

and

$$\bigcup_{k=1}^{\infty} A_k = \left\{ x \in \mathbb{R}^I : (x(i_1), \ldots, x(i_n)) \in \bigcup_{k=1}^{\infty} C_k \right\}. \tag{2.10}$$

Furthermore if

$$A = \left\{ x \in \mathbb{R}^I : (x(i_1), \ldots, x(i_n)) \in C \right\},$$

then

$$\mathbb{R}^I \backslash A = \left\{ x \in \mathbb{R}^I : (x(i_1), \ldots, x(i_n)) \in \mathbb{R}^n \backslash C \right\}. \tag{2.11}$$

From (2.9), (2.10) and (2.11) we see that countable unions, countable intersections or complements on finite-dimensional sets amounts to performing said operations to the underlying Borel sets. Thus all finite-dimensional sets can be generated by sets on the form

$$A = \left\{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_n)) \in C_1 \times \cdots \times C_n\right\} = \bigcap_{k=1}^{n} \{x \in \mathbb{R}^I : x(i_k) \in C_k\},$$

where $C_1, \dots, C_n$ are intervals, and therefore $\mathbb{B}_I^0 \subseteq \mathbb{B}_I$. □

The above lemma allows us to derive a more useful relationship between between $\mathbb{B}_I^0$ and $\mathbb{B}_I$.

Lemma 2.6. *Let I be a non-empty set. If I is infinite, then $\mathbb{B}_I^0$ is an algebra and $\sigma(\mathbb{B}_I^0) = \mathbb{B}_I$. If $I = \{i_1, \dots, i_n\}$ for some $n \in \mathbb{N}$ then $\mathbb{B}_I^0 = \mathbb{B}_I$.*

Proof. This proof is a more detailed version of the arguments in Billingsley [1995, p. 485-486].

Due to Lemma 2.5 it is obvious that $\sigma(\mathbb{B}_I^0) = \mathbb{B}_I$.

First we show that $\mathbb{B}_I^0$ is an algebra. In order to do this, we need to show that

(i) $\mathbb{R}^I \in \mathbb{B}_I^0$

(ii) $A \in \mathbb{B}_I^0$ implies that $\mathbb{R}^I \backslash A \in \mathbb{B}_I^0$

(iii) $A, B \in \mathbb{B}_I^0$ implies that $A \cup B \in \mathbb{B}_I^0$.

For any $i \in I$ we have that

$$\mathbb{R}^I = \{x \in \mathbb{R}^I : x(i) \in \mathbb{R}\},$$

and since $\mathbb{R} \in \mathbb{B}$, $\mathbb{R}^I$ is a finite-dimensional set, proving (i).

Now let $A = \{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_n)) \in C\}$ be a finite-dimensional set. Then

$$\mathbb{R}^I \backslash A = \{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_n)) \in \mathbb{R}^n \backslash C\},$$

proving (ii).

To prove (iii), suppose that

$$A = \left\{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_n) \in C)\right\},$$

and

$$B = \left\{x \in \mathbb{R}^I : (x(j_1), \dots, x(j_m) \in D)\right\}.$$

Then choose $t_1, \dots, t_r \in I$ such that $\{t_1, \dots, t_r\} = \{i_1, \dots, i_n\} \cup \{j_1, \dots, j_m\}$. Then there exists permutations π_1 and π_2 such that

$$t_{\pi_1^{-1}(1)} = i_1, \dots, t_{\pi_1^{-1}(n)} = i_n, \quad t_{\pi_2^{-1}(1)} = j_1, \dots, t_{\pi_2^{-1}(m)} = i_m.$$

Then we have that $(x(i_1), \dots, x(i_n)) \in C$ if and only if $(x(t_1), \dots, x(t_r)) \in \psi_{\pi_1, r-m}^{-1}(C)$, where $\psi_{\pi_1, r-m}$ is defined as in Section 2.1, and $(x(j_1), \dots, x(j_m)) \in D$ if and only if $(x(t_1), \dots, x(t_r)) \in \psi_{\pi_2, r-n}^{-1}(D)$. Thus

$$\begin{aligned} A \cup B &= \left\{x \in \mathbb{R}^I : (x(t_1), \dots, x(t_r)) \in \psi_{\pi_1, r-m}^{-1}(C)\right\} \cup \left\{x \in \mathbb{R}^I : (x(t_1), \dots, x(t_r)) \in \psi_{\pi_2, r-n}^{-1}(D)\right\} \\ &= \left\{x \in \mathbb{R}^I : (x(t_1), \dots, x(t_r)) \in \psi_{\pi_1, r-m}^{-1}(C) \cup \psi_{\pi_2, r-n}^{-1}(D)\right\}, \end{aligned}$$

which is a finite dimensional set, which proves (iii).

Now let $I = \{i_1, \dots, i_n\}$ for some $n \in \mathbb{N}$, and let $\{i_k\}_{k\in\mathbb{N}}$ be a sequence of elements in I and let $\{A_k\}_{k\in\mathbb{N}}$ be a sequence of Borel sets. Obviously $\{i_k\}_{k\in\mathbb{N}}$ must have repetitions, so let $N_l = \{k \in \mathbb{N} : i_k = i_l\}$. Then

$$\begin{aligned}
\bigcap_{k=1}^{\infty} \{x \in \mathbb{R}^I : x(i_k) \in A_k\} &= \bigcap_{l=1}^{n} \left(\bigcap_{k\in N_l} \{x \in \mathbb{R}^I : x(i_l) \in A_k\} \right) \\
&= \bigcap_{l=1}^{n} \left\{ x \in \mathbb{R}^I : x(i_l) \in \bigcap_{k\in N_l} A_k \right\} \\
&= \left\{ x \in \mathbb{R}^I : (x(i_1), \dots, x(i_n)) \in \bigcap_{k\in N_1} A_k \times \cdots \times \bigcap_{k\in N_n} A_k \right\},
\end{aligned}$$

and since $\bigcap_{k\in N_1} A_k \times \cdots \times \bigcap_{k\in N_n} A_k \in \mathbb{B}_n$, this is a finite-dimensional set. Since $\mathbb{B}_I^0$ is an algebra, which is closed under countable intersections, it is also a σ-algebra, and since $\sigma(\mathbb{B}_I^0) = \mathbb{B}_I$ it follows that $\mathbb{B}_I^0 = \mathbb{B}_I$. □

In the proof of Kolmogorov's Theorem, we construct a stochastic process on a probability space with $\mathbb{B}_I$ as a σ-algebra. To construct the probability measure, we use the above lemma and the following theorem.

Theorem 2.7. *Let $\mathcal{G}$ be an algebra, let P be probability measure on $\mathcal{G}$ and let $\mathcal{F}$ be the σ-algebra generated by $\mathcal{G}$. Then there exists a unique probability measure, $\tilde{P}$, on $\mathcal{F}$ which satisfies*

$$\tilde{P}(A) = P(A),$$

for all $A \in \mathcal{G}$.

The theorem above is a major result, the proof of which is rather involved. It can be found in Billingsley [1995, Section 3], but it is beyond the scope of this project. Now we have nearly laid all the necessary groundwork to prove Kolmogorov's Theorem. The last thing we need is a technical result, here stated as a lemma.

Lemma 2.8. *Let μ be a measure on $\mathbb{B}_n$ where $\mu(A) < \infty$ for all bounded $A \in \mathbb{B}_n$. Then the following statements hold*

(i) For any $A \in \mathbb{B}_n$ and $\varepsilon > 0$, there exists a closed set $C \in \mathbb{B}_n$ and an open set $G \in \mathbb{B}_n$ such that $C \subset A \subset G$ and $\mu(G \backslash C) < \varepsilon$.

(ii) For any $A \in \mathbb{B}_n$ with $\mu(A) < \infty$ it holds that $\mu(A) = \sup\{\mu(K) : K \subseteq A, K \text{ is compact}\}$.

The proof Lemma 2.8 involves *semirings* and *outer measures*. This would be a significant departure from the focus of the report, and thus we do not show it here, but it can be found in Billingsley [1995, Section 12]. We now have everything we need to state and prove Kolmogorov's Theorem.

Theorem 2.9 (Kolmogorov's Theorem). *Let I be a non-empty set and let $\{\mu_{i_1,\dots,i_n}\}_{i_1,\dots,i_n\in I}$ be a class of probability measures, which satisfy the symmetry (see (2.1)) and compatibility (see (2.2)) properties. Then, on some probability space $(\Omega,\mathcal{F},P)$, there exists a stochastic process with index-set I with $\{\mu_{i_1,\dots,i_n}\}_{i_1,\dots,i_n\in I}$ as its finite-dimensional distributions.*

Before we show the proof, we show why the set $\mathbb{R}^I$ and σ-algebra $\mathbb{B}_I$ are quite natural to use in the proof. First let ζ be defined as in (2.4). Then it is clearly a stochastic process on $(\mathbb{R}^I,\mathbb{B}_I)$, since for any $i\in I$ and any $A\in\mathbb{B}$ we have

$$\zeta_i^{-1}(A)=\{x\in\mathbb{R}^I:\zeta(x,i)\in A\}=\{x\in\mathbb{R}^I:x(i)\in A\}\in\mathbb{B}_I,$$

so ζ_i is $\mathbb{B}_I$-$\mathbb{B}$-measurable for any $i\in I$. Now suppose we have some stochastic process, Z, on an index-set I and a probability space $(\Omega,\mathcal{F},P)$ with finite-dimensional distributions $\{\mu_{i_1,\dots,i_n}\}_{i_1,\dots,i_n\in I}$. Now let $\eta:\Omega\to\mathbb{R}^I$ be a function such that

$$\zeta\left(\eta(\omega),i\right)=Z(\omega,i).$$

Then for $i_1,\dots,i_n$ and an $A\in\mathbb{B}_n$ we have that

$$\begin{aligned}\eta^{-1}\left(\left\{x\in\mathbb{R}^I:\left(\zeta(x,i_1),\dots,\zeta(x,i_n)\right)\in A\right\}\right)&=\left\{\omega\in\Omega:\left(\zeta(\eta(\omega),i_1),\dots,\zeta(\eta(\omega),i_n)\right)\in A\right\}\\&=\left\{\omega\in\Omega:\left(Z(\omega,i_1),\dots,Z(\omega,i_n)\right)\in A\right\}.\end{aligned}\tag{2.12}$$

Since Z is a stochastic process, it holds that for any $i\in I$, $Z_i(\omega)=Z(\omega,i)$ is an $\mathcal{F}$-$\mathbb{B}$-measurable function. Thus the set on the left-hand-side of (2.12) is in $\mathcal{F}$. Therefore η is $\mathcal{F}$-$\mathbb{B}_I^0$-measurable and it follows from Billingsley [1995, Theorem 13.1] that η is also a $\mathcal{F}$-$\mathbb{B}_I$-measurable function. This means that the measure P_η, which satisfies that for $B\in\mathbb{B}_I$, $P_\eta(B)=P\left(\eta^{-1}(B)\right)$, is well-defined and

$$P_\eta\left(\left\{x\in\mathbb{R}^I:\left(\zeta(x,i_1),\dots,\zeta(x,i_n)\right)\in A\right\}\right)=P\left(\left\{\omega\in\Omega:\left(Z(\omega,i_1),\dots,Z(\omega,i_n)\right)\in A\right\}\right)=\mu_{i_1,\dots,i_n}(A).$$

This shows that if there exists a stochastic process, Z, on *any* probability space, then there also exists a stochastic process on $(\mathbb{R}^I,\mathbb{B}_I,P_\eta)$ for a suitably chosen η. So in the case where η is the identity we get that $\Omega=\mathbb{R}^I$, $\mathcal{F}=\mathbb{B}_I$ and $P_\eta=P$. For this reason, ζ is called the canonical process for an index set I. In the following we do not construct η, but we use Theorem 2.7 to construct a probability measure, such that the canonical process is a stochastic process on $(\mathbb{R}^I,\mathbb{B}_I)$ with said probability measure.

Proof of Theorem 2.9. This proof is a more detailed version of the proof given in [Billingsley, 1995, p. 489-490].

We prove the theorem by construction. So let I be an index-set and let $\{\mu_{i_1,\dots,i_n}\}_{i_1,\dots,i_n\in I}$ be a class of probability measures. We show that there exists a probability measure, P, such that the the canonical process is a stochastic process on $(\mathbb{R}^I,\mathbb{B}_I,P)$ with $\{\mu_{i_1,\dots,i_n}\}_{i_1,\dots,i_n\in I}$ as its finite dimensional distributions. First we construct such a probability measure.

Let

$$A=\left\{x\in\mathbb{R}^I:\left(x(i_1),\dots,x(i_n)\right)\in C\right\},\tag{2.13}$$

be a finite-dimensional set. Then we define a mapping, $P : \mathbb{B}_I^0 \to [0,1]$, such that

$$P(A) = \mu_{i_1,\dots,i_n}(C). \tag{2.14}$$

Firstly we note that, by definition,

$$P\left(\left\{x \in \mathbb{R}^I : \big(\zeta(i_1,x),\dots,\zeta(i_n,x)\big) \in C\right\}\right) = P\left(\left\{x \in \mathbb{R}^I : \big(x(i_1),\dots,x(i_n)\big) \in C\right\}\right) = \mu_{i_1,\dots,i_n}(C),$$

so if P is a probability measure, we can use Theorem 2.7 and extends P to a probability measure, $\tilde{P}$, on all of $\mathbb{B}_I$. Then $(\mathbb{R}^I, \mathbb{B}_I, \tilde{P})$ is a probability space and ζ is a stochastic process with index-set I and with $\{\mu_{i_1,\dots,i_n}\}_{i_1,\dots,i_n \in I}$ as finite dimensional distributions on said probability space. Thus it is now sufficient to show that P is a probability measure on $\mathbb{B}_I^0$.

First we must show that the formulation in (2.14) is consistent. More specifically, we may have that a finite-dimensional set has more than one representation (as an example, note that $\{x \in \mathbb{R}^I : x(i_1) \in C\} = \{x \in \mathbb{R}^I : (x(i_1), x(i_2)) \in C \times \mathbb{R}\}$), and in such a case, P must attain the same value for both representations. So now let A be defined as in (2.13) and let

$$B = \left\{x \in \mathbb{R}^I : \big(x(j_i),\dots,x(j_m)\big) \in D\right\}, \tag{2.15}$$

where $D \in \mathbb{B}_n$ and $j_1,\dots,j_m$ are distinct elements in I. We now show that if A and B are two representations of the same set, that is, if $A = B$, then $P(A) = P(B)$. Now let $\{t_1,\dots,t_r\} \subseteq I$ be a set such that $\{i_1,\dots,i_n\} \subset \{t_1,\dots,t_r\}$ and $\{j_1,\dots,j_m\} \subset \{t_1,\dots,t_r\}$. Then there exists permutations π_1 and π_2 such that $t_{\pi_1^{-1}(1)} = i_1,\dots,t_{\pi_1^{-1}(n)} = i_n$ and $t_{\pi_2^{-1}(1)} = j_1,\dots,t_{\pi_2^{-1}(m)} = j_m$. Then we have that

$$A = \left\{x \in \mathbb{R}^I : \big(x(t_1),\dots,x(t_r)\big) \in \psi_{\pi_1,r-m}^{-1}(C)\right\}, \tag{2.16}$$

and

$$B = \left\{x \in \mathbb{R}^I : \big(x(t_1),\dots,x(t_r)\big) \in \psi_{\pi_2,r-n}^{-1}(D)\right\}, \tag{2.17}$$

where $\psi_{\pi_1,r-m}$ and $\psi_{\pi_2,r-n}$ are defined as in Section 2.1. Now we show that $\psi_{\pi_1,r-m}^{-1}(C)$ and $\psi_{\pi_2,r-n}^{-1}(D)$ coincide when $A = B$. Suppose there is some point $z = (z_1,\dots,z_r)$ such that $z \in \psi_{\pi_1,r-m}^{-1}(C)$ but $z \notin \psi_{\pi_2,r-n}^{-1}(D)$. But since $t_1,\dots,t_r$ are distinct, there is an $x \in \mathbb{R}^I$ such that

$$\big(x(t_1),\dots,x(t_r)\big) = (z_1,\dots,z_r),$$

and then $x \in A$ but $x \notin B$, which is a contradiction, and thus $\psi_{\pi_1,r-m}^{-1}(C) = \psi_{\pi_2,r-n}^{-1}(D)$. Finally, note that (2.1) and (2.2) implies (2.3). Thus we get

$$P(A) = \mu_{i_1,\dots,i_n}(C) = \mu_{t_1,\dots,t_r}\left(\psi_{\pi_1,r-m}^{-1}(C)\right) = \mu_{t_1,\dots,t_r}\left(\psi_{\pi_2,r-n}^{-1}(D)\right) = \mu_{j_1,\dots,j_m}(D) = P(B),$$

which proves the consistency of P.

Clearly, for any $n \in \mathbb{N}$ and $i_1,\dots,i_n \in I$, we have

$$P\left(\mathbb{R}^I\right) = P\left(\left\{x \in \mathbb{R}^I : \big(x(i_1),\dots,x(i_n)\big) \in \mathbb{R}^n\right\}\right) = \mu_{i_1,\dots,i_n}(\mathbb{R}^n) = 1,$$

so all we need to prove now is that P is countably additive for pairwise disjoint sets. By Billingsley [1995, Example 2.10] we have countable additivity, if we have finite additivity and any sequence $A_1 \supseteq A_2 \supseteq \cdots$ in $\mathbb{B}_I^0$ where $A_k \to \emptyset$ for $k \to \infty$ satisfies $P(A_k) \to 0$ for $k \to \infty$.

We first prove finite additivity. Let A and B be defined as in (2.13) and (2.15), and let $A \cap B = \emptyset$. Then, as before, there exists a set $\{t_1, \dots, t_r\} \subseteq I$ and permutations π_1 and π_2 such that A and B can be expressed as in (2.16) and (2.17). Then suppose there is some $z \in \mathbb{R}^r$ such that $z \in \psi^{-1}_{\pi_1, r-m}(C)$ and $z \in \psi^{-1}_{\pi_2, r-n}(D)$. But since there exists an $x \in \mathbb{R}^I$ such that $(x(t_1), \dots, x(t_r)) = z$, we have that $x \in A$ and $x \in B$, but then $A \cap B$ is non-empty, which is a contradiction, so $\psi^{-1}_{\pi_1, r-m}(C) \cap \psi^{-1}_{\pi_2, r-n}(D) = \emptyset$. Thus we get

$$\begin{aligned} P(A \cup B) &= P\left(\left\{x \in \mathbb{R}^I : (x(t_1), \dots, x(t_r)) \in \psi^{-1}_{\pi_1, r-m}(C) \cup \psi^{-1}_{\pi_2, r-n}(D)\right\}\right) \\ &= \mu_{t_1, \dots, t_r}\left(\psi^{-1}_{\pi_1, r-m}(C) \cup \psi^{-1}_{\pi_2, r-n}(D)\right) = \mu_{t_1, \dots, t_r}\left(\psi^{-1}_{\pi_1, r-m}(C)\right) + \mu_{t_1, \dots, t_r}\left(\psi^{-1}_{\pi_2, r-n}(D)\right) \\ &= \mu_{i_1, \dots, i_n}(C) + \mu_{j_1, \dots, j_m}(D) = P(A) + P(B), \end{aligned}$$

which proves finite additivity.

Now let $A_1 \supseteq A_2 \supseteq \cdots$ be a sequence of sets in $\mathbb{B}^0_I$ where $A_k \to \emptyset$ for $k \to \infty$. We now show that this implies that $P(A_k) \to 0$ for $k \to \infty$. To do so we show the contra-positive, that is, we show that if for some $\varepsilon > 0$, $P(A_k) > \varepsilon$ for all $k \in \mathbb{N}$, we have that $\bigcap_{k=1}^{\infty} A_k \neq \emptyset$. Since A_k is a finite-dimensional set for all $k \in \mathbb{N}$, there must also exist a sequence of Borel sets (not necessarily of the same dimension), $\{C_k\}_{k \in \mathbb{N}}$ such that

$$A_k = \left\{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_{a_k})) \in C_k\right\},$$

where $a_{k+1} \leq a_k$ for all $k \in \mathbb{N}$. Now note that for an arbitrary $k \in \mathbb{N}$ we have that $\mu_{i_1, \dots, i_{a_k}}(C_k) \leq 1 < \infty$, so by Lemma 2.8 (ii) it follows that there exists a compact set D_k such that $D_k \subseteq C_k$ and $\mu_{i_1, \dots, i_{a_k}}(C_k \backslash D_k) < \varepsilon / 2^{k+1}$. Let $B_k = \{x \in \mathbb{R}^I : (x(i_1), \dots, x(i_{a_k})) \in D_k\}$. Then

$$P(A_k \backslash B_k) = \mu_{i_1, \dots, i_{a_k}}(C_k \backslash D_k) < \varepsilon / 2^{k+1}.$$

Therefore we get

$$P\left(A_k \backslash \bigcap_{l=1}^{k} B_l\right) = P\left(\bigcup_{l=1}^{k} A_k \backslash B_l\right) \leq P\left(\bigcup_{l=1}^{k} A_l \backslash B_l\right) \leq \sum_{l=1}^{k} P\left(A_l \backslash B_l\right) < \sum_{l=1}^{k} \frac{\varepsilon}{2^{l+1}} < \frac{\varepsilon}{2},$$

where the second and third relations are properties possessed by measures, which follow from finite additivity. This means that $P(A_k) - P\left(\bigcap_{l=1}^{k} B_l\right) < \varepsilon/2$, and since $P(A_k) > \varepsilon$ for all $k \in \mathbb{N}$, it follows that $P\left(\bigcap_{l=1}^{k} B_l\right) > \varepsilon/2$. This means that $\bigcap_{l=1}^{k} B_l$ is non-empty for any $k \in \mathbb{N}$.

Using the result shown above, we now prove that $\bigcap_{k=1}^{\infty} A_k$ is non-empty. Let $k \in \mathbb{N}$ and let $x_k \in \bigcap_{l=1}^{k} B_l$. Now for $p \leq k$ we have that

$$x_k \in \bigcap_{l=1}^{k} B_l \subseteq \bigcap_{l=1}^{p} B_l \subseteq B_p,$$

and thus $\left(x_k(i_1), \dots, x_k(i_{a_p})\right) \in D_p$. Since $x_1(i_{a_p}), \dots, x_{p-1}(i_{a_p})$ are all finite, and the sequence $\{x_l(i_{a_p})\}_{l \geq p}$ is in D_p, which is bounded, we get that the sequence $\{x_l(i_{a_p})\}_{l \in \mathbb{N}}$ is bounded for any p (the same argument can used to show that the sequence $\{x_l(i_r)\}_{l \in \mathbb{N}}$ for any r such that $a_{p-1} < r < a_p$). By Billingsley [1995, A14, p. 539] there exists an increasing sequence $\{n_l\}_{l \in \mathbb{N}}$ of positive integers such that the limit of the sequence $\{x_{n_l}(i_r)\}_{l \in \mathbb{N}}$ exists for any $r \in \mathbb{N}$.

Now let $x : I \to \mathbb{R}$ be a function such that $x(i_r) = \lim_{l \to \infty} x_{n_l}(i_r)$ for all r. But then $\left(x_{n_l}(i_1), \dots, x_{n_l}(i_{a_p})\right) \in D_p$ for all $l \in \mathbb{N}$ such that $n_l \geq p$, so the sequence $\left\{\left(x_{n_l}(i_1), \dots, x_{n_l}(i_{a_p})\right)\right\}_{l : n_l \geq p}$

is in D_p, and since D_p is compact

$$\lim_{l\to\infty}\left(x_{n_l}(i_1),\dots,x_{n_l}(i_{a_p})\right)=\left(x(i_1),\dots,x(i_{a_p})\right)\in D_p,$$

which implies that $x\in B_p$ and thus $x\in A_p$. Since these arguments were carried out for an arbitrary p, we get that

$$x\in\bigcap_{p=1}^{\infty}A_p,$$

which proves the Theorem. □

2.3 Generalised Stochastic Processes

This section is based on [Schäffler, 2018, Chapter 2.1-2.3].

As we see in Definition 2.1 we can choose the index set of a stochastic process to be any non-empty set. Specifically we may choose $I=\mathcal{T}_d$, to obtain random variables indexed over test functions. This is used to define the notion of generalised stochastic processes, but first we need to define the notion of continuity for test-function-indexed stochastic process.

Definition 2.10 (Continuity for Test-function-indexed Stochastic Process). *Let Z be a stochastic process with index set $\mathcal{T}_d$. Then Z is said to be continuous if*

$$\lim_{k\to\infty}\mathbb{E}\left[g\left(Z(\varphi_{1k}),\dots,Z(\varphi_{mk})\right)\right]=\mathbb{E}\left[g\left(Z(\varphi_1),\dots,Z(\varphi_m)\right)\right],$$

for any bounded function $g:\mathbb{R}^m\to\mathbb{R}$ and for any m sequences of test-functions such that $\varphi_{ik}\to\varphi_i$ for $k\to\infty$ and $i=1,\dots,m$ (here $\varphi_{ik}\to\varphi_i$ refers to convergence in the sense of Definition 1.2).

This makes it possible to define generalised stochastic processes

Definition 2.11 (Generalised Stochastic Processes). *Let Z be a stochastic process with index-set $\mathcal{T}_d$. If Z is continuous and*

$$Z(a\varphi+b\psi)=aZ(\varphi)+bZ(\psi),$$

almost surely for all $a,b\in\mathbb{R}$ and $\varphi,\psi\in\mathcal{T}_d$, then Z is said to be a generalised stochastic process.

It is clear from Definition 2.1 that generalised stochastic process are *not* generalisations of stochastic processes. Their name arise for different reasons. For a fixed $\omega\in\Omega$ we have that a generalised stochastic process $Z(\omega,\cdot)$ is a functional, and thus if Z has continuous paths, $Z(\omega,\cdot)$ is a generalised function for any $\omega\in\Omega$. In fact it turns out, that for any generalised stochastic process, Z, there exists another generalised stochastic process, Z^*, with has continuous paths (in the sense of Definition 1.5) and $Z(\cdot,\varphi)=Z^*(\cdot,\varphi)$ almost surely for all $\varphi\in\mathcal{T}_d$ [Dawson, 1970, Theorem 2.1].

We define the mean- and covariance-functional for a generalised stochastic process as we would have for a random field,

$$m(\varphi)=\mathbb{E}\left[Z(\varphi)\right],\quad C(\varphi,\psi)=\mathbb{C}\mathrm{ov}\left[Z(\varphi),Z(\psi)\right].$$

Furthermore we note that, as usual, the covariance-functional is symmetric, bi-linear and positive semi-definite, meaning that

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C(\varphi_i, \varphi_j) \geq 0,$$

for all $a_1, \ldots, a_m \in \mathbb{R}$ and $\varphi_1, \ldots, \varphi_m \in \mathcal{T}_d$.

If all finite dimensional distributions of Z are Gaussian, then Z is said to be a *Gaussian generalised stochastic process*. With this we can show the following theorem.

Theorem 2.12. *Let $m : \mathcal{T}_d \to \mathbb{R}$ be a continuous linear functional, and let $C : \mathcal{T}_d^2 \to \mathbb{R}$ be a continuous, symmetric, bi-linear and positive semi-definite functional. Then there exists a Gaussian generalised stochastic process, Z, which satisfies*

$$\mathbb{E}\left[Z(\varphi)\right] = m(\varphi)$$

and

$$\mathbb{C}\text{ov}\left[Z(\varphi), Z(\psi)\right] = C(\varphi, \psi).$$

Proof. In order to prove the results, we need to show the following:

(i) The functionals $m(\cdot)$ and $C(\cdot,\cdot)$ induces a class of finite-dimensional distribution functions. That is for any $\varphi_1, \ldots, \varphi_n$, the normal-distribution with mean-vector and covariance-matrix given by

$$\mu_{\varphi_1,\ldots,\varphi_n} = \begin{bmatrix} m(\varphi_1) \\ \vdots \\ m(\varphi_n) \end{bmatrix}, \quad \Sigma_{\varphi_1,\ldots,\varphi_n} = \begin{bmatrix} C(\varphi_1,\varphi_1) & \cdots & C(\varphi_1,\varphi_n) \\ \vdots & \ddots & \vdots \\ C(\varphi_n,\varphi_1) & \cdots & C(\varphi_n,\varphi_n) \end{bmatrix},$$

is a valid distribution. In other words it must be shown that $\Sigma_{\varphi_1,\ldots,\varphi_n}$ is positive semidefinite for any choice of $\varphi_1, \ldots, \varphi_n$.

(ii) There exists a stochastic process, Z, with index-set $\mathcal{T}_d$ such that

$$\Big(Z(\varphi_1), \ldots, Z(\varphi_n)\Big)^\top \sim N\left(\mu_{\varphi_1,\ldots,\varphi_n}, \Sigma_{\varphi_1,\ldots,\varphi_n}\right), \tag{2.18}$$

for any $\varphi_1, \ldots, \varphi_n \in \mathcal{T}_d$.

(iii) $Z(a\varphi + b\psi) = aZ(\varphi) + bZ(\psi)$ almost surely for all $\varphi, \psi \in \mathcal{T}_d$ and $a, b \in \mathbb{R}$.

(iv) Z is continuous.

Choose arbitrary test-functions $\varphi_1, \ldots, \varphi_n \in \mathcal{T}_d$ and define a random vector $Z_{\varphi_1,\ldots,\varphi_n} : \Omega \to \mathbb{R}^n$ such that

$$Z_{\varphi_1,\ldots,\varphi_n} \sim N\Big(\mu_{\varphi_1,\ldots,\varphi_n}, \Sigma_{\varphi_1,\ldots,\varphi_n}\Big),$$

where $\mu_{\varphi_1,\ldots,\varphi_n}$ and $\Sigma_{\varphi_1,\ldots,\varphi_n}$ are defined as in (2.18).

We get that for any $a \in \mathbb{R}^n$

$$a^\top \Sigma_{\varphi_1,\ldots,\varphi_n} a = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\varphi_i, \varphi_j) \geq 0,$$

and thus $\Sigma\varphi_1, \ldots, \varphi_n$ is a positive semi-definite matrix, which proves (i). Furthermore we get that (ii) holds by Theorem 2.9.

To prove (iii) choose arbitrary $\varphi, \psi \in \mathcal{T}_d$ and $a, b \in \mathbb{R}$. To prove that $Z(a\varphi + b\psi) = aZ(\varphi) + bZ(\psi)$ almost surely we consider the random variable $Z(a\varphi + b\psi) - (aZ(\varphi) + bZ(\psi))$. First we note that

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left[Z(a\varphi + b\psi) - (aZ(\varphi) + bZ(\psi))\right] \\ &= \mathbb{V}\mathrm{ar}\left[Z(a\varphi + b\psi)\right] + \mathbb{V}\mathrm{ar}\left[aZ(\varphi) + bZ(\psi)\right] - 2\mathbb{C}\mathrm{ov}\left[Z(a\varphi + b\psi), aZ(\varphi) + bZ(\psi)\right]. \end{aligned} \tag{2.19}$$

We now express each term in (2.19) in terms of the covariance function of Z. By construction we get

$$\mathbb{V}\mathrm{ar}\left[Z(a\varphi + b\psi)\right] = C(a\varphi + b\psi, a\varphi + b\psi). \tag{2.20}$$

The second term in(2.20) can be expanded as follows

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[aZ(\varphi) + bZ(\psi)\right] &= a^2\mathbb{V}\mathrm{ar}\left[Z(\varphi)\right] + b^2\mathbb{V}\mathrm{ar}\left[Z(\psi)\right] + 2ab\mathbb{C}\mathrm{ov}\left[Z(\varphi), Z(\psi)\right] \\ &= a^2C(\varphi, \varphi) + b^2C(\psi, \psi) + 2abC(\varphi, \psi) \\ &= C(a\varphi + b\psi, a\varphi + b\psi). \end{aligned} \tag{2.21}$$

Finally, for the third in (2.20) term we get

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left[Z(a\varphi + b\psi), aZ(\varphi) + bZ(\psi)\right] &= a\mathbb{C}\mathrm{ov}\left[Z(a\varphi + b\psi), Z(\varphi)\right] + b\mathbb{C}\mathrm{ov}\left[Z(a\varphi + b\psi), Z(\psi)\right] \\ &= aC(a\varphi + b\psi, \varphi) + bC(a\varphi + b\psi, \psi) \\ &= C(a\varphi + b\psi, a\varphi + b\psi). \end{aligned} \tag{2.22}$$

Now by inserting (2.20), (2.21) and (2.22) into (2.19) we get that

$$\mathbb{V}\mathrm{ar}\left[Z(a\varphi + b\psi) - (aZ(\varphi) + bZ(\psi))\right] = 0,$$

which means that $Z(a\varphi + b\psi) - (aZ(\varphi) + bZ(\psi))$ is constant almost surely. By examining the mean we get

$$\mathbb{E}\left[Z(a\varphi + b\psi) - (aZ(\varphi) + bZ(\psi))\right] = m(a\varphi + b\psi) - m(a\varphi) - m(b\psi) = 0,$$

by linearity of m, and thus $Z(a\varphi + b\psi) - (aZ(\varphi) + bZ(\psi)) = 0$ almost surely, proving (iii).

To show (iv) we note that if we have sequences of test-functions $\{\varphi_{ik}\}_{k\in\mathbb{N}}$, $i = 1, \ldots, n$, such that $\varphi_{ik} \to \varphi_i$ for $i = 1, \ldots, n$, then

$$(m(\varphi_{nk}), \ldots, m(\varphi_{nk}))^\top \to (m(\varphi_1), \ldots, m(\varphi_n))^\top, \text{ for } k \to \infty,$$

and

$$\begin{bmatrix} C(\varphi_{1k}, \varphi_{1k}) & \cdots & C(\varphi_{1k}, \varphi_{nk}) \\ \vdots & \ddots & \vdots \\ C(\varphi_{nk}, \varphi_{1k}) & \cdots & C(\varphi_{nk}, \varphi_{nk}) \end{bmatrix} \to \begin{bmatrix} C(\varphi_1, \varphi_1) & \cdots & C(\varphi_1, \varphi_n) \\ \vdots & \ddots & \vdots \\ C(\varphi_n, \varphi_1) & \cdots & C(\varphi_n, \varphi_n) \end{bmatrix}, \text{ for } k \to \infty,$$

due to the continuity of m and C. This means that

$$(Z(\varphi_{ik}), \ldots, Z(\varphi_{nk}))^\top \xrightarrow{d} (Z(\varphi_1), \ldots, Z(\varphi_n))^\top, \text{ for } k \to \infty,$$

and thus by Portmanteau's Lemma [Rongfeng, Theorem 1.3] it follows that

$$\mathbb{E}\left[g(Z(\varphi_{ik}), \ldots, Z(\varphi_{nk}))\right] \to \mathbb{E}\left[g(Z(\varphi_1), \ldots, Z(\varphi_n))\right], \text{ for } k \to \infty,$$

for any continuous bounded function $g : \mathbb{R}^n \to \mathbb{R}$. This shows that Z is continuous by Definition 2.10, and thus it is a generalised stochastic process. □

Derivative of a Brownian Motion

We now show an application of generalised stochastic processes, by differentiating a Brownian Motion. First we introduce Brownian Motion as a stochastic process on $\mathbb{R}^+$.

Definition 2.13 (Brownian Motion). *Let B be a stochastic process with index-set $\mathbb{R}^+$. B is said to be a Brownian Motion if the following criteria are satisfied*

(i) For any $t, s \in \mathbb{R}^+$, $B(t) - B(s) \sim N(0, |t - s|)$.

(ii) $B(0) = 0$ almost surely.

(iii) For any $t, s \in \mathbb{R}^+$, $\mathbb{C}\text{ov}\left[B(t), B(s)\right] = \min(t, s)$.

From (iii) if is easily verified that for any $r, s, t, u \in \mathbb{R}^+$ with $r < s \leq t < u$, $B(s) - B(r)$ and $B(u) - B(t)$ are independent. The most intuitively straightforward way to define the derivative of a Brownian Motion is

$$B'(\omega, t) = \lim_{\varepsilon \to 0} \frac{B(\omega, t - \varepsilon) - B(\omega, t + \varepsilon)}{2\varepsilon}. \tag{2.23}$$

This is known as path-wise differentiation. Since

$$\mathbb{E}\left[\frac{B(\omega, t - \varepsilon) - B(\omega, t + \varepsilon)}{2\varepsilon}\right] = 0,$$

and

$$\mathbb{V}\text{ar}\left[\frac{B(\omega, t - \varepsilon) - B(\omega, t + \varepsilon)}{2\varepsilon}\right] = \frac{1}{2\varepsilon},$$

we would expect B' to be a zero-mean process with infinite variance, and since Brownian Motions have independent increments, we would also have that $B'(t)$ and $B'(s)$ are independent for any $t, s \in \mathbb{R}^+$. Of course these considerations are heuristic, and unfortunately they do not hold in practice. This is due to a well-known results, which states that all paths of a Brownian Motion are continuous but non-differentiable at all points. This means that the limit in (2.23) is undefined. However all is not lost, since, as we shall see, it is possible to express a Brownian Motion as a generalised stochastic process. This makes it possible to give meaning to the notion of a derivative of a Brownian Motion similarly to how we "defined" the derivatives of non-differentiable locally integrable functions in Section 1.2. To achieve this we define the derivative of a generalised stochastic process similarly to how we defined the derivative of a generalised function.

Definition 2.14 (Derivative of a Generalised Stochastic Process). *Let Z be a generalised stochastic process with index-set $\mathcal{T}_1$. Then the generalised stochastic process, Z', defined*

$$Z'(\omega, \varphi) = -Z(\omega, \varphi'), \text{ for all } \omega \in \Omega \text{ and } \varphi \in \mathcal{T}_d, \tag{2.24}$$

is called the derivative of Z.

Note that the right-hand-side of (2.24) is always well-defined, since $\mathcal{T}_d \subset \mathcal{C}_d^\infty$. Before proceeding with Brownian Motions we first show some properties of the derivative of a generalised stochastic process. First let Z be a generalised stochastic process. Then we have

$$\mathbb{E}\left[Z'(\varphi)\right] = -\mathbb{E}\left[Z(\varphi')\right] = -m(\varphi') = m'(\varphi)$$

for any $\varphi \in \mathcal{T}_1$ by Definition 1.8, since m is a generalised function. Now let C' be the covariance function of Z'. Then

$$\begin{aligned} C'(\varphi,\psi) = \mathbb{C}\mathrm{ov}\left[Z'(\varphi), Z'(\psi)\right] &= \mathbb{E}\left[\Big(Z'(\varphi) - m'(\varphi)\Big)\Big(Z'(\psi) - m'(\psi)\Big)\right] \\ &= \mathbb{E}\left[\Big(-Z(\varphi') + m(\varphi')\Big)\Big(-Z(\psi') + m(\psi')\Big)\right] \\ &= \mathbb{E}\left[\Big(Z(\varphi') - m(\varphi')\Big)\Big(Z(\psi') - m(\psi')\Big)\right] \\ &= \mathbb{C}\mathrm{ov}\left[Z(\varphi'), Z(\psi')\right] = C(\varphi',\psi'). \end{aligned}$$

As mentioned previously, Brownian Motions are path-wise non-differentiable at all points. So the best we can do is to define a generalised stochastic process similar to how we defined a regular generalised function, and differentiate that. Specifically we define (with abuse of notation)

$$B(\omega,\varphi) = \int_0^\infty B(\omega,t)\varphi(t)\mathrm{d}t, \text{ for all } \omega \in \Omega.$$

This is a Riemann integral, and is thus defined as the limit of a sum of increasingly narrow increments. The terms of the sum can be written as $B(\omega,t_i)\varphi(t_i - t_{i-1})$, and are zero-mean Gaussian, and thus $B(\cdot,\varphi)$ is also Gaussian with mean zero. Furthermore

$$\begin{aligned} \mathbb{C}\mathrm{ov}\left[B(\varphi), B(\psi)\right] &= \mathbb{E}\left[\left(\int_0^\infty B(t)\varphi(t)\mathrm{d}t\right)\left(\int_0^\infty B(s)\psi(s)\mathrm{d}s\right)\right] \\ &= \mathbb{E}\left[\int_0^\infty\int_0^\infty B(t)B(s)\varphi(t)\psi(s)\mathrm{d}t\mathrm{d}s\right] \\ &= \int_0^\infty\int_0^\infty \mathbb{E}\left[B(t)B(s)\right]\varphi(t)\psi(s)\mathrm{d}t\mathrm{d}s \\ &= \int_0^\infty\int_0^\infty \min(t,s)\varphi(t)\psi(s)\mathrm{d}t\mathrm{d}s. \end{aligned}$$

With this we can state the following result.

Proposition 2.15. *Let B be a generalised stochastic process on $\mathcal{T}_1$ such that*

$$\mathbb{E}\left[B(\varphi)\right] = 0 \text{ for all } \varphi \in \mathcal{T}_1 \tag{2.25}$$

and

$$\mathbb{C}\mathrm{ov}\left[B(\varphi), B(\psi)\right] = \int_0^\infty\int_0^\infty \min(t,s)\varphi(t)\psi(s)\mathrm{d}t\mathrm{d}s.$$

Then $B'(\varphi)$ satisfies

$$\mathbb{E}\left[B'(\varphi)\right] = 0, \tag{2.26}$$

and

$$\mathbb{C}\mathrm{ov}\left[B'(\varphi), B'(\psi)\right] = \int_0^\infty\int_0^\infty \varphi(t)\psi(s)\mathrm{d}t\mathrm{d}s. \tag{2.27}$$

Proof. This proof follows the arguments of the proof of [Schäffler, 2018, Theorem 2.4].

Firstly (2.26) follows trivially from (2.25) since $\varphi \in \mathcal{T}_1$ implies that $\varphi' \in \mathcal{T}_1$ and $B'(\varphi) = -B(\varphi')$.

Now we show (2.27). We do this by deriving an expression for $\mathbb{C}\mathrm{ov}\left[B(\varphi), B(\psi)\right]$ and then use the fact that

$$\mathbb{C}\mathrm{ov}\left[B'(\varphi), B'(\psi)\right] = \mathbb{C}\mathrm{ov}\left[B(\varphi'), B(\psi')\right].$$

We note that

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left[B(\varphi), B(\psi)\right] &= \int_0^\infty \int_0^\infty \min(t,s)\varphi(t)\psi(s)\mathrm{d}t\mathrm{d}s \\ &= \int_0^\infty \int_0^s t\varphi(t)\psi(s)\mathrm{d}t\mathrm{d}s + \int_0^\infty \int_0^t s\varphi(t)\psi(s)\mathrm{d}s\mathrm{d}t \\ &= \int_0^\infty \varphi(t)\int_0^t s\psi(s)\mathrm{d}s\mathrm{d}t + \int_0^\infty \psi(s)\int_0^s t\varphi(t)\mathrm{d}t\mathrm{d}s. \end{aligned} \tag{2.28}$$

Now we consider the first term, and perform integration by parts where $u(t) = \int_0^t s\psi(s)\mathrm{d}s$ and $v'(t) = \varphi(t)$. First we note that defining u and v' this way implies $u'(t) = t\psi(t)$ and $v(t) = \int_0^t \phi(s)\mathrm{d}s$. Thus we get

$$\begin{aligned}\int_0^\infty \varphi(t)\int_0^t s\psi(s)\mathrm{d}s\mathrm{d}t &= \left[\int_0^t \varphi(s)\mathrm{d}s \int_0^t s\psi(s)\mathrm{d}s\right]_{t=0}^\infty - \int_0^\infty t\psi(t)\int_0^t \varphi(s)\mathrm{d}s\mathrm{d}t \\ &= \int_0^\infty \varphi(t)\mathrm{d}t\int_0^\infty t\psi(t)\mathrm{d}t - \int_0^\infty t\psi(t)\int_0^t \varphi(s)\mathrm{d}s\mathrm{d}t.\end{aligned}$$

Now define $\Phi(t) = \int_0^t \varphi(s)\mathrm{d}s$, then

$$\begin{aligned}\int_0^\infty \varphi(t)\int_0^t s\psi(s)\mathrm{d}s\mathrm{d}t &= (\Phi(\infty) - \Phi(0))\int_0^\infty t\psi(t)\mathrm{d}t - \int_0^\infty t\psi(t)(\Phi(t) - \Phi(0))\mathrm{d}t \\ &= \int_0^\infty (\Phi(\infty) - \Phi(t))\, t\psi(t)\mathrm{d}t. \end{aligned} \tag{2.29}$$

Defining $\Psi(t) = \int_0^t \psi(s)\mathrm{d}s$ and performing similar computations on the second term of (2.28), we get

$$\int_0^\infty \psi(s)\int_0^s t\varphi(t)\mathrm{d}t\mathrm{d}s = \int_0^\infty (\Psi(\infty) - \Psi(t))t\varphi(t)\mathrm{d}t. \tag{2.30}$$

Now inserting (2.29) and (2.30) into (2.28) yields

$$\mathbb{C}\mathrm{ov}\left[B(\varphi), B(\psi)\right] = \int_0^\infty t\Big(\psi(t)\Big(\Phi(\infty) - \Phi(t)\Big) + \varphi(t)\Big(\Psi(\infty) - \Psi(t)\Big)\Big)\mathrm{d}t.$$

Once again we perform integration by parts. Here $u(t) = t$ and $v'(t) = \psi(t)\Big(\Phi(\infty) - \Phi(t)\Big) + \varphi(t)\Big(\Psi(\infty) - \Psi(t)\Big)$ and thus $v(t) = -(\Phi(\infty) - \Phi(t))(\Psi(\infty) - \Psi(t))$. Thus we obtain

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left[B(\varphi), B(\psi)\right] &= \left[-t(\Phi(\infty) - \Phi(t))(\Psi(\infty) - \Psi(t))\right]_{t=0}^\infty + \int_0^\infty (\Phi(\infty) - \Phi(t))(\Psi(\infty) - \Psi(t))\mathrm{d}t \\ &= \lim_{t\to\infty}\Big(-t\big(\Phi(\infty) - \Phi(t)\big)\big(\Psi(\infty) - \Psi(t)\big)\Big) + \int_0^\infty (\Phi(\infty) - \Phi(t))(\Psi(\infty) - \Psi(t))\mathrm{d}t.\end{aligned}$$

We now consider the limit in the above expression.

$$\begin{aligned}&\lim_{t\to\infty} -t\Big(\Phi(\infty) - \Phi(t)\Big)\Big(\Psi(\infty) - \Psi(t)\Big) = \lim_{t\to\infty} -\frac{\Big(\Phi(\infty) - \Phi(t)\Big)\Big(\Psi(\infty) - \Psi(t)\Big)}{1/t} \\ &= \lim_{t\to\infty} -\frac{\psi(t)\Big(\Phi(\infty) - \Phi(t)\Big) + \varphi(t)\Big(\Psi(\infty) - \Psi(t)\Big)}{1/t^2},\end{aligned}$$

where the second equality follows from L'Hospital's rule. Since $\varphi, \psi \in \mathcal{T}_d$ they have bounded supports. This means that $t^* = \max(\text{Supp}(\varphi) \cup \text{Supp}(\psi))$ is finite and well-defined, and that the expression in the limit above is constantly equal to zero for $t > t^*$. Thus

$$\lim_{t\to\infty} -\frac{\psi(t)\Big(\Phi(\infty) - \Phi(t)\Big) + \varphi(t)\Big(\Psi(\infty) - \Psi(t)\Big)}{1/t^2} = 0,$$

which in turn implies that

$$\mathbb{C}\text{ov}\left[B(\varphi), B(\psi)\right] = \int_0^\infty (\Phi(\infty) - \Phi(t))(\Psi(\infty) - \Psi(t))\mathrm{d}t.$$

Now the fact that $\mathbb{C}\text{ov}\left[B'(\varphi), B'(\psi)\right] = \mathbb{C}\text{ov}\left[B(\varphi'), B(\psi')\right]$ implies that

$$\mathbb{C}\text{ov}\Big[B'(\varphi), B'(\psi)\Big] = \int_0^\infty \varphi(t)\psi(t)\mathrm{d}t,$$

which proves the result. □

Chapter 3

Set- and measure-indexed Stochastic Processes

3.1 Generalised Random Fields

This section is based on Mondal [2015].

In this section we introduce the notion of *generalised random fields*. These objects are used to model more complex structures than traditional random fields. When modelling a dataset with random fields, it is assumed that the data is sampled at infinitesimally sized points. Of course this is an abstraction, and in reality the area sampled has a non-zero area. In order to facilitate a model that takes this into account, we introduce set-indexed random fields.

Definition 3.1 (Set-indexed Random Fields). *Let $(X, \mathcal{G}, \mu)$ be measure space and denote $\mathcal{G}_\mu = \{A \in \mathcal{G} : \mu(A) < \infty\}$. Then a stochastic process, Z, with index-set $\mathcal{G}_\mu$, is called a set-indexed random field over $\mathcal{G}$ and μ.*

Suppose that Y_{RF} is an integrable random field on $\mathbb{R}^d$. Then we can define a set-indexed random field over the Borel σ-algebra and the Lebesgue measure as follows. For any bounded set $A \subset \mathbb{R}^d$ we define

$$Y(\omega, A) = \int_A Y_{\mathrm{RF}}(\omega, x)\mathrm{d}x. \tag{3.1}$$

Of course Definition 3.1 also allows other constructions, and in practice we do not use Y_{RF}, but it is used here for illustrative purposes. Henceforth, unless otherwise specified, we only consider set-indexed random fields over the Borel sigma-algebra and the Lebesgue measure. In this chapter we seek to describe the so-called *De Wijs process*. This can only be defined on *contrasts* of set-indexed random fields. In order to handle this theoretically we need a class of stochastic processes indexed on measures. First denote $\mathbb{M}(X, \mathcal{G})$ as the set of all signed measures on the measurable space $(X, \mathcal{G})$. Then denote $\mathcal{M} = \{\mu \in \mathbb{M}(\mathbb{R}^d, \mathbb{B}_d) : \mu(\mathbb{R}^d) = 0\}$. It is easy to check that $\mathcal{M}$ is a vector space, which gives rise to the following definition.

Definition 3.2 (Generalised Random Fields). *Let Z be a stochastic process with index-set $\mathcal{S}$. Z is called a generalised random field if $\mathcal{S}$ is a subspace of $\mathcal{M}$, and if it satisfies that*

$$Z(a\mu + b\nu) = aZ(\mu) + bZ(\nu), \tag{3.2}$$

almost surely for all $\mu, \nu \in \mathcal{S}$ and $a, b \in \mathbb{R}$.

We note that the requirement in the above definition, that the index-set, $\mathcal{S}$, is a vector-space, is a reasonable one, since otherwise the left-hand-side of (3.2) might be undefined for some choices of $\mu, \nu \in \mathcal{S}$.

A generalised random field can be used to model contrasts over set indexed random fields as follows. Define a generalised random field as

$$Z(\omega, \mu) = \int Y_{\mathrm{RF}}(\omega, x)\mathrm{d}\mu(x).$$

When a generalised random field has this form, we call Y_{RF} the *underlying random field*. Now let Y be defined as in (3.1) and let $A, B \in \mathbb{B}_d$ with $|A| = |B|$, where $|\cdot|$ is the Lebesgue measure. Suppose μ is a signed measure on $(\mathbb{R}^d, \mathbb{B}_d)$ with density $1_A - 1_B$ with respect to the Lebesgue measure. Then we have

$$\mu(\mathbb{R}^d) = \int 1_A(x)\mathrm{d}x - \int 1_B(x)\mathrm{d}x = |A| - |B| = 0,$$

so $\mu \in \mathcal{M}$. Now we get

$$\begin{aligned} Z(\omega, \mu) &= \int Y_{\mathrm{RF}}(\omega, x)\mathrm{d}\mu(x) = \int Y_{\mathrm{RF}}(\omega, x)(1_A(x) - 1_B(x))\mathrm{d}x \\ &= \int_A Y_{\mathrm{RF}}(\omega, x)\mathrm{d}x - \int_B Y_{\mathrm{RF}}(\omega, x)\mathrm{d}x = Y(\omega, A) - Y(\omega, B). \end{aligned}$$

If we were to restrict Z to the set of measures $\mu \in \mathcal{M}$ with a test-function as density with respect to the Lebesgue measure, we would have a generalised stochastic process where $Z(\mu) = Z(\varphi)$ with an abuse of notation, when φ is the density for μ. That is

$$Z(\omega, \varphi) = \int Y_{\mathrm{RF}}(\omega, x)\varphi(x)\mathrm{d}x, \tag{3.3}$$

for all $\varphi \in \mathcal{T}_d$. It is easily seen that the space of measures in $\mathcal{M}$ with a test-function as density with respect to the Lebesgue measure is vector space. Thus it is possible to construct a process that is both a generalised stochastic process and a generalised random field.

Some generalised random fields have certain useful properties given by the finite-dimensional distributions. A generalised random field where all finite dimensional distributions are Gaussian is called a Gaussian generalised random field. Now let Z be a generalised random field with index-set $\mathcal{S}$ and let μ be an arbitrary measure in $\mathcal{S}$. Then for $h \in \mathbb{R}^d$ define

$$\mu^h(A) = \mu(A + h),$$

where $A + h = \{a + h \in \mathbb{R}^d : a \in A\}$. Then, if for any $h \in \mathbb{R}^d$ and $n \in \mathbb{N}$ it holds for any $\mu_1, \ldots, \mu_n \in \mathcal{S}$ that $\Big(Z(\mu_1), \ldots, Z(\mu_n)\Big)$ has the same distribution as $\Big(Z(\mu_1^h), \ldots, Z(\mu_n^h)\Big)$, Z is said to be stationary. Similarly we may define

$$\mu^R(A) = \mu(RA),$$

where R is a rotation matrix and $RA = \{Ra \in \mathbb{R}^d : a \in A\}$. Then if $\Big(Z(\mu_1), \dots, Z(\mu_n)\Big)$ has the same distribution as $\Big(Z(\mu_1^R), \dots, Z(\mu_n^R)\Big)$ for any rotation matrix, any $n \in \mathbb{N}$ and any $\mu_1, \dots, \mu_n \in \mathcal{S}$, Z is said to be isotropic. Finally define

$$\mu^b(A) = \mu(bA),$$

where $b \in \mathbb{R}$ and $bA = \{ba \in \mathbb{R}^d : a \in A\}$. Then if $\Big(Z(\mu_1), \dots, Z(\mu_n)\Big)$ has the same distribution as $\Big(Z(\mu_1^b), \dots, Z(\mu_n^b)\Big)$ for any $b \in \mathbb{R}$, any $n \in \mathbb{N}$ and any $\mu_1, \dots, \mu_n \in \mathcal{S}$, Z is said to be self-similar.

The usual covariance function may not be sufficient to model the covariance-structure of generalised random fields. Instead a different but related object is used.

Definition 3.3 (Generalised Covariance Function). *Let Z be a generalised random field with index-set $\mathcal{S}$, where $\mathbb{E}\left[Z(\mu)\right] = 0$ for all $\mu \in \mathcal{S}$. Then a symmetric function $K(\cdot, \cdot)$ which satisfies*

$$\mathbb{C}\text{ov}\left[Z(\mu), Z(\nu)\right] = \int\int K(x, y)\mathrm{d}\mu(x)\nu(y),$$

for $\mu, \nu \in \mathcal{S}$ is called a generalised covariance function for Z.

It turns out that the notion of generalised covariance functions for intrinsic random fields of order k, which are examined in Jensen and Fitzhugh [2018], is a special case of Definition 3.3. We can define Λ_k as the space of allowable discrete measures of order k, i.e signed measures on the form

$$\lambda(A) = \sum_{i=1}^{m} \lambda_i \delta_{x_i}(A),$$

where $x_1, \dots, x_m \in \mathbb{R}^d$ are distinct points, δ_{x_i} is the Dirac-measure centred at x_i and $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ are chosen such that

$$\sum_{i=1}^{m} \lambda_i P(x_i) = 0,$$

for any polynomial P with $\deg(P) \leq k$. We now define a stochastic process with index-set Λ_k as

$$Z(\lambda) = \sum_{i=1}^{m} \lambda_i Y_{\text{RF}}(x_i),$$

where Y_{RF} is a random field over $\mathbb{R}^d$. Since Λ_k is a subspace of $\mathcal{M}$, Z is a generalised random field. It is worth noting that the notion of intrinsic random fields of order k defined using random fields over $\mathbb{R}^d$ (see Jensen and Fitzhugh [2018, Definition 1.12]), is equivalent to requiring that the covariance of Z is invariant under translations. In other words, for $\lambda, \mu \in \Lambda_k$ with $\lambda(A) = \sum_{i=1}^m \lambda_i \delta_{x_i}(A)$ and $\mu(A) = \sum_{i=1}^m \mu_i \delta_{y_i}(A)$ we get that

$$\mathbb{C}\text{ov}\left[Z(\lambda), Z(\mu)\right] = \int\int K(y - x)\mathrm{d}\lambda(y)\mathrm{d}\mu(x) = \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \mu_j K(y_j - x_i).$$

This is equivalent to Jensen and Fitzhugh [2018, Definition 1.16].

In this project we do not prove that the generalised covariance function of a generalised random field exists, but for a proof of the special case of a generalised random field with index-set Λ_k and a $k+1$ times mean square differentiable underlying random field over $\mathbb{R}$, see Jensen and Fitzhugh [2018, Theorem 1.18].

Relation between index-sets

In this subsection we discuss the relationship between measure-indexed stochastic processes, generalised stochastic processes, set-indexed random fields and generalised random fields. For this purpose we define the notion of stochastic processes with multiple index-sets.

Definition 3.4 (Stochastic Processes with Multiple Index-sets). *Let Z be a stochastic process on the probability space $(\Omega, \mathcal{F}, P)$, where Z has index-set I and let J be a non-empty set. Then Z is said to also be stochastic process with index-set J, if there is a one-to-one correspondence between I and a subset of J.*

(1) **Generalised Stochastic Processes are Measure-indexed Stochastic Processes:**

Let Z be a generalised stochastic process. Then for any $\varphi \in \mathcal{T}_d$ we can define a measure, μ, with density φ with respect to the Lebesgue measure. This amounts to a one-to-one correspondence between $\mathcal{T}_d$ and the set $\{\mu \in \mathbb{M}(\mathbb{R}^d, \mathbb{B}_d) : \mu(B) = \int_B \varphi(x)\mathrm{d}x, \text{ where } \varphi \in \mathcal{T}_d\}$. Thus by Definition 3.4 Z is also a measure-indexed stochastic process.

(2) **Set-indexed Random Fields are Measure-indexed Stochastic Processes:**

Let $(X, \mathcal{G}, \nu)$ be a measure-space and let Z be a set-indexed random field with index-set $\mathcal{G}_\nu = \{A \in \mathcal{G} : \nu(A) < \infty\}$. Then for any $A \in \mathcal{G}_\nu$ we can define a measure with density 1_A with respect to ν. Thus we have a one-to-one correspondence between $\mathcal{G}_\nu$ and

$$\left\{\mu \in \mathbb{M}(X, \mathcal{G}) : \mu(B) = \int_B 1_A \mathrm{d}\nu,\ A \in \mathcal{G}_\nu\right\}.$$

Thus Z is also a measure-indexed stochastic process.

(3) **Generalised Random Fields are Measure-indexed Stochastic Processes:**

This is trivial.

(4) **Set-indexed Random Fields are Generalised Stochastic Processes:**

Let $(X, \mathcal{G}, \nu)$ be a measure-space and let Z be a set-indexed random field with index-set $\mathcal{G}_\nu$. Now for each $A \in \mathcal{G}_\nu$ we can define a test-function, φ, with support A. Then we can construct a set $T \subset \mathcal{T}_d$ which, for each $A \in \mathcal{G}_\nu$, contains exactly one test-function with support A. Then we have a one-to-one correspondence between $\mathcal{G}_\nu$ and T. Thus Z is also a generalised stochastic process.

(5) **Generalised Stochastic Processes can be Generalised Random Fields:**

Let Z be a generalised stochastic process with index-set $\mathcal{T}_d^0 = \{\varphi \in \mathcal{T}_d : \int_{\mathbb{R}^d} \varphi(x)\mathrm{d}x = 0\}$. Then we have a one-to-one correspondence between $\mathcal{T}_d^0$ and the set $\{\mu \in \mathcal{M} : \mu(B) = \int_B \varphi(x)\mathrm{d}x, \text{ where } \varphi \in \mathcal{T}_d\}$, and thus Z is also a generalised random field. That is, only generalised stochastic processes where the index-set is restricted to $\mathcal{T}_d^0$ are generalised random fields.

(6) **Set-indexed Random Fields can be Generalised Random Fields:**

Let $(X, \mathcal{G}, \nu)$ be a measure-space and let Z be a set-indexed random field. For any $A \in \mathcal{G}_\nu$ we can define a measure μ with 1_A as density with respect to ν. But since

$$\mu(X) = \int \mathrm{d}\mu = \int 1_A \mathrm{d}\nu = \nu(A),$$

we cannot construct a one-to-one correspondence between $\mathcal{G}_\nu$ and a subset of $\{\mu \in \mathbb{M}(X, \mathcal{G}) : \mu(X) = 0\}$. So instead we restrict the index-set of Z to $\mathcal{G}_{\nu,0} = \{A \in \mathcal{G}_\nu : \nu(A) = 0\}$. Then we have a one-to-one correspondence between the sets $\mathcal{G}_{\nu,0}$ and

$$M_0(X, \mathcal{G}, \nu) = \left\{\mu \in \mathbb{M}(X, \mathcal{G}) : \mu(X) = 0, \mu(B) = \int_B 1_A \mathrm{d}\nu,\ A \in \mathcal{G}_{\nu,0}\right\}$$

and Z is a generalised random field. However in some cases further restrictions may be necessary, since there may be sets, $A, \tilde{A} \in \mathcal{G}_{\nu,0}$, where $\int_B 1_A \mathrm{d}\nu = \int_B 1_{\tilde{A}} \mathrm{d}\nu$ for all $B \in \mathcal{G}$.

Note that the most common and practically useful setting for set-indexed random fields is when $X = \mathbb{R}^d$, $\mathcal{G} = \mathbb{B}_d$ and ν is the Lebesgue measure. Then the set-indexed random field, Z, discussed above, would be defined only on countable unions of points in $\mathbb{R}^d$. This object is not useful as we show here. For any $A \in \mathcal{G}_{\nu,0}$ we would have $\mathbb{R}^d \ni x_1, x_2, \ldots$ such that $A = \bigcup_{i=1}^\infty \{x_i\}$. Then for any measure in $M_0(\mathbb{R}^d, \mathbb{B}_d, \nu)$ we get

$$\mu(B) = \int_B \mathrm{d}\mu = \int_B 1_A(x)\mathrm{d}x = \int_{B \cap \bigcup_{i=1}^\infty \{x_i\}} \mathrm{d}x = \int_{\bigcup_{i=1}^\infty B \cap \{x_i\}} \mathrm{d}x = \sum_{i=1}^\infty \left|B \cap \{x_i\}\right|,$$

and since

$$B \cap \{x_i\} = \begin{cases} \emptyset & \text{when } x_i \notin B \\ \{x_i\} & \text{when } x_i \in B \end{cases},$$

we get that $\left|B \cap \{x_i\}\right| = 0$ for all $i \in \mathbb{N}$. Thus the only way to construct a one-to-one correspondence is to restrict the index-set of Z to $\{\emptyset\}$, in which case Z is only a single random variable.

The relations described above are summarised in Figure 3.1.

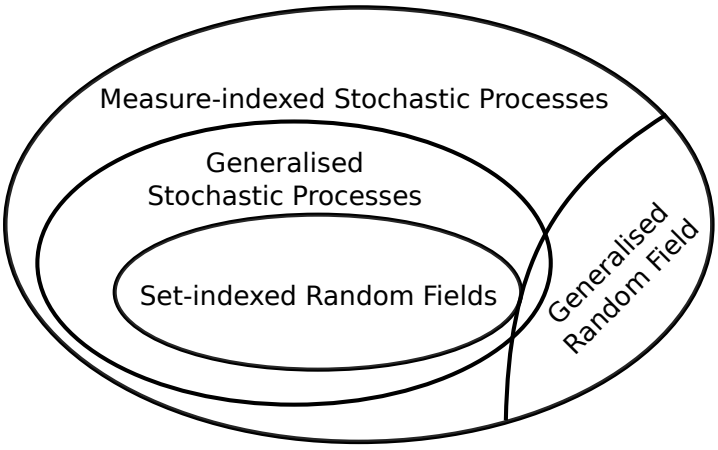

Figure 3.1: This Figure shows the relation between measure-indexed stochastic processes, generalised random fields, generalised stochastic processes and set-indexed random fields.

3.2 De Wijs Process

This section is based on Mondal [2015].

We can now define the central process of this project.

Definition 3.5 (De Wijs Process). *Let Z be a zero-mean Gaussian generalised random field with index-set $\mathcal{W}$, which is the set of all measures $\mu \in \mathcal{M}$ that satisfy*

$$\int\int |\log\|x-y\|\,|\mathrm{d}\mu^+(x)\mathrm{d}\mu^+(y) < \infty,$$

where $\mu^+(A) = |\mu(A)|$ for all $A \in \mathbb{B}_d$. Z is called the De Wijs process if it is stationary, isotropic, self-similar and has generalised covariance function $K(x,y) = -\log\|x-y\|$, that is

$$\mathbb{C}\mathrm{ov}\left[Z(\mu), Z(\nu)\right] = -\int\int \log\|x-y\|\,\mathrm{d}\mu(x)\mathrm{d}\nu(y),$$

for all $\mu, \nu \in \mathcal{W}$.

We note that by Definition 3.2 the index set of a generalised random field must be a subspace of $\mathcal{M}$. This means that in order for the De Wijs process to be well defined, we must show that $\mathcal{W}$ is a subspace of $\mathcal{M}$. Obviously $\mathcal{W} \subset \mathcal{M}$, so we need to show that $\mathcal{W}$ is a vector-space. For this purpose we state part of Mattner [1997, Corollary 2.5], here formulated as a lemma.

Lemma 3.6. *Let f be an infinitely differentiable function such that $(-1)^n f^{(n)}$ is non-constant and non-negative everywhere for $n \geq k$ for some non-negative integer k. Then the set*

$$M^f(\mathbb{R}^d) = \left\{\mu \in M(\mathbb{R}^d) : \int\int |f(\|x-y\|^2)|\mathrm{d}\mu^+(x)\mathrm{d}\mu^+(y) < \infty, \text{ and } \int x^\alpha \mathrm{d}\mu(x) = 0, \text{ for } |\alpha| \leq k-1\right\}$$

where $M(\mathbb{R}^d)$ is the set of all signed radon measures on $\mathbb{R}^d$, is a vector space.

We do not prove this result in this project. Applying Lemma 3.6 gives the following proposition.

Proposition 3.7. *The De Wijs process is well-defined or, equivalently, $\mathcal{W}$ is a vector space.*

Proof. The idea for this proof originates in [Mondal, 2015, p. 5]

Let $f(r) = -\log\sqrt{r}$. Thus we have that $f(r) < 0$ for $r > 1$, but since

$$f^{(n)} = (-1)^n \frac{1}{2(n-1)!r^n},$$

we have that $(-1)^n f^{(n)}(r) > 0$ for all $r \in (0,\infty)$ when $n \geq 1 = k$. Thus the requirement that μ satisfies

$$\int x^\alpha \mathrm{d}\mu(x) = 0, \text{ for } |\alpha| \leq k-1$$

can be restated as $\mu(\mathbb{R}^d) = 0$, and thus we get that $\mathcal{W} = M^f(\mathbb{R}^d)$, which is a vector space by Lemma 3.6. □

We now state Mattner [1997, Corollary 2.4]

Proposition 3.8. *Let f be an infinitely differentiable function such that $(-1)^n f^{(n)}$ is non-constant and non-negative everywhere for $n \geq k$ for some non-negative integer k, and let μ be a measure in $M(\mathbb{R}^d)$ that is not constantly zero, which satisfies*

$$\int_{\mathbb{R}^d} x^\alpha \mathrm{d}\mu(x) = 0 \tag{3.4}$$

for $|\alpha| \leq k-1$. Then we get that

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(\|x-y\|^2)\mathrm{d}\mu(x)\mathrm{d}\mu(y) > 0. \tag{3.5}$$

We note that the function $-\log\sqrt{r}$ satisfies the conditions of Proposition 3.8 with $k=1$. We also note that when $k=1$ the condition in (3.4) amounts to $\mu(\mathbb{R}^d)=0$, thus the measures satisfying this condition are all non-zero Radon measures in $\mathcal{M}$. Applying the proposition to $-\log\sqrt{r}$ then implies that the functional

$$C(\mu,\nu) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} -\log(\|x-y\|)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$$

satisfies

$$C(\mu,\mu) \geq 0 \tag{3.6}$$

for all $\mu \in \mathcal{W}$, with $C(\mu,\mu)=0$ if and only if $\mu(A)=0$ for all $A \in \mathbb{B}_d$. In other words, C is positive definite. This means that $\mathcal{W}$ is a normed inner product space, with inner product $\langle\mu,\nu\rangle = C(\mu,\nu)$ and norm

$$\|\mu\|_{\mathcal{W}} = \langle\mu,\mu\rangle^{1/2}.$$

It is now straight-forward to show the existence of the de Wijs process. First define

$$\mathbb{C}\mathrm{ov}\left[Z(\mu),Z(\nu)\right] = C(\mu,\nu), \tag{3.7}$$

which is positive definite. Thus by using the same arguments as in (i), (ii) and (iii) in the proof of Theorem 2.12 we can show that there exists a generalised random field, Z, such that for any $\mu_1,\ldots,\mu_n \in \mathcal{W}$ we have

$$\begin{bmatrix} Z(\mu_1) \\ \vdots \\ Z(\mu_n) \end{bmatrix} \sim N\left(\begin{bmatrix} m(\mu_1) \\ \vdots \\ m(\mu_n) \end{bmatrix}, \begin{bmatrix} C(\mu_1,\mu_1) & \cdots & C(\mu_1,\mu_n) \\ \vdots & \ddots & \vdots \\ C(\mu_n,\mu_1) & \cdots & C(\mu_n,\mu_n) \end{bmatrix}\right),$$

where C is given as in (3.7), $m(\mu_1)=\cdots=m(\mu_n)=0$ and where $Z(a\mu+b\nu)=aZ(\mu)+bZ(\nu)$ almost surely. In order to show that the de Wijs process exists, it suffices to show that C is preserved under translation, rotation and scaling. Denoting $\mu_h(A)=\mu(A+h)$ for a $h\in\mathbb{R}^d$ we get

$$\begin{aligned} C(\mu_h,\nu_h) &= -\int\int \log\|x-y\|\,\mathrm{d}\mu_h(x)\mathrm{d}\nu_h(y) \\ &= -\int\int \log\|(\tilde{x}-h)-(\tilde{y}-h)\|\,\mathrm{d}\mu_h(\tilde{x}-h)\mathrm{d}\nu_h(\tilde{y}-h) \\ &= -\int\int \log\|\tilde{x}-\tilde{y}\|\,\mathrm{d}\mu(\tilde{x})\mathrm{d}\nu(\tilde{y}) = C(\mu,\nu), \end{aligned}$$

where the second equality comes from performing the change of variable $\tilde{x} = x + h$. Denoting $\mu_R(A) = \mu(RA)$ where R is a rotation matrix, it follows from similar arguments that $C(\mu_R, \nu_R) = C(\mu, \nu)$. Finally we show that the covariance function is invariant under changes of scale. Denoting $\mu_a(A) = \mu(aA)$ for $a \in \mathbb{R}$ we get

$$\begin{aligned} C(\mu_a, \nu_a) &= -\int\int \log\|x - y\| \, \mathrm{d}\mu_a(x)\mathrm{d}\nu_a(y) \\ &= -\int\int \log\|(\tilde{x} - \tilde{y})/a\| \, \mathrm{d}\mu_a(\tilde{x}/a)\mathrm{d}\nu_a(\tilde{y}/a) \\ &= -\int\int \log\|(\tilde{x} - \tilde{y})\| \, \mathrm{d}\mu(\tilde{x})\mathrm{d}\nu(\tilde{y}) + \int\int \log|a|\mathrm{d}\mu(\tilde{x})\mathrm{d}\nu(\tilde{y}) \\ &= -\int\int \log\|(\tilde{x} - \tilde{y})\| \, \mathrm{d}\mu(\tilde{x})\mathrm{d}\nu(\tilde{y}) = C(\mu, \nu), \end{aligned}$$

where the second equality comes from performing the change of variable $\tilde{x} = ax$ and the second to last equality comes from the fact that

$$\int\int \log|a|\mathrm{d}\mu(\tilde{x})\mathrm{d}\nu(\tilde{y}) = \log|a| \int\int \mathrm{d}\mu(\tilde{x})\mathrm{d}\nu(\tilde{y}) = 0,$$

since

$$\int \mathrm{d}\mu(\tilde{x}) = \mu(\mathbb{R}^d) = 0.$$

This proves that the de Wijs Process exists.

3.3 Restricted Maximum Likelihood Estimate for the WWN Process

In this section we consider the De Wijs Plus White Noise process, which is studied in Clifford and McCullagh [2006].

Definition 3.9 (De Wijs Plus White Noise Process). *Let Z be a generalised random field with index-set $\mathcal{W}$, such that when $\mu \in \mathcal{W}$ has a density wrt. the Lebesgue measure on the form $\sum_{i=1}^k a_i 1_{A_i}(x)$ where $A_1, \ldots, A_k \in \mathbb{B}_d$, then*

$$Z(\mu) = \sum_{i=1}^k a_i Y(A_i),$$

where Y is a set-indexed random field. If Z furthermore has generalised covariance function on the form

$$K(\|x - y\|) = \sigma_0^2 \delta_{x-y} - \sigma_1^2 \log\|x - y\|,$$

where $\sigma_0^2, \sigma_1^2 > 0$, then Z is said to be a De Wijs Plus White Noise (WWN) process.

In this section we seek to describe the estimation of the covariance parameters σ_0^2 and σ_1^2. Suppose we have data on the form $Y = (Y(A_1), \ldots, Y(A_n))^\top$ where $A_1, \ldots, A_n \in \mathbb{B}_d$ and $A_i \cap A_j = \emptyset$ for $i \neq j$. Furthermore let Q be an $n \times N$ matrix. Then suppose

$$Z = Q^\top Y = \begin{bmatrix} q_{11}Y(A_1) + \cdots + q_{n1}Y(A_n) \\ \vdots \\ q_{1N}Y(A_1) + \cdots + q_{nN}Y(A_n) \end{bmatrix} = \begin{bmatrix} Z(\mu_1) \\ \vdots \\ Z(\mu_N) \end{bmatrix},$$

where $Z(\cdot)$ is a WWN process, and μ_i has density $\sum_{j=1}^n q_{ji} 1_{A_j}$ wrt. the Lebesgue measure for $i = 1, \dots, n$. Since the distribution of Z is known, we can perform maximum likelihood estimation on $Q^\top Y$.

Now suppose that $\mathbb{E}[Y] \in \mathcal{X}$. Then we may choose Q such that the columns form an orthonormal basis of $\mathcal{X}^\top$ and thus $\mathbb{E}\left[Q^\top Y\right] = 0$. Then for $\dim(\mathcal{X}) = p$, we get $N = n - p$. Performing likelihood estimation on $Q^\top Y$ corresponds to using restricted maximum likelihood estimation (REML) as described in [Waagepetersen, 2019, slide 17]. Usually REML is used in order to eliminate the mean. In this case the main purpose is to ensure that the distribution is well-defined, however since the mean is not of interest here, eliminating it is still useful.

As mentioned previously the covariance matrix of Y is not defined. However it is possible to define a matrix, Φ, that mimics the behaviour of a covariance matrix of Y, in the sense that

$$\mathbb{V}\mathrm{ar}[Z] = \mathbb{V}\mathrm{ar}\left[Q^\top Y\right] = Q^\top \Phi Q. \tag{3.8}$$

As we shall see, this is satisfied by the matrix, which satisfies that $\Phi_{ij} = B(A_i, A_j)$, where

$$B(A_i, A_j) = \int_{A_i} \int_{A_j} \sigma_0^2 \delta_{x-y} - \sigma_1^2 \log\|x - y\| \, \mathrm{d}x\mathrm{d}y, \tag{3.9}$$

where δ_x is Dirac's delta function, which satisfies $\int_A \delta_x \mathrm{d}x = 1[x \in A]$. To write an expression of Φ, we first compute the integral in (3.9). We first note that $\int_{A_i} \delta_{x-y} \mathrm{d}x = 1[y \in A_i]$ and thus

$$\int_{A_i} \int_{A_j} \sigma_0^2 \delta_{x-y} \mathrm{d}x\mathrm{d}y = \sigma_0^2 \int_{A_i \cap A_j} \mathrm{d}y = \sigma_0^2 |A_i \cap A_j| = \sigma_0^2 |A_i| 1[i = j],$$

where the last equality comes from the fact that the sets are disjoint. Then the second term in (3.9) can be expressed as

$$\sigma_1^2 \int_{A_i} \int_{A_j} -\log\|x - y\| \, \mathrm{d}x\mathrm{d}y = \sigma_1^2 |A_i||A_j| \mathrm{Ave}_{A_i \times A_j} \left(-\log\|x - y\|\right),$$

where $\mathrm{Ave}_{A_i \times A_j} \left(-\log\|x - y\|\right)$ refers to the average value of the negative log of the distance between the points x and y, where $x \in A_i$ and $y \in A_j$. Now define V as the matrix where $V_{ij} = \mathrm{Ave}_{A_i \times A_j} \left(-\log\|x - y\|\right)$, and suppose $|A_1| = \cdots = |A_n|$ and denote this volume as $|A|$. Then we get

$$\Phi = \sigma_0^2 |A| I + \sigma_1^2 |A|^2 V. \tag{3.10}$$

Suppose we have measures $\mu_1, \dots, \mu_{n-p}$ such that μ_i has density $\sum_{k=1}^n q_{ki} 1_{A_k}(x)$ wrt. the Lebesgue measure. Then we have

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[Z(\mu_i), Z(\mu_j)\right] &= \int \int K(\|x - y\|) \mathrm{d}\mu_i(x) \mathrm{d}\mu_j(y) \\
&= \int \int K(\|x - y\|) \left(\sum_{k=1}^n q_{ki} 1_{A_k}(x) \right) \left(\sum_{l=1}^n q_{lj} 1_{A_l}(y) \right) \mathrm{d}x\mathrm{d}y \\
&= \sum_{k=1}^n \sum_{l=1}^n q_{ki} q_{lj} \int_{A_k} \int_{A_l} K(\|x - y\|) \mathrm{d}x\mathrm{d}y \\
&= \sum_{k=1}^n \sum_{l=1}^n q_{ki} q_{lj} B(A_k, A_l) = \left[Q^\top \Phi Q\right]_{ij}.
\end{aligned}$$

This shows that (3.8) holds when Φ is given as in (3.10). Before describing restricted maximum likelihood estimation we must verify that the $\mu_1, \dots, \mu_{n-p} \in \mathcal{W}$ and that $Q^\top \Phi Q$ is positive definite. The former is a consequence of the following proposition.

Proposition 3.10. *Let $A_1, \dots, A_n$ be bounded sets in $\mathbb{B}_d$ satisfying $A_i \cap A_j = \emptyset$ for $i \neq j$ and $|A_1| = \cdots = |A_n|$, and let the signed measure μ satisfy $\mu(B) = \sum_{i=1}^n q_i |A_i \cap B|$ for $q_1, \dots, q_n \in \mathbb{R}$, where $\sum_{i=1}^n q_i = 0$, for any $B \in \mathbb{B}_d$. Then $\mu \in \mathcal{W}$.*

Proof. In order to prove the proposition, we must show that

$$\int \int |\log\|x-y\|| \, |\mathrm{d}\mu^+(x)\mathrm{d}\mu^+(y) < \infty,$$

where $\mu^+(B) = |\mu(B)|$ for all $B \in \mathbb{B}_d$. Now define $\tilde{\mu}(B) = \sum_{i=1}^n |q_i||B \cap A_i|$ and note that $\mu^+(B) \leq \tilde{\mu}(B)$ for any $B \in \mathbb{B}_d$, so since $|\log\|x-y\|| > 0$ for all $x, y \in \mathbb{R}^d$ we get

$$\begin{aligned}
&\int \int |\log\|x-y\|| \, |\mathrm{d}\mu^+(x)\mathrm{d}\mu^+(y) \leq \int \int |\log\|x-y\|| \, |\mathrm{d}\tilde{\mu}(x)\mathrm{d}\tilde{\mu}(y) \\
=&\int \int |\log\|x-y\|| \sum_{i=1}^n \sum_{j=1}^n |q_i q_j| 1_{A_i}(x) 1_{A_j}(y) \mathrm{d}x\mathrm{d}y \\
=&\sum_{i=1}^n \sum_{j=1}^n |q_i q_j| \int_{A_i} \int_{A_j} |\log\|x-y\|| \, |\mathrm{d}x\mathrm{d}y \\
=&C + \sum_{i=1}^n q_i^2 \int_{A_i} \int_{A_i} |\log\|x-y\|| \, |\mathrm{d}x\mathrm{d}y,
\end{aligned}$$

where C is the sum of all the terms where $i \neq j$. Since $A_i \cap A_j = \emptyset$ for $i \neq j$, $|\log\|x-y\||$ is finite on $A_i \times A_j$ for $i \neq j$. Furthermore since $A_1, \dots, A_n$ are bounded it follows that $C < \infty$. Since $q_1, \dots, q_n$ are real numbers it is now sufficient to show that

$$\int_{A_i} \int_{A_i} |\log\|x-y\|| \, |\mathrm{d}x\mathrm{d}y < \infty,$$

for any $i \leq n$. Now let $D_i = \{(x, y) \in A_i \times A_i : \|x-y\| < 1\}$. Then

$$\int_{A_i} \int_{A_i} |\log\|x-y\|| \, |\mathrm{d}x\mathrm{d}y = \int_{(A_i \times A_i) \setminus D_i} \log\|x-y\| \, \mathrm{d}x\mathrm{d}y + \int_{D_i} -\log\|x-y\| \, \mathrm{d}x\mathrm{d}y. \tag{3.11}$$

Once again we note that $\log\|x-y\|$ is finite on $(A_i \times A_i) \setminus D_i$ and thus $\int_{(A_i \times A_i) \setminus D_i} \log\|x-y\| \, \mathrm{d}x\mathrm{d}y < \infty$, so we only need to show that $\int_{D_i} -\log\|x-y\| \, \mathrm{d}x\mathrm{d}y$ is finite. Note that

$$\int_{D_i} -\log\|x-y\| \, \mathrm{d}x\mathrm{d}y = \int_{A_i} \int_{B(y,1) \cap A_i} -\log\|x-y\| \, \mathrm{d}x\mathrm{d}y \leq \int_{A_i} \int_{B(y,1)} -\log\|x-y\| \, \mathrm{d}x\mathrm{d}y,$$

where the inequality holds since $-\log\|x-y\| > 0$ when $\|x-y\| < 1$. Now we perform the change of variable $\tilde{x} = x - y$ and get

$$\int_{A_i} \int_{B(y,1)} -\log\|x-y\| \, \mathrm{d}x\mathrm{d}y = \int_{A_i} \int_{B(0,1)} -\log\|\tilde{x}\| \, \mathrm{d}\tilde{x}\mathrm{d}y = |A_i| \int_{B(0,1)} -\log\|\tilde{x}\| \, \mathrm{d}\tilde{x}.$$

Expressing this integral in terms of polar coordinates yields

$$|A_i| \int_{B(0,1)} -\log\|\tilde{x}\| \, \mathrm{d}\tilde{x} = |A_i| \int_{[0,2\pi]^{d-1}} \int_0^1 -r^{d-1} \log r \mathrm{d}r\mathrm{d}\theta = -|A_i|(2\pi)^{d-1} \int_0^1 r^{d-1} \log r \mathrm{d}r.$$

The remaining integral may be solved by using integration by parts, where $u(r) = \log r$ and $v'(r) = r^{d-1}$. Thus we get

$$\int_0^1 r^{d-1} \log r \mathrm{d}r = \left[\frac{r^d}{d} \log r \right]_0^1 - \int_0^1 \frac{r^{d-1}}{d} \mathrm{d}r = -\lim_{r \to 0} \left(\frac{r^d}{d} \log r \right) - \frac{1}{d^2}.$$

Finally we note that

$$\lim_{r\to 0}\left(\frac{r^d}{d}\log r\right) = \frac{1}{d}\lim_{r\to 0}\frac{\log r}{r^{-d}} = -\frac{1}{d^2}\lim\frac{1/r}{r^{-(d+1)}} = -\frac{1}{d^2}\lim_{r\to 0} r^d = 0,$$

where the second equality comes from L'hospital's rule. Thus by (3.11) we have

$$\int_{A_i}\int_{A_i} \left|\log\|x-y\|\right| \mathrm{d}x\mathrm{d}y < \infty,$$

which proves the proposition.

□

Now in order to perform maximum likelihood estimation of $Q^\top Y$, which is normally distributed with mean 0 and covariance matrix $Q^\top \Phi Q$, we must check that $Q^\top \Phi Q$ is positive definite. Let $a \in \mathbb{R}^{n-p}$ be a non-zero vector, and note that

$$Q^\top \Phi Q = \sigma_0^2|A|Q^\top Q + \sigma_1^2|A|^2 Q^\top V Q = \sigma_0^2|A|I + \sigma_1^2|A|^2 Q^\top V Q,$$

since $Q^\top Q = I$ due to the fact that the columns of Q forms an orthonormal basis. Thus the first term of $a^\top Q^\top \Phi Q a$ is $\sigma_0^2|A|\|a\|^2 > 0$, so if $\sigma_1^2|A|^2 a^\top Q^\top V Q a > 0$ then $Q^\top \Phi Q$ is positive definite. First we expand the expression

$$\sigma_1^2|A|^2 a^\top Q^\top V Q a = \sigma_1^2 \sum_{i=1}^{n-p}\sum_{j=1}^{n-p} a_i a_j |A|^2 [Q^\top V Q]_{ij} = \sigma_1^2 \sum_{i=1}^{n-p}\sum_{j=1}^{n-p} a_i a_j \sum_{k=1}^{n}\sum_{l=1}^{n} q_{ki} q_{lj} |A|^2 V_{kl}. \quad (3.12)$$

Now recall that

$$|A|^2 V_{kl} = \int_{A_k}\int_{A_l} -\log\|x-y\| \,\mathrm{d}x\mathrm{d}y = \int\int -\log\|x-y\| \, 1_{A_k}(y) 1_{A_l}(x) \mathrm{d}x\mathrm{d}y.$$

Thus by (3.12) we get

$$\begin{aligned}
\sigma_1^2|A|^2 a^\top Q^\top V Q a &= \sigma_1^2 \sum_{i=1}^{n-p}\sum_{j=1}^{n-p} a_i a_j \sum_{k=1}^{n}\sum_{l=1}^{n} q_{ki} q_{lj} \int\int -\log\|x-y\| \, 1_{A_k}(y) 1_{A_l}(x) \mathrm{d}x\mathrm{d}y \\
&= \sigma_1^2 \sum_{i=1}^{n-p}\sum_{j=1}^{n-p} a_i a_j \int\int -\log\|x-y\| \left(\sum_{k=1}^{n} q_{ki} 1_{A_k}(y)\right)\left(\sum_{l=1}^{n} q_{lj} 1_{A_l}(x)\right) \mathrm{d}x\mathrm{d}y \\
&= \sigma_1^2 \int\int -\log\|x-y\| \left(\sum_{i=1}^{n-p}\sum_{k=1}^{n} a_i q_{ki} 1_{A_k}(y)\right)\left(\sum_{j=1}^{n-p}\sum_{l=1}^{n} a_j q_{lj} 1_{A_l}(x)\right) \mathrm{d}x\mathrm{d}y.
\end{aligned}$$

Now define the measure $\nu(B) = \sum_{i=1}^{n-p}\sum_{k=1}^{n} a_i q_{ki} |B \cap A_k|$. Clearly this is a Radon measure, and since $|A| = |A_1| = \cdots = |A_n|$ we also get that

$$\nu(\mathbb{R}^d) = \sum_{i=1}^{n-p}\sum_{k=1}^{n} a_i q_{ki} |A_k| = |A| \sum_{i=1}^{n-p} a_i \sum_{k=1}^{n} q_{ki} = 0,$$

since $\sum_{k=1}^{n} q_{ki} = 0$ due to the fact that $\mu_i \in \mathcal{M}$. Thus ν is a signed Radon measure in $\mathcal{M}$ which is not constantly zero. Thus by (3.6)

$$\begin{aligned}
\sigma_1^2|A|^2 a^\top Q^\top V Q a &= \sigma_1^2 \int\int -\log\|x-y\| \left(\sum_{i=1}^{n-p}\sum_{k=1}^{n} a_i q_{ki} 1_{A_k}(y)\right)\left(\sum_{j=1}^{n-p}\sum_{l=1}^{n} a_j q_{lj} 1_{A_l}(x)\right) \mathrm{d}x\mathrm{d}y \\
&= \sigma_1^2 \int\int -\log\|x-y\| \,\mathrm{d}\nu(x)\mathrm{d}\nu(y) > 0.
\end{aligned}$$

Now the REML estimate is obtained, by computing the maximum likelihood of $Q^\top Y$. We get the log-likelihood function

$$\ell(\sigma_0^2, \sigma_1^2) = -\frac{1}{2}\log\left|\sigma_0^2|A|I + \sigma_1^2|A|^2 Q^\top V Q\right| - \frac{1}{2} Z^\top \left(Q^\top \Phi Q\right)^{-1} Z. \tag{3.13}$$

In this case it is more useful to perform the following reparametrisation.

$$\Phi = \sigma_0^2\left(|A|I + \frac{\sigma_1^2}{\sigma_0^2}|A|^2 V\right) = \sigma_0^2\left(|A|I + \tau|A|^2 V\right) = \sigma_0^2 W(\tau),$$

where $\tau = \sigma_1^2/\sigma_0^2$. Thus we obtain the log-likelihood

$$\ell(\sigma_0^2, \tau) = -\frac{n-p}{2}\log\sigma_0^2 - \frac{1}{2}\log|Q^\top W(\tau) Q| - \frac{1}{2\sigma_0^2} Z^\top \left(Q^\top W(\tau) Q\right)^{-1} Z. \tag{3.14}$$

Differentiating (3.14) wrt. σ_0^2 yields

$$\frac{\partial \ell(\sigma_0^2, \tau)}{\partial \sigma_0^2} = -\frac{n-p}{2\sigma_0^2} + \frac{1}{2\sigma_0^4} Z^\top \left(Q^\top W(\tau) Q\right)^{-1} Z,$$

and thus the estimate

$$\hat{\sigma}_0^2(\tau) = \frac{Z^\top \left(Q^\top W(\tau) Q\right)^{-1} Z}{n-p}$$

is obtained. Finally inserting $\hat{\sigma}_0^2(\tau)$ into (3.14) yields

$$\ell(\hat{\sigma}_0^2(\tau), \tau) \propto -\frac{n-p}{2}\log\left(Z^\top \left(I + \tau|A|Q^\top V Q\right)^{-1} Z\right) - \frac{1}{2}\log\left|I + \tau|A|Q^\top V Q\right|. \tag{3.15}$$

We compute an estimate of τ by maximising the above numerically.

Chapter 4

Fitting the WWN Process

In this chapter we apply the theory described in Chapter 3 to a practical dataset.

4.1 Data Introduction

In this project we seek to apply the WWN process to a particular dataset. The dataset in question contains measurements of various properties of the forest soil on Barro Colorado Island (BCI) in Panama. Specifically, the properties are the concentration of aluminium, boron, calcium, copper, iron, potassium, magnesium, manganese, sodium, phosphorus, sulphur, zinc, ammonium, nitrate, nitrogen, mineralised ammonium, mineralised nitrate, mineralised nitrogen as well as the pH-value of the soil. Throughout the report these soil properties are, somewhat inaccurately, collectively referred to as mineral concentrations, even when pH is included.

The samples are obtained using a so called LaMotte soil sampling tube [Harms et al., 2004], which is a round tube, extracting a cylinder shaped sample with a diameter of 2.5 cm and a depth of 30.5 cm. The points are sampled on a 1000 by 500 metre area. the dataset contains 200 points located on a regularly spaced grid, with a spacing of 50 metres and offset by 25 metres in both the x- and the y-direction. Furthermore 100 points are sampled off the grid, so the dataset contains 300 points in total. The locations of the sample can be seen in the plot below.

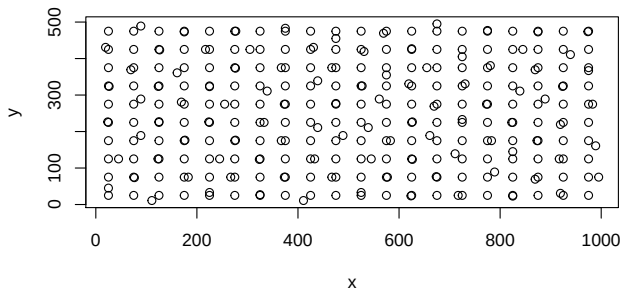

Figure 4.1: Locations of the measurements of soil data.

At each sampling point, three different samples are obtained using the sampling tube, all within a radius of 1 meter of each other. Note that the exact location of the three samples are not reported in the dataset. The size of the sampled areas are $3 \cdot (1.25 \text{ cm})^2 \cdot \pi \approx 14.7\text{cm}^2$ in total.

For more detailed information about how the data has been obtained, the reader may see Harms et al. [2004].

4.2 Computing the Covariance Matrix

In order to perform the estimations described in Section 3.3, we need to compute the entries of the matrix V. In practice this needs to be approximated. Specifically, for every pair of integers (i, j) where $i, j \in \{1, \dots, n\}$ we need to approximate $\mathrm{Ave}_{A_i \times A_j}(-\log(\|x - y\|))$. How this should be done varies depending on how large the areas sampled are. In the BCI dataset the sampled areas are rather small. So let $z_i \in \mathbb{R}^d$ be the centre of A_i. These are the locations that are reported in the dataset. We may then define

$$V_{ij} = -\log(\|z_i - z_j\|).$$

This presents a problem when $i = j$, and thus other methods must be used to define the diagonal.

The mineral concentrations in the dataset are sampled at three points, at most one metre from the location reported in the dataset. More precisely a set A_i, centred at z_i, is actually the union of the areas A_i^1, A_i^2, A_i^3 centred at the points z_{i1}, z_{i2}, z_{i3}, which satisfy $\|z_i - z_{ij}\| \leq 1$ for $j = 1, 2, 3$. These points are not reported in the dataset, but for the purposes of obtaining well-defined values on the diagonal of V, we assume that $\|z_i - z_{ij}\| = 1$ for $j = 1, 2, 3$ and that $\|z_{i1} - z_{i2}\| = \|z_{i1} - z_{i3}\| = \|z_{i2} - z_{i3}\|$. Now we note that

$$\begin{aligned} |A|^2 \mathrm{Ave}_{A_i \times A_i}(-\log\|x - y\|) &= \int_{A_i}\int_{A_i} -\log\|x - y\| \,\mathrm{d}x\mathrm{d}y = \sum_{k=1}^{3}\sum_{l=1}^{3}\int_{A_i^k}\int_{A_i^l} -\log\|x - y\| \,\mathrm{d}x\mathrm{d}y \\ &= \frac{|A|^2}{9}\sum_{k=1}^{3}\sum_{l=1}^{3}\mathrm{Ave}_{A_i^k \times A_i^l}(-\log\|x - y\|) \end{aligned} \tag{4.1}$$

For the terms where $k \neq l$ we may once again use the fact that the areas are small and use the approximation

$$\mathrm{Ave}_{A_i^k \times A_i^l}(-\log\|x - y\|) \approx -\log\|z_{ik} - z_{il}\|\,.$$

By Figure 4.2 we get

$$\mathrm{Ave}_{A_i^k \times A_i^l}(-\log\|x - y\|) \approx -\log\left(2\sin\frac{\pi}{3}\right). \tag{4.2}$$

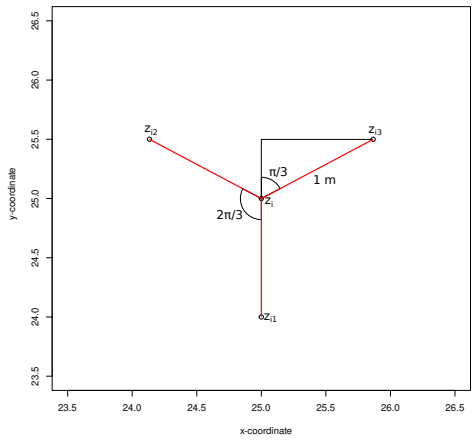

Figure 4.2: This plot shows the assumed locations of the actual sample-points in relation to the points reported in the dataset.

This approach clearly cannot be used to approximate the terms where $k = l$. However it turns out that $\text{Ave}_{B\times B}(-\log\|x-y\|)$ can be computed exactly when B is a square. Specifically, by Clifford [2005, p. 156 and p. 160], it holds that

$$\text{Ave}_{B\times B}(-\log\|x-y\|) = \frac{25-4(\pi+\log 2)}{12} - \log b,$$

when B has side-length b. Thus if we assume that A_i^k is a square of side-length $\sqrt{|A|/3}$, we obtain

$$\text{Ave}_{A_i^k\times A_i^k}(-\log\|x-y\|) = \frac{25-4(\pi+\log 2)}{12} - \frac{1}{2}\log|A| + \frac{1}{2}\log 3. \tag{4.3}$$

Using (4.1), (4.2), (4.3), and the fact that $|A| = 0.00147$ we define

$$V_{ii} = -\frac{2}{3}\log\left(2\sin\frac{\pi}{3}\right) + \frac{1}{3}\left(\frac{25-4(\pi+\log 2)}{12} - \frac{1}{2}\log 0.00147 + \frac{1}{2}\log 3\right). \tag{4.4}$$

Finally we note that due to the small size of $|A|$, it may be difficult to get a sense of the relative sizes of the noise and the signal. To see this note that when $|A| = 0.00147$ the covariance matrix is

$$|A|\sigma_0^2 I + |A|^2\sigma_1^2 V = 0.00147\sigma_0^2 I + 0.00000216\sigma_1^2 V.$$

This means that, even when the two terms are of similar size, σ_1^2 becomes much larger than σ_0^2. A way to remedy this is to scale the locations of all points by a factor of $1/\sqrt{|A|}$. If $|A_s|$ is the size of the scaled areas, then

$$|A_s| = \left|A/\sqrt{|A|}\right| = |A|/|A| = 1.$$

Since V is determined by the distances between points, it is also effected by scaling, but the approach described above still applies. Let V^s the matrix obtained by performing the steps above after scaling, and let z_i^s be the centre-point after scaling the locations. Then the off-diagonal elements are

$$V_{ij}^s = -\log\left\|z_i^s - z_j^s\right\| = -\log\left\|(z_i - z_j)/\sqrt{0.00147}\right\| = -\log\|z_i - z_j\| + \frac{1}{2}\log 0.00147,$$

and the diagonal elements are

$$V_{ii}^s = -\frac{2}{3}\log\left(2\sin\frac{\pi}{3}/\sqrt{0.00147}\right) + \frac{1}{3}\left(\frac{25-4(\pi+\log 2)}{12} + \frac{1}{2}\log 3\right). \tag{4.5}$$

Now we have

$$\Phi = \sigma_0^2 I + \sigma_1^2 V^s.$$

All data-analysis performed in the following sections is done using the scaled points, where V refers to V^s and σ_0^2 and σ_1^2 refers to parameters in the model where V^s is used.

4.3 Fitting the WWN Process to the BCI Soil Data

In this section we fit a WWN process to the dataset described in Section 4.1. The measurements of concentrations in the dataset corresponds to the vector $Y = (Y(A_1), \ldots, Y(A_n))^\top$ in Section 3.3. We assume that Y has constant mean, and thus we must choose a matrix Q, whose columns form a linearly independent basis of the orthogonal complement to the span of 1_{300}. The function `Null()` in the R-package `MASS` computes the left nullspace of a matrix. That is, the first two lines in R-code 4.1

```
X <- matrix(1,300,1)
Q <- Null(X)
Q <- Q[c(2:300,1),]
```

Code 4.1: R-code for automatic constructing a contrast matrix

returns a matrix, Q, such that $Q^\top X = 0$ where $X = 1_{300}$. The third line simply performs row exchanges, which gives the contrasts a slightly more useful interpretation. The result of the parameter estimation is invariant under the choice of Q, so any choice satisfying $Q^\top X = 0$ will do, so the row-exchanges does not effect the estimations. But since the resulting matrix Q has the form

$$Q = \begin{bmatrix} I - 0.00315 \cdot 1_{299} 1_{299}^\top \\ -0.0577 \cdot 1_{299}^\top \end{bmatrix},$$

it follows that for $Z(\mu_i) = [Q^\top Y]_{i\cdot}$, $Y(A_i)$ has a weight close to 1 (specifically 0.99685) and $Y(A_j)$ for $j \neq i$ has a weight close to zero (specifically -0.00315 for $Y(A_j)$ where $300 \neq j \neq i$ and -0.0577 for $Y(A_{300})$). Thus we can in some sense consider $Z(\mu_i)$ to be 'located' at z_i, the centre of the set A_i. The profile-likelihood of τ is implemented below.

```
reml <- function(tau,z,A,V,Q){
  q <- length(z)
  cov.mat <- eye(q)+tau*A*t(Q) %*% V %*% Q
  d <- as.numeric(determinant(cov.mat)$modulus)
  cov.mat.inv <- solve(cov.mat)
  l <- -q/2*log(t(z) %*% cov.mat.inv %*% z)-0.5*d
  return(l)
}
```

Code 4.2: Here the function to compute the profile-likelihood of τ is shown.

Analysis of Aluminium Data

We now apply the above functions to the aluminium concentrations. We obtain the log-likelihood shown in Figure 4.3.

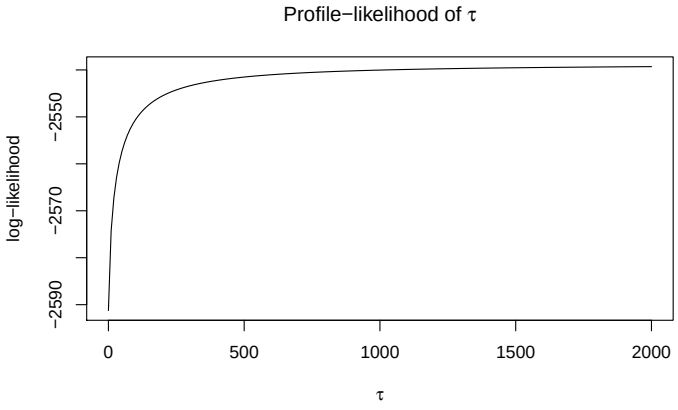

Figure 4.3: Plot of the profile-likelihood as a function of τ.

Here we encounter a problem, since the likelihood function does not have a unique maximum. In fact the likelihood function increases asymptotically and thus it has no maximum at all. A potential explanation is that $\sigma_0^2 = 0$, which cannot be detected in this setup.
To remedy this, a reparametrisation may be performed.

$$\begin{aligned}\Phi &= \sigma_0^2|A|I + \sigma_1^2|A|^2V = (\sigma_0^2+\sigma_1^2)|A|I + \sigma_1^2(|A|^2V - |A|I)\\ &= (\sigma_0^2+\sigma_1^2)\Big(|A|I + \frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2}(|A|^2V - |A|I)\Big)\\ &= \gamma_0|A|\Big(I + \gamma_1(|A|V - I)\Big),\end{aligned}$$

where $\gamma_0 = \sigma_0^2 + \sigma_1^2$ and $\gamma_1 = \sigma_1^2/(\sigma_0^2+\sigma_1^2)$. Now proceeding as in Section 3.3 we obtain

$$\hat{\gamma}_0(\gamma_1) = \frac{Z^\top\Big(I + \gamma_1(|A|Q^\top VQ - I)\Big)^{-1}Z}{|A|(n-p)},$$

and

$$\ell(\hat{\gamma}_0(\gamma_1),\gamma_1) = -\frac{n-p}{2}\log\left(Z^\top\Big(I+\gamma_1(|A|Q^\top VQ - I)\Big)^{-1}Z\right) - \frac{1}{2}\log\Big|I+\gamma_1(|A|Q^\top VQ - I)\Big|.$$

Here we have $\gamma_1 \in [0,1]$, where $\gamma_1 = 1$ means that $\sigma_0^2 = 0$ and $\gamma_1 = 0$ means that $\sigma_1^2 = 0$. Since $\gamma_1 = \frac{\sigma_1^2}{\sigma_1^2+\sigma_0^2} \Leftrightarrow \gamma_1\sigma_0^2 + (\gamma_1-1)\sigma_1^2 = 0$, we get the linear equation system

$$\begin{bmatrix}1 & 1\\ \gamma_1 & \gamma_1 - 1\end{bmatrix}\begin{bmatrix}\sigma_0^2\\ \sigma_1^2\end{bmatrix} = \begin{bmatrix}\gamma_0\\ 0\end{bmatrix}. \tag{4.6}$$

Solving (4.6) yields the unique solution $(\sigma_0^2,\sigma_1^2) = \Big(\gamma_0(1-\gamma_1),\gamma_0\gamma_1\Big)$, and thus estimating γ_0 and γ_1 is sufficient.

For the aluminium-concentration-data we obtain the profile likelihood of γ_1 seen in Figure 4.4.

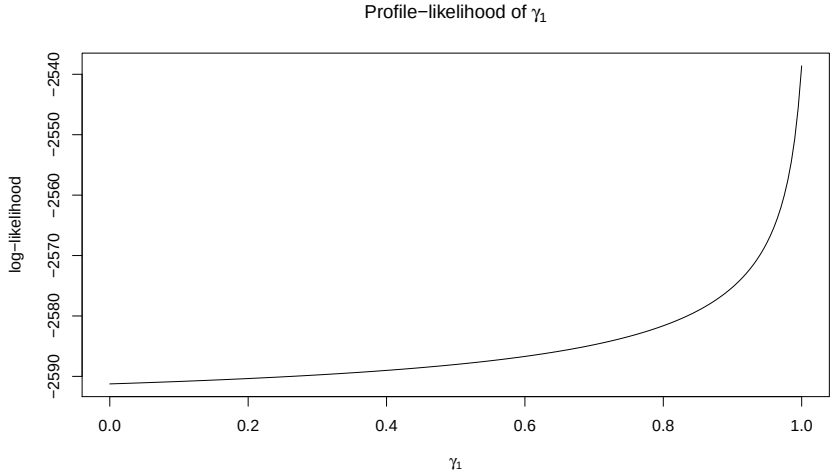

Figure 4.4: Plot of the profile-likelihood as a function of γ_1.

Performing REML-estimation yields $\hat{\gamma}_1 = 1$ as expected and $\hat{\gamma}_0 = 15,129$, and thus we get $\hat{\sigma}_0^2 = 0$ and $\hat{\sigma}_1^2 = 15,129$. This suggests that this data is noise-free, which seems unlikely. This behaviour and potential explanations are discussed in further detail in Section 4.4. For now we analyse a more well-behaved dataset.

Analysis of the Mineralised Nitrate Data

We now consider the concentrations of mineralised nitrate. The profile-likelihood of γ_1 for the mineralised nitrate concentrations in the BCI dataset can be seen in Figure 4.5.

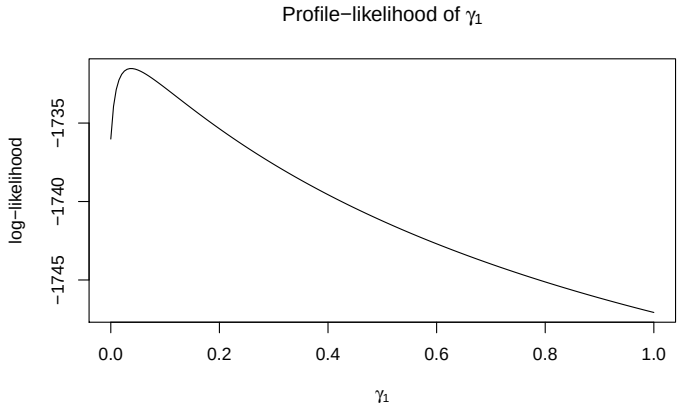

Figure 4.5: Plot of the profile-likelihood as a function of γ_1 for the mineralised nitrate concentrations .

Performing the estimations for the BCI dataset, we obtain the estimates

$$\hat{\gamma}_1 = 0.0373, \quad \hat{\gamma}_0 = 299.8$$

and thus

$$\hat{\sigma}_0^2 = 299.8 \cdot 0.9627 = 288.6, \quad \hat{\sigma}_1^2 = 299.8 \cdot 0.0373 = 11.19.$$

One way to get a rough idea about how good the fit is, is by simulating contrasts using the estimated values, and comparing them to the contrasts found in the data. That is we simulate

$Z_s \sim N\left(0, \hat{\sigma}_0^2 I + \hat{\sigma}_1^2 Q^\top V Q\right)$, and compare this to the data $Z = Q^\top Y$. The results can be seen in Figure 4.6.

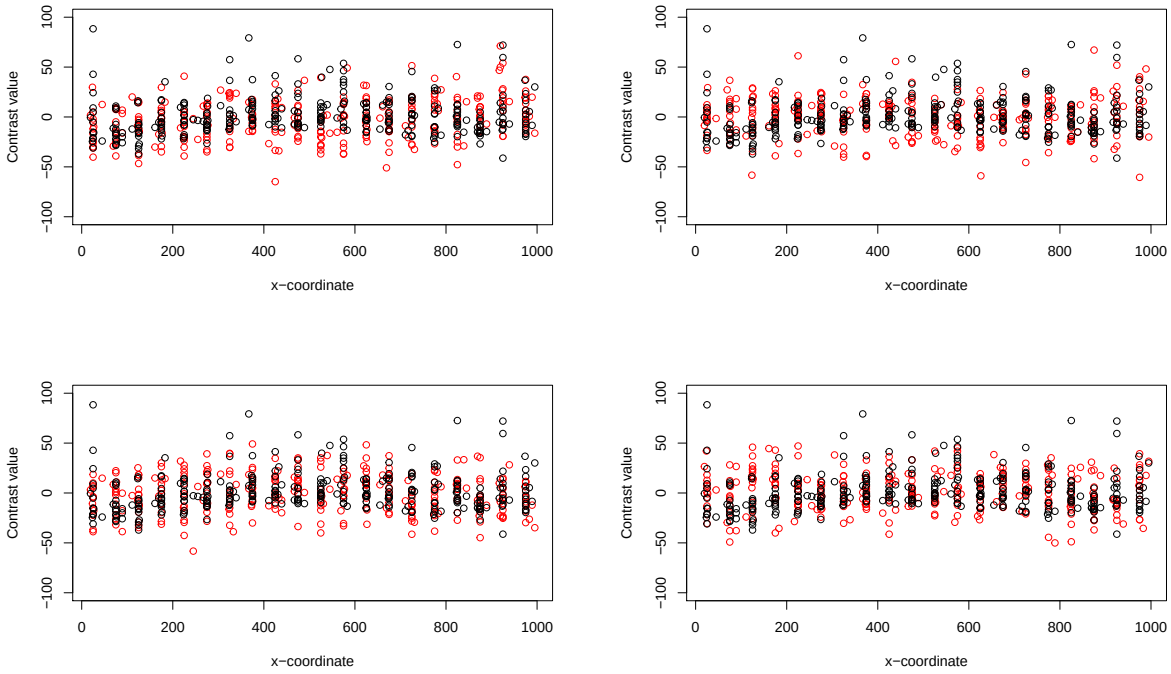

Figure 4.6: Black dots are the contrasts from the data, red dots are contrasts simulated from a WWN process with the fitted values, $\hat{\sigma}_0^2 = 288.6$ and $\hat{\sigma}_1^2 = 11.19$.

Here we see that the data and the simulations are reasonably similar. They tend to have outliers at different locations, but that is to be expected due to random variation. Overall these simulations seems consistent with the hypothesis that the data follows the fitted model.

4.4 Simulation Study

In this section we perform a simulation study of the WWN process. This means that we specify some parameter values and perform many simulations with the specified parameter-values, estimate the parameters given the simulations and finally examine the behaviour of the estimated values. First we use the values $\sigma_0^2 = 10$ and $\sigma_1^2 = 20$.

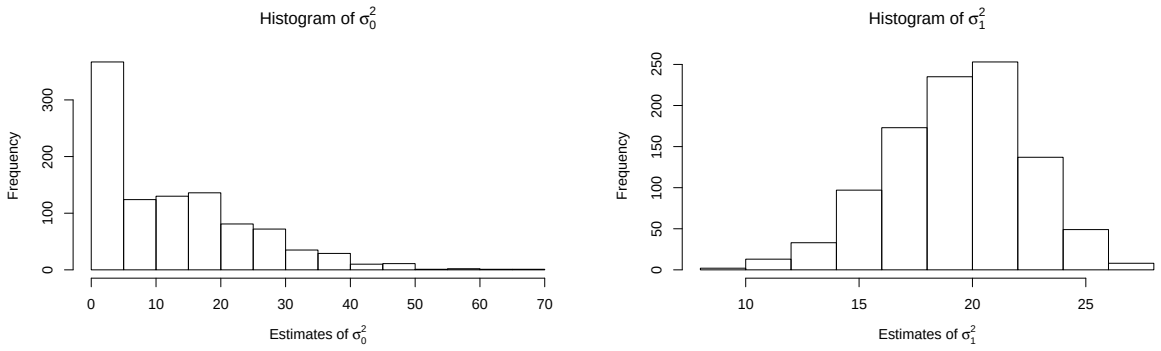

Figure 4.7: Histograms of the estimated values of σ_0^2 and σ_1^2 based on simulations where the true values are $\sigma_0^2 = 10$ and $\sigma_1^2 = 20$.

On the right in Figure 4.7 we see that σ_1^2 is estimated reasonably accurately most of the time. Indeed the mean of the estimates is 19.45, which is quite close to the correct value, and the standard deviation is 3.13. Things are much worse for the estimations of σ_0^2. Here we see that it is rarely accurate and most of the time substantially underestimated. In fact $\hat{\sigma}_0^2 = 0$ for about 25% of the estimates. This behaviour may explain what we see in Section 4.3 for the aluminium data. It is also worth noting that, in some cases, σ_0^2 is estimated to be much larger than the true value. In fact for 9 % of the simulations we get $\hat{\sigma}_0^2 > 30$. However Figure 4.8 shows that these cases coincide with the cases where σ_1^2 is underestimated.

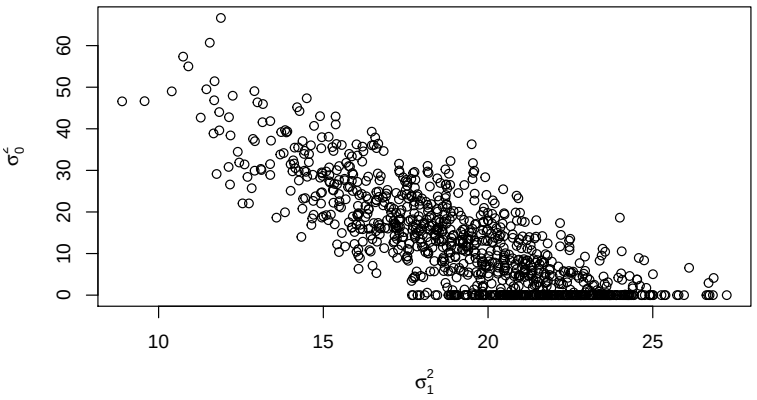

Figure 4.8: This plot shows the correlation between the estimated values for σ_0^2 and σ_1^2, based on simulated data with the true parameter-values $\sigma_0^2 = 10$ and $\sigma_1^2 = 20$.

There are several potential explanations for this behaviour. First note that while we have shown in Section 3.3 that $Q^\top VQ$ is positive definite in the theoretical setting, in practice we have to approximate the entries in V. We note that if the approximated covariance matrix $Q^\top \Phi Q = \sigma_0^2 I + \sigma_1^2 Q^\top VQ$ was not positive definite, then R would return an error when simulation is attempted, but it could be the case that $Q^\top \Phi Q$ is positive definite, but not $Q^\top VQ$. To check this we simply evaluate the eigenvalues, $\lambda_1 > \cdots > \lambda_{n-1}$, of $Q^\top VQ$. Here we get

$$\lambda_1 = 150, \quad \lambda_{n-1} = 1.85,$$

so obviously the matrix is positive definite. Furthermore we also have that the condition number is $150/1.85 = 79.8$, so it is also well-conditioned. This rules out numeric issues as an explanation.

Another explanation may be that for the chosen parameter-values, the term $\sigma_1^2 V$ dominates the term $\sigma_0^2 I$ to the point where the noise term is lost in some of the estimations. This would be consistent with what we see for the mineralised nitrate concentrations in Section 4.3, where $\hat{\sigma}_0^2 > 0$. Here $\hat{\sigma}_0^2$ was much larger than $\hat{\sigma}_1^2$.

The noise term may be dominated by the signal term, if the diagonal entries in $Q^\top V Q$ are much larger than 1. The diagonal elements of $Q^\top V Q$ are not identical, and they vary between 6.54 and 7.78. This means that the true values of diagonal elements of $Q^\top V Q$ varies between $10 + 6.54 \cdot 20 = 140.8$ and $10 + 7.78 \cdot 20 = 165.6$, of which the noise term only contributes 10.

To test whether the behaviour in Figure 4.7 is caused by the noise term contributing too little in relation to the signal term, we perform the simulations for values, where σ_0^2 is much larger than σ_1^2. Specifically let $\sigma_0^2 = 50$ and $\sigma_1^2 = 5$. Then the diagonal elements of $Q^\top V Q$ varies from 82.7 to 88.9 of which the noise term contributes 50, that is the noise term contributes more than the signal term. When performing these simulations, the following histograms are obtained.

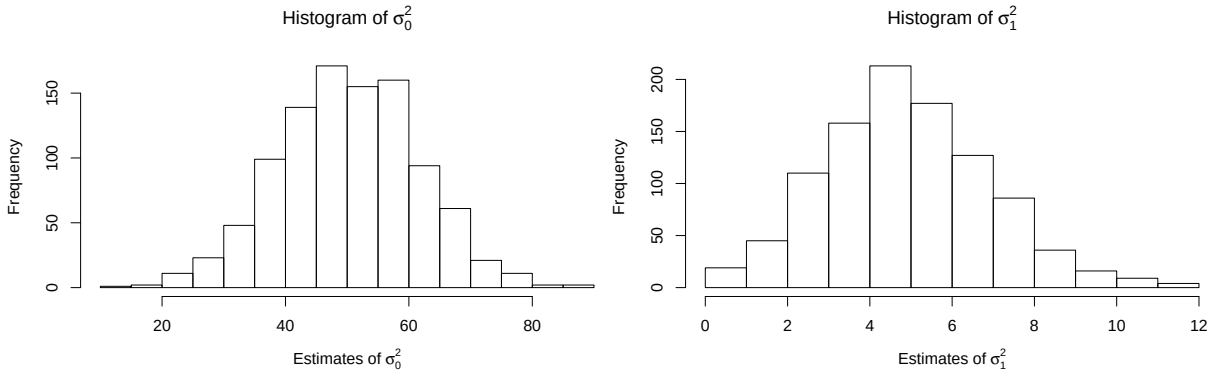

Figure 4.9: Histograms of the estimated values of σ_0^2 and σ_1^2 for simulated data with true values $\sigma_0^2 = 50$ and $\sigma_0^2 = 5$.

What we see in Figure 4.9 backs up the hypothesis. We see that the histogram of estimates of σ_0^2 is concentrated around the true value. Indeed the mean value is 50.22. Furthermore $\hat{\sigma}_0^2 = 0$ was not even obtained a single time. In fact we do see three cases where $\hat{\sigma}_1^2 = 0$ is obtained, but overall σ_1^2 tends to be estimate fairly accurately too (the mean of the estimates of σ_1^2 is 4.93).

The above simulations seem to suggest that, due to the fact that the diagonal of $Q^\top V Q$ is much larger than the diagonal of I, for some datasets the signal term dominates to the point, where the noise term cannot be detected. But all this means that we cannot determine the size of the noise term. It does not necessarily make the resulting models worthless. On Figure 4.7 we see that while the estimations of σ_0^2 tended to yield poor results, the results for σ_1^2 were still reasonably accurate. This suggests that, in general, estimates of σ_0^2 for the WWN Process for the BCI dataset cannot be trusted, but estimates of σ_1^2 are, generally, useful. Furthermore if the reason we obtain $\hat{\sigma}_0^2 = 0$ is that the noise term contributes so little that it becomes undetectable, then this is unlikely to be cause for concern, since the overall covariance is likely still reasonably accurately estimated. To check this we return to the simulations for $\sigma_0^2 = 10$ and $\sigma_1^2 = 20$ and check the estimates of $\mathbb{V}\text{ar}\left[Z(\mu_1)\right] = \sigma_0^2 + \sigma_1^2 [Q^\top V Q]_{11}$. Since $[Q^\top V Q]_{11} = 7.78$, the true value is $10 + 20 \cdot 7.78 = 165.6$. On Figure 4.10 we see the estimated values of $\mathbb{V}\text{ar}\left[Z(\mu_1)\right]$ based on

the simulations.

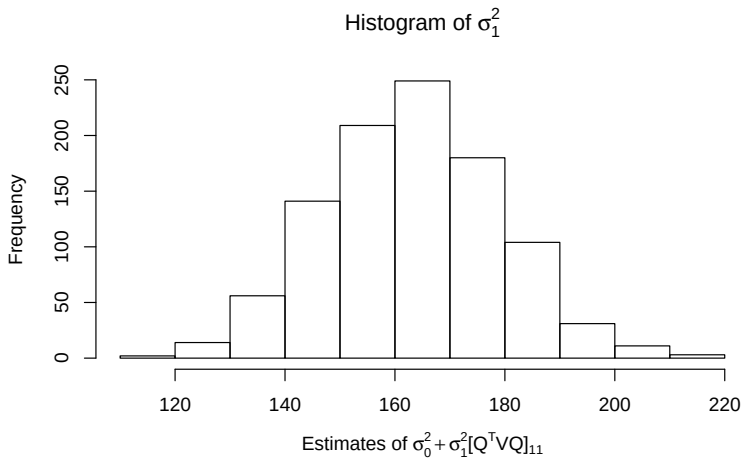

Figure 4.10: This figure show a histogram of the estimated values of $\sigma_0^2 + \sigma_1^2[Q^\top VQ]_{11}$ based on simulated data with the true parameter-values $\sigma_0^2 = 10$ and $\sigma_1^2 = 20$.

We see the histogram over the estimated values of $\sigma_0^2 + [Q^\top VQ]_{11}$ centres around the true value. In fact about 96 % of the simulated values are within 20 % of the true value. This confirms that, assuming that a WWN Process is an appropriate model for a given dataset, the covariance-structure is accurately estimated, even if the noise-term cannot be computed.

Chapter 5

Model Validation

In this chapter we evaluate how well the WWN Process fits the BCI dataset.

5.1 Computing the Variogram

In this section we seek to compute the empirical and the theoretical variogram for a WWN Process with given parameter-values. This is done with the intention of using the variogram as a summary statistic. First we define the variogram.

Definition 5.1 (Variogram). *Let Y be a set-indexed random field. Then the function $\gamma : \mathbb{B}_d^2 \to \mathbb{R}$ given by*

$$\gamma(A_i, A_j) = \mathbb{V}\mathrm{ar}\left[Y(A_i) - Y(A_j)\right]$$

for any $A_i, A_j \in \mathbb{B}_d$ is said to be the variogram of Y.

Now let $A_1, \ldots, A_n$ be the observed sets in the dataset and let $Y = \left(Y(A_1), \ldots, Y(A_n)\right)^\top$. At first glance it would not appear to be possible to compute an empirical variogram for the BCI dataset, since for each pair of sets A_i, A_j only one observation of $Y(A_i) - Y(A_j)$ is available, and thus a sample variance cannot be computed. However as with random fields over $\mathbb{R}^d$ we may assume that the variogram depends only on the distance between sets. What exactly is meant by distance in the case of sets becomes clear later on. First we compute the theoretical variogram. Suppose that $Z = Q^\top Y$, where Q is a projection-matrix, has covariance function

$$\mathbb{C}\mathrm{ov}\left[Z(\mu_i), Z(\mu_j)\right] = \int\int \sigma_0^2 \delta_{x-y} - \sigma_1^2 \log\|x - y\| \, \mathrm{d}\mu_i(x) \mathrm{d}\mu_j(y).$$

This makes it possible to compute the variogram of Y by letting μ_{ij} be the measure with $1_{A_i} - 1_{A_j}$ as density with respect to the Lebesgue measure, since $Y(A_i) - Y(A_j) = Z(\mu_{ij})$ and thus

$$\gamma(A_i, A_j) = \mathbb{V}\mathrm{ar}\left[Y(A_i) - Y(A_j)\right] = \mathbb{V}\mathrm{ar}\left[Z(\mu_{ij})\right] = \mathbb{C}\mathrm{ov}\left[Z(\mu_{ij}), Z(\mu_{ij})\right].$$

Now we can compute an expression for the variogram using the covariance function of Z. Since μ_{ij} has density $1_{A_i}(x) - 1_{A_j}(x)$ with respect to the Lebesgue measure, we get

$$\begin{aligned}
&\mathbb{C}\text{ov}\left[Z(\mu_{ij}), Z(\mu_{ij})\right] = \int\int \sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\,\mathrm{d}\mu_{ij}(x)\mathrm{d}\mu_{ij}(y) \\
=&\int\int \left(\sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\right)\left(1_{A_i}(x) - 1_{A_j}(x)\right)\left(1_{A_i}(y) - 1_{A_j}(y)\right)\mathrm{d}x\mathrm{d}y \\
=&\int_{A_i}\int_{A_i} \sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\,\mathrm{d}x\mathrm{d}y - \int_{A_i}\int_{A_j} \sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\,\mathrm{d}x\mathrm{d}y \\
+&\int_{A_j}\int_{A_j} \sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\,\mathrm{d}x\mathrm{d}y - \int_{A_j}\int_{A_i} \sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\,\mathrm{d}x\mathrm{d}y \\
=&2|A|\sigma_0^2 - 2\sigma_1^2\text{Ave}_{A_i\times A_j}\left(-\log\|x-y\|\right) + 2\sigma_1^2\text{Ave}_{A_i\times A_i}\left(-\log\|x-y\|\right). \qquad (5.1)
\end{aligned}$$

Approximating the average log-distances as in Section 4.2 yields

$$\gamma(A_i, A_j) \approx 2|A|\sigma_0^2 + 2\sigma_1^2\left(\log\|z_i - z_j\| + V_{ii}\right),$$

where z_i and z_j are the centre-points of A_i and A_j respectively. This gives rise to the distance-function for sets, $d(A_i, A_j) = \|z_i - z_j\|$ (this is not a metric, since we may have $d(B, C) = 0$ for $B \neq C$, for example if $C \subset B$, but B and C has the same centre). Now we have that the variogram, at least approximately, only depends on the distance between sets, since V_{ii} is a constant that does not depend on i. If $h = d(A_i, A_j)$ we have, with an abuse of notation, that

$$\gamma(A_i, A_j) \approx \gamma(h) = 2|A|\sigma_0^2 + 2\sigma_1^2\left(\log h + V_{ii}\right).$$

Non-parametric Estimation of the Variogram

As mentioned previously, the purpose for introducing the variogram in this project, is to use it as a summary statistic. That is, we seek to asses the goodness of fit of a fitted model, by comparing the theoretical variogram (computed above) based on the fitted values, to a non-parametric empirical estimate of the variogram. As seen above we may assume that the variogram depends only on the distance between sets. As such we can use a modified versions of the non-parametric variogram estimations used for random fields on $\mathbb{R}^d$. The most straightforward way to estimate the variogram non-parametrically is simply by estimating the variances at each distance of interest. This gives the estimator

$$\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{(i,j)\in N(h)} (Y(A_i) - Y(A_j))^2,$$

where $N(h) = \{(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,n\} : d(A_i, A_j) = h\}$. This is the sample variogram for random fields [Kim, 2015, p. 7] modified to be used for set-indexed random fields. The sample variogram of a stationary random field on $\mathbb{R}^d$ is unbiased. For the sample variogram of a set-indexed random field, the unbiasedness is only approximate, since

$$\mathbb{E}\left[\hat{\gamma}(h)\right] = \frac{1}{|N(h)|}\sum_{(i,j)\in N(h)} \mathbb{E}\left[(Y(A_i) - Y(A_j))^2\right] \approx \frac{1}{|N(h)|}\sum_{(i,j)\in N(h)} \gamma(h) = \gamma(h). \qquad (5.2)$$

The disadvantage of this estimator is that it is highly inaccurate, unless there is a large number of pairs of observations at the exact same distance from one another. This means that it is only usable for datasets where the sample points are located on a grid. For situations where this is

not the case, *binning* is used. Binning means that we seek to estimate the variogram for the distances $h_1, \ldots, h_m$, where $h_{i+1} - h_i = \delta$ for $i = 1, \ldots, m-1$, and use all pairs of observations where the distance is between $h_i - \delta/2$ and $h_i + \delta/2$ to compute $\hat{\gamma}(h_i)$ [Kim, 2015, p. 17]. This gives the estimator

$$\hat{\gamma}(h_k) = \frac{1}{|N_\delta(h_k)|} \sum_{(i,j) \in N_\delta(h_k)} \left(Y(A_i) - Y(A_j)\right)^2,$$

where $N_\delta(h_k) = \{(i,j) \in \{1, \ldots, n\} \times \{1, \ldots, n\} : h_k - \delta/2 < d(A_i, A_j) \leq h_k + \delta/2\}$. This estimator is not unbiased, but it does include all data. In many geostatistical datasets binning is a necessity, since there is at most one pair of observations at any given distance, making estimation without binning impossible. For the BCI dataset either estimator can be used, but each has disadvantages. The obvious disadvantage of using binning is that not all pairs used in the estimation of the variogram at a given distance, are actually spaced apart by that distance, and as mentioned the estimator is biased. Estimation without binning is approximately unbiased, as shown in (5.2), but since one third of the data is not on a regularly spaced grid, it must be left out of the estimation.

When using the estimate without binning, and comparing to the theoretical variogram with the parameters estimated in Section 4.3 for mineralised nitrate, Figure 5.1 is obtained.

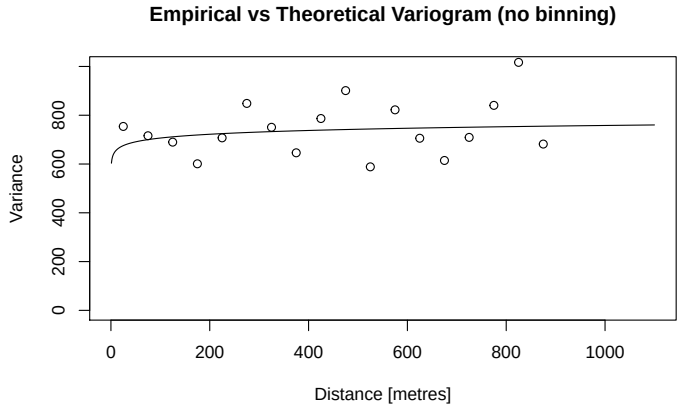

Figure 5.1: The solid line is the theoretical variogram given a WWN process with parameters $\hat{\sigma}_0^2 = 288.6$ and $\hat{\sigma}_1^2 = 11.19$. The dots are values of the variogram estimated from the data non-parametrically without binning.

The sample variogram fluctuates around the theoretical variogram, which is to be expected from a good fit. Unfortunately the sample variogram does not exhibit the same initial rapid increase that the theoretical variogram does, but this mostly seems to be happening at distances smaller than what we have available in the data anyway.

When using binning with $h_1 = 25$ and $\delta = 50$, Figure 5.2 is obtained.

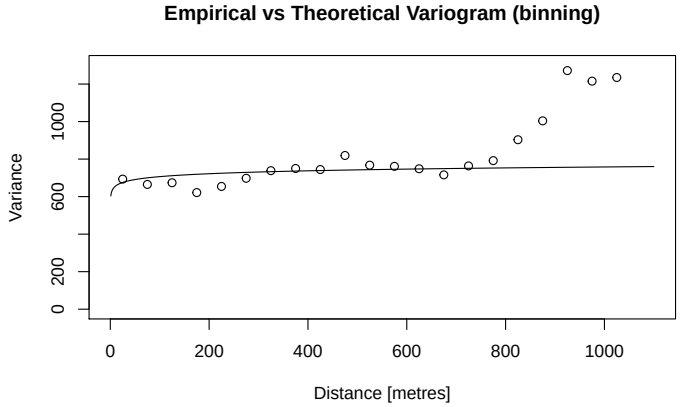

Figure 5.2: The solid line is the theoretical variogram given a WWN process with parameters $\hat{\sigma}_0^2 = 288.6$ and $\hat{\sigma}_1^2 = 11.19$. The dots are values of the variogram estimated from the data non-parametrically with binning.

We see that the sample variogram follows the theoretical variogram nicely up to a distance of 800 metres, after which it starts increasing. It should be noted that at these distances the number of observations available to compute the sample variogram starts to decrease.

We cannot determine visually if the behaviour for $h > 800$ on Figure 5.2 is due to random fluctuations, too few observations or a real discrepancy between the model and the data. A helpful tool in this regard would be a confidence interval, but since we do not know the distribution of the fitted variogram, an expression of the confidence interval cannot be obtained. Instead we use simulations. This is done by performing N simulations of WWN Process with the parameter values fitted in Sections 4.3 as the true values. Then the empirical variogram is computed for each simulation.

Suppose we have N variograms, $\gamma_1(\cdot,\cdot), \ldots, \gamma_N(\cdot,\cdot)$, each computed from a distinct simulation and let $\gamma_0(\cdot,\cdot)$ be the empirical variogram of the data. Under the null-hypothesis (that the fitted values are true) we have

$$P\Big(\gamma_0(A_i, A_j) \geq \max\{\gamma_1(A_i, A_j), \ldots, \gamma_N(A_i, A_j)\}\Big) = \frac{1}{N+1}.$$

This is because all the variograms are identically distributed, so all of them are equally likely to be the largest. Similarly

$$P\Big(\gamma_0(A_i, A_j) \leq \min\{\gamma_1(A_i, A_j), \ldots, \gamma_N(A_i, A_j)\}\Big) = \frac{1}{N+1}.$$

Setting $N = 39$ thus yields a points-wise 95 % envelope.

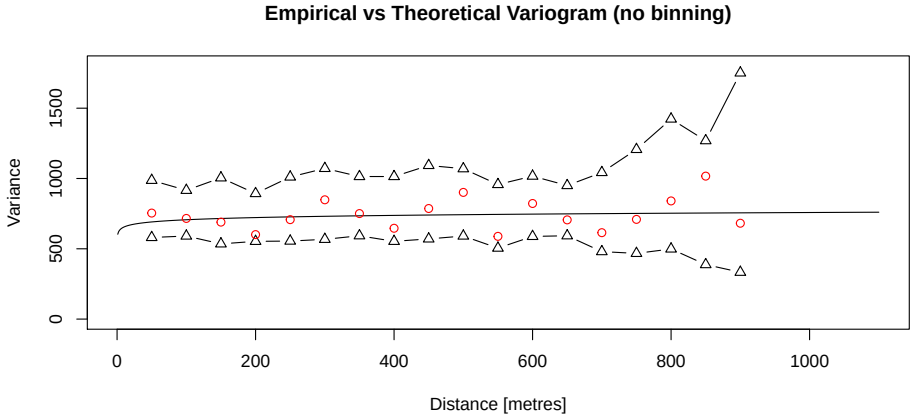

Figure 5.3: The solid line is the theoretical variogram given a WWN process with parameters $\hat{\sigma}_0^2 = 288.6$ and $\hat{\sigma}_1^2 = 11.19$. The red dots are values of the variogram estimated from the data non-parametrically without binning, and the black triangles are the lower and upper edges of a 95 % envelope.

On Figure 5.3 we see that all the empirically estimated values for the variogram are inside the envelope. Since it is only estimated at 18 points, this is reasonable, even though we would expect that one point would be outside the envelope. Under the null hypothesis there is a probability of $0.95^{18} \approx 0.4$ that all points are within the envelope, so it is fairly probable.

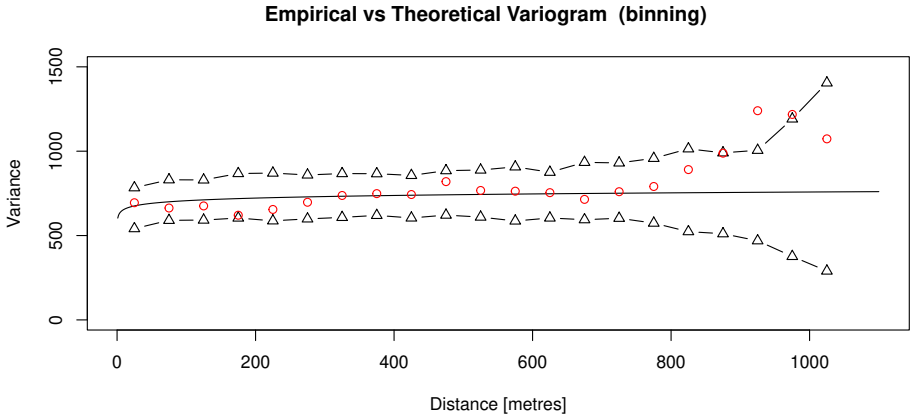

Figure 5.4: The solid line is the theoretical variogram given a WWN process with parameters $\hat{\sigma}_0^2 = 288.6$ and $\hat{\sigma}_1^2 = 11.19$. The red dots are values of the variogram estimated from the data non-parametrically with binning, and the black triangles are the lower and upper edges of a 95 % envelope.

On Figure 5.4 we see that two points at the end are outside the envelope, but one of them is only just barely outside.

In summary the empirical variogram fits fairly well to the theoretical variogram. Under the null hypothesis, one point should be outside the envelope. Without binning all points are inside the envelope and with binning, two points are outside the envelope. This behaviour does not seem inconsistent with the model, and as such it cannot be rejected on this basis.

5.2 Kriging on Set-indexed Random Fields

In practical applications, predicting a value not present in the dataset is often useful. In this section we seek to perform kriging when working with set-indexed random fields. The covariance structure is only defined for contrast-measures, but predicting the value of a contrast is rarely useful in practice. Instead we seek to predict values of the underlying set-indexed random field. Now suppose we have have $A_1, \dots, A_n \in \mathbb{B}_d$ where $|A_1| = \cdots = |A_n|$ and we denote this volume $|A|$. Then for some other $A_0 \in \mathbb{B}_d$ we wish to predict $Y(A_0)$ based on $Y = (Y(A_1), \dots, Y(A_n))^\top$. We do this using *kriging*, which means we seek to use the Best Linear Unbiased Predictor (BLUP). That is we seek a predictor $\hat{Y}(A_0) = \lambda^\top Y$, where $\lambda \in \mathbb{R}^n$, such that $\mathbb{E}\left[\hat{Y}(A_0) - Y(A_0)\right] = 0$ and the variance is minimised.

First we focus on minimising the variance and, as we shall see, the unbiasedness gets taken care of in the process. First we note that we cannot choose the kriging weight in an unconstrained manner. This is because the covariance structure of the underlying set-indexed random field of a WWN Process is not well-defined, and thus a closed form of $\mathbb{E}\left[\left(\hat{Y}(A_0) - Y(A_0)\right)^2\right]$ does not, in general, exist. What we do instead is to demand certain restrictions on the kriging weights, such that the covariance-matrix is well-defined. Specifically we demand that the kriging weights are chosen such that the measure μ satisfying $\mu(B) = \sum_{i=1}^n \lambda_i |B \cap A_i| - |B \cap A_0|$ is in $\mathcal{W}$. Then $\hat{Y}(A_0) - Y(A_0) = Z(\mu)$, which has known covariance-structure. Due to the way μ is constructed $\mu \in \mathcal{W}$ holds by Proposition 3.10 if $\mu(\mathbb{R}^d) = 0$. Since $\mu(\mathbb{R}^d) = |A| \sum_{i=1}^n \lambda_i - |A_0|$, this holds if $\sum_{i=1}^n \lambda_i = |A_0|/|A|$. Obviously if A_0 is chosen to have the same volume as the sets in the dataset, then $\sum_{i=1}^n \lambda_i = 1$ is required. If the set-indexed random field is sufficiently well-behaved, this also takes care of unbiasedness. If $|A_0| = |A|$ and $\mathbb{E}\left[Y(B)\right] = m$ for all $B \in \mathbb{B}_d$ and some $m \in \mathbb{R}$, then

$$\mathbb{E}\left[\sum_{i=1}^n \lambda_i Y(A_i) - Y(A_0)\right] = m \sum_{i=1}^n \lambda_i - m = m - m = 0,$$

and the estimator is unbiased. If $|A_0| \neq |A|$, then

$$\mathbb{E}\left[\sum_{i=1}^n \lambda_i Y(A_i) - Y(A_0)\right] = m \sum_{i=1}^n \lambda_i - m = m|A_0|/|A| - m \neq 0.$$

This may be solved by slightly altering the assumptions on the mean structure. Instead of assuming that the mean is constant, we may assume that $\mathbb{E}\left[Y(B)\right] = |B|m$, for some $m \in \mathbb{R}$. If all the sets in the dataset has the same size it changes very little, except that unbiasedness is ensured if $\sum_{i=1}^n \lambda_i = |A_0|/|A|$. This assumption is also quite sensible for many datasets encountered in practice, such as crop yield data. It makes sense that an area twice as large is expected to yield twice as much of a given crop. Unfortunately this assumption does not make sense for the BCI dataset, since the mineral concentrations are measured in mg per kg of oven dried soil, so there is no reason to assume that larger areas have higher concentrations. For this reason we now restrict ourselves to the case where $|A_0| = |A|$.

In order to minimise the variance of $\lambda^\top Y - Y(A_0)$ we note that, while we do not have a covariance function for the set-indexed random field, the function

$$B(A_i, A_j) = \int_{A_i} \int_{A_j} \delta_{x-y} - \log\|x - y\| \, \mathrm{d}x\mathrm{d}y$$

comes close in the sense that a matrix, Φ, where $\Phi_{ij} = B(A_i, A_j)$, satisfies $\mathbb{V}\mathrm{ar}\left[Q^\top Y\right] = Q^\top \Phi Q$. As we saw in Section 3.3 we get that $\Phi = \sigma_0 |A| I + \sigma_1^2 |A|^2 V$ when $Y = (Y(A_1), \dots, Y(A_n))^\top$. If

we use $\begin{bmatrix}-1 & \lambda^\top\end{bmatrix}$ in place of Q and $\tilde{Y} = (Y(A_0), Y(A_1), \ldots, Y(A_n))^\top$ we get

$$\mathbb{V}\mathrm{ar}\left[\begin{bmatrix}-1 & \lambda^\top\end{bmatrix}\tilde{Y}\right] = \begin{bmatrix}-1 & \lambda^\top\end{bmatrix}\begin{bmatrix}B(A_0,A_0) & \phi^\top \\ \phi & \Phi\end{bmatrix}\begin{bmatrix}-1 \\ \lambda\end{bmatrix} = B(A_0,A_0) - 2\lambda^\top\phi + \lambda^\top\Phi\lambda,$$

where $\phi = \big(B(A_0,A_1),\ldots,B(A_0,A_n)\big)^\top$. By definition we have

$$B(A_0,A_i) = \int_{A_0}\int_{A_i} \sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\,\mathrm{d}x\mathrm{d}y,$$

and from Section 3.3 we know that

$$\int_{A_0}\int_{A_i} \delta_{x-y}\mathrm{d}x\mathrm{d}y = |A_0\cap A_i| = 0,$$

and

$$\int_{A_0}\int_{A_i} -\log\|x-y\|\,\mathrm{d}x\mathrm{d}y = |A|^2\mathrm{Ave}_{A_0\times A_i}(-\log\|x-y\|) \approx -|A|^2\log\|z_0-z_i\|\,,$$

where z_0 is the centre of A_0. Thus we get

$$B(A_0,A_i) \approx -\sigma_1^2|A|^2\log\|z_0-z_i\|\,. \tag{5.3}$$

We now use Lagrange multipliers to minimise the variance of $\lambda^\top Y - Y(A_0)$ under the constraint that $\lambda^\top 1_n = 1$. We must minimise

$$L(\lambda,\alpha) = B(A_0,A_0) - 2\lambda^\top\phi + \lambda^\top\Phi\lambda + 2\alpha(1_n^\top\lambda - 1).$$

We get the derivatives

$$\frac{\partial L}{\partial\lambda} = -2\phi + 2\Phi\lambda + 2\alpha 1_n$$

and

$$\frac{\partial L}{\partial\alpha} = 2(1_n^\top\lambda - 1).$$

Thus we get that the kriging weight may be obtained by solving the linear equation system

$$\begin{bmatrix}\Phi & 1_n \\ 1_n^\top & 0\end{bmatrix}\begin{bmatrix}\lambda \\ \alpha\end{bmatrix} = \begin{bmatrix}\phi \\ 1\end{bmatrix}. \tag{5.4}$$

Then we get that the kriging variance is

$$\begin{aligned}\mathbb{E}\left[\Big(Z(\lambda^\top)Y - Y(A_0)\Big)^2\right] &= B(A_0,A_0) - 2\lambda^\top\phi + \lambda^\top\Phi\lambda \\ &= B(A_0,A_0) - \lambda^\top\phi + \lambda^\top(\Phi\lambda - \phi) \\ &= B(A_0,A_0) - \lambda^\top\phi - \alpha\lambda^\top 1_n \\ &= B(A_0,A_0) - \lambda^\top\phi - \alpha,\end{aligned}$$

where the last two equalities comes from (5.4). In order to compute this we need an expression for $B(A_0,A_0)$. We first note that

$$B(A_0,A_0) = \int_{A_0}\int_{A_0} \sigma_0^2\delta_{x-y} - \sigma_1^2\log\|x-y\|\,\mathrm{d}x\mathrm{d}y = \sigma_0^2|A| + \sigma_1^2|A|^2\mathrm{Ave}_{A_0\times A_0}(-\log\|x-y\|).$$

In Section 4.2 we discussed how to compute the average value of $-\log\|x-y\|$ over a set on the form $A\times A$, but only when that set consists of three smaller sets all one metre from a

centre-point. We have no such assumptions about the shape of A_0, only that it has the same volume as $A_1, \dots, A_n$. However when performing kriging, one is likely more interested in the value over a single square. In that case we simply define A_0 as a square with side-length $\sqrt{|A|}$ and we get

$$\text{Ave}_{A_0 \times A_0}(-\log\|x-y\|) = \frac{25 - 4(\pi + \log 2)}{12} - \frac{1}{2}\log|A|.$$

Finally we also note that due to (5.3) we use the approximation

$$\phi \approx -\sigma_1^2 |A|^2 v,$$

where $v = (\log\|z_1 - z_0\|, \dots, \log\|z_n - z_0\|)^\top$.

From (5.4) we see that kriging for set-indexed random fields is analogous to intrinsic kriging for an intrinsic random field of order 0. The only difference is that for intrinsic kriging, the generalised covarince function would be used instead of $B(\cdot,\cdot)$ (for more information see [Jensen and Fitzhugh, 2018, Section 2.1]).

The code for the above theory is shown below

```
Setkriging <- function(pts,new.pts,data,A,sigma0,sigma1){
  pts <- as.matrix(pts)
  new.pts <- as.matrix(new.pts)
  data <- as.numeric(data)
  n <- nrow(pts)

  V <- Vmatrix(pts)
  V <- V+t(V)
  diag(V) <- 2/3*(-log(2*sin(pi/3)*sqrt(A/0.00147)))
             +1/3*((25-4*(pi+log(2)))/12-0.5*log(A)+0.5*log(3))

  big.mat <- matrix(1,n+1,n+1)
  big.mat[1:n,1:n] <- sigma0*A*eye(n)+sigma1*A^2*V
  big.mat[n+1,n+1] <- 0

  v <- apply(pts, 1, function(x) log(Norm(x-new.pts)))
  phi <- -sigma1*A^2*v
  big.vec <- c(v,1)

  lambdalpha <- solve(big.mat, big.vec)
  lambda <-  lambdalpha[-(n+1)]
  alpha <-  lambdalpha[n+1]

  y <- as.numeric(lambda %*% data)
  krig.variance <- sigma0*A + sigma1*A^2*( (25-4*(pi+log(2)))/12
                   - 0.5*log(A) ) - lambda %*% phi + alpha

  out <- list(pred=y, var=krig.variance)
  return(out)
}
```

Code 5.1: Here a function for kriging using realisations of a set-indexed random field is shown.

Note that the function `Vmatrix` used in R-code 5.1 is simply a function which computes the

above-diagonal elements of the matrix V given a set of locations. The argument `A` is the desired size of the sample areas, so the points are scaled accordingly.

Cross-validation

We now perform a 10-fold cross-validation on the BCI Dataset. In order to get an idea of how good the results of the cross-validation are, we need some naive prediction method, to use as a benchmark. Here we predict using the sample mean, that is, if $\hat{\mu}$ is the sample mean of the data, we predict that $\hat{Y}(A_0) = \hat{\mu}$ for all $A_0 \in \mathbb{B}_d$. If the WWN Process does not provide better predictions than this, it is not a useful model for prediction purposes.

Mineral	RMSE	Benchmark	$\hat{\sigma}_0$ (full data)	$\hat{\sigma}_1$ (full data)
Al	260	336	0	15128
B	0.648	0.760	0	0.080
Ca	833	1022	0	139395
Cu	2.15	2.88	0.055	1.07
Fe	48.8	65.2	0	530
K	89.9	110	0	1593
Mg	151	190	0	4559
Mn	132	186	0	4059
Na	38.4	72.5	0	448
P	2.01	2.45	0	0.840
S	6.98	8.04	5.72	9.55
Zn	3.93	5.10	0	3.23
NH4	7.00	11.8	0	11.8
NO3	4.87	6.21	9.96	2.78
total N	9.84	12.5	9.12	17.44
min.NH4	10.2	15.8	20.2	18.7
min.NO3	19.0	19.2	289	11.2
min N	23.0	25.3	304	40.1
pH	0.290	0.427	0	0.019

Table 5.1: Results of 10-fold cross-validation for all minerals. Benchmark refers to the RMSE of predictions using the sample mean.

Note that on Table 5.1, the values for $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ in the last two columns are computed based on the full data, but for every prediction in the cross-validations, the parameters estimated based only on the training data are used.

We see that the WWN process performs better than the benchmark for all minerals, but for mineralised nitrate, which we analysed in previous the sections, it only performs marginally better than the benchmark. However, this is to be expected since $\hat{\sigma}_0^2$ is considerably larger than $\hat{\sigma}_1^2$ (at least for the full data). To see this note that if $\sigma_0^2 > 0$ and $\sigma_1^2 = 0$ then the only solution to (5.4) is $\lambda_1 = \cdots = \lambda_n = 1/n$ which would make kriging identical to using the empirical mean, which is the benchmark. For the other minerals the WWN process yields RMSEs that range between 10 % and 45 % under the RMSE of the benchmark. Finally, we also see that there is no relation between the performance of the WWN process, and whether or not $\hat{\sigma}_0^2 = 0$ for the full data.

5.3 Comparison to Other Models

In this section we compare the WWN Process to models with other generalised covariances. Specifically we compare it to the power model, which has a generalised covariance on the form

$$K_{\text{Pow}}(\|x-y\|) = \sigma_0^2\delta_{x-y} - \sigma_1^2\|x-y\|^{2\nu},$$

and the exponential model, which has a generalised covariance on the form

$$K_{\exp}(\|x-y\|) = \sigma_0^2\delta_{x-y} + \sigma_1^2\exp\left(-\nu\|x-y\|\right).$$

The covariance can be computed using the same approach used for the WWN process in Section 3.3. Thus, for the power model, we get

$$\begin{aligned}\mathbb{C}\text{ov}\left[Z(\mu_i), Z(\mu_j)\right] &= \int\int \sigma_0^2\delta_{x-y} - \sigma_1^2\|x-y\|^{2\nu}\,\mathrm{d}\mu_i(x)\mathrm{d}\mu_j(y)\\ &= |A|1[i=j]\sigma_0^2 + |A|^2\sigma_1^2\text{Ave}_{A_i\times A_j}\left(-\|x-y\|^{2\nu}\right),\end{aligned}$$

so if $Z = \left(Z(\mu_1),\dots,Z(\mu_{n-1})\right)^\top$ follows a power model, then

$$\mathbb{V}\text{ar}\left[Z\right] = \sigma_0^2|A|I + \sigma_1^2|A|^2Q^\top V(\nu)Q, \tag{5.5}$$

where $V_{ij}(\nu) = \text{Ave}_{A_i\times A_j}\left(-\|x-y\|^{2\nu}\right)$. The form of the covariance matrix for data following an exponential model would be identical to (5.5) except here $V_{ij}(\nu) = \text{Ave}_{A_i\times A_j}\left(\exp\left(-\nu\|x-y\|\right)\right)$. In Section 4.2 the diagonal of V for the WWN Process was computed using a result in Clifford [2005], which gives an explicit closed form of the integral $\int_{A_i}\int_{A_i} -\log\|x-y\|\,\mathrm{d}x\mathrm{d}y$ when A_i is a square. A similar result for the power model exists, but it is extremely long and unwieldy, so instead we simply use Monte Carlo approximation. For the exponential model no explicit expression is obtained, and thus, once again, we use Monte Carlo approximation to compute the diagonal of V.

Before proceeding further, we show that the matrix in (5.5) is positive definite. We first note that the argument used in Section 3.3 to show that the covariance matrix obtained from the WWN Process can also be used here, if

$$\int\int -\|x-y\|^{2\nu}\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) > 0, \tag{5.6}$$

for all $\mu\in\mathcal{M}$ for the power model and

$$\int\int \exp\left(-\nu\|x-y\|\right)\mathrm{d}\mu(x)\mathrm{d}\mu(y) > 0, \tag{5.7}$$

for all $\mu\in\mathcal{M}$ for the exponential model. We start by showing (5.6). By Proposition 3.8 it is sufficient to show that $f(r) = -r^\nu$ satisfies that $(-1)^n f^{(n)}(r)\geq 0$ for all $n\leq 1$. Note that

$$f^{(n)}(r) = -\nu(\nu-1)\cdots(\nu-(n-1))r^{\nu-n},$$

and thus if we restrict the model to the case where $\nu\in(0,1)$, we get

$$\begin{aligned}(-1)^n f^{(n)}(r) &= (-1)^{n+1}\nu(\nu-1)\cdots(\nu-(n-1))r^{\nu-n}\\ &= (-1)^2\nu(1-\nu)\cdots(n-1-\nu)r^{\nu-n}\geq 0,\end{aligned}$$

where the inequality follows from the fact that $0 < \nu < 1$. Now we show (5.7). Once again we seek to use Proposition 3.8, so it is sufficient to show that $g(r) = \exp\left(-\nu\sqrt{r}\right)$ satisfies $(-1)^n g^{(n)}(r) \geq 0$ for $n \geq 1$. Note that

$$g^{(n)}(r) = \sum_{i=1}^{n} (-1)^n \frac{a_{i,n}\nu^i \exp(-\nu\sqrt{r})}{2nr^{(2n-1)/2}},$$

where $a_{i,n}$ is a positive integer, which depends on i and n. Thus, if we restrict the model to the case where $\nu > 0$, we get $(-1)^n g^{(n)}(r) \geq 0$ for $r \geq 0$.

The estimate of γ_0 can be computed as for the WWN Process, and we get

$$\hat{\gamma}_0(\gamma_1, \nu) = \frac{Z^\top \left(I + \gamma_1(|A|Q^\top V(\nu)Q - I)\right)^{-1} Z}{|A|(n-p)}.$$

Then we can write the profile log-likelihood of γ_1 and ν as

$$\ell(\hat{\gamma}_0(\gamma_1, \nu), \gamma_1, \nu) = -\frac{n-p}{2} \log\left(Z^\top \left(I + \gamma_1(|A|Q^\top V(\nu)Q - I)\right)^{-1} Z\right) - \frac{1}{2} \log\left|I + \gamma_1(|A|Q^\top V(\nu)Q - I)\right|.$$

This is then maximised numerically over (γ_1, ν). In practice this is simply done by using `optim` rather than `optimise` in R. Note that for both the exponential and the power model, we use the unscaled locations, so $|A| = 0.00147$. This is because otherwise the power- and exponential models may obtain very large or very small values values respectively, even for modest values of ν. For instance if $\nu = 0.7$, then for i, j such that $d(A_i, A_j) = 100$, the power model yields $\mathbb{C}\text{ov}\left[Z(\mu_i), Z(\mu_j)\right] = V_{ij}(0.7) = -(100/\sqrt{0.00147})^{1.4} \approx -60,657$ for the scaled locations, but only $V_{ij}(0.7) = -100^{1.4} \approx -631$ for the unscaled locations. It is even more necessary to use unscaled data for the exponential model, where we get $V_{ij}(0.1) = \exp(-0.1 \cdot 100) = 5.34 \cdot 10^{-5}$ for the unscaled data, but $V_{ij}(0.1) = \exp(-0.1 \cdot 100/\sqrt{0.00147}) = 5.34 \cdot 10^{-114}$ for the scaled data. This means that the log-likelihood is less susceptible to small changes of ν, and thus the numerical estimates become more stable.

Before applying the exponential- or power model, we need to address a practical issue. While, *in theory*, the matrix $Q^\top V(\nu)Q$ is positive definite for all $\nu > 0$ for the exponential and $\nu \in (0, 1)$ for the power model, it is not always the case in practice. This is because $V(\nu)$ has to approximated, so when ν is close the to boundary of the area it is defined on (close to 0 for the exponential model and close to 1 for the power model) the matrix $Q^\top V(\nu)Q$ is not positive definite. This happens for $\nu > 0.91$ for the power model and $\nu < 1.65$ for the exponential model. We cannot entirely solve this problem, but we can substantially mitigate it, by only applying these models on the grid. Here $Q^\top V(\nu)Q$ becomes non-positive-definite for $\nu > 0.998$ for the power model and $\nu < 0.0175$ for the exponential model.

Applying The Power Model

In order to compute the profile log-likelihood of (ν, γ_1) for the power model, the following R-code is used.

```r
reml <- function(param,z,A){
  q <- length(z)
  V <- Vmatrix(nu=exp(param[1]),cov = "power",A=A)
  cov.mat <- eye(q)+param[2]*(A*t(Q) %*% V %*% Q-eye(q))
  d <- as.numeric(determinant(cov.mat)$modulus)
  cov.mat.inv <- solve(cov.mat)
  l <- -q/2*log(t(z) %*% cov.mat.inv %*% z)-0.5*d
  return(-l)
}
```

Code 5.2: R-code used to compute the profile-likelihood function for (ν, γ_1). The matrix Q is a projection-matrix that must be specified before running this function.

R-code 5.2 returns the negative profile log-likelihood, since `optim` minimises rather than maximises. The function `Vmatrix` returns the matrix $V(\nu)$ and the argument `cov` accepts the inputs `"power"`, `"exp"` and `"log"` (if `cov="log"`, then `nu=NULL` is required as an input). The reason that the argument `nu` takes `exp(param[1])` rather than simply `param[1]` is to ensure that $\hat{\nu}$ is positive. That is, rather than optimise ν over $(0, 0.998]$, we optimise $\log \nu$ over $(-\infty, \log 0.998]$.

When performing the estimation for the mineralised nitrate dataset we obtain the estimates

$$\hat{\sigma}_0^2 = 236,285, \quad \hat{\sigma}_1^2 = 7,215,022, \quad \hat{\nu} = 0.114.$$

As an initial indication of the fit, we check simulations of a generalised random field with a power covariance matrix, with the estimated parameters.

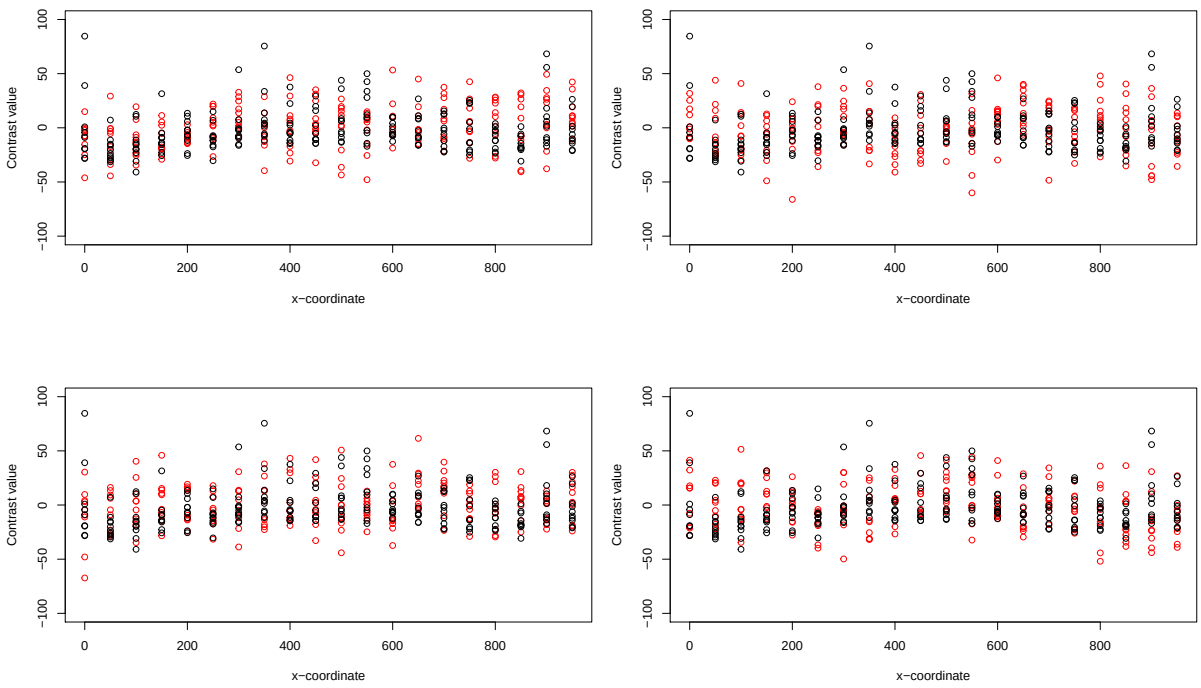

Figure 5.5: Black dots are the contrasts from the data, red dots are contrasts simulated from a generalised random field following the power model with the fitted values $\hat{\sigma}_0^2 = 236,285$, $\hat{\sigma}_1^2 = 7,215,022$ and $\hat{\nu} = 0.114$.

We see that the simulations and the data behave reasonably similarly. Any difference between them seem consistent with random variation. Next we use the variogram as a summary statistic, and we note that, using the same arguments as in Section 5.1, we obtain

$$\gamma(A_i, A_j) \approx \gamma(h) = 2|A|\sigma_0^2 + 2\sigma_1^2(h^{2\nu} + V_{ii}(\nu)),$$

where $d(A_i, A_j) = h$. Below we examine the variogram for the mineralised nitrate data using the power model.

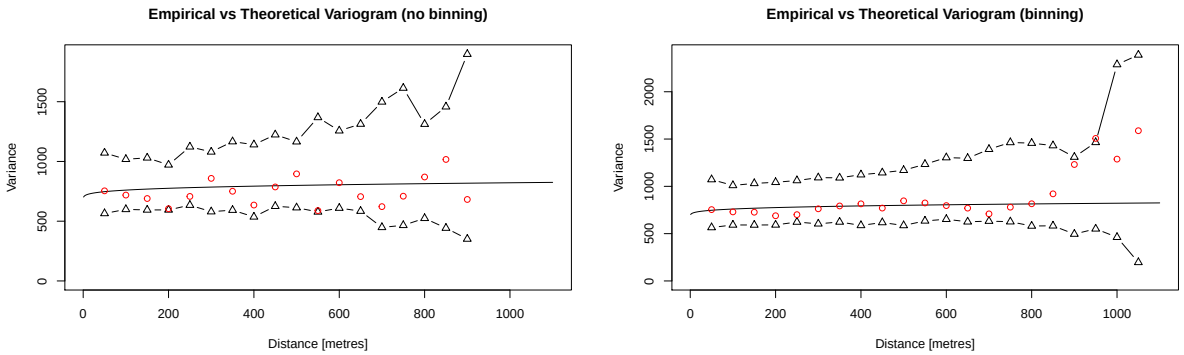

Figure 5.6: On the left we see the empirical variogram without binning (red dots) vs the theoretical variogram (solid line) for the power model, with an upper and lower simulated point-wise 95% envelope (triangles). On the right we see the same, except the empirical variogram is computed with binning.

On Figure 5.6 on the left we see that without binning, all points are inside the envelope, but two of them are just barely above the lower edge of it. When using binning, there is one point outside the envelope. This is consistent with the behaviour expected if the model is true.

Now we perform a 10-fold cross-validation using the power model.

Mineral	RMSE	Benchmark	$\hat{\sigma}_0^2$ (full grid)	$\hat{\sigma}_1^2$ (full grid)	$\hat{\nu}$ (full grid)
Al	277	331	14154517	11630728482	0.132
B	0.656	0.760	187	8967	0.225
Ca	909	1025	316944749	54758210743	0.140
Cu	2.25	2.83	0	904719	0.148
Fe	51.9	59.8	89960	735732409	0.093
K	96.9	114	4104121	317335842	0.187
Mg	159	184	16349713	210671	0.829
Mn	141	183	0	3758205970	0.145
Na	37.1	70.9	0	7229105	0.622
P	2.13	2.35	1778	290426	0.149
S	8.02	8.05	42402	302	0.669
Zn	4.08	5.13	10638	81.0	0.873
NH4	7.46	11.7	11610	2519861	0.258
NO3	5.11	6.13	10596	445270	0.266
total N	10.5	12.7	36046	4516762	0.213
min.NH4	9.76	15.0	52342	24847	0.672
min.NO3	19.8	19.9	236286	7214990	0.114
min N	24.1	26.0	355613	26955	0.664
pH	0.327	0.423	26.3	11674	0.143

Table 5.2: Results for a 10-fold cross validation for all minerals using the power model. The column 'benchmark' refers to the RMSE when predicting using the sample mean.

As with Table 5.1, we note that the parameter-estimates shown in Table 5.2 are based on the whole grid, but the predictions in the cross-validation uses parameters estimated based only on the training data.

From Table 5.2 we see that, with the exceptions of sulphur (S) and mineralised nitrate (min.NO3), where the power model and benchmark perform identically, the power model always outperform the benchmark. This is not better than what we saw for the WWN process, but it should be noted that results on Table 5.2 and those on Table 5.1 are not directly comparable, since one was performed on the full dataset and one was only performed on the grid.

Applying The Exponential Model

Now we perform the same analysis using the exponential model, as we just did using the power model. First we apply it to the mineralised nitrate dataset, where we obtain the estimates

$$\hat{\sigma}_0^2 = 245,532, \quad \hat{\sigma}_1^2 = 24,288,605, \quad \hat{\nu} = 0.0175.$$

We recall that due to numerical issues, we had to require $\nu \geq 0.0175$, so the fact that $\hat{\nu} = 0.0175$ is somewhat problematic. It suggests that a better exponential model may be obtainable, but it is inaccessible with the approximations we have performed.

Once again we start by simulating data with the above parameters with the exponential model, and compare them to the data.

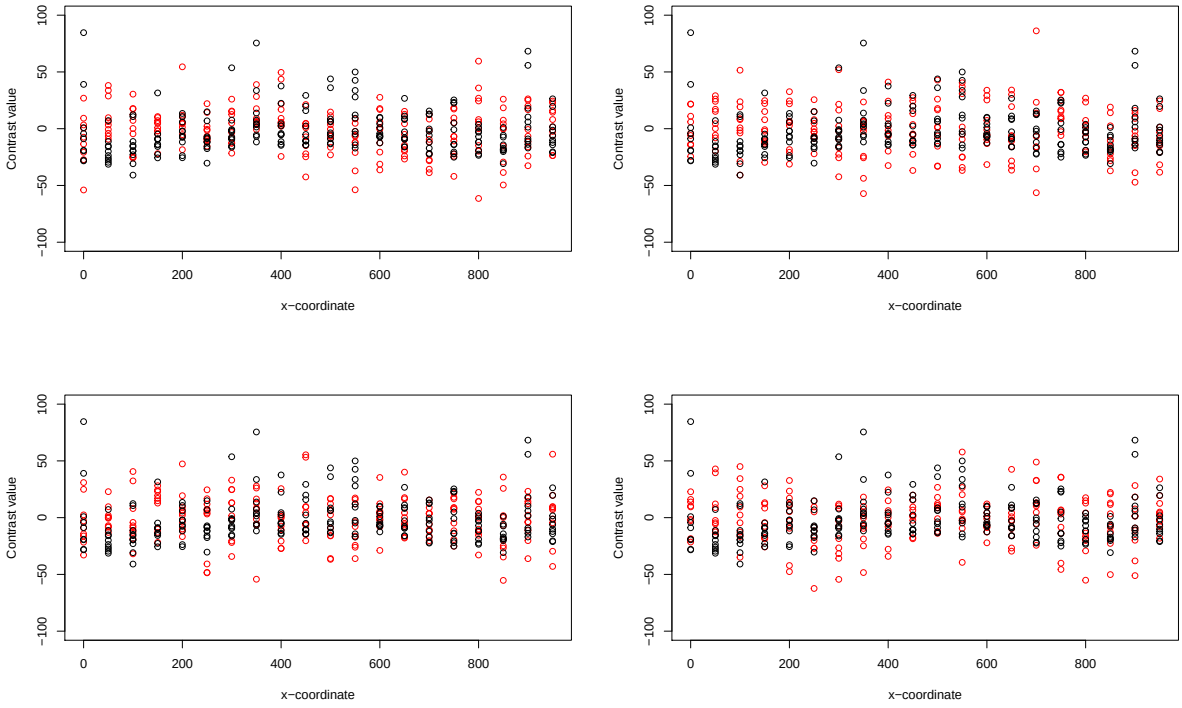

Figure 5.7: Black dots are the contrasts from the data, red dots are contrasts simulated from a generalised random field following an exponential model with the fitted values $\hat{\sigma}_0^2 = 245,532$, $\hat{\sigma}_1^2 = 24,288,605$ and $\hat{\nu} = 0.0175$.

Overall the simulations look fairly reasonable, although they do tend to have more negative outliers than the data. Next we use the variogram as a summary statistic, and note that, analogously to the computations for the WWN Process in Section 5.1, we get

$$\gamma(A_i, A_j) \approx \gamma(h) = 2|A|\sigma_0^2 + 2\sigma_1^2(V_{ii}(\nu) - \mathrm{e}^{-\nu h}),$$

where $d(A_i, A_j) = h$. Now we check how well the exponential model fits the mineralised nitrate data by using the variogram.

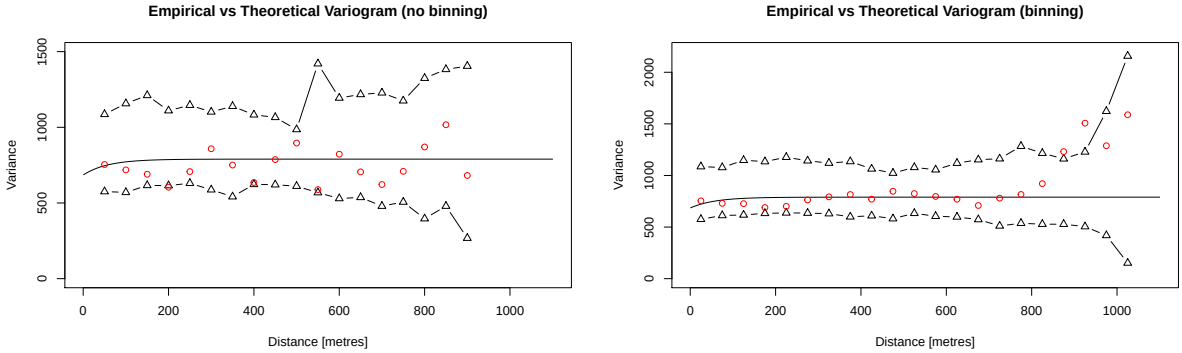

Figure 5.8: On the left we see the empirical variogram without binning (red dots) vs the theoretical variogram (solid line) for the exponential model, with an upper and lower simulated point-wise 95% envelope (triangles). On the right we see the same, except the empirical variogram is computed with binning.

When using binning, only one point is outside the envelope, as expected. When using binning two points are outside the envelope rather than the expected one point. This does not go against the model strongly enough to reject it.

We now perform a 10-fold cross-validation of the exponential model on all minerals.

Mineral	RMSE	Benchmark	$\hat{\sigma}_0^2$ (full grid)	$\hat{\sigma}_1^2$ (full grid)	$\hat{\nu}$ (full grid)
Al	339	331	36768006	36587310350	0.018
B	0.658	0.760	192	179101	0.018
Ca	964	1025	369958933	340263857007	0.018
Cu	2.23	2.83	2541	2275644	0.018
Fe	50.0	59.8	1270753	941880406	0.018
K	96.7	114	4481644	3768725875	0.018
Mg	187	184	14032896	8116555976	0.018
Mn	161	183	9856203	9860475081	0.018
Na	32.8	70.9	0	3004879366	0.019
P	2.14	2.35	2375	1339147	0.018
S	8.06	8.05	41109	3070009	0.018
Zn	4.25	5.13	8971	6425664	0.018
NH4	7.80	11.7	31148	31306097	0.018
NO3	5.06	6.13	12403	9959397	0.018
total N	10.5	12.7	53909	43283070	0.018
min.NH4	10.3	15.0	53691	49066775	0.018
min.NO3	19.9	19.9	245536	24284150	0.018
min N	24.4	26.0	324493	122431174	0.018
pH	0.337	0.423	52.6	47528	0.018

Table 5.3: Results for a 10-fold cross validation for all minerals using the exponential model. The column 'benchmark' refers to the RMSE when predicting using the sample mean.

Once again, we note that the parameter-estimates shown in Table 5.3 are based on the whole grid, but the predictions in the cross-validation uses parameter-estimates based only on the training data.

Comparing Table 5.3 to Table 5.2 we see that the exponential and power model perform very similarly on most minerals. That said, for the exponential model we get $\hat{\nu} = 0.0175$ for all but one mineral (rounded up to 0.018 on Table 5.3). Once again this suggests that the exponential model may have yielded superior results if our approximations of $\int_{A_i}\int_{A_j} \exp\left(-\nu\|x-y\|\right) \mathrm{d}x\mathrm{d}y$ had allowed for smaller ν.

Finally we compare the exponential model, the power model and the WWN process directly. To do so cross-validations on the grid is performed with the WWN process.

Mineral	RMSE (WWN)	RMSE (power)	RMSE (exp)	Benchmark
Al	280	277	339	331
B	0.703	0.656	0.658	0.760
Ca	990	909	964	1025
Cu	2.37	2.25	2.23	2.83
Fe	49.8	51.9	50.0	59.8
K	104	96.9	96.7	114
Mg	182	159	187	184
Mn	146	141	161	183
Na	47.9	37.1	32.8	70.9
P	2.32	2.13	2.14	2.35
S	8.08	8.02	8.06	8.05
Zn	4.91	4.08	4.25	5.13
NH4	7.55	7.46	7.81	11.7
NO3	5.00	5.11	5.06	6.13
total N	10.2	10.5	10.5	12.7
min.NH4	9.70	9.76	10.3	15.0
min.NO3	20.0	19.8	19.9	19.9
min N	25.9	24.1	24.4	26.0
pH	0.324	0.327	0.337	0.423

Table 5.4: Results for a 10-fold cross validation for all minerals using the WWN process, the power mode, and the exponential model on the grid. The column 'benchmark' refers to the RMSE when predicting using the sample mean.

On Table 5.4 we see that the power model has the lowest RMSE for most minerals. However it should be noted that the WWN process does not have the parameter ν and thus fitting it is substantially faster. Moreover it is much easier to check that $Q^\top VQ$ is positive definite, since V does not depend on any parameters, thus giving WWN greater ease of use. With these advantages in mind, the WWN process is probably best, unless the performance advantage of the power or exponential model is substantial. For example for sodium (Na), the exponential model is preferred (RMSE of 32.8 compared to 47.9 for the WWN process), but for aluminium (Al) the WWN process is preferred even though the power model performs marginally better (RMSE 277 for the power model versus 280 for the WWN process).

Chapter 6

Discussion and Conclusion

In this chapter, the results from the previous chapters are discussed and summarised.

6.1 Estimations and Numerical approximations

In Section 4.2 it is shown how the matrix V is specified in this project. Specifically we have that $V_{ij} = \text{Ave}_{A_i \times A_j}(-\log\|x-y\|)$, which we do not know exactly, so approximations are necessary. When $i \neq j$ this is done by noting $\text{Ave}_{A_i \times A_j}(-\log\|x-y\|) \approx -\log\|z_i - z_j\|$, where z_i and z_j are the centres of A_i and A_j respectively. However, we know that the set A_i actually consists of three areas, each at most one metre from z_i, centred at z_i^1, z_i^2 and z_i^3, so it should be more accurate to use the approximation

$$\text{Ave}_{A_i \times A_j}(-\log\|x-y\|) \approx \sum_{k=1}^{3}\sum_{l=1}^{3} -\log\left\|z_i^k - z_j^l\right\|.$$

In practice it makes no difference which approximation is used when $i \neq j$. To see this let $z_i = (25, 25)$ and $z_j = (75, 25)$, and let z_i^1, z_i^2, z_i^3 and z_j^1, z_j^2, z_j^3 be distributed around z_i and z_j as in Figure 4.2. Then we get that

$$\left|\left(\sum_{k=1}^{3}\sum_{l=1}^{3} -\log\left\|z_i^k - z_i^l\right\|\right) - (-\log\|z_i - z_j\|)\right| < 10^{-9},$$

so clearly this has no impact on the results obtained by the WWN Process.

The approximations made to compute V_{ii} is a bigger concern, since the diagonal entries are more affected by the locations of z_i^1, z_i^2, z_i^3. This is a problem, since these locations are not reported in the dataset, so the necessary information is simply not available. We assumed that they are located as shown in Figure 4.2, but this is merely an assumption. Another approach could be to randomly sample the points in a unit disc centred around the point reported in the dataset. This might make the diagonal of V more realistic, since it would be non-constant, like for the real locations (it is highly unlikely that samples are taken a the same locations relative to their centres for all data-points). However this does not solve the underlying problem, which is that the real locations of soil samples are unknown, so it is unlikely that it would lead to better results.

In Section 5.3 we compared the WWN process with the exponential model and the power model. A problem was encountered in that for some parameter-values the approximations performed to compute the matrix $V(\nu)$ results in $Q^\top V(\nu) Q$ not being positive definite. For

the power model this is largely mitigated by only using the grid, instead of the whole data, but for the exponential model this still presents a problem. Specifically it was found that for $\nu < 0.0175$, $Q^\top V(\nu)Q$ is not positive definite, but for all but one mineral we get $\hat{\nu} = 0.0175$. This suggests that a lower value for ν may have yielded better results, but then a different approximation would have been necessary. All in all this raises questions of the generality for the approximations performed in this report as it pertains to the exponential and power model. If $Q^\top V(\nu)Q$ is not positive definite for all ν when applied to the whole BCI dataset, what about other datasets? Clearly if the power and exponential models are to be used on general datasets, other approximations to compute $V(\nu)$ must be used.

In Section 4.3 we found that when performing REML estimation on the aluminium dataset, we get $\hat{\sigma}_0^2 = 0$. We then found in Section 4.4 that this is likely due to the fact that the term $\sigma_1^2|A|^2 Q^\top VQ$ dominates on the diagonal, to the point where the effect of $\sigma_0^2|A|I$ becomes undetectable. We then saw that $\hat{\sigma}_0^2 > 0$ is obtained for the mineralised nitrate data. Specifically $\hat{\sigma}_0^2 = 288.6$, and since $\hat{\sigma}_1^2 = 11.19$ this is consistent with the conclusion from Section 4.4; that the term $\sigma_0^2|A|I$ is only detectable, when σ_0^2 is substantially larger than σ_1^2. However it may also be an example of what we see on Figure 4.7, where σ_0^2 is sometimes estimated to be far larger than σ_1^2 even if the true value of σ_1^2 is actually slightly larger than the true value of σ_0^2. Whether or not this is the case is impossible to determine.

6.2 Comparison to Traditional Geostatistics

It is not entirely clear if the BCI dataset is even suitable for the set-indexed random field setup. The idea behind using sets rather than points to index the data, is that in practice geostatistical data is not sampled on an infinitesimally sized area. For some datasets, such as crop yields, this is crucial, but in the BCI dataset, the sampled areas are so small and so far apart, that the abstraction of considering the soil samples to be from infinitely small areas is not an unreasonable one. To investigate whether it is even beneficial to use the set-index setup, we compare the cross-validations of kriging with the WWN Process to intrinsic kriging for an IRF-0 with polynomial generalised covariance function (see Jensen and Fitzhugh [2018, Section 2.1-2.2]), that is, the generalised covariance function is on the form $K(\|x-y\|) = -b\|x-y\|$ (note that the implementation we have available does not include a nugget effect). This is almost a special case of the power model, where $\sigma_0^2 = 0$ and $\sigma_1^2 = b/|A|^2$. The only difference between them for kriging purposes is the diagonal of Φ. An IRF-0 is indexed on points in $\mathbb{R}^d$ so here we can simply take $\Phi_{ii} = -b\|x_i - x_i\| = 0$, but since the WWN process is indexed on measures, the diagonal of Φ is an approximation of $\int_{A_i}\int_{A_i} -\log\|x-y\|\,\mathrm{d}x\mathrm{d}y$ (see Section 5.3).

Mineral	RMSE (WWN)	RMSE (IRF-0)	Benchmark
Al	260	290	336
B	0.648	0.697	0.760
Ca	833	869	1022
Cu	2.15	2.38	2.88
Fe	48.8	52.4	65.2
K	89.9	94.2	110
Mg	151	166	190
Mn	132	138	186
Na	38.4	29.4	72.5
P	2.01	2.11	2.45
S	6.98	7.02	8.04
Zn	3.93	4.76	5.10
NH4	7.00	7.53	11.8
NO3	4.87	5.46	6.21
total N	9.84	10.3	12.5
min.NH4	10.2	11.8	15.8
min.NO3	19.0	22.5	19.2
min N	23.0	26.7	25.3
pH	0.290	0.294	0.427

Table 6.1: Results of 10-fold cross-validation for all minerals for both the WWN Process and for an IRF-0 with polynomial generalised covariance function. Benchmark refers to prediction using the sample mean.

We see on Table 6.1 that the WWN process consistently outperforms the IRF-0 with the exception of sodium (Na), where the RMSE of the WWN process is about 30 % larger than that of the IRF-0. It should also be noted that the IRF-0 achieves a better performance relative to the benchmark on sodium, than any other model considered in this report does on any mineral. We also note that, while the WWN process technically outperforms the IRF-0 on all other minerals, they perform essentially identically on Sulphur (S) and pH (the RMSEs of the WWN process are respectively 0.5 % and 1.3 % lower than that of the IRF-0). There are even two minerals (mineralised nitrate and mineralised nitrogen), where the IRF-0 performs worse than the benchmark.

6.3 Summary of Results for Predictions

When analysing the BCI dataset the WWN Process, the kriging predictions are worse than those for an IRF-0 for sodium (Na), the two performed essentially identically on sulphur (S) and pH, and on the remaining 16 minerals the WWN performed better than the IRF-0. The best predictions obtained from the WWN process was on sodium, where the RMSE of the kriging predictions using the WWN process was 46 % lower than the RMSE of predictions using the sample mean. When predicting only on the grid the WWN process achieves better predictions than the exponential and power model on five minerals (iron (Fe), nitrate (NO3), total nitrogen, mineralised ammonium (min.NH4) and pH), and there are two minerals where the WWN process is outperformed considerably by either the power or exponential model (calcium (Ca) and sodium (Na)). Then there are four minerals (magnesium (Mg), phosphorous (P),

zinc (Zn) and mineralised nitrogen (Min N)) where the WWN process only barely outperforms the sample mean, but the power model outperforms the sample mean substantially. On two of the minerals (sulphur (S) and mineralised nitrate (min.NO3)), none of the model outperformed the sample mean. On the six remaining minerals (aluminium (Al), boron (B), copper (Cu), potassium (K), manganese (Mn) and ammonium (NH4)) the power model performs better than the WWN process, but not by much and the WWN process may still be preferred due to its simplicity.

6.4 Is the WWN Process a Loi du Terroir?

Suppose we have a generalised random field with a generalised covariance function on the form

$$K(\|x-y\|) = \sigma_0^2\delta_{x-y} + \sigma_1^2(\lambda\|x-y\|)^\nu \mathcal{K}_\nu(\lambda\|x-y\|),$$

where $\mathcal{K}_\nu$ is a Bessel function. This is called the *Matérn class* of generalised covariance functions, and it turns out that for $\nu = 0.5$ it reduces to the exponential model, when $\nu > 0$ and $\lambda \to 0$ it reduces to the power model, and when $\nu \to 0$ and $\lambda \to 0$ it reduces to the WWN process [Clifford and McCullagh, 2006, p. 2120-2121]. This means that the power model, the exponential model and the WWN process are all from the same model class. David Clifford and Peter McCullagh noticed that on crop yield data, the WWN process is as good as the full Matérn model regardless of the climate and the crop [Clifford and McCullagh, 2006, p. 2142]. Due to this, they deemed the WWN process to be a *loi du terroir* (law of the soil) [McCullagh, 2003]. In this project we can conclude that this notion of a loi du terroir does not extend to forest soil data. To see this observe Table 5.4. Here we see that for sodium (Na), the RMSE of the kriging predictions obtained by the exponential model is 32 % lower than the RMSE obtained by the WWN process on the grid.

Bibliography

Herbert Amann and Joachim Escher. *Analysis II*. Birkhauser, 2008.

J. Besag and D. Mondal. First-order intrinsic autoregressions and the de wijs proce. *Biometrika*, 92(4):909–920, 2005.

P. Billingsley. *Probability and Measures*. Wiley-Interscience, 1995. ISBN 0-471-00710-2.

D. Clifford, A. B. McBratney, J. Taylor, and B. M. Whelan. Generalized analysis of spatial variation in yield monitor data. *Journal of Agricultural Science*, 144(1):45–51, 2006.

David Clifford. Computation of spatial covariance matrices. *Journal of Computational and Graphical Statistics*, 14(1):155–167, 2005.

David Clifford and Peter McCullagh. Evidence for conformal invariance of crop yields. *Proceedings of the Royal society*, 462(2071):2119–2143, 2006.

D. A. Dawson. Generalized Stochastic Integrals and Equations. *Transactions of the American Mathematical Society*, 147:473–506, 1970.

Kyle Harms, James W. Dalling, Joseph B. Yavitt, and Robert Stallard. Forest dynamics plot soil sampling protocols, September 2004.

Ib Thorsgaard Jensen and Nicholas Fitzhugh. Intrinsic random fields of order k - theory and implementation. 2018.

S. Kim. The estimation of the variogram in geostatistical data with outliers, 2015.

L. Mattner. Strict definiteness of integrals via complete monotonicity of derivatives. *Transactions of the American Mathematical Society*, 349(8):3321–3342, 1997. ISSN 00029947.

P. McCullagh. How oats and beans and barley grow, October 2003. URL http://www.stats.ox.ac.uk/~bithell/seminars/mccullagh.abstract.pdf.

D. Mondal. Applying Dynkin's isomorphism: An alternative approach to understand the Markov property of the de Wijs process. *Bernoulli*, 21(3):1289–1303, aug 2015. doi: 10.3150/13-bej541.

S. Rongfeng. Lecture 5. Lecture notes on the convergence of random variables.

S. Schäffler. *Generalized Stochastic Processes*. Birkhauser, 2018. ISBN 978-3-319-78767-1.

Rasmus Waagepetersen. Maximum Likelihood Estimation for Linear Mixed Models. February 2019. URL http://people.math.aau.dk/~rw/Undervisning/MM_BI/Handouts/handout2-2x2.pdf.

Appendix A

Danish Summary

Dette resumé eksisterer for at overholde de formelle krav forskrevet i studieordningens sektion 5.1.

Kapitel 1 omhandler genereliserede funktioner. Først introduceres test-funktioner, som er uendelige-ofte differentiable funktioner med begrænset support. Herefter indføres funktionaler og derefter genereliserede funktioner, som er kontinuerte afbildninger fra mængden af test-funktioner over i de reelle tal.

I Kapitel 2 introduceres stokastiske processer med en meget generel formulering. Her anses de som en samling a stokastiske variable, indekseret over en vilkårlig ikke-tom mængde. Herefter følger nogle målteoretiske resultater, som til sidst munder ud i et bevis for *Kolmogorovs Sætning*, som siger, at for enhver indeksmængde og enhver passende valgt samling af sandsynlighedsmål, eksisterer der en stokastisk proces over indeksmængden med de valgte sandsynlighedsmål som dens endelig-dimensionelle fordelinger. Herefter introduceres genereliserede stokastiske processer, som er stokastiske processer indekseret over mængden af testfunktioner. Hovedresultatet herom omhandler konstruktion af Gaussiske stokastiske processer, og anvender Kolmogorovs Sætning. Desuden vises det også hvordan der kan gives mening til differentialet af en Brownsk bevægelse, som eksempel på en anvendelse af genereliserede stokastiske processer.

I Kapitel 3 introduceres to nye typer stokastiske processer, de mængdeindekserede stokastiske funktioner (*set-indexed random fields* på engelsk) samt de genereliserede stokastiske funktioner (*generalised random fields* på engelsk). Disse to typer er tæt forbundne, da genereliserede stokastiske funktioner som oftest anvendes til at modellere kontraster af mængdeindekserede stokastiske funktioner. Herefter introduceres De Wijs processen, og det bevises at den eksisterer. Herefter introduceres De Wijs Plus Hvid Støj (WWN) processen, og det beskrives hvordan dens parametre kan estimeres ved brug af *restricted maximum likelihood* (REML).

Kapitel 4 omhandler praktisk anvendelse af WWN processen. Først introduceres datasættet, som indeholder koncentrationer af 18 forskellige mineraler samt pH værdi (for simpelhedens skyld refereres de alle til som mineraler, selv når pH er inkluderet) fra jordprøver opsamlet i 300 forskellige lokationer inden for et 50 hektar stort område af en regnskov i Panama. Herefter beskrives nogle numeriske approksimationer, som foretages for at udregne kovariansmatricen af en WWN proces. Herefter fittes en WWN proces data for aluminium og mineraliseret nitrat. Fittet til aluminium har nogle tilsyneladende problematiske egenskaber, som undersøge nærmere i et simulationsstudie.

I Kapitel 5 foretages modelkontrol af WWN processen. Først introduceres variogrammet, som anvendes som summary statistic. Ud fra variogrammet ses det, at fittet til data for min-

eraliseret nitrat er plausibelt. Herefter introduceres kriging af mængdeindekserede stokastiske funktioner, analogt med intrinsic kriging inden for klassisk geostatistik. Herefter foretages en 10-folds krydsvalidering på alle mineraler. Resultaterne herfor sammenlignes med prædiktion ved brug af den empiriske middelværdi (altså prædiktion med modellen, hvor middelværdien er konstant og der ikke er nogen rumlig korrelation), og det konkluderes at WWN processen er bedre end den empiriske middelværdi for alle mineraler. Til sidst introduceres to andre kovariansmodeller (potensmodellen og den eksponentielle model), som begge er inden for samme modelklasse som WWN processen, men med flere parametre. Disse fittes også til data, og variogrammet konstrueres. Disse modeller er lige så plausible som WWN processen for mineraliseret nitrat. Herefter foretages også en 10-folds krydsvalidering på potensmodellen og den eksponentielle model. Det konkluderes at WWN processen er at foretrække i de fleste tilfælde, men der er også tilfælde hvor enten potensmodellen eller den eksponentielle model giver bedre resultater end WWN processen. Peter McCullaghs kaldte WWN processen for en *loi du terroir*, da den skulle være lige så god som den fulde model inden for dens modelklasse på al udbyttedata for afgrøder, uafhængig hvilken plante og hvilket klima der er tale om. Resultaterne i denne rapport antyder, at dette ikke kan genereliseres til skovbundsdata.

I Kapitel 6 diskuteres og opsummeres resultaterne fra de tidligere kapitler. Desuden foretages der her også en sammenligning mellem WWN processen, og metoder indenfor klassisk geostatistik. Her konkluderes det at i 16 ud af 19 tilfælde giver WWN processen bedre prædiktioner end en IRF-0 med polynomiel generaliseret kovariansfunktion. I 2 ud af 19 er prædiktionerne for begge modeller lige gode og i ét enkelt tilfælde giver IRF-0-modellen bedre prædiktioner.