Thesis for the Degree of Doctor of Engineering

On flexible random field models for spatial statistics: Spatial mixture models and deformed SPDE models

Anders Hildeman

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2019

On flexible random field models for spatial statistics
Anders Hildeman

ISBN: 978-91-7905-134-1
Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4601
ISSN 0346-718X

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg, Sweden
Phone: +46 (0)31 772 1000
Author e-mail: `hildeman@chalmers.se`

Cover: Rough seas outside Gothenburg, Sweden, 2018. An example of a sea state characterized by a large significant wave height and a relatively short mean wave period.

Typeset with LaTeX.
Department of Mathematical Sciences
Printed in Göteborg, Sweden 2019

On flexible random field models for spatial statistics: Spatial mixture models and deformed SPDE models

Anders Hildeman

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Spatial random fields are one of the key concepts in statistical analysis of spatial data. The random field explains the spatial dependency and serves the purpose of regularizing interpolation of measured values or to act as an explanatory model.

In this thesis, models for applications in medical imaging, spatial point pattern analysis, and maritime engineering are developed. They are constructed to be flexible yet interpretable. Since spatial data in several dimensions tend to be large, the methods considered for estimation, prediction, and approximation are focused on reducing computational complexity.

The novelty of this work is based on two main ideas. First, the idea of a spatial mixture model, i.e., a stochastic partitioning of the spatial domain using a latent categorically valued random field. This makes it possible to explain discontinuities in otherwise smoothly varying random fields. It also introduces a different perspective—that of a spatial classification problem. This idea is used to model the spatial distribution of tissue types in the human head; an application important in reducing cell damage due to ionizing radiation in medical imaging. The idea is also used to introduce an extension of the popular log-Gaussian Cox process. This extension adds an extra layer of a latent random partitioning of the spatial domain. Using this model, it is possible to classify spatial domains based on observed point patterns.

The second main idea of this thesis is that of spatially deforming a solution to a stochastic partial differential equation. In this way, a random field with a needed degree of non-stationarity and anisotropy can be acquired. A coupled system of two such stochastic partial differential equations is used to model the joint distribution of significant wave heights and wave periods in the north Atlantic. The model is used to assess risks in naval logistics.

Keywords: Spatial statistics, Point processes, Substitute-CT, Gaussian random field, Stochastic partial differential equation, Significant wave height

Acknowledgements

I would like to thank my supervisor David Bolin for introducing me to the interesting field of spatial statistics and teaching me the SPDE-approach, since you are a leading expert in this field. When I studied for my master's degree I could not choose between focusing on statistics or numerical solutions to PDEs, therefore I kind of did both. Due to the SPDE-approach I kind of did both for my PhD as well. I am also grateful for you finding the time when I have questions, and introducing me to interesting research topics.

I would also like to thank my co-supervisor Igor Rychlik for his great ideas and vast knowledge—your insights in ocean wave modeling and its effect on naval vessels have been an invaluable contribution to this thesis. Jonas Wallin for your ideas and quick comprehension of new problems as well as collaboration on Paper I and II. Jun Yu for your hospitality during my visit to Umeå and collaboration on Paper I. Janine Illian for your expert insight in point process theory, the importance of interdisciplinary work, and collaboration on Paper II.

A grateful thanks to Milo Viviani and Efthymios Karatzas for our friendship and fantastic jam sessions with 3lele. Thank you to all of you in the lunch group for making the days at the office much better. During these years I discovered, unexpectedly, how fun it is to teach. I would like to thank the people who have supported me in this, most notably Johan Tykesson, Reimond Emanuelsson, and Johan Jonasson.

Attending conferences has given me an important understanding of contemporary ideas and modern tools in my area of research. These visits would not have been possible without the travel grants I was awarded. Therefore I would like to thank Wilhelm och Martina Lundgrens vetenskapsfond, Stiftelsen GS Magnussons fond, ÅForsk, SVeFUM, and Oscar Ekmans stipendiefond.

I would like to thank all my friends and family for being who you are, caring for me, and making life enjoyable. A particular thanks to Herman Lundgren for proofreading this thesis. Last, but certainly not least, I would like to thank Karin Mellqvist for all the strong support when I needed it.

Anders Hildeman
Gothenburg, April, 2019

List of appended papers

Paper I **A. Hildeman**, D. Bolin, J. Wallin, A. Johansson, T. Nyholm, T. Asklund, and J. Yu.
Whole-brain substitute CT generation using Markov random field mixture models.
Preprint

Paper II **A. Hildeman**, D. Bolin, J. Wallin, J. Illian.
Level set Cox processes.
Spatial Statistics, 28: 169-193, doi:10.1016/j.spasta.2018.03.004

Paper III **A. Hildeman**, D. Bolin, I. Rychlik
Spatial modeling of significant wave height using stochastic partial differential equations.
Preprint

Paper IV **A. Hildeman**, D. Bolin, I. Rychlik
Joint spatial modeling of significant wave height and wave period using the SPDE approach.
Preprint

My contribution to the appended papers:

Paper I: I participated in the development of the model. I conducted the analysis and drafted the manuscript by myself and, after consultation, produced the final manuscript. I co-developed the code together with J. Wallin and D. Bolin.

Paper II: I co-developed the model together with the other authors of the paper. I developed the code and conducted the analysis. I produced the theoretical results of Appendices A and B together with J. Wallin and D. Bolin. I drafted the manuscript and we finalized it together.

Paper III: I co-developed the model together with the other authors and I produced most of the theoretical results. I developed the code and conducted the analysis. I drafted the manuscript and we finalized it together.

Paper IV: I co-developed the model together with the other authors. I produced the theoretical results. I developed the code and conducted the analysis. All authors wrote the paper together.

Publications not included in this thesis:

- O. Eliasdottir, **A. Hildeman**, M. Longfils, O. Nerman, J.Lycke.
  A nationwide survey of the influence of month of birth on the risk of developing multiple sclerosis in Sweden and Iceland.
  Journal of Neurology (2018), 265 (1): 108-114. doi:10.1007/s00415-017-8665-y

- O. Andersen, **A. Hildeman**, M. Longfils, H. Tedeholm, B. Skoog, W. Tian, J. Zhong, S. Ekholm, L. Novakova, B. Runmarker, O. Nerman, S.E. Maier.
  Diffusion tensor imaging in multiple sclerosis at different final outcomes.
  Acta Neurologica Scandinavica (2018), 137 (2): 165-173. doi:10.1111/ane.12797

List of abbreviations

CDF	Cumulative Distribution Function
CSR	Complete Spatial Randomness
CT	Computed Tomography
EM	Expectation Maximization
EMG	EM-Gradient
FEM	Finite Element Method
GMM	Gaussian Mixture Model
GRF	Gaussian Random Field
H_s	Significant wave height
INLA	Integrated Nested Laplace Approximation
LGCP	Log Gaussian Cox Process
LHS	Left Hand Side
LSCP	Level Set Cox Process
MALA	Metropolis Adjusted Langevin Algorithm
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Metropolis Hastings
ML	Maximum Likelihood
MRI	Magnetic Resonance Imaging
NIG	Normal Inverse Gaussian
PC	Penalized Complexity
PDE	Partial Differential Equation
PDF	Probability Distribution Function
PET	Positron Emission Tomography
RHS	Right Hand Side
SDE	Stochastic Differential Equation
SPDE	Stochastic Partial Differential Equation
T_1	Mean wave period
T_p	Peak wave period
T_z	Mean zero-level crossing wave period

Contents

Part I

Introduction

Chapter 1

Introduction

The thesis you are currently holding in your hand (or reading in the soothing light of your screen) is a work made up of four articles in the field of spatial statistics. In order to set the stage for presenting this work you need to know the background and main concepts on which the effort was based. The remainder of this chapter is devoted to a brief introduction to the field of spatial statistics. Chapter 2 introduces the important concept of random fields, Chapter 3 introduces the basics of spatial point processes, and Chapter 4 introduces the basics of stochastic partial differential equations. The main philosophy behind the parameter estimation and statistical inference methods used are explained in Chapter 5. The models presented in this thesis were developed to solve problems arising in several separated fields of study. These fields have their own methods, technology, and nomenclature. Chapter 6 give an overview of the most important problems and concepts associated to the particular applications considered in this thesis. Chapter 7 presents brief summaries of the papers and finally, Chapter 8 discusses possible future extensions to the work of this thesis.

Spatial statistics is a subfield of statistics that arose from problems in the industrial sectors in the early 1800s. The purpose of spatial statistics is to draw conclusions or aid in decision making based on observed spatial data. The word *spatial* here referring to data that can be compared using geometrical concepts such as distance, direction, and/or neighborhood structure. The methodology originated from the fields of forestry, agriculture, and mining (Gelfand et al., 2010).

In agriculture the yield of cereal was being studied. It was recognized that spatial variations in yield could be attributed partly to soil constituents or

other known covariates. The remaining variation usually showed some spatial dependency that needed to be accounted for.

In forestry, the distribution of trees was being studied. Spatially repulsive effects, such as the competition for sunlight and other resources, explained why trees did not grow infinitely dense. At the same time, spatially attractive effects due to pollination paths and seed dispersal explained why trees did not grow far apart from each other. In order to model the distribution of trees, such effects had to be represented by models and inference needed to be drawn based on data.

In mining, engineers needed to predict the prevalence of certain minerals in the ground based on samples. The samples were typically acquired by drilling holes in the ground. Sampling was costly and they needed as much information as possible from the smallest possible sample sizes.

The main philosophy behind the methodology of spatial statistics can be summed up in Tobler's first law of geography, i.e., "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p.236). Therefore, the methods are concerned with quantifying and modeling spatial dependency structures. Spatial data can be sorted into three main categories:

- Data sampled on a continuous spatial domain.

  Between any two points $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, in some continuous space, $\mathcal{D}$, there are an infinite number of other points. The data consist of values at some of these points. The interest of the analyst is how these measurements relate to the values on the entire spatial domain made up of an uncountable number of locations. Examples of such data are surface air temperatures and water salinity.

- Data sampled on a discrete spatial domain.

  The spatial domain only has a countable number of points. The data consist of values at some of these points. Example of such data sets are observed values associated with spatial regions such as countries, digital images (that are made up of a discrete set of pixels), experimental designs with "blocked" regions. For data on a discrete spatial domain there is usually some logic to the discretization that is not directly associated to geometrical distances. The discretization might instead be due to regions of varying natural resources, policies, or risks. Hence, it is often of more interest to measure proximity using the neighborhood structure and number of paths between two locations instead of the typical metrics of geometrical distance.

- Spatial point pattern data.

  For point patterns, the location of events are studied. The spatial domain concerned is most often continuous. The big difference compared to the two other types of data is that the randomness is not in the values at the locations but in which locations were chosen. That is, the data is a countable collection of points; spread out over a (usually) continuous spatial region. Typical examples of point patterns are locations of trees in a forest, locations of robberies in a city, or locations of earthquakes in a geographical region.

Statistical analysis of spatial data is typically needed to answer one or more of the following questions:

- What are the values at unobserved points in space? (Spatial prediction / Kriging / interpolation)
- What are the parameter values of the spatial model explaining the data? (Model estimation)
- Is the assumed model reasonable? (Model validation)

Spatial prediction refers to prediction of values at unobserved points in space given the values at some observed ones, i.e., interpolation/extrapolation. Spatial prediction was of interest to the South African mining engineer Danie Gerhardus Krige who pioneered research in this field. In spatial statistics such conditional prediction problems are hence, as a homage to Krige, referred to as *Kriging*. Often predictions are more than just point values, instead the analyst wants to know the whole conditional distribution given the observed data. From conditional distributions, important point estimates such as the expected value, median, or mode can be acquired. Additionally, estimates of the uncertainty such as the standard deviation or interquartile range can be acquired from the conditional distribution and give important information about the prediction error of the corresponding point estimate.

Model estimation is the act of fitting the parameters of a model to the observed data. This is typically needed in order to draw conclusions about the underlying process that generated the spatial data. For instance, a parametric model representing tree growth in a forest might have a parameter representing the repulsive effect between trees. Estimation of this particular parameter gives information about the extent of the repulsive effect among this particular species of tree.

Model validation examines a model's ability to explain the observed data. Since conclusions are drawn based on data and some model assumptions, it is

important to assess whether these assumptions are reasonable given observed data. Validating a model is an important part of accepting or rejecting a theory in any scientific field. Hence, having methodology to validate a spatial model is of great importance. Moreover, if the model does not explain the data well, the Kriging estimates and model estimation might not give any useful information.

In order to perform meaningful spatial analysis, some model of spatial dependency is assumed, either explicitly or implicitly. The assumed model is often simplistic in order to make model estimation reliable and computationally feasible. However, the true, but usually unknown, mechanism of spatial dependencies might not be so simple. Therefore, some degree of model misspecification will often be present. An interesting phenomena is that a true but complex model can often be less useful than a simplification. This is because a simple model often has analytical expressions of important characteristics, easier interpretation of parameters, a lighter computational footprint, and can be estimated using smaller sample sizes and/or with greater robustness. Due to these issues, statistical modeling is a constant balancing act between what is possible and what is required. Closing this gap is one of the main aims of research in spatial statistics. Particularly, this thesis has focused on adding flexibility while keeping a low computational cost and robust estimates. An effort has also been made on developing models in which parameters are easily interpreted and convey a message; a property that is important in communicating research results, especially in interdisciplinary work.

The archetypal spatial model is the mixed effects generalized linear model. Here, $Y(\boldsymbol{s})$ is an observable random variable associated to the spatial location $\boldsymbol{s}$. The mean of Y is dependent on some covariates $\{B_j(\boldsymbol{s})\}_{j=1}^K$ as well as some spatially varying random effect $X(\boldsymbol{s})$, i.e.,

$$\mathbb{E}\left[Y(\boldsymbol{s})\middle|\{B_j(\boldsymbol{s})\}_{j=1}^K, X(\boldsymbol{s})\right] = g^{-1}\left(\beta_0 + \sum_{j=1}^K \beta_j B_j(\boldsymbol{s}) + X(\boldsymbol{s})\right),$$

The link function, g, adds an extra layer of flexibility since the conditional mean does only need to be a linear model after transformation.

What makes this model stand out compared to a typical generalized linear model is the spatially dependent random effect $X(\boldsymbol{s})$. Typically, X is used to model unknown covariates and/or interactions between values at separate locations. The distribution of $Y(\boldsymbol{s})$ given the conditional expectation models independent randomness between measurements at separate locations, typically measurement noise.

Spatial variations can act on different scales. For instance, looking at crop yield, there might be large scale variations due to regions with different weather and there might be small scale variations due to the distance to some nearby stream. Often, it is not possible to model all scales of variability simultaneously. Therefore, depending on the scope of the problem, the short scale variability might be included in the spatially independent noise of $Y|X$, or the long scale variability in the baseline β_0.

Chapter 2

Random fields

In statistics, conclusions are drawn based on incomplete information using concepts from probability theory. Probability theory concerns processes where the outcome of an action is not determinstic, i.e., the same action can result in different outcomes under exactly the same surrounding conditions. We will call such an action an *experiment* and the outcome of the experiment a *realization*. A real-valued random variable is a mapping between a realization and a real value, i.e., $X : \Omega \to \mathbb{R}$, where X is the random variable, $X(\omega)$ a real value, $\omega \in \Omega$ a realization, and Ω is the set of all possible realizations. A random field is a mapping between a realization and a, possibly infinite, set of random variables, $\mathrm{X}(\boldsymbol{s}, \omega)$, indexed in space. Here $\boldsymbol{s}$ denotes a point in the spatial domain $\mathcal{D}$.

Intuitively, we can think of a realization of a random field as a real-valued function in $\mathcal{D}$. Hence, a random field is a random function with the domain $\mathcal{D}$. An example of two different realizations of the same random field on a bounded and continuous domain in $\mathbb{R}^2$ can be seen in Figure 2.1. Note how the two images show similar qualities even though they are, pointwise, completely different.

A random field can have a discrete spatial domain, or a continuous spatial domain. We will refer to a random field on a spatially discrete domain as a *spatially discrete random field* and the contrary as a *spatially continuous random field*. Likewise, the image of the random variables, $X(\boldsymbol{s})$, (all possible values attainable) at a point $\boldsymbol{s}$ can also be continuous or discrete. We will refer to a random field where $X(\boldsymbol{s})$ can only take on a countable number of values for any fixed $\boldsymbol{s}$ as a *discrete random field*. From here on we omit the dependence on the sample space in the notation, i.e., $\mathrm{X}(\boldsymbol{s}, \omega) = \mathrm{X}(\boldsymbol{s})$.

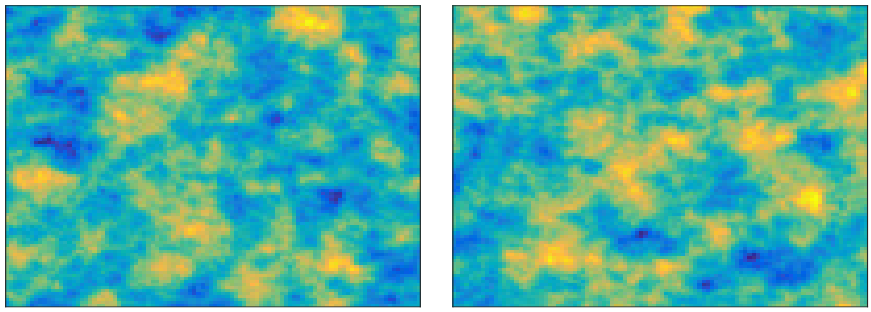


Figure 2.1: Two realizations of the same stationary Gaussian random field on a bounded domain in $\mathbb{R}^2$.

As was mentioned in Chapter 1, spatial statistics concerns analysis of data observed on a spatial domain. For the case of the first two types of data (continuous- and discrete-domain spatial data), the quantity of interest is the values at points in space. In other words, the data can be seen as observations (or partial observations) of realizations of random fields. Also in the third type of data (spatial point patterns) a realization is often dependent on some underlying random fields, see Section 3. Therefore, the concept of random fields is a vital part of spatial statistical methodology.

Two important functions used to characterize random fields are the *mean function* and *covariance function*.

Definition 2.0.1 (Mean function)**.** *The mean function, $\mu(\boldsymbol{s})$, of a random field, $X(\boldsymbol{s})$, is defined as*

$$\mu(\boldsymbol{s}) := \mathbb{E}\left[X(\boldsymbol{s})\right].$$

Definition 2.0.2 (Covariance function)**.** *The covariance function,* $\mathrm{Cov}(\boldsymbol{s}_1, \boldsymbol{s}_2)$*, of a random field, $X(\boldsymbol{s})$, is defined as*

$$\mathrm{Cov}(\boldsymbol{s}_1, \boldsymbol{s}_2) := \mathbb{E}\left[X(\boldsymbol{s}_1)X(\boldsymbol{s}_2)\right] - \mu(\boldsymbol{s}_1)\mu(\boldsymbol{s}_2).$$

The mean function is a *first order characteristic* since it only concerns the behavior of X at one location in $\mathcal{D}$ at a time. The covariance function instead relates the value at two locations with each other and is hence a second order characteristic.

An important concept of random fields is *stationarity*.

Definition 2.0.3 (Strongly stationary random field)**.** *Let X be a random field on the spatial domain $\mathcal{D}$. Furthermore, assume that translations are defined on $\mathcal{D}$, i.e., $\boldsymbol{s}_2 = \boldsymbol{s}_1 + \boldsymbol{t}$.*

The random field, X, is strongly stationary if the vector $[X(\boldsymbol{s}_1), ..., X(\boldsymbol{s}_n)]$ is equal in distribution to the vector $[X(\boldsymbol{s}_1 + \boldsymbol{t}), ..., X(\boldsymbol{s}_n + \boldsymbol{t})]$ for any finite n, any set of locations, and for any translation, $\boldsymbol{t}$, that keep the locations within the spatial domain, $\mathcal{D}$.

In other words, strong stationarity is when the joint distribution between a set of points is only dependent on their relative positions and not on their absolute positions. Random fields that are stationary have important useful properties. However, stationarity is a strong restriction and many real world problems can not be modeled by truly stationary random fields. However, for a small region the most random fields are approximately stationary.

A slightly less restrictive and related property of a random field is that of *weak stationarity*, also known as *second order stationarity.*

Definition 2.0.4 (Weakly stationary random field)**.** *A random field is weakly stationary if*

$$\mathrm{Cov}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{Cov}(\boldsymbol{s}_1 + \boldsymbol{t}, \boldsymbol{s}_2 + \boldsymbol{t}), \text{ and } \mu(\boldsymbol{s}_1) = \mu, \forall\, \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{D}\,.$$

A strongly stationary random field with finite variance is weakly stationary. A weakly stationary field do not, however, need to be strongly stationary.

Just as stationarity concerns translations, *isotropy* concerns rotations.

Definition 2.0.5 (Isotropic random field)**.** *The random field is isotropic if $[X(\boldsymbol{s}_1), ..., X(\boldsymbol{s}_n)]$ is equal in distribution to $[X(\boldsymbol{r}\,\boldsymbol{s}_1), ..., X(\boldsymbol{r}\,\boldsymbol{s}_n)]$ for any rotation, $\boldsymbol{r}$, any finite n, and any set of locations.*

An important property of a random field that is both weakly stationary and isotropic is that it will have a covariance function that only depends on the distance between the two points considered.

Another property of a random field that is of great concern both in Paper I and Papers III and IV is the Markov property. There are three slightly different definitions of the Markov property, the local, global, and pairwise (Rue and Held, 2005). We here only present the global Markov property since it can be defined both for spatially discrete and spatially continuous random fields.

Definition 2.0.6 (Global Markov property)**.** *X is globally Markov if $X(A)$ and $X(B)$ are independent conditioned on $X(C)$ for any two subdomains $A, B \subset \mathcal{D}$ separated by a domain $C \subset \mathcal{D}$. That is, if the values at points in A and the values at points in B are independent conditioned on the values of all points in C.*

This implies that if we want to predict values of X at locations in A and we know the values of X at all locations in C, there is no additional information in knowing the values at locations in B. For many applications this is a natural property arising from the propagation of information in the physical system that is being modeled. However, the Markov property can be computationally beneficial and making a Markov approximation of a non-Markov system can—if done properly—be very attractive. This is a major part of both Paper I and Papers III and IV.

2.1 Spatially continuous random fields

A spatially continuous random field is a random field on $\mathcal{D}$ for which $\mathcal{D}$ is a continuous spatial domain. Typically $\mathcal{D}$ is a Riemannian manifold and in most applications of spatial statistics just some subset of $\mathbb{R}^2$ or $\mathbb{R}^3$.

An important theorem applicable to weakly stationary random fields on $\mathbb{R}^d$ is *Bochners theorem* (Stein, 1999).

Theorem 2.1.1 (Bochners theorem)**.** *A complex-valued function $C(\boldsymbol{s}), \boldsymbol{s} \in \mathbb{R}^d$ is a covariance function for a weakly stationary mean square continuous complex-valued random field if and only if it can be represented as*

$$C(\boldsymbol{s}) = \int e^{i\omega\cdot\boldsymbol{s}} dF(\omega),$$

where F is a positive finite measure.

The theorem states that the covariance function is related to a *spectral measure*, F, through a Fourier transform. Hence, it is possible to model covariance structures using spectral methods. When F is absolutely continuous with respect to the Lebesgue measure, the Radon-Nikodym derivative of F with respect to the Lebesgue measure exists and is known as the *spectral density*. Often the spectral density can have an expression that is easier to work with than the covariance function. It might also be computationally advantageous to generate or analyze data using the spectral density. This is used frequently in ocean wave modeling and is of importance to Paper III and IV, see Section 6.3.

2.1.1 Gaussian random fields

A Gaussian random field (GRF) is a random field such that any finite set of points on the spatial domain has a joint Gaussian distribution. A multivariate Gaussian distribution can be characterized by the mean value and

covariance matrix. Likewise, a GRF can be characterized solely by the mean- and covariance-functions. Often it is easier to work with a centered GRF, i.e., $\mu(\boldsymbol{s}) \equiv 0$. Such a field can easily be attained by subtracting the mean function from the original random field. Since the dependency structure of a GRF is completely determined by the covariance function, a stationary covariance function will lead to a stationary GRF (if the GRF is centered).

2.1.2 Matérn covariance

In applications, the amount of data and computing power is limited. A robust estimate of an arbitrary covariance function cannot be achieved since data is finite while the degrees of freedom of arbitrary covariance functions are infinite. Therefore it is common to assume that the covariance function is of some parametric family with only a small number of parameters. One such popular parametric class of stationary and isotropic covariance functions is the Matérn class (Matérn, 1986; Stein, 1999). This class can be parametrized by the marginal variance σ^2, a smoothness parameter ν, and a correlation dampening parameter, κ. The smoothness parameter, ν, controls the differentiability of the covariance function at the origin. For a Gaussian random field this controls the smoothness of the realizations of the field itself in the sense that the field is almost surely Hölder continuous with ν as the corresponding Hölder constant. Let r be the practical *correlation range* of the random field, i.e., the distance between two points for which their correlation is 0.1. Then the dampening, κ, is proportional to the inverse of r, $\kappa \propto r^{-1}$. Increasing κ makes points a fixed distance apart less correlated while decreasing κ has the opposite effect. A good approximation is that $\sqrt{8\nu}/\kappa$ correspond to the distance between points for which the correlation is 0.13. The marginal variance, σ^2, is the variance of the marginal distribution of $X(\boldsymbol{s})$ for any fixed $\boldsymbol{s} \in \mathcal{D}$.

The Matérn covariance function is very popular in spatial statistics due to its flexibility using only three easily interpretable parameters. Both the exponential and Gaussian covariance functions are special cases of it and Stein (1999) famously proclaimed "Use the Matérn model" due to its ability to model the local smoothness of a Gaussian random fields using the ν parameter. The Matérn covariance function is defined as

$$\mathbb{C}(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^\nu K_\nu(\kappa h),$$

where $h = \| \boldsymbol{s}_2 - \boldsymbol{s}_1 \|$, Γ is the gamma function, and K is the modified Bessel function of the second kind. The spectral density of the Matérn covariance

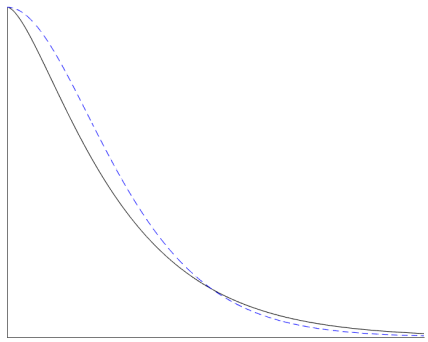


Figure 2.2: A Matérn covariance function as function of distance for $\nu = 1$ (solid line) and $\nu = 2$ (dashed line).

function is

$$\gamma(\omega) = \sigma^2 \frac{\Gamma(\nu + d/2)}{\Gamma(\nu)\pi^{d/2}} \frac{\kappa^{2\,\nu}}{(\kappa^2 + \omega^2)^{\nu + d/2}},$$

where d is the dimensionality of the spatial domain.

In Figure 2.2 a Matérn covariance function is plotted for two different values of ν but with the same dampening and marginal variance. As can be seen, a larger smoothness parameter increases correlation for points close to each other but decreases correlation for points far away.

In Figure 2.3 realizations of three different Matérn Gaussian random fields can be seen. Notice the difference when changing the correlation range as well as when changing the smoothness parameter.

2.1.3 Gaussian white noise

A concept of great importance to this thesis is that of Wiener noise. In its most general definition it can be defined as follows (Adler and Taylor, 2007).

Definition 2.1.2 (Wiener noise)**.** *Let $(\mathcal{D}, \mathcal{A}, \nu)$ be a σ-finite measure space and $A, B \in \mathcal{A}$. Then, a Wiener noise satisfies*

1. $W(A) \sim \mathbb{N}(0, \nu(A))$
2. $A \cap B = \emptyset \Rightarrow W(A \cup B) = W(A) + W(B)$ *a.s.*

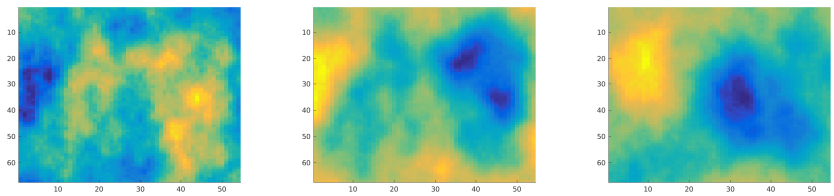


Figure 2.3: Realizations of three distinctively different Matérn fields. Left: $\nu = 1, r = 0.2$. Middle: $\nu = 2, r = 0.2$, Right: $\nu = 2, r = 0.4$.

3. $A \cap B = \emptyset \Rightarrow W(A)$ *and* $W(B)$ *are independent.*

The measure space $(\mathcal{D}, \mathcal{A}, \nu)$ is for most practical considerations a subset of $\mathbb{R}^d$ with associated Borel σ-algebra and Lebesgue measure or the corresponding measure space when mapping this to a Riemannian manifold. As can be noted, the Wiener noise is a random measure on $\mathcal{D}$ with a centered normal distribution for which the variance is equal to the spatial measure of the subset chosen. Also, the covariance between $W(A)$ and $W(B)$ is equal to the spatial measure of the intersection $A \cap B$. Since the Wiener measure of disjoint subsets of $\mathcal{D}$ has zero correlation, the Wiener measure of two small balls arbitrarily close but disjoint in $\mathcal{D}$ must be independent of each other.

An alternative interpretation of the Wiener noise is as the Radon-Nikodym derivative of W with respect to ν. With this interpretation, W is a random field. However, this random field do not have pointwise meaning and should be interpreted in a distributional sense, i.e., as a generalized random field (Stein, 1999).

As a generalized random field, the Wiener noise is defined by how functionals act on it. Considering the typical setting of a L^2-space on $\mathcal{D}$ for which a functional, $f(W)$, is defined through Riesz representation theorem as $\langle f, W \rangle_{L^2(\mathcal{D})}$, the functionals with respect to a Wiener noise have the properties

$$f(W) \sim \mathbb{N}\left(0, \int_{\mathcal{D}} f(\boldsymbol{s}) d\nu(\boldsymbol{s})\right)$$
$$\mathbb{E}\left[f(W)g(W)\right] = \langle f, g \rangle_{L^2(\mathcal{D})}.$$

In short, the Wiener noise is defined by how it acts on square integrable function on $\mathcal{D}$. The usage of the Wiener noise in this thesis will be closely connected to a Gaussian random field with a Matérn covariance function. This connection is revealed in Chapter 4. Any Gaussian random field can be

generated as a convolution between the square root of the covariance function and a Wiener noise. This is used in Paper II to efficiently sample from a Gaussian random field using the method of Lang and Potthoff (2011).

2.2 Spatially discrete random fields

Spatially discrete random fields only have a countable number of spatial locations in $\mathcal{D}$. Such a space occurs either in applications where the space is inherently discrete or where the space has been discretized for some reason. In Paper I we consider a spatial domain that in reality is continuously indexed. However, the data are measurements of activity over regions rather than at points. Hence, the measurements need to be modeled on a discrete spatial domain.

2.2.1 Gibbs random fields

For a spatially discrete random field, the spatial domain can be expressed as an undirected graph, i.e., a set of nodes and edges between neighboring nodes. For an undirected graph, a *clique* is a set of nodes that are all neighbors of each other. A *maximal clique* is a clique which cannot be made any larger without losing the clique property. A specific class of discrete random fields are the Gibbs random fields.

Definition 2.2.1 (Gibbs random field)**.** *Any probability distribution of values at nodes on an undirected graph defined as*

$$\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}\right) \propto e^{-\sum_{c \in \mathcal{C}} E_c(\boldsymbol{x}_c)} \tag{2.1}$$

is a Gibbs random field.

Here, $\mathcal{C}$ is the set of all maximal cliques and E_c is a strictly positive function representing the energy associated with the configuration of the cliques (Murphy, 2012, Chapter 19). The higher the energy in the clique, the less likely it is to occur. The proportionality constant of the Gibbs distribution will be denoted as W and is known as the *partition function*. The partition function is simply the sum of the right hand side of Equation (2.1) over all possible clique configurations. A Gibbs random field might involve a large number of nodes and hence a very large number of possible configurations. This often makes the partition function infeasible to compute.

The Hammersley-Clifford theorem is an important result that relates Gibbs random fields to Markov random fields. We here recite the theorem as stated in (Winkler, 2003).

Theorem 2.2.2 (Hammersley-Clifford)**.** *Let a neighborhood system, $\mathcal{N}$, on $\mathcal{D}$ be given. Then the following holds:*

1. *A random field is a Markov field with respect to $\mathcal{N}$ if and only if it is a Gibbs field for $\mathcal{N}$.*
2. *For a Markov random field, X with neighborhood system $\mathcal{N}$,*

$$\begin{aligned}&\mathbb{P}\left(X(\boldsymbol{s}) = x(\boldsymbol{s}), \boldsymbol{s} \in A | X(\boldsymbol{t}) = x(\boldsymbol{t}), \forall \boldsymbol{t} \in \mathcal{D} \setminus A\right)\\ &= \mathbb{P}\left(X(\boldsymbol{s}) = x(\boldsymbol{s}), \boldsymbol{s} \in A | X(\boldsymbol{t}) = x(\boldsymbol{t}), \forall \boldsymbol{t} \in \mathcal{N}(A)\right),\end{aligned}$$

   for every subset A of $\mathcal{D}$.

A neighborhood system, $\mathcal{N}$, here refers to a collection of sets such that $\boldsymbol{s} \notin \mathcal{N}(\boldsymbol{s})$ and $\boldsymbol{s} \in \mathcal{N}(\boldsymbol{t})$ if and only if $\boldsymbol{t} \in \mathcal{N}(\boldsymbol{s})$ (Winkler, 2003).

The Hammersley-Clifford theorem states that all Gibbs random fields are equivalent to a *Markov random field* (MRF) and vice versa. Through the Hammersley-Clifford theorem a Gibbs field can be defined by conditional probabilities. Since neighborhoods usually involve a smaller number of nodes, the normalizing constant of such conditional distributions is often attainable although the partition function of the corresponding Gibbs distribution is not.

In Paper I a Gibbs random field is used. This random field was defined by the conditional probability on the form

$$\mathbb{P}\left(X_i = k | \boldsymbol{X}_{-i}\right) = \frac{exp\left(-\alpha_k - \beta_k f_{ik}(\boldsymbol{X}_{-i})\right)}{W(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{X}_{-i})},$$

where f_{il} denotes the number of points in the *neighborhood* of node i that have the value l. The value of X_i can be referred to as the class that node i belongs to. The β-parameters control the amount of attraction/repulsion between points of classes. The α-parameters control the marginal probabilities of classes, i.e., they are equivalent but not identical with the, unconditional, probability of X_i belonging to a certain class. The three dimensional neighborhood structure used in Paper I can be seen in Figure 2.4. In this paper, the spatial domain of the discretely indexed random field was on a lattice grid. In the figure, the white ball denotes a point at node i. The black balls correspond to the first order neighborhood of node i, that is the points that have the smallest euclidean distance to $\boldsymbol{s}_i$ on the lattice.

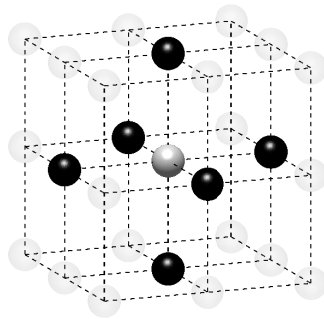
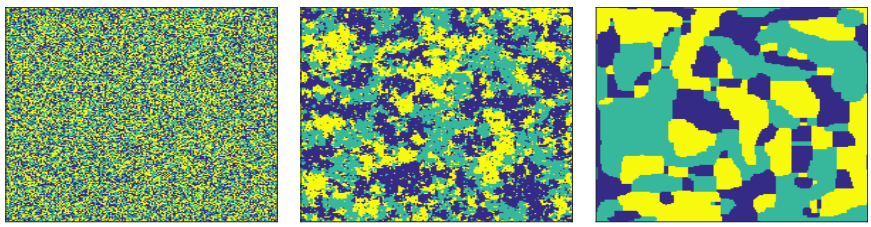


Figure 2.4: A first order neighborhood structure on a regular lattice in three dimensions.



(a) $\beta = 0$ (b) $\beta = -1$ (c) $\beta = -10$

Figure 2.5: Example of realizations of a 3-class Potts field using three different values of the attraction parameter.

Figure 2.5 shows three realizations of such a random field on a two dimensional lattice having three different classes (here illustrated by the colors blue, green, and yellow). The first figure was generated without any spatial interaction, $\beta_k = 0$, the second with an attractive effect, $\beta_k = 1$, and the third with an even stronger attractive effect, $\beta_k = 10$. As can be seen, the β_k parameters control the average size of the class regions.

2.2.2 Gaussian Markov random fields

A Gaussian random field on a discretely indexed domain can be stated as a multivariate Gaussian distribution. For a Gaussian random field on a finite spatial domain this distribution can be characterized by the probability

density function

$$f(\boldsymbol{x}) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

Here, $\boldsymbol{\mu}$ is the mean vector whose elements denotes the mean value for each of the n points in $\mathcal{D}$. The corresponding covariance matrix between each pair of points in $\mathcal{D}$ is denoted by Σ. The inverse of the covariance matrix, i.e., the precision matrix $Q = \Sigma^{-1}$, explains the conditional dependence between points in the Gaussian random field.

The PDF of the Gaussian random field has similarities with that of a Gibbs field, see Equation (2.1). In particular, if Q is non-zero only for pairs of points which are neighbors to each other, the Gaussian random field will be a Gibbs field. By the Hammersley-Clifford theorem it will hence be a Markov random field.

In spatial statistics, the computations that are of main concern when working with a Gaussian distribution is computing the conditional mean, conditional variance, and likelihood. The computational difficulties with these tasks can basically be reduced to evaluating the determinant of Q, matrix multiplications with Q, and solving a linear system with Q. The precision matrices for non-degenerate Gaussian distributions are positive definite and symmetric. Hence, Q can be factorized using the Cholesky decomposition $Q = LL^T$ where L is a lower triangular matrix. Having the precision matrix expressed by L is beneficial since solving a linear system of a triangular matrix has a computational complexity of $\mathcal{O}(n^2)$, as compared to $\mathcal{O}(n^3)$ for general matrices. Also, the determinant equals the square of the product of the diagonals of L. The only problem is that computing the Cholesky triangle, L, generally has a computational complexity of $\mathcal{O}(n^3)$.

Rue and Held (2005) made a strong point when showing that for a Gaussian Markov random field, the computational complexity of the Cholesky factorization is greatly reduced. Considering a spatial domain in two dimensions, the computational cost of the Cholesky factorization is reduced to $\mathcal{O}(n^{3/2})$ which makes a big difference when considering a spatial domain of many points. This property is one of the key benefits of the models proposed in Papers III and IV, see Section 4.3.

2.3 Spatial mixture models

A finite mixture model (Everitt and Hand, 1981) can be defined in two different but equivalent ways. Let us start by defining K classes; each associated

with a random variable X_k with corresponding probability distributions D_k. Assume further a random variable, Z, with probability distribution D_0, on a discrete sample space, $\{1, 2, ..., K\}$. The K different values that Z can assume correspond to the K classes. The random variable Y will be distributed according to a finite mixture model if it is generated by first acquiring a realization z from Z, then assigning Y the value from a realization of X_z. Hence

$$Y = \sum_{k=1}^{K} \mathbb{I}(Z = k) X_k.$$

The finite mixture model can be viewed as a doubly stochastic model since it requires evaluation of random variables in two steps. If a probability density function (or probability mass function) exists, the mixture distribution can equivalently be defined by

$$f_Y(x) = \sum_{k=1}^{K} \pi_k f_k(x),$$

where f_Y is the PDF (or PMF) of Y, $\pi_k = \mathbb{P}(Z = k)$, and f_k is the PDF (or PMF) of X_k.

Typically, the first definition is used when the properties of the latent variable Z is of interest, which is the case for classification problems. The second definition is more common when a complex probability distribution should be approximated by a set of simple ones. For instance explaining a multimodal distribution as a superposition of unimodal ones as in Figure 2.6.

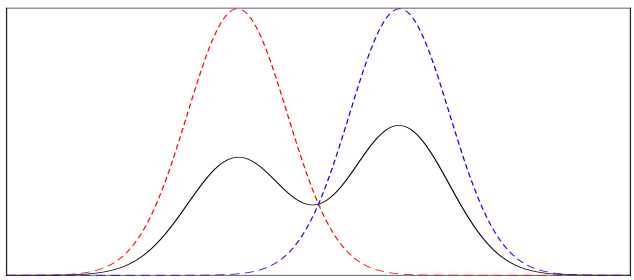


Figure 2.6: An example of a PDF for a finite mixture distribution (black) defined as the superposition of two Gaussian distributions (blue and red). The probability of being a member of the blue class is slightly larger than that of the red class, as can be seen by the right mode being larger.

From here on out, finite mixture models will simply be referred to as mixture models.

In Papers I and II, mixture models were incorporated in spatial models where Z is no longer a random variable but instead a random field, $Z(\boldsymbol{s})$. Likewise, $\{X_k\}_k$ are no longer random variables but random fields as well, $\{X_k(\boldsymbol{s})\}_k$. This is a natural extension of the mixture model definition to a spatial model since the marginal distribution for a fixed point in space is a regular mixture model. Typically, such models can be used to classify regions of a spatial domain or to acquire non-linear prediction functions.

In Paper I, a spatial mixture model was used to model the distribution of voxel values in medical images. A Gibbs model was used to model the latent classification for each voxel. Given this classification, each voxel was assigned a value from the distribution of the corresponding class. Figure 7.2 shows an example of the classification of the spatial region into 4 different classes.

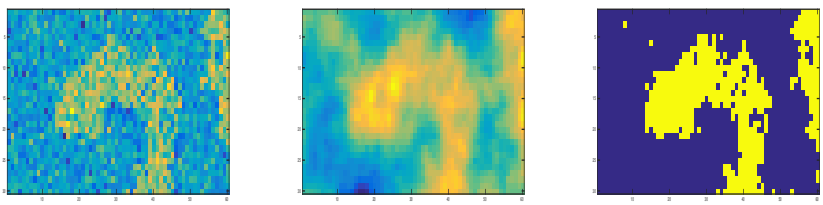
In Paper II, a spatial mixture model was used to model the distribution of the intensity function of a Cox process. The latent classification field, $Z(\boldsymbol{s})$, was acquired from level sets of a Gaussian random field using the level set inversion approach of Iglesias et al. (2016) and Dunlop et al. (2016). Compared to the model of Paper I, this model has the advantage that it defines a classification field in a continuous spatial domain. In a geometric level set inversion problem, *level set functions* define a partition of the spatial domain through level sets of the function. That is, $A_k = \{\boldsymbol{s} : c_{k-1} < \mathrm{X}(\boldsymbol{s}) \leq c_k\}$, where $\{A_k\}_k$ is the partition and $\{c_k\}_k$ are threshold values. The aim of the inversion problem is to estimate the partitioning of the domain given observations,

$$Y(\boldsymbol{s}_i) = \sum_{k=1}^{K} a_k \mathbb{I}\left(X(\boldsymbol{s}_i) \in]c_{k-1}, c_k]\right) + \epsilon_i,$$

where ϵ_i are Gaussian i.i.d. random noise and a_k are parameters.

Figure 2.7 show the observed field, Y, the underlying (latent) field, X, and the classification field, Z, acquired from thresholding X in a realization of the level set model.

In Paper II, the Gaussian noise model of Dunlop et al. (2016) is replaced by a Poisson likelihood yielding an extension of the popular log-Gaussian Cox process.



(a) (b) (c)

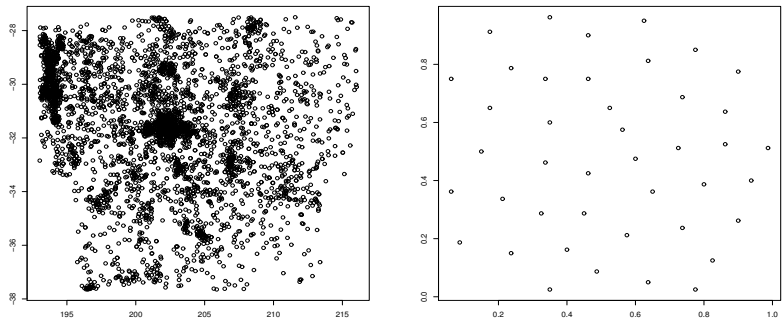
Figure 2.7: (a) Observed data corrupted by noise, Y. (b) Corresponding level set function, X. (c) Classification field.

Chapter 3

Spatial point processes

A spatial point pattern is a countable set of locations, $Y = \{x_1, x_2, ...\}$, $x_i \in \mathcal{D}$ on some continuous spatial domain, $\mathcal{D}$. We can refer to the locations as events, in the sense that they correspond to locations where something occurs. Often the point pattern is observed in an observational window, W. That is, the point pattern exists on $\mathcal{D}$ but is only observed on $W \subseteq \mathcal{D}$. Here, we consider two types of point patterns, the finite and the infinite. An infinite point pattern consist of an infinite number of events and is typically defined on an open domain such as $\mathbb{R}^d$. Practically, it is impossible to observe such a pattern in its full domain, i.e., the observational window will be a strict subset of $\mathcal{D}$. A finite point pattern on the other hand will have a bounded spatial domain including all of the events. Practically, for a finite point pattern, the observational window is often the whole spatial domain while for an infinite point pattern, the observational window is never the whole spatial domain. Point patterns occur in a vast number of applications, e.g., locations of galaxies as seen in Figure 3.1a (Drinkwater et al., 2004; Baddeley and Turner, 2005), locations of cell centers observed under optical microscopy as seen in Figure 3.1b (Baddeley and Turner, 2005; Ripley, 1977), and location of trees as seen in Figure 7.3 and used in Paper II.

A point pattern can be defined as a counting measure, N, on the spatial domain $\mathcal{D}$, where $N(A)$ counts the number of points in the spatial region $A \subseteq \mathcal{D}$. Often, point patterns can be seen as a realization from some stochastic model. Using the examples, whenever a new cell colony is grown, a new cluster of galaxies are observed, or a new region of forest is surveyed there will be a new point pattern observed. Of course, under similar conditions we expect



(a) (b)

Figure 3.1: (a) Observations of galaxies in the Shapley supercluster. (b) Location of centres of observed biological cells observed under optical microscopy.

corresponding point patterns to have similar structures, even though the set of locations are different. Hence, we need to characterize the stochastic model from which the observed patterns emanated. A point process is a stochastic model of point patterns in the same way as a random variable is a stochastic model of real values. Since a point pattern could be described as a counting measure, a point process can be described as a random counting measure. Statistical analysis of point patterns corresponds to analyzing the properties of the point process that generated these point patterns.

3.1 The Poisson process

Historically, the most important point process is the *homogeneous Poisson process*. This is the model of *complete spatial randomness* (CSR), i.e., an unstructured point pattern. That is, events occur independently of each other and the number of points in a chosen region are distributed according to a Poisson random variable with intensity parameter proportional to the spatial measure of the chosen region. In this work, we will only consider spatial domains in euclidean spaces with corresponding Lebesgue measure, $\mathcal{L}$, i.e, $\mathbb{E}[N(A)] \propto \mathcal{L}(A)$, where $A \subseteq \mathcal{D}$, and $\mathcal{D}$ is the spatial domain. This definition yields that the point process for CSR is defined by a random counting measure $N(A) \sim Pois(\lambda \cdot \mathcal{L}(A))$, where $\lambda \geq 0$ and A is any measurable subset of $\mathcal{D}$.

Historically, most methods in point process statistics have focused on differentiating between CSR and structured patterns. A structured pattern can either differ from CSR due to interaction between points and/or by spatial dependencies due to some available or unknown covariates. The difference between the two effects lies in the generative process more than the actual observed pattern. For example, assume that a seed is planted in a spatial region. The seed grows into a tree and then a new seed is planted. If the second seed is planted too close to the first tree, the plant will be shaded. The shade inhibits its possibility of growing into a large tree itself. This is an example of a repulsive interaction between points. On the other hand, the possibility of the plant growing into a large tree might also depend on the topography and soil constituents of the spatial region. Planting a seed close to a stream or in a dry desert will affect its chances as well. This is an example of spatial dependency.

Definition 3.1.1 (Intensity measure)**.** *The intensity measure,* Λ*, of a point process is a deterministic measure defined as the expected value of the random counting measure, i.e.,*

$$\Lambda(A) = \mathbb{E}\left[N(A)\right], A \in \mathcal{D}\,.$$

If Λ is absolutely continuous with respect to the spatial measure, it can be described by the *intensity function* λ as $\Lambda(A) = \int_A \lambda(\boldsymbol{s}) d\,\boldsymbol{s}$. For the homogeneous Poisson process, $\lambda(\boldsymbol{s}) = \lambda, \forall\, \boldsymbol{s} \in \mathcal{D}$, i.e., a constant intensity. The *inhomogeneous Poisson process* is a point process which behaves as a homogeneous Poisson process on infinitesimal subregions of $\mathcal{D}$. Due to the additivity of Poisson distributed random variables, the counting measure of an inhomogeneous Poisson process is Poisson distributed as $N(A) \sim Pois(\Lambda(A))$. Any Poisson process (homogeneous or not) is characterized solely by the intensity measure, Λ.

The inhomogeneous Poisson process can model some types of spatial dependencies. If an intensity function exists, covariates can be included in the model by letting λ be a function of the covariate values. In Paper II, a log-linear relationship is considered where $\log \lambda(\boldsymbol{s}) = \sum_j B_j(\boldsymbol{s})\beta_j$ for covariates B_j and coefficients β_j. However, the inhomogeneous Poisson process assumes no interaction between points, a feature inherited from the CSR model due to the additivity of Poisson random variables. Hence, it is an important but restricted special case of point processes.

3.2 Cox processes

A further extension of the Poisson process is that to a *Cox process*. In this process, $\lambda(\boldsymbol{s})$ is itself modeled as a random object, i.e., a positive random field on $\mathcal{D}$. Conditioned on a given realization of $\lambda(\boldsymbol{s})$, the point process is an inhomogeneous Poisson process with λ as its intensity function. Hence, the model is doubly stochastic in the sense that it defines a generative process based on two steps of random objects. A Cox process can also be considered as a Bayesian model of a Poisson process, where the latent intensity field, λ, is given a prior probability distribution.

A popular Cox process model is the *log-Gaussian Cox process* (LGCP) for which $\lambda(\boldsymbol{s}) = e^{\mathrm{X}(\boldsymbol{s})}$, and X is a Gaussian random field. The popularity of the LGCP model is partly due to the marriage between the two most used and studied spatial stochastic processes, the Gaussian random fields and the Poisson processes. The popularity of the LGCP is also partly due to its versatility. It can model point patterns under uncertainty about covariates, i.e., it is unknown how and which covariates that affect the probability of events. The uncertainty about the covariates is explained by the randomness of λ. It can also model clustering effects, i.e., attractive interaction effects. Regions with higher intensity in λ would correspond to clustered regions with a higher probability of observing many events. The structure of the random field, X, could in this sense explain to what extent points tend to be clustered.

A Cox process is however not enough to characterize all point processes. For instance, repulsive interaction effects such as trees competing over sunlight cannot be explained by such a model.

3.3 Characterizations of point processes

Just as moments, PDF's, and CDF's characterize a random variable, point processes can be characterized by some similar concepts. One such characterization is through the moment measures. The moment measures characterize the k-th order moments of $N(A)$, analogously to how the intensity measure was defined.

Definition 3.3.1 (k-th moment measure)**.** *The k-th moment measure of a spatial point process is defined as*

$$\mu^{(k)}(A_1 \times ... \times A_k) = \mathbb{E}\left[N(A_1)...N(A_k)\right].$$

Here, $A_1, ..., A_k$ are arbitrary measurable spatial regions on $\mathcal{D}$.

Note that $\Lambda(A) = \mu^{(1)}(A)$, the first order moment measure does not characterize interactions but higher order moments do.

Just as with random fields, the concepts of stationarity and isotropy are defined for point processeses.

Definition 3.3.2 (Stationarity)**.** *A point process with counting measure $N(A)$ is said to be stationary if,*

$$\mathbb{P}\left(N(A_1) = n_1, ..., N(A_k) = n_k\right) = \mathbb{P}\left(N(B_1) = n_1, ..., N(B_k) = n_k\right),$$

for any finite set $\{A_l\}_{l=1}^k$ where $B_l = A_l + \boldsymbol{t} = \{\boldsymbol{s} : \boldsymbol{s} - \boldsymbol{t} \in A_l\}$, i.e., a translation of A_l. (Illian et al., 2008)

Definition 3.3.3 (Isotropy)**.** *A point process is isotropic if,*

$$\mathbb{P}\left(N(A_1) = n_1, ..., N(A_k) = n_k\right) = \mathbb{P}\left(N(B_1) = n_1, ..., N(B_k) = n_k\right),$$

for any finite set $\{A_l\}_{l=1}^k$ where $B_l = \{\boldsymbol{s} : R_\theta\, \boldsymbol{s} \in A_l\}$, i.e. a rotation with angle θ of the points of A_l around the origin. (Illian et al., 2008)

The concept of ergodicity is also an important one. For an ergodic point process, the dependency between $N(A)$ and $N(B)$ will be negligible if the closest points in the two regions are sufficiently far away. This property means that if an ergodic point pattern is observed on a sufficiently large observational window, W, subregions far away from each other will have points distributed as if from different realizations of the underlying point process. The implications being that, as long as the observational window is large enough and the point process is ergodic and stationary, one point pattern is enough for statistical analysis of the underlying process—since it acts as having observed several independent realizations of point patterns from the same point process. Historically, point pattern data have been scarse and spatial statisticians have often been forced to work with single replicates of point patterns. To draw any conclusion from such a dataset the ergodicity property is necessary. Nowadays, more often datasets have an abundance of replicates and ergodicity becomes less important.

A point pattern is a set of countable point locations in a sample space of uncountable point locations. For analysis, it can often be of interest to consider probability distributions conditioned on one or more events at specific locations. This shifts the viewpoint from "an absolute frame of reference outside the process under study, to a frame of reference inside the process" (Daley and Vere-Jones, 2003). Such probabilities can be modeled using Palm

distributions. The Palm distribution of a point process is a probability distribution of point locations conditioned on that one of the points of a realization is located at a location, o. We will denote the expectation with respect to the Palm distribution as $\mathbb{E}_o$, in contrast to the regular expectation with regards to the absolute frame of reference, $\mathbb{E}$. The following definition holds for a stationary point process.

Definition 3.3.4 (Palm expectation)**.**

$$\mathbb{E}_o\left[f(Y)\right] = \frac{1}{\lambda \mathcal{L}(W)} \mathbb{E}\left[\sum_{x \in Y \cap W} f(Y - x)\right],$$

where Y is a point process, W is the observational window, and f is some real valued function of a point pattern. (Illian et al., 2008)

In words, the Palm expectation gives the expected value of $f(Y)$ conditioned on that one of the points of realizations are observed in o.

In point process literature, some functional characteristics have been given particular attention. Here, a functional characteristic refers to a function that characterizes some aspect of the point process. Originally they were mainly used to test if point patterns behaved as CSR. Nowadays they are commonly used also to evaluate the goodness-of-fit of more general point process models, i.e., compare if the estimate of the characteristic from an observed point pattern is similar to that of the model. In Paper II, a point pattern is compared to simulations from several assumed models. Evaluation of the model's performance is based on the similarity of the functional characteristics between the real pattern and the simulated ones.

In the case of a stationary point process, Ripley's K-function (Ripley, 1977) (or estimates thereof) has been used extensively in order to investigate departures from complete spatial randomness.

Definition 3.3.5 (Ripley's K-function)**.** *For a stationary and isotropic point process with counting measure $N(A)$, the K-function is defined as,*

$$K(r) = \frac{1}{\lambda} \mathbb{E}_o\left[N(b(o, r) \setminus \{o\})\right],$$

where $b(o, r)$ is the ball with center in point o and radius r.

In words, $K(r)$ is the expected number of other points found inside a ball of radius r normalized with the intensity and conditioned on that there is a point in the center of the ball. For the CSR model, $K(r) = b_d r^d$, where b_d is

the volume of the unit ball in $\mathbb{R}^d$ and d is the spatial dimension of the point pattern. Hence, by estimating the K-function from the point pattern, it is possible to study the deviations from the theoretical K-function of the CSR model. For a point process with attractive spatial interaction (clustering), $K(r) > b_d r^d$. Likewise, a point process with repulsive spatial interaction (regularization), $K(r) < b_d r^d$.

A variant of the K-function that represents the same information but is easier to interpret is Besag's L-function (Ripley, 1977, Besags comments),

$$L(r) = \left(\frac{K(r)}{b_d} \right)^{1/d}.$$

The L-function is a modification of K such that for the CSR model, $L(r) = r$ and estimations tend to be homoscedastic with respect to r. A further modification as $L^*(r) = L(r) - r$ transforms the L-function into the centered L-function for which the CSR model would have $L^*(r) \equiv 0$.

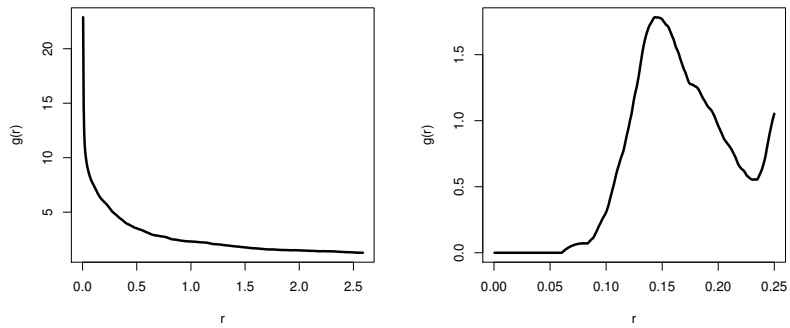
The pair correlation function, $g(r)$, is another functional characteristic which also relates to the K-function by,

$$g(r) = \frac{K'(r)}{db_d r^{d-1} r},$$

where K' denotes the derivative of K. For the CSR model, $g(r) \equiv 1$. Values of $g(r)$ larger than 1 means that there is a clustering effect at distance, r, while $g(r) < 1$ means that there is a repelling effect. Typically, a point process might have attractive effects on some intervals and repulsive effects on others. Taking the example with tree locations, a repulsive effect exists for points very close to each other due to the competition for sun. However, at medium distances there should be an attractive effect since the seed dispersal has a limited range.

Figure 3.2 shows estimates of the pair correlation function for the two point patterns shown in Figure 3.1. The galaxy dataset show a clustering effect on short distances seen by $g(r) > 1$, while the cell data seem to be regularly spaced, seen by the peak above 1 at the range of $0.11 - 0.20$. This fits intuitively with the visual perception of the two point patterns seen in Figure 3.1.

In the setting of Paper II, we have used estimates of the pair correlation function in order to compare our point pattern with simulations from the fitted models. In that setting, we did neither assume isotropy nor stationarity of the point process. However, we still expect the fitted model to yield estimated functions similar to the ones estimated from the actual point pattern. Hence,



(a) (b)

Figure 3.2: Estimated pair correlation functions. (a) Estimated g for the Shapley galaxy supercluster. (b) Estimated g for the cell data.

even though the interpretation of the functional characteristics is not clear in the non-stationary case, estimates can still be used for comparison. For details about estimating the functional characteristics mentioned above, see Illian et al. (2008).

Chapter 4

Stochastic differential equations

Let $X(t)$ be a differentiable function defined on the interval $[0, T]$ and $\mu(x, t)$ a function of x and t. If

$$\frac{dX(t)}{dt} = \mu\left(X(t), t\right), X(0) = x_0, \forall t \in [0, T],$$

then $X(t)$ is a solution to an ordinary differential equation (ODE) of first order with the initial value condition $X(0) = x_0$. ODEs often describe the evolution in time of a variable given some initial condition, i.e. a temporal system. Since this thesis concerns models for spatial statistics, we can also consider $X(t)$ to be a function in one-dimensional space, not necessarily time.

In real world applications there is often some type of random noise present. This noise can be due to measurement errors in equipment used to gather data or uncertainties about the exact domain of the study. There can also be some stochasticity inherent to the actual system of study. For an ODE, this randomness leads to the solution, $X(t)$, being a stochastic process rather than a deterministic function. This random behavior can be modeled by considering the differential to be a function of some random process, i.e.,

$$dX(t) = \mu\left(X(t), t\right) dt + \sigma\left(X(t), t\right) dB(t). \tag{4.1}$$

Here, $dB(t)$ denotes a random process (or generalized random process) and σ denotes a function characterizing the influence of $dB(t)$. Such a differential equation with a random component is known as a *stochastic differential equation* (SDE).

The solution to equation (4.1), when such exist, is no longer a deterministic function but a stochastic process itself. SDEs provide an alternative way of characterizing a stochastic process, as compared to, e.g., autocorrelation functions and spectral densities.

The solution to a SDE can be interpreted in several ways, the most intuitive being the strong solution.

Definition 4.0.1 (Strong solution to SDE)**.** *$X(t)$ is a strong solution to the SDE of Equation* (4.1) *if the integrals $\int_0^t \mu\left(X(t), t\right) dt$ and $\int_0^t \sigma\left(X(t), t\right) dB(t)$ exists for all $t \in [0, T]$ and*

$$X(t) = X(0) + \int_0^t \mu\left(X(t), t\right) dt + \int_0^t \sigma\left(X(t), t\right) dB(t).$$

The stochastic integral $\int_0^t \sigma\left(X(t), t\right) dB(t)$ is an integral where the integrand is a stochastic process and which is integrated with respect to a stochastic measure induced by $dB(t)$. Furthermore, integrating $dB(t)$ with respect to $\sigma \equiv 1$ gives rise to the stochastic process $B(t)$ defined as

$$B(t) = B(s) + \int_s^t dB(t).$$

For an exact definition of a stochastic integral, see (Klebaner, 2012, Itô- and Stratonovich-calculus).

Often SDEs are defined with respect to a Brownian motion process, i.e., B is a Brownian motion. A Brownian motion has the property

$$B(t) - B(s) \sim \mathbb{N}\left(0, t - s\right), s \leq t,$$

and is continuous everywhere but nowhere differentiable. Hence, $dB(t)$ is defined as a Wiener noise in one dimension when $B(t)$ is a Brownian motion.

Since the solution to a SDE is a stochastic process, $X(t)$ is a random variable for any fixed t. The distribution of $X(t)$ for large fixed t:s is often of interest.

Definition 4.0.2 (Invariant probability distribution of stochastic process)**.** *An SDE is said to have an invariant probability distribution, π, if $X(s) \sim \pi$ implies that $X(t) \sim \pi, \forall t > s$.*

Note, not all SDEs have an invariant probability distribution. One important SDE that does have one is

$$dX(t) = \frac{1}{2} \nabla \log f(X(t)) + dB(t), \tag{4.2}$$

where $f(x)$ is a twice continuously differentiable PDF of a probability distribution. In fact, the invariant probability distribution of Equation (4.2) is the distribution characterized by the PDF f. That is, with any feasible initial value, the marginal distribution of X for time T will converge to a random variable with PDF f as $T \to \infty$. This is utilized in the MCMC method MALA, see Section 5.3.1.

Often we cannot compute the solutions to the SDEs explicitly, instead we have to approximate them numerically. A common method used to acquire approximations of sample paths of SDEs is the *Euler-Maruyama algorithm*. It is the SDE equivalent to Eulers method (using Itô calculus).

Definition 4.0.3. *Euler-Maryuama method*

Consider a diffusion process such as in Equation (4.1)*. Decide time steps* $t_i : t_i = t_{i-1} + \Delta t$*, then*

$$X_{t+1} = \mu(X_t, t)\Delta t + \sigma(X_t, t)\Delta B_t,$$

where $\Delta B_t \overset{D}{=} B(t+1) - B(t)$.

4.1 Partial differential equations

The ODE gave rise to a solution which was a function in one dimension. This is enough when considering the evolution of some value over time—spatial problems on the other hand are often concerned with observational domains in two- or three-dimensions. Modeling spatial systems in dimensions higher than one can be achieved with differential equations that include differential operators with respect to several variables, i.e., *partial differential equations* (PDE).

A PDE can be a differential equation in both spatial variables and time. It is common to make a distinction between these two classes of variables; since time is causal and spatial dimensions are not, i.e., spatial dimensions are *acausal*. In this thesis we are concerned with purely spatial PDEs with respect to the Laplacian and gradient operators. Most of all, we are interested in the stationary dampened heat equation,

$$\left(\kappa^2 - \Delta\right) X(\boldsymbol{s}) = F(\boldsymbol{s}), \forall\, \boldsymbol{s} \in \mathcal{D},$$

with some boundary value conditions. Here, $\boldsymbol{s}$ denotes a point in a d-dimensional space and Δ is the Laplacian operator, i.e., $\Delta f(\boldsymbol{s}) := \sum_{i=1}^{d} \frac{\partial^2}{\partial \boldsymbol{s}_i^2} f(\boldsymbol{s})$. The boundary value conditions considered are typically: the value of X at the

boundary of $\mathcal{D}$ (Dirichlet), the value of the projection of ∇X on to the normal vector at the boundary (Neumann), or a mixture of the two (Robin).

This PDE models heat transfer in materials which are able to absorb heat to a certain degree. The RHS, $F(\boldsymbol{s})$, is known as the *source term* and models heat sources and heat sinks, κ^2 models the dampening, where a large κ^2 correspond to a material with a high degree of heat absorption.

The strong solution does not exist for all PDEs. This is often because the strong solution has to be a smooth function while abrupt changes in material constants or heat sources often occur in real world problems. However, in the physical world we often observe solutions to PDEs, even when the strong solution does not exist!

The problem is that the strong solution is interpreted pointwise, i.e., the partial differential equation should hold for every point in the spatial domain. In reality, this is not how the solution to most physical systems should be interpreted. Instead, the differential equation does not need to hold for every point but it should hold in a distributional sense, such a solution is known as the *weak solution*. We present the weak solution for the dampened heat equation on a subdomain $\mathcal{D} \subseteq \mathbb{R}^d$ with respect to the Hilbert space $L^2(\mathcal{D})$ and its inner product $\langle \cdot, \cdot \rangle$. First, assume that the strong solution, X, exists and is smooth enough such that $\left|\left\langle \left(\kappa^2 - \Delta\right) X(\boldsymbol{s}), \phi \right\rangle\right| < \infty$ for some class of functions $\phi \in V$. Then, by Green's first identity,

$$\begin{aligned}\left\langle \left(\kappa^2 - \Delta\right) X(\boldsymbol{s}), \phi \right\rangle &= \left\langle \kappa^2 X(\boldsymbol{s}), \phi \right\rangle + \langle -\Delta X(\boldsymbol{s}), \phi \rangle = \left\langle \kappa^2 X(\boldsymbol{s}), \phi \right\rangle \\ &\quad + \langle \nabla X(\boldsymbol{s}), \nabla \phi \rangle - \langle \boldsymbol{n} \cdot \nabla X(\boldsymbol{s}), \phi \rangle_\Gamma =: a(X, \phi),\end{aligned}$$

where $\langle \cdot, \cdot \rangle_\Gamma$ denotes the inner product on the boundary of $\mathcal{D}$, $\boldsymbol{n}(\boldsymbol{s})$ the normal vector to the boundary at point $\boldsymbol{s} \in \partial \mathcal{D}$, and V is a function space of differentiable functions. Note that $a(\cdot, \cdot)$ is a bilinear form, i.e., linear in both its first and second argument. The bilinear form describes the differential operator in a distributional sense. The weak solution to the PDE would be $X \in U$ which fulfills

$$a(X, \phi) = \langle F, \phi \rangle, \quad \forall \phi \in V. \tag{4.3}$$

That is, the function $X(\boldsymbol{s})$ in the function space U for which the LHS equals the RHS for any choice of ϕ within V. Here, U is known as the *trial space* and V as the *test space*. The choice of test- and trial spaces depends on a and F, since both $a(X, \phi)$ and $\langle F, \phi \rangle$ has to be well defined and bounded for every choice of $\phi \in V$ and $X \in U$. The boundary conditions can be incorporated into the weak formulation, either by constraining the test space further

(for Dirichlet boundary conditions) or implicitly through a modification of the bilinear form (for Neumann boundary conditions). Note that if a strong solution exists it is also a weak solution.

We here considered the dampened heat equation since this will be of importance in Papers III and IV. For other PDEs the weak solution can be defined in a similar fashion where $a(\cdot,\cdot), U$, and V will depend on the PDE, the boundary conditions, and the spatial domain, $\mathcal{D}$.

4.2 Finite element method

PDEs occur in many forms and on all kinds of spatial domains. Often, an explicit solution is not available. Instead, we are forced to resort to numerical methods to approximate the true solution. One class of numerical approximations to solutions of PDEs are the *finite element methods* (FEM). These approximations are based on the weak formulation of the problem. The test- and trial spaces are in most problems infinite dimensional and therefore hard to handle practically. The Galerkin method can be used to reduce dimensionality. Here, the test- and trial-spaces are reduced to the same finite dimensional subspace, i.e., $U_h = V_h \subset V$. In words, instead of forcing the equality of Equation (4.3) to hold for all possible choices of $\phi \in V$, it is only required to hold for all ϕ in a subspace, V_h. Also, the solution should be found within the class of functions, V_h.

Since V_h is a finite dimensional space, the weak solution of Equation (4.3) is reduced to a system of linear equations,

$$\sum_{i=1}^{N_U} z_i a(\phi_i, \phi_j) = \langle F, \phi_j \rangle \,, \forall j \in \{1, ..., N\} \Leftrightarrow KZ = F_h. \tag{4.4}$$

Here, $\{\phi_i\}_{i=1}^N$ are basis functions of V_h and $\{z_i\}_{i=1}^N$ are the coefficients yielding the approximate solution, $X_h = \sum_{i=1}^N z_i \phi_i$. The solution, X_h, exists and is unique if $a(\cdot,\cdot)$ is bounded, symmetric, and coercive and V_h is a closed subspace of the Hilbert space considered (Brenner and Scott, 2008, theorem 2.5.6).

Obviously, the true solution will fulfill Equation (4.4) if $X \in V_h$. If $X \notin V_h$, the approximate solution, X_h, will not be identical to X and $\|X - X_h\| > 0$. Galerkin's method has an important orthogonality property,

$$a(X - X_h, \phi) = a(X, \phi) - a(X_h, \phi) = \langle F, \phi \rangle - \langle F, \phi \rangle = 0.$$

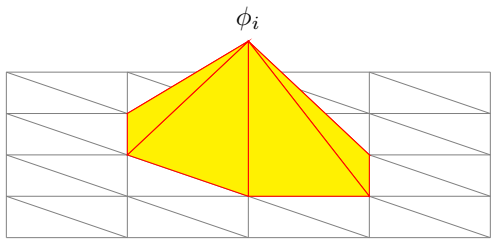


Figure 4.1: Example of a test function for node i on a 2-dimensional mesh.

Furthermore, even though $\|X - X_h\|$ will not necessarily minimize the error among all functions in V_h, the error will be proportional to the minimum error (Brenner and Scott, 2008, Céa's lemma 2.8.1).

In the finite element method, the function space, V_h, and corresponding basis, $\{\phi_i\}_i^N$, are chosen in a clever way. The aim is to acquire a sparse system matrix, K, and a V_h that can approximate reasonably smooth solutions. For this thesis we are concerned with V_h being the space of continuous and piecewise linear functions on $\mathcal{D}$—other piecewise polynomial functions are often considered in FEM applications and are constructed similarly.

By dividing the spatial domain, $\mathcal{D}$, into a triangular mesh (in two dimensions, tetrahedral mesh in three dimensions and so on), V_h can be characterized as the space spanned by a certain type of basis functions—each basis function exclusively identified by a node on the mesh. For a node indexed by i, the basis function, ϕ_i, is defined as a function that is linear in each triangular sub domain, has value 1 in node i and value 0 in all other nodes in the mesh. Such a function is sketched in Figure 4.1, it is easy to see why they are often referred to as "pyramid functions". Since each basis function has compact support, and hence only overlaps with a small number of other basis functions, K will be a sparse matrix. By making the mesh finer, V_h increases in dimensionality but will also approximate smooth functions in V better.

The main computational costs associated with FEM are the cost of solving the linear system of equations in Equation (4.4) and the cost of creating the triangular mesh. Both operations will have a computational cost that depends on the number of nodes in the mesh. The cost of solving the linear system will also be affected by the sparsity of K. A smaller number of nodes and a sparser K leads to less computations. However, a mesh with more nodes yields better accuracy. As can be seen, there is a trade-off between low computational cost and high accuracy.

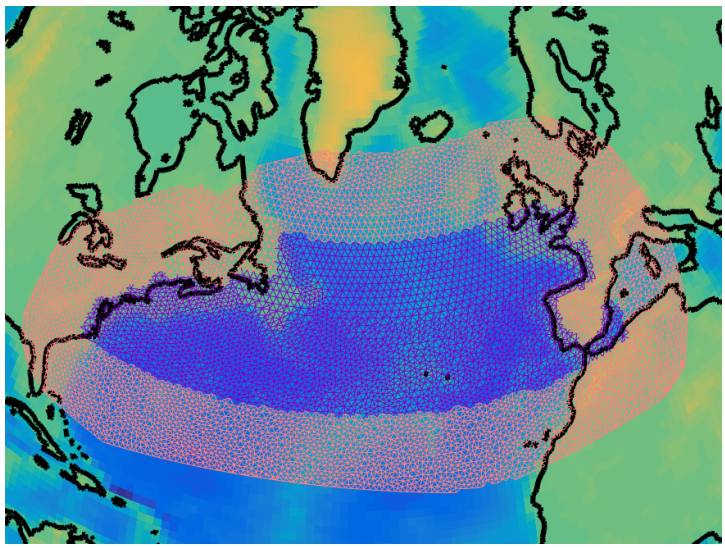


Figure 4.2: Mesh of the north Atlantic ocean considered in Paper IV. Region of interest (blue triangles) and extension region (pink triangles).

4.3 The SPDE approach to Matérn fields

In the same way as a SDE was acquired by introducing stochasticity into an ODE, a stochastic partial differential equation (SPDE) is acquired by introducing stochasticity into a PDE. In this thesis we are concerned with the SPDE acquired from feeding the dampened heat equation of fractional power with Wiener noise,

$$\mathcal{L}^{\alpha/2} X(\boldsymbol{s}) := \left(\kappa^2 - \Delta\right)^{\alpha/2} X(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \tag{4.5}$$

where $\mathcal{W}(\boldsymbol{s})$ is a Wiener noise, as explained in Chapter 2. For integer valued powers, $\alpha/2 \in \mathbb{Z}^+$, the power operator should be interpreted as a composition of the same operator over and over, e.g., $\mathcal{L}^2 X(\boldsymbol{s}) := \mathcal{L}\left(\mathcal{L} X(\boldsymbol{s})\right)$. On $\mathbb{R}^d$ this definition can be extended to any $\alpha > d/2$ using the Fourier transform,

$$\mathcal{L}^{\alpha/2} X(\boldsymbol{s}) := \mathcal{F}^{-1}\left[\lambda(\boldsymbol{\omega})^{\alpha/2} \hat{X}(\boldsymbol{\omega})\right](\boldsymbol{s}),$$

where $\mathcal{F}$ denotes the Fourier transform, λ is the spectrum of $\mathcal{L}$ and $\hat{X}(\boldsymbol{\omega})$ is the Fourier transform of $X(\boldsymbol{s})$. It should be noted that for the differential

operator of Equation (4.5), $\lambda(\boldsymbol{\omega}) = \left(\kappa^2 + \|\boldsymbol{\omega}\|^2\right)$ (Whittle, 1954). Hence, on $\mathbb{R}^d$ we can solve Equation (4.5) using the Fourier method,

$$\begin{aligned}\left(\kappa^2 + \|\boldsymbol{\omega}\|^2\right)^{\alpha/2} \mathcal{F}\left[X\right](\boldsymbol{\omega}) &= \mathcal{F}\left[\mathcal{W}\right](\boldsymbol{\omega}) \\ &\Leftrightarrow X(\boldsymbol{s}) = \mathcal{F}^{-1}\left[\left(\kappa^2 + \|\boldsymbol{\omega}\|^2\right)^{-\alpha/2} \hat{\mathcal{W}}\right](\boldsymbol{s}),\end{aligned}$$

where $\hat{\mathcal{W}}$ is the Fourier transform of the Wiener noise—which is also a Wiener noise. From Sections 2.1.2 and 2.1.3 we see that the spectral density of the random field is $\left(\kappa^2 + \|\boldsymbol{\omega}\|^2\right)^{-\alpha} (2\pi)^{-d}$. Hence, X is a centered Gaussian random field with a Matérn covariance function with parameters, $\nu = \alpha - d/2, \kappa = \kappa$, and $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu+d/2)(4\pi)^{d/2}\kappa^{2\nu}}$.

The interpretation of the correlation structure by the operator $\mathcal{L}$ is interesting since $\mathcal{L}$ models dampened diffusion. In other words, the family of Matérn Gaussian random fields have a correlation structure which behaves qualitatively the same as heat dissipation under dampening. Since $\mathcal{L}$ acts locally, the family of Matérn correlation structures can be generalized to non-stationary random fields by letting κ be spatially varying, this is used in Paper III and IV. The SPDE approach can also be used to define Matérn covariance on arbitrary Riemannian manifolds by a generalization of the Laplacian, the Laplace-Beltrami operator.

Lindgren et al. (2011) realized that the dampening, κ, is decorrelating points far away in space in a solution to Equation (4.5). Therefore, by extending $\mathcal{D}$ to a much larger spatial domain, $\hat{\mathcal{D}}$, such that the boundary of $\hat{\mathcal{D}}$ will be far away from any point in $\mathcal{D}$, the boundary conditions will have virtually no impact on X inside $\mathcal{D}$—yielding a Matérn correlation structure on $\mathcal{D}$.

Using the finite element method and the properties of the Wiener noise, Lindgren et al. (2011) constructed a finite dimensional approximation of the solution to Equation (4.5) for $\alpha = 2$. The FEM solution, see Equation (4.4), corresponds to the system of linear equations, $KZ = \langle \mathcal{W}, \phi_j \rangle$. This yields,

$$X_h \sim \mathbb{N}(0, K^{-1}CK^{-T}), \quad C_{ij} = \mathbb{E}[\langle \mathcal{W}, \phi_i \rangle \langle \mathcal{W}, \phi_j \rangle] = \langle \phi_i, \phi_j \rangle.$$

Furthermore, Bolin and Kirchner (2018) introduced a method for acquiring FEM solutions for arbitrary $\alpha > d/2$ using the K- and C matrices, this is used in Paper IV.

The methods where FEM is used to acquire a finite dimensional approximation of a spatially continuous random field has become known as the "SPDE approach" in the spatial statistics community. Figure 4.2 portrays a mesh created for the SPDE approach in Paper IV. The blue colored triangles belong

to the spatial domain of interest. The pink colored triangles belong to the extension regions used to get rid of boundary condition artifacts (Lindgren et al., 2011).

Compared to standard, covariance-function-based, methods of working with Matérn fields, the SPDE approach provides some useful properties such as

- Generalizing Matérn GRFs to arbitrary smooth Riemannian manifolds embedded in $\mathbb{R}^d$.
- Allowing non-stationary and anisotropic models.
- Control of the approximation error.
- X_h is a Gaussian Markov random field, see Section 2.2.2.

Chapter 5

Estimation and inference

Statistical inference is the art of drawing conclusions based on the available data with the aid of some probabilistic model. In spatial statistics, this is usually associated with estimating parameter values of a model or acquiring some prediction based on such parameters. There are two main inference philosophies, the Bayesian and the frequentist.

From a frequentist's perspective there exists some true parameter values of the model. The aim is to find the best estimate of these parameters given the observed data. Once the parameters have been estimated, prediction can be made using the model. From a Bayesian perspective, the parameters themselves are not absolute but considered to be random variables. Instead of finding the "true" parameter values, Bayes theorem can be applied in order to acquire a conditional distribution of the parameter values given the observed data.

Which choice, Bayesian or frequentist, depends on the purpose of the analysis, computational restrictions, and to some degree personal preference. Broadly speaking, Bayesian methods should be chosen when some information is known about the parameters, and we want this information to guide the analysis. Bayesian methods also often give a clearer understanding about uncertainties in the parameter estimation. However, it can be easier to introduce involuntary bias in a Bayesian setting, because of the need of specifying a prior distribution. Also, the methods are often more computationally costly than a frequentist approach.

5.1 Maximum likelihood estimation using the EMG algorithm

Maximum likelihood (ML) estimation is a common frequentist approach to parameter estimation. The ML estimates are obtained as, $\hat{\Theta}_{ML} = \arg\max_\theta L(\Theta; \boldsymbol{x})$, where L is the likelihood function, Θ are the parameters, and $\boldsymbol{x}$ the observed data. The ML estimators are consistent, asymptotically unbiased and asymptotically most efficient among all estimators (Olofsson and Andersson, 2012). Sometimes it is possible to find explicit analytical solutions to ML estimators but often numerical methods are required. The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is an iterative method for finding a local maximum of the likelihood function. This method is commonly used for finding $\hat{\Theta}_{ML}$ when no analytic solution is available due to missing information such as latent variables. Mixture models, as presented in Section 2.3, could be viewed as latent models where the classification values are the missing information. Hence, the EM algorithm is often utilized to find ML estimates of mixture models.

The EM algorithm starts with some initial parameters values, $\Theta^{(0)}$. Then, in each iteration, an E-step is performed followed by a M-step. The E-step corresponds to computing the expected value of the latent variables given the current parameter values. The ensuing M-step maximizes the likelihood conditioned on the latent variables being equal to their expectation found in the E-step. The method has been shown to converge for a very general class of problems (Wu, 1983).

In Paper I, the EM algorithm could not be applied since the M-step was not computationally feasible to perform, or even approximate. Instead, the EM gradient (EMG) algorithm (Lange, 1995) was utilized. This method is based on the same idea as EM but the M-step is replaced with one step of the Newton-Raphson method, i.e,

$$\Theta^{(i+1)} = \Theta^{(i)} + H^{-1}(\Theta^{(i)})\mathbb{E}\left[\nabla \log L(\Theta^{(i)})\right],$$

where H is the Hessian matrix. That is, the M-step is replaced by one step of an iterative optimization algorithm. Any strict local maximum of the likelihood locally attracts the EM and EMG algorithm at the same rate of convergence. Hence, EMG can be a good alternative when the M-step of the EM-algorithm is unavailable.

5.2 Bayesian inference

The maximum likelihood approach to parameter estimation yields a point estimate of the "true" value of the assumed model. From the Bayesian perspective there is no "true" value. Instead, probability distributions of possible parameter values has to be incorporated, the so called *prior distribution*. Then, given data and the prior distribution it is possible to acquire a probability distribution of the parameter values. A strong point in the Bayesian approach is that prior knowledge about parameter values can be included in the estimation of parameter values. On the other hand, some choice of prior has to be chosen. If the assumption of the prior do not hold, the estimated parameter values might be strongly biased.

Given the data and the prior distribution, inference can be drawn from the, so called, *posterior* probability distribution. The posterior probability distribution is the probability distribution of the parameters conditioned on the observed data. From Bayes' theorem,

$$f(\Theta|\boldsymbol{X} = \boldsymbol{x}) = \frac{f(\boldsymbol{X} = \boldsymbol{x}|\Theta)f(\Theta)}{f(\boldsymbol{X} = \boldsymbol{x})} \propto f(\boldsymbol{X} = \boldsymbol{x}|\Theta)f(\Theta),$$

where f denotes PDF:s, $\boldsymbol{x}$ the data, and Θ the parameters. Bayes' theorem can be generalized to handle more abstract probability spaces where no PDF:s exist, see for instance (Stuart, 2010).

The posterior probability distribution is not only a point estimate but a whole probability distribution. Hence, more information is given since questions about uncertainties in the parameter estimation can be answered as well as several choices of point estimates (mean, median, mode, etc). How to choose the prior distribution depends on what is known about the problem. If nothing can be assumed there are two philosophies, either to choose an uninformative prior or to chose a prior that penalizes the complexity of the model (Simpson et al., 2017). The first philosophy, choosing an uninformative prior, will let the data explain the posterior distribution as much as possible in lack of known information. The second philosophy, penalized complexity prior (PC prior), assumes that a simpler model is better since it is more easily understood and is less prone to overfitting. In lack of information indicating the opposite, the simpler model should be preferred according to this philosophy.

5.3 Monte Carlo simulation

Through Bayes' theorem, the posterior distribution is known as a function of marginal and conditional distributions. the normalizing constant, $f(\boldsymbol{X} = \boldsymbol{x})$, is often not explicitly available. Obtaining it through analytical integration is usually impossible and Θ can be high dimensional, i.e., also numerical integration becomes troublesome. In spatial statistics in particular, Θ often includes latent random fields and is high dimensional, or even infinite dimensional. The high dimensionality makes numeric integration computationally infeasible but Monte Carlo (MC) integration is often a viable alternative.

Monte Carlo integration is a method of approximating expected values by simulating samples from the probability distribution and computing the sample mean as a proxy for the true expectation. For instance, the probability of finding $\Theta \in A$ can be written as an expectation and be estimated using MC simulation as

$$\mathbb{P}\left(\Theta \in A | \boldsymbol{X} = \boldsymbol{x}\right) = \mathbb{E}\left[\mathbb{I}\left(\Theta \in A\right) | \boldsymbol{X} = \boldsymbol{x}\right] \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(x^{(i)} \in A\right),$$

where $x^{(i)}$ are sampled from the distribution of $\Theta | \boldsymbol{X} = \boldsymbol{x}$. Hence, the posterior distribution can be approximated arbitrarily well if it is possible to generate a large enough sample from it.

MC simulation is not only useful for Bayesian inference. For instance, in Paper I, MC simulation was used in the EMG algorithm to approximate expectations related to the Markov random field. However, in Bayesian statistics, the problem with approximating the normalizing constant is so common that Bayesian inference has become more or less synonymous with MC simulations.

5.3.1 The Metropolis-Hastings algorithm

We saw that MC simulation can be used to approximate the posterior distribution as long as it is possible to sample from the true posterior distribution. This is a rather strong requirement. For Bayesian analysis, it is often the case that we are not able to sample from the posterior directly. However, it is often possible to create Markov chains with the correct stationary distribution.

A Markov chain Monte Carlo (MCMC) simulation is a MC simulation where the sample points are generated dependent on each other. Each sample point is dependent on the closest former sample point, i.e., the sample is generated from a Markov chain. This Markov chain is constructed such that its stationary probability distribution equals the *target distribution*. Here, target

distribution refers to the probability distribution which we would want to sample from in a MC simulation; typically a posterior probability distribution.

It might sound strange to generate highly dependent data in order to estimate expectations. Of course, if it was possible, we would like to generate independent data instead, then we would have just a regular MC simulation. However, if the dependency between consecutive sample points diminish fast enough compared to the sample size of the MCMC simulation, the MCMC integration will produce consistent estimates of the true expectation.

The main archetype of MCMC algorithms is the Metropolis-Hasting (MH) algorithm. It is based on proposing a new sampled value, y, distributed according to some given probability distribution conditioned on the most recently sampled value, $x^{(i-1)}$. This probability distribution is known as the *proposal probability distribution* and we denote its PDF as $q(y|x^{(i-1)})$. The value y is not necessarily chosen as the next sampled value in the Markov chain, $x^{(i)}$. Instead, it has to go through a trial where it is chosen as $x^{(i)}$ with a probability α. If not chosen, $x^{(i)} = x^{(i-1)}$. This trial is known as the *accept/reject step* and α is known as the *acceptance probability* and is given by

$$\alpha = \min\left\{\frac{f(y)}{f(x^{(i-1)})}\frac{q(x^{(i-1)}|y)}{q(y|x^{(i-1)})}, 1\right\}.$$

Here, f denotes the PDF of the target distribution. Of course, f is often known only up to a normalizing constant. However, it turns out that this provides no obstacle since the normalizing constants are canceled out in the expression anyway, i.e., we use the expression without the normalizing constant instead of f in the formula. The first ratio in α weights how probable y is compared to $x^{(i-1)}$ with respect to the target distribution. Since there might be a higher probability of realizing y conditioned on $x^{(i-1)}$ than vice versa, the second ratio balances this.

Due to the Markov structure, the initial value of the samples, $x^{(0)}$, will affect the distribution at later iterations. The first couple of iterations can be highly dependent on $x^{(0)}$ and the iterations until the dependency on the initial value has become insignificant are known as the burnin phase. How many iterations the Markov chain spends in the burnin phase varies upon the choice of initial value, the target distribution, and the proposal distribution. It is important that the chain is run for sufficiently many iterations as to leave the burnin phase. Furthermore, after leaving the burnin phase, it has to have been run for enough further iterations to yield a reasonable MC estimate. This creates a common dilemma since it is not always obvious when the Markov chain leaves the burnin phase. Typically, the burnin phase is identified visually

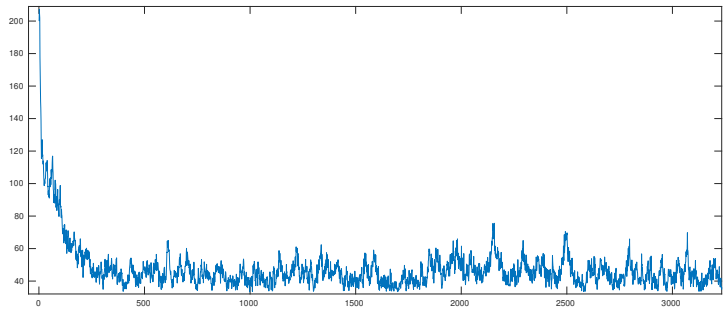


Figure 5.1: Example of a parameter path during MCMC simulation.

using plots of the parameter paths for some of the parameters. When a parameter path no longer show a clear trend, as it does in the beginning, it is assumed that it has passed the burnin phase. Figure 5.1 shows an example of a parameter path during a MCMC simulation. By visual inspection we would conclude that the Markov chain passed the burnin phase after about 400 iterations. The samples from the burnin phase are typically removed before computing the expectations—this is important since the burnin could otherwise bias the estimation.

An efficient MCMC chain should have as low dependency between consecutive samples as possible in order to make efficient use of the number of iterations available. This is known as quick mixing, as compared to slow mixing where there are significant dependencies between samples in the Markov chain even when separated by a large number of iterations. Quick mixing requires small dependencies on the prior sample in the proposal distribution while still allowing for a high acceptance probability. These are usually competing requirements that are hard to satisfy simultaneously.

A common proposal distribution for the MH algorithm is the Gaussian proposal centered at x, i.e., $q(y|x) \propto \exp\left(-\frac{(y-x)^T\Sigma^{-1}(y-x)}{2\delta}\right)$ for some chosen covariance Σ. Here, δ controls the stochastic step length. Note that this is a symmetric proposal since $q(y|x) = q(x|y)$. MH methods with this specific class of proposal distributions are known as random walk Metropolis-Hastings algorithms, since the proposals would have behaved like a random walk if the accept/reject step had not been present.

Gibbs sampler

An important special case of the MH algorithm, that even predates MH, is the *Gibbs sampler*. Suppose that the random variable is two-dimensional, $x = [x_1, x_2]$. If the conditional probabilities are known, i.e., the distribution of $x_1|x_2$ and $x_2|x_1$, it is possible to use these conditional distributions as proposals. Hence, the acceptance probability of the MH algorithm becomes $\alpha = \min\left\{\frac{f(x_1)}{f(x_2)}\frac{f(x_2|x_1)}{f(x_1|x_2)}, 1\right\} = \min\left\{\frac{f(x_1,x_2)}{f(x_1,x_2)}, 1\right\} = 1$. Since the acceptance probability is always 1, the accept/reject step is not necessary. That means that if we sample first from $x_1^{(i)}|x_2^{(i-1)}$ and then from $x_2^{(i)}|x_1^{(i)}$, in each iteration, the corresponding sample path will be a realization of a Markov chain with stationary distribution equivalent to the target distribution. This holds for x, x_1, and x_2 of arbitrary dimensionality.

It is also possible to mix the Gibbs samplers and general MH algorithms such that disjoint subsets of parameters for the target distribution are updated separately using the conditional distributions and the Gibbs sampler. However, sampling from these conditional distributions does not need to be done explicitly. Instead, one iteration of the MH algorithm can be used to sample from the conditional distributions. This is known as Metropolis-within-Gibbs MCMC and is utilized in Paper II.

MALA

The Metropolis adjusted Langevin algorithm (MALA) (Roberts and Tweedie, 1996) is a special case of the MH algorithm that, compared to the regular MH algorithm with symmetric Gaussian proposals, make use of the target distribution in designing the proposal distribution. This is achieved by using the gradient of the target PDF. It is based on the SDE of Equation (4.2), i.e.,

$$dX(t) = \Sigma\nabla \log f(X(t))dt + \sqrt{2}\Sigma^{\frac{1}{2}}dB(t), \tag{5.1}$$

where ∇ is the gradient operator with respect to the dimensions of $X(t)$, $W(t)$ is a Brownian motion of corresponding dimensionality, and Σ is the covariance operator of the proposal distribution. The solution to equation (5.1) has the target distribution as its invariant probability distribution. Hence, if the sample path of the SDE would be available, taking samples at distances sufficiently far apart would correspond to independent sampling from the target distribution. The MALA algorithm uses the Euler-Maryuama method to acquire a discretization of a sample path from the SDE. However, the discretization introduces errors and an accept/reject step is necessary to enforce sampling from the correct target distribution.

The proposals generated from the regular random walk MH algorithm can be considered as being Euler-Maryuama time discretizations of a Brownian motion. The Brownian motion does not have the target distribution as its stationary distribution and therefore it will yield more rejections than the MALA algorithm, for comparable steps lengths. This means that the MALA algorithm mixes better than the Random walk MH. However, sometimes the computational overhead of computing the gradient is so high that the regular random walk MH performs better anyway.

5.3.2 Crank-Nicholson MCMC

Cotter et al. (2013) remarked that the Euler-Maryuama scheme used for MALA is not stable with respect to the step size and the number of dimensions of the random variable. With increased dimensionality, the step length needs to be decreased in order to keep a constant acceptance probability. That corresponds to a mixing of the MCMC chain that becomes slower with increased dimensionality. This can be a problem when approximating an infinite dimensional model by a finite dimensional approximation. Cotter et al. (2013) noticed that if the target probability measure, μ^Y, is absolutely continuous with respect to a Gaussian probability measure, μ_0, the SDE,

$$dX(t) = -\mathcal{K}\mathcal{Q}X(t)dt + \gamma\mathcal{K}\nabla \log f(X(t))dt + \sqrt{2\mathcal{K}}dW(t),$$

has the probability measure μ_0 as a stationary solution if $\gamma = 0$ and μ^Y if $\gamma = 1$. Here, $\mathcal{K}$ can be chosen either as the covariance operator of μ_0 or the identity operator. $\mathcal{Q}$ is the precision operator of μ_0 and f is the Radon-Nikodym derivative $\frac{d\mu}{d\mu_0}$.

Note that Cotter et al. (2013) are speaking in terms of function spaces, operators, and Radon-Nikodym derivatives instead of random vectors, matrices, and PDFs. This is because the framework of Cotter et al. (2013) is mainly concerned with infinite dimensional random objects, or high dimensional approximations thereof.

By discretizing this SDE using a Crank-Nicholson approximation on the linear part of the drift, stability is achieved and the discretization errors for a chosen step length are no longer dependent on the number of dimensions. This scheme can be written as

$$\left(I + \frac{1}{2}\mathcal{K}\mathcal{Q}\right) X(t_i) = \left(I - \frac{1}{2}\mathcal{K}\mathcal{Q}\right) X(t_{i-1}) + \gamma\mathcal{K}\nabla \log f(X(t_{i-1}))\delta + \sqrt{2\mathcal{K}\delta}\epsilon.$$

The Crank-Nicholson MCMC method is a MH algorithm and the ensuing accept/reject step is identical to the regular MH algorithm. The novelty with the method, as compared to random walk MH and MALA for instance, is that the discretization is based on the Crank-Nicholson finite difference method and that it puts MCMC methods into a more general functional analytic framework.

Note that $\mathcal{K}$ could be chosen either as the covariance operator of μ_0 or the identity operator. If it is possible to draw samples from μ_0, the choice should be made as $\mathcal{K} = \mathcal{C}$, i.e., setting $\mathcal{K}$ to the covariance operator of μ_0. However, if it is not possible to draw samples from μ_0, $\mathcal{K}$ should be chosen as the identity operator. In this case, $(I + \frac{1}{2}\mathcal{Q})$ has to be inverted, which is not always possible.

Just as with MALA compared to a random walk MH, choosing $\gamma = 1$ requires evaluation of the gradient but will lead to a higher acceptance probability. The Crank-Nicholson MCMC scheme is particularly well-suited to Bayesian spatial modeling including continuous Gaussian random fields. This is because the posterior distribution of spatially continuous random fields are generally not Gaussian even though their prior distribution is. Since these posterior random fields are generally high dimensional, the step length's invariance to the number of dimensions in the Crank-Nicholson MCMC algorithms is important. This was utilized in Paper II in order to acquire an efficient posterior sampler.

Chapter 6

Applications

Although this thesis is mainly focused on methods for statistical analysis of spatial data, it is hard not to mention some applications for where it can be used. The first paper as well as the third and fourth are specifically focused on solving problems related to applications outside of the field of spatial statistics. The following chapter will present some prerequisites that are needed in order to make sense out of these appended papers.

6.1 Computed tomography

A standard X-ray projection image is acquired by exposing the region of interest of a patient (or inanimate object) to X-ray radiation. A detector or film is placed such that the patient is blocking the straight lines between the detector and the emitter of the radiation. By measuring the amount of radiation at the detector and comparing with the amount emitted, it is possible to compute the amount of attenuation of X-ray radiation as it passed through the region of interest. From the spatial extent of the detector surface a two dimensional image of the X-ray attenuation in the region of interest is acquired.

Computed tomography (CT) imaging is a technique for acquiring three-dimensional internal images of electron density within living organisms (or inanimate objects). The method relies on acquiring X-ray projection images at different angles. The Radon transform (Radon, 1986) relates the projected attenuation at the detector for a X-ray beam to the two-dimensional and spatially varying attenuation constant in a slice of the scanned region of interest. Hence, by a carefully considered set of angles of X-ray projection images, it is

possible to acquire a discretized Radon transform image of two dimensional slices of the patient. Such images can in turn be transformed to Euclidean space in order to approximate the spatially varying attenuation constants, slice by slice.

The X-radiation is ionizing, i.e., it can ionize atoms and molecules along their path by separating electrons. This ionization occurs either through Compton scattering or the photoelectric effect (Haidekker, 2013). Such ionization causes molecules to react with its surrounding, possibly breaking apart. This in turn can lead to damaged cells and mutated DNA, increasing the risk of cancer. This ionizing property can also be used in radiation therapy. Then, the ionizing property of the high-energy X-ray photons are used to damage cancerous tumors. However, for the most part, the ionization is an unwanted side effect when scanning living beings. Since a CT scan will require a large number of X-ray projection images it will expose the patient to a considerable dose of X-radiation and hence a higher risk of dangerous ionization effects. This issue is the motivation to the work of Paper I.

6.2 Magnetic resonance imaging

Magnetic resonance imaging (MRI) is another three-dimensional and non-invasive medical imaging technology. It measure tissue proton density and magnetization properties. MRI is of great importance in medicine due to its ability to detect differences in tissue even when the density of the tissue does not change considerably (Farncombe and Iniewski, 2014). Hence, MRI has outstanding soft tissue contrast compared to CT, which could mainly differentiate between tissues of varying density such as soft tissue, bone, and cavities. In Paper I we are concerned with acquiring CT-equivalent information from a MR scan in order to avoid the ionizing radiation inherent to CT imaging.

The MR scanner uses a strong magnet to produce a static magnetic field and the patient is placed inside this magnetic field. The magnetic spins of protons, mostly from hydrogen atoms, will align either parallel or orthogonal to the external magnetic field. Since the parallel orientation has a lower energy level than the orthogonal, more protons will align parallel to the external magnetic field. The relative amount of the total number of protons that are oriented parallel to the magnetic field is called the *spin excess*.

Besides the strong magnet, the MR scanner is equipped with coils able to transmit RF pulses. By transmitting pulses with a certain frequency and duration, the spin excess will flip from a parallel alignment to an alignment of

chosen degree, θ, i.e., the *flip angle*. That is, a majority of the protons will be oriented such that their spin vector has an angle of θ with the magnetization vector of the external magnetic field. Moreover, an effect known as *phase coherence* will make most of the spins direct themselves in the same direction also in the plane orthogonal to the direction of the external magnetic field. The spins will then rotate in their orthogonal direction while keeping the angle, θ, from the external magnetic field, i.e., precessing around the axis of the external field.

Once the RF pulse has been transmitted, the system will gradually return to the original state. This occurs through two processes.

1. The phase coherence will gradually relax, i.e., less alignment among the spin excess in the plane orthogonal to the external field. This process is known as dephasing or $T2$-relaxation.

2. The spin excess will gradually go back to an alignment with the external field. This is known as $T1$-*relaxation*.

$T2$-relaxation occurs at a smaller time scale (milliseconds) than that of $T1$-relaxation. The time constant of the $T2$-relaxation is called $T2$ and, correspondingly, the time constant of the $T1$-relaxation is called $T1$ (Haidekker, 2013). Both $T1$ and $T2$ are dependent on the tissue examined. Therefore, knowing $T1$ and/or $T2$ give important information about the tissue.

It is possible to measure both $T1$ and $T2$ since the precessing excess spin will emit RF signals as it relaxes. These signals can be measured by the MR scanner. Since the MR scanner measures this relaxation in three dimensions a three-dimensional image is acquired.

There are several parameters that can be modified in the MR scanner in order to acquire slightly different images. For instance, the flip angle can be chosen by modifying the RF pulse accordingly. In Paper I we are using images from two different flip angles as well as both the T2 and T1 relaxations from each flip angle. Hence, we acquired four three-dimensional images for each patient scanned. In Paper I we are using this joint set of images in order to acquire CT-equivalent information from the MR scanner.

6.3 Sea states

Papers III and IV of this thesis are concerned with modeling the spatial probability distribution of ocean waves on a large region of the north Atlantic ocean. In order to understand the papers, this section present some key concepts of ocean waves and their characterization.

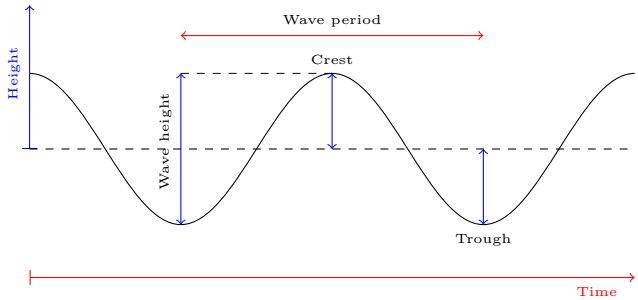


Figure 6.1: Qualitative sketch of the time series generated by a single sinusoidal wave as it propagates through a fixed point in space.

6.3.1 Local sea states

Ocean waves are dynamic motions of surface elevation. They are created by the friction between the water surface and wind. The waves transfer energy through the water away from the disturbance that introduced the energy into the system. The pattern of wave evolution can be seen as a function in time and space of the sea surface elevation, $W(t, \boldsymbol{s})$. As such, we can define local characteristics of the surface elevation, for instance wave crests (local maxima), wave troughs (local minima), and points of zero elevation (zero contour curves).

The simplest case of an ocean wave would be a monochromatic plane wave. Such a wave can be characterized as

$$W(\boldsymbol{s}, t) = A \cos\left(\omega t - k\, \boldsymbol{s} \cdot \hat{\boldsymbol{x}} + \epsilon\right),$$

where $\hat{\boldsymbol{x}}$ is the unit vector in the direction of propagation of the wave, ω the angular frequency, k the wave number, A the amplitude, and ϵ the phase. Figure 6.1 shows a qualitative sketch of a monocromatic plane wave observed as the sea elevation at a fixed point in space and as a function of time.

Due to gravity, the spatial propagation speed of the wave (*phase velocity*) is controlled by the angular frequency. This means that the wave number is $k = \frac{g}{\omega}$, where $g \approx 9.81$ is the gravitational acceleration. The phase velocity becomes $V = \frac{\omega}{k} = \frac{\omega^2}{g}$. Note that the phase velocity is dependent on the angular frequency.

In reality, a sea surface cannot be modeled by a single monochromatic plane wave. However, it can be decomposed into a sum of individual monochromatic

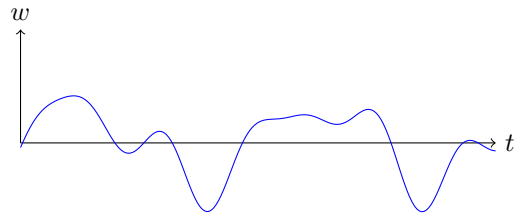


Figure 6.2: A polycromatic waveform.

plane waves, i.e., a *polychromatic wave*. For a polychromatic wave, the sea surface elevation can be described as

$$W(\boldsymbol{s},t) = \sum_{i=1}^{N} A_i \cos\left(\omega_i t - \frac{\omega_i^2}{g}\,\boldsymbol{s}\cdot\hat{\boldsymbol{x}}_i + \epsilon_i\right).$$

Here, N is the number of monochromatic waves that are being superimposed. Note that the propagation direction, amplitude, angular frequency, and phase all can vary between the individual monochromatic waves. An example of a polychromatic waveform observed at a fixed point in space can be seen in Figure 6.2. As can be seen, for a polycromatic wave, the concepts of wave height and wave period is less clear. Furthermore, since the propagation speed of monochromatic waves depend on their angular frequency, a polychromatic wave, including sinusoids with varying angular frequency, will change shape as time evolve.

The number of terms, N, might be large and we might ask ourselves how we can characterize this apparently chaotic behavior of ever-changing waveforms? The answer being, as a stochastic process, i.e., $W(\boldsymbol{s},t)$ is a stochastic process in space and time. The probability distribution of $W(\boldsymbol{s},t)$ is known as the *sea state*. A sea state typically concerns the behavior of W for a smaller region in space during a smaller interval of time. Due to the changes in wind, current, tides, etc., the probability distribution of W will not remain the same when considering other time intervals or spatial regions.

If the water is deep enough, the surface elevation can be modeled by a Gaussian process. This is due to the fact that deep water allows the individual monochromatic waves to superimpose on each other without interacting. Since the sea surface elevation can be explained as a large sum of non-interacting sinusoidal waves of independent random amplitude, phase, and direction, the central limit theorem implies Gaussianity. As a Gaussian stochastic process, the properties can be completely characterized by a first order mean function

and a second order covariance function. In this thesis we are mainly concerned with the implications of the sea state to naval logistics. In most of these cases, the mean value is not of any importance and we might for practical purposes consider a centered Gaussian process—that is, we consider the mean function to be zero everywhere. By considering a sufficiently small region in space and time, the Gaussian process will be approximately stationary. Hence, for a local sea state, W is completely characterized by its spectral density, S.

An important consequence of the dependency between propagation speed and angular frequency of monochromatic waves is that the spatio-temporal spectral density can be characterized by a function of only two variables, the angular frequency, ω, and the direction, θ (Åberg et al., 2008). We will refer to the spectrum on this form as the directional spectrum, $S(\omega, \theta)$. The spectral moments of W in the spatial direction 0° are then defined as

$$m_{ij} = \int_0^\infty \int_0^{2\pi} \left(\frac{\omega^2}{g} \cos\theta \right)^i \omega^j S(\omega, \theta) d\theta d\omega.$$

If we would want to, the angle 0° can be redefined to any angle, τ, by just translating $\cos\theta$ to $\cos(\theta - \tau)$. The spectral moments give important characterizations of the distribution of W as will become apparent in Papers III and IV.

The *temporal spectrum* characterizes the stochastic process of sea elevation in time for a fixed point in space, i.e., $W(0, t)$ for a stationary process. The temporal spectrum is defined as

$$S(\omega) = \int_0^\infty S(\omega, \theta) d\theta.$$

6.3.2 Significant wave height

For the monochromatic wave of Figure 6.1 the wave height was defined as the elevation between the crest and trough. When characterizing a polychromatic wave, this definition is not as useful since the wave height of individual waves will be random. An important statistical quantity often used to characterize the distribution of individual wave heights for a given sea state is the *significant wave height*. The significant wave height was originally defined as the mean wave height among the highest third of the individual waves, denoted as $H_{1/3}$. This definition was intended to mathematically express the average wave height as estimated by a "trained observer". Even though $H_{1/3}$ is clearly defined it is not always easy to compute. Hence, when stochastic modeling

was popularized, another definition became more practical. This newer definition, denoted as H_s, is four times the marginal standard deviation of the stochastic process W. That is,

$$H_s = 4\sigma = 4\sqrt{\mathrm{Var}[W(0,0)]} = 4\sqrt{m_{00}}.$$

Under the narrow-band approximation (a spectral density with its main energy in a narrow interval of frequencies) $H_{1/3} \leq H_s$. For most practical problems the inequality is close to sharp and $H_s \approx H_{1/3}$. Therefore the two definitions are often used interchangeably.

The significant wave height is measured in units of height (typically meters). Many important properties of the sea state can be derived simply from H_s. In Paper III it is used to compute risks associated with naval logistics.

6.3.3 Wave period

Just as the significant wave height was a quantity summarizing the random behavior of the wave heights, it is also possible to summarize the distribution of individual wave periods. There are several such quantities in use, three of these are the *peak wave period* (T_p), the *mean zero-level crossing wave period* (T_z), and the *mean wave period* (T_1). The peak wave period is defined as the period which maximizes the spectral density, i.e.,

$$T_p := 2\pi \left(\arg\max_{\omega} S(\omega) \right)^{-1}.$$

The mean wave period is defined as the first moment of the normalized period spectrum, i.e.,

$$T_1 = 2\pi \frac{\int_0^\infty \omega^{-1}\, S(\omega) d\omega}{\int_0^\infty S(\omega) d\omega}.$$

The mean zero-level crossing period is defined as the mean time between two consecutive zero upcrossings (or alternatively zero downcrossings). This corresponds to

$$T_z = 2\pi \sqrt{\frac{\int_0^\infty S(\omega) d\omega}{\int_0^\infty \omega^2\, S(\omega) d\omega}}.$$

All three quantities are measured in units of time (typically seconds). The notation T will, from here on, be used to specify a quantity of wave period

distribution, without specifying which. Considering H_s together with a quantity related to the wave periods gives more information than only considering H_s. For instance, a very high wave is often as most dangerous when the wave period yields wave lengths of the same order as the length of the ship. In Paper IV, the joint information of both H_s and T_1 are taken into account in order to evaluate the risk of capsizing.

6.3.4 Wave velocity and direction

Both H_s and T can be defined from the temporal process $W(0,t)$, i.e., without concern of the wave direction. However, waves do have directional properties. For a monochromatic wave, the concept of wave direction is clear. The crests of a wave, or any other identifiable point of the wave, will move in a straight line; we also know the propagation speed of the wave, i.e., the phase velocity. For a polychromatic wave, the concepts of direction and speed become more ambiguous. Just as with the wave period, several definitions of local wave velocity are in use (Longuet-Higgins, 1957; Baxevani et al., 2003). Let us consider one such definition, the zero-level curve velocity,

$$V(\boldsymbol{s},t) = -\left[\frac{W_t(\boldsymbol{s},t)}{W_x(\boldsymbol{s},t)}, \frac{W_t(\boldsymbol{s},t)}{W_y(\boldsymbol{s},t)}\right],$$

where W_i denotes the partial derivative in direction i of the stochastic process W. It should be noted that if W is a monochromatic wave, the velocity is deterministic and corresponds to the phase velocity, g/ω. However, except for monochromatic waves, the velocity varies randomly over time since it depends on the stochastic process W.

A common approximation is that of a *long crested sea*, i.e., a unidirectional waveform. For such a sea, the directional spectrum is reduced to $S(\omega,\theta) = S(\omega)\delta(\theta-\theta_0)$, where θ_0 is the direction of the waves and δ is the Dirac delta function. For a long crested sea, the sea state can be characterized by the temporal spectrum together with one single direction, θ_0. A long crested sea is generally a good approximation for severe sea states—the reason being that big waves are created by strong wind blowing consistently in one direction for a long time. Hence, most of the energy of the wave state will be focused in one single direction, or a narrow band of directions.

The *mean velocity* is a vector valued quantity characterizing the mean speed and direction of the waves,

$$V_m(\boldsymbol{s},t) := \mathbb{E}\left[V(\boldsymbol{s},t)\right].$$

For a long crested sea, the mean wave velocity in the direction of wave propagation can be given as a function of the spectral moments, i.e.,

$$V_m := -\frac{m_{11}}{m_{20}}.$$

6.3.5 Parametric spectral densities

In the general case, the spectral density, $S(\omega, \theta)$, can be infinite-dimensional. Quantities such as H_s, T and V_m give limited information about the sea state. However, in many cases, these quantities are all that is needed to characterize the sea state completely. Typically, sea states are approximated by some parametric family of spectral densities. There exists several such families but one of the most popular, which is both relatively simple and widely applicable, is the Bretschneider spectrum (Bretschneider, 1959; Ochi, 1998). It is a parametric family of temporal spectrums on the form

$$S(\omega) = c\omega^{-5} \exp(-1.25\omega_p^4/\omega^4), \qquad c = 0.3125\, H_s^2 \omega_p^4, \qquad \omega_p = 2\pi\,/T_p.$$

When this spectrum characterizes W, all three definitions of wave period are proportional to each other, $T_p = 1.408 \cdot T_z$ and $T_p = 1.2965 \cdot T_1$. Note that the Bretschneider spectrum is completely characterized by H_s and T.

A Bretschneider spectrum only characterizes the temporal spectrum but there are many applications where the directional distribution is of great importance. Luckily, many of those applications concern severe seas, i.e., where the waves are relatively high. We already know that the assumption of a long crested sea generally is a good approximation for such sea states. Hence, only the temporal spectrum and one single wave direction is needed.

When the wind has blown with a constant speed for long enough time and over a long enough stretch of water (*fetch*), the waves will not grow any bigger. This is called a *fully developed sea*. The amount of time and size of the fetch needed to reach full development depends on the wind speed. The significant wave height and mean wave period of the fully developed sea also depends on the wind speed. Hence, for a fully developed sea there is a one-to-one relationship between the wind speed, the significant wave height, and any of the three definitions of wave period. This relationship can be used to simplify the Bretschneider spectrum and remove the dependency on T. This spectrum is known as the *Pierson-Moskovitz* spectrum and is a Bretschneider spectrum where $\omega_p = 0.4\sqrt{g/H_s}$. While the Bretschneider spectrum can model a wide variety of sea states on deep open ocean, the Pierson-Moskovitz spectrum is restricted only to fully developed seas.

6.3.6 Risks in naval logistics

In Papers III and IV we are concerned with three types of risks. These are *fatigue damage*, *extreme wave loads*, and *broaching-to*.

A ship traversing the ocean is subjected to wear due to collisions with waves. These collisions will create microscopic cracks in the hull of the ship. With time and further exposure to the wave environment such cracks will grow while new will form. If not repaired, the cracks will sooner or later grow large enough for the hull to fail, ultimately sinking the ship. This type of wear damage is called fatigue. A ship will accumulate a certain amount of fatigue damage on any journey. However, the accumulated fatigue damage will vary in severity depending on the sea states encountered en route.

In Paper III, the probability distribution of the significant wave height was modeled spatially. From this model it is possible to compute the distribution of accumulated fatigue damage on a planned journey. Knowing this distribution makes it possible to plan the maintenance intervals as well as predicting the life length and costs of maintaining a ship. Moreover, it can also aid in making decisions about which route or mission for the ship to undertake.

If a ship encounters very high waves, severe damage to the hull can occur from more direct forces than that of fatigue damage. Hence, it is also important to know the risks of encountering extreme wave loads. In Paper III, a formula for computing the probability of encountering a significant wave height above a certain threshold value (exceedance probability) is derived. This can aid in deciding the route a ship should undertake and could save lives.

Finally, a phenomenon known as "broaching-to" can cause a ship to capsize. Broaching-to occurs when a ship traveling at a certain speed is overtaken from behind by a wave with an unfortunate combination of velocity, height, and period. This can cause the ship to start sliding down the wave (surfriding), loosing control of steering, and making the ship "trip" on its own keel and capsize. It can also cause the ship to turn quickly, angling it perpendicularly to the oncoming waves, and hence leaving it more exposed to danger. In Paper IV, we analyze the risk of capsizing due to broaching-to using the joint spatial model of both H_s and T_1.

Chapter 7

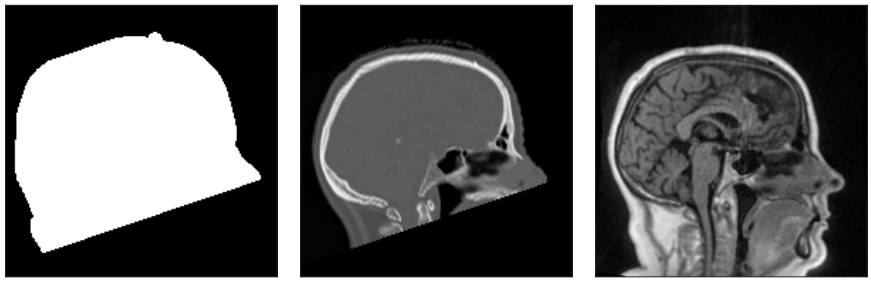
Summary of papers

The papers included in this thesis are concerned with three types of applications of spatial statistics. These are generation of substitue-CT images from MRI, point processes for latent partitioning of spatial domains, and spatial modeling of sea states. This chapter presents a brief summary of each of the four papers.

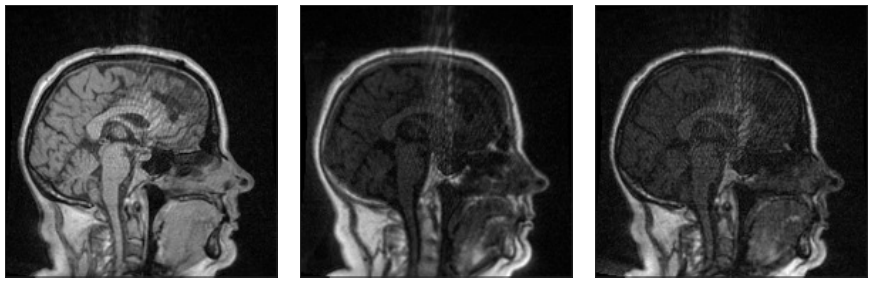
7.1 Paper I: whole-brain substitute CT generation using Markov random field mixture models

This paper models medical images from CT scans and MRI scans (of several modalities) using a joint probabilistic model. The model is developed to predict CT images conditioned on the MR images. The incentives of this work are the need for CT-equivalent information in medical applications while reducing ionization damage. Two of these applications are PET imaging and dose planning of radiation therapy.

Johansson et al. (2011) showed that it is possible to acquire a substitute CT (s-CT) image from magnetic resonance imaging (MRI) using statistical methods. Their model considered the value at a voxel for a set of aligned images (several MR images of different modalities and a CT image) distributed as a multivariate Gaussian mixture model (GMM). The set of images portrayed the same scene and contained four MR modalities (two flip angles and both T1 and T2 relaxation for each flip angle). Hence, the GMM for each



(a) Binary mask (b) CT (c) First echo, 10°



(d) Second echo, 10° (e) First echo, 30° (f) Second echo, 30°

Figure 7.1: A two-dimensional profile slice of the three-dimensional image of one of the subjects in the CT/MRI data. Binary data mask (panel a), CT image (panel b), four MRI UTE sequences (panels c-f).

voxel was five-dimensional. An example of the dataset for one patient can be seen in Figure 7.1, the images are three-dimensional but we can only show one two-dimensional slice from the full three-dimensional image.

By considering a dataset of patients where all five images were available, the parameters of the GMM could be estimated using the EM method. A substitute-CT image was generated, given the set of the four MR images, as the conditional mean of the probability distribution.

The model of Johansson et al. (2011) did not consider any spatial structure of the images. Instead, they assumed each voxel to be independent of the others. Clearly, both the MR- and CT-images have a dependency between neighboring voxels. Particularly, neighboring voxels have a high probability of portraying the same tissue type. Paper I extends on the Gaussian mixture model by assuming a Markov random field for the class memberships of the voxels, yielding a spatial mixture model as described in Section 2.3.

Paper I also extends the model by replacing the Gaussian distributions in the mixture model by the more flexible multivariate normal inverse Gaussian (NIG) distribution. Hence, allowing for varying kurtosis and skewness.

Unfortunately, for datasets of realistic size, the normalizing constant of the Markov random field is too expensive to compute. Hence, a probability measure of the joint distribution is only available up to a normalizing constant. This makes ML estimation complicated since the likelihood function cannot be computed explicitly. As an alternative, Paper I considered a pseudolikelihood, $\tilde{L}(\Theta; \boldsymbol{x})$, where the joint likelihood is approximated as a product of all conditional probabilities. Furthermore, even with the pseudolikelihood, the M-step of the EM algorithm becomes computationally too costly on medical images of realistic size. However, the gradient of the pseudolikelihood can be approximated through Gibbs sampling. Therefore, the parameters can be estimated using an EM gradient algorithm.

Variations of the proposed method are evaluated and compared with the original model of Johansson et al. (2011) using cross-validation. The study is performed using a dataset of brain scans from 14 different patients. The conclusion is that the spatial classification model clearly increases the predictive ability. However, adding the NIG distribution did not show much improvement. An example of classification for a slice of a head is shown in Figure 7.2. The classes represent: soft tissue (turquoise), bone (blue), air (red), and a class representing small scale mixing between both bones and soft tissue (green).

7.2 Paper II: Level set Cox processes

A popular point process model of non-interacting point observations, with spatially varying and unknown intensity, is the log-Gaussian Cox process (LGCP). A common assumption is assuming that the covariance function of the latent Gaussian field is a member of a parametric family of stationary covariance structures and the mean function includes a finite number of fixed effects. Such a model is viable in many cases but the regularizing assumptions can also be too strong. An example of this can be seen in Figure 7.3 which shows a point pattern of observed locations of the tree *Beilschmiedia pendula* in a region of Barro Colorado island, Panama. The figure show a pattern that seems to be made up of two mutual exclusive subsets of the observational window. One region of low intensity, and one region of higher, and spatially varying, intensity.

Suppose that we knew the partition of the observational window into these

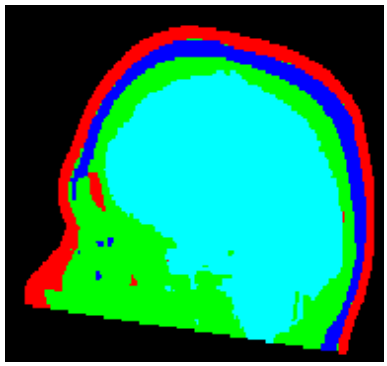


Figure 7.2: Example classification using the proposed model. Here showing a two-dimensional slice of a three-dimensional classification field on one of the subjects in the CT/MRI data set. Each color denotes a class in the mixture model.

two classes. It would then be reasonable to model the two regions separately as two different point processes on two disjoint spatial domains. Typically, a homogeneous Poisson process could be used for the low intensity regions and a LGCP model for the region with higher and spatially varying intensity. This is the idea of Paper II, where the latent structure of the LGCP model is extended with an extra level with a classification field. The classification field models the unknown partitioning of the spatial domain into the different types of classes. Conditioned on the latent classification field, each region is modeled separately by a LGCP model with a simple latent Gaussian random field structure, in this paper a Matérn covariance. The model can be viewed as a Cox process where the logarithm of the intensity surface is distributed as a spatial mixture model between several classes of Gaussian random fields, i.e.,

$$\log \lambda(\boldsymbol{s}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{s})\, \mathrm{X}_k(\boldsymbol{s}).$$

The spatially dependent classification probabilities, $\pi_k(\boldsymbol{s})$, are defined by excursion sets of a Gaussian random field, i.e., the level set approach described in Section 2.3. The model is named the *level set Cox process* (LSCP) due to this latent classification based on the level set approach.

The LSCP model is a latent Gaussian model since the intensity surface is

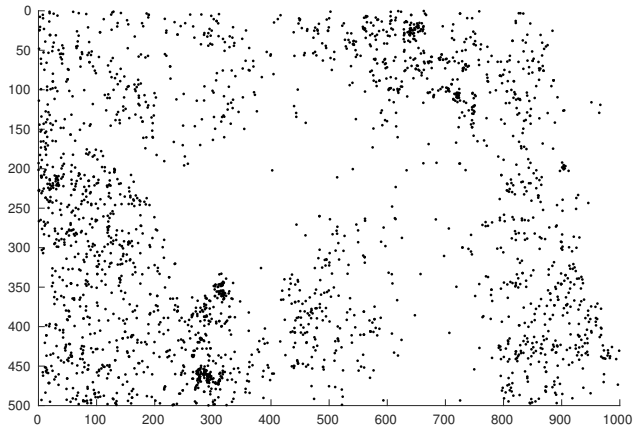


Figure 7.3: Observations of the tree *Beilschmiedia pendula* on a 1000×500 square metres area of the Barro Colorado island of Panama.

completely defined by the realizations of latent Gaussian random fields, one field for each mixture class as well as one field for the classification. Compared to the Markov random field model of Paper I, the level set approach is defined on a continuous spatial domain which makes the LSCP model continuous in space.

Having a continuous point process model is important since point observations, most often, are observed in continuous space. However, in order to practically work with the model, some finite-dimensional representation of continuous space is required. In Paper II, the observational domain, $\mathcal{D}$, is discretized into a finite number of subregions on an equidistant lattice grid, forming a partition of $\mathcal{D}$. It is shown that the posterior probability measures of the latent Gaussian fields of the finite-dimensional model converges to the posterior measure of the continuous model under refinement of the lattice grid.

The data of Figure 7.3 concerns the distribution of trees in a forest. The interest of a biologist would typically be to understand the distribution of tree density, i.e., the distribution of the intensity function, $\lambda(\boldsymbol{s})$. Most of that information could be conveyed in the correlation range and dependence on soil constituents. Hence, the biologist would want to know the values of the parameters of the Matérn correlation and the fixed effects. To get an understanding of this, Paper II uses a Bayesian approach with PC priors.

The posterior distributions of both parameters, latent Gaussian fields, and the intensity surface can be acquired by Monte Carlo simulations. Since the

model is, a priori, latent Gaussian and high dimensional, the Crank-Nicholson MCMC method (Cotter et al., 2013) is ideal, see Section 5.3.2. Also, since the finite-dimensional approximation of the random fields is a grid on a rectangular observational window, a spectral approach using fast Fourier transforms (Lang and Potthoff, 2011) further speeds up inference.

The dataset of Figure 7.3 is used to evaluate different incarnations of the proposed LSCP model. This example highlights the flexibility and potential of the model. An important conclusion is that the standard LGCP model yields biased inference on model parameters. A result that is relevant since this particular dataset has been widely used in the point process literature, often analyzed using LGCP models.

7.3 Paper III: Spatial modeling of significant wave height using SPDEs

The theory of modeling sea states by the spectral density only holds if the distribution is stationary. In Section 6.3 we stated that this is approximately true as long as we are only concerned with the sea state over a small region in space and a small interval of time. If we would be interested in large regions, the assumption of stationarity does not hold anymore. On the other hand, modeling the sea elevation spatially for large regions does not give much more information than modeling the governing parameters of the local sea states spatially. That is, instead of modeling the spatio-temporal stochastic process of the elevation of the sea surface, we are satisfied with modeling the parameters of the sea state spectrum spatio-temporally.

Remember that, for a fully developed sea, the significant wave height gave all information about the distribution of W at a fixed point in space. Since H_s is the single most important sea state parameter for most practical applications, Paper III is dedicated to modeling H_s spatially. It has been shown that the logarithm of significant wave height at a fixed point in space in the north Atlantic can be approximated by a normal distribution. It turns out that the log-normality holds also for multivariate distributions of several points in the north Atlantic simultaneously, i.e., $\log H_s(\boldsymbol{s})$ is a Gaussian random field.

When modeling H_s on scales as large as the north Atlantic, the spatial discretization needs to be high-dimensional. Therefore, it is important to use a model that scales well with respect to dimensionality. Due to this, the SPDE approach was used, see Section 4.3. The method is also advantageous since the discretized model still models a continuous random field, and the

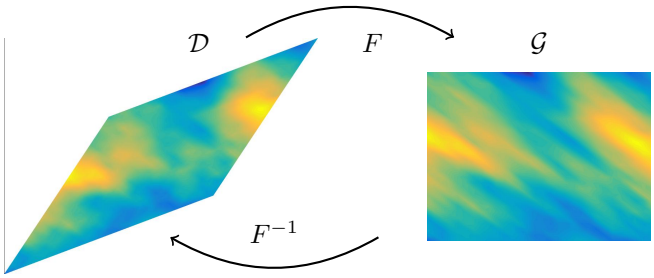


Figure 7.4: Realization of anisotropic Gaussian random field using deformation method.

discretization does not necessarily have to increase with the number of spatial locations of observations. However, the standard SPDE approach demands a Matérn correlation structure, which is stationary and isotropic.

Data show that the true Gaussian random field is neither stationary nor isotropic. Hence, the standard SPDE approach is not a reasonable model for this problem. In Paper III we modify the SPDE approach such that we can model a Gaussian random field that is both anisotropic and non-stationary while maintaining the beneficial properties of the standard SPDE approach (Lindgren et al., 2011). The modification is based on the deformation method (Sampson and Guttorp, 1992). That is, to consider a diffeomorphism between the spatial domain of the north Atlantic, $\mathcal{G}$, and some spatial domain on a related Riemennian manifold, $\mathcal{D}$. Even though the random field $X(\boldsymbol{s}) = \log H_s(\boldsymbol{s})$ is neither stationary nor isotropic on $\mathcal{G}$, it will be on $\mathcal{D}$. That is, given a differentiable bijective function $F : \mathcal{D} \to \mathcal{G}$, $\tilde{X}(\tilde{\boldsymbol{s}}) := X(F^{-1}(\boldsymbol{s}))$ is a Matérn Gaussian random field.

By letting $\tilde{X}$ be a Matérn Gaussian random field modeled by the SPDE of Equation (4.5) with unit dampening, we acquire a SPDE with X as its solution. This SPDE is

$$\left[\kappa(\boldsymbol{s})^{\frac{2}{\alpha}-2}\left(\kappa(\boldsymbol{s})^2 - \nabla \cdot H(\boldsymbol{s})\nabla\right)\right]^{\alpha/2}(\tau(\boldsymbol{s})X(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \tag{7.1}$$

where $\kappa(\boldsymbol{s}) = \sqrt{|J[F^{-1}](\boldsymbol{s})|}$, $J[F^{-1}](\boldsymbol{s})$ is the Jacobian matrix of $F^{-1}(\boldsymbol{s})$, and $H = \kappa(\boldsymbol{s})^2 J[F^{-1}]^{-1}(\boldsymbol{s}) J[F^{-1}]^{-T}(\boldsymbol{s})$. What is interesting to note is that the Jacobian of F^{-1} defines both κ and H.

We parametrize the functions $\kappa(\boldsymbol{s})$ and $H(\boldsymbol{s})$ as low-dimensional basis expansions using cosine functions. The parameters are estimated on data with

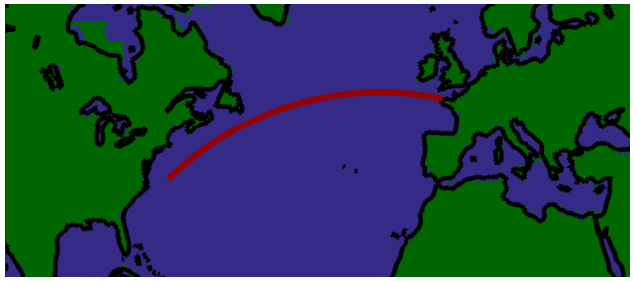


Figure 7.5: Atlantic route.

the maximum likelihood method using a numerical quasi-Newton method.

To evaluate the model, we use data of significant wave height for the month of April for the years 1979-2017. The fitted model was shown to agree well with the data. The model was also used to simulate the distribution of fatigue damage accumulated by a ship traversing the north Atlantic route, see Figure 7.5. It was shown that the spatial dependency was important to represent the correct fatigue damage probability distribution. Also, an upper bound on the exceedance probability of encountered H_s on the journey could be formulated. The bound was derived using Rice's method and the proposed model. This upper bound was shown to agree well with observations for threshold values above 5 meters.

7.4 Paper IV: Joint spatial modeling of significant wave height and wave period using SPDEs

Paper IV extends the work of Paper III. Here, not only the significant wave height, H_s, but also the wave period, T, is modeled spatially. It turns out that also the data of $\log T$ is explained well by a Gaussian random field. Hence, the univariate spatial models of H_s and T independently are equivalent to that of Paper III. That is, both $\log H_s$ and $\log T$ independently can be modeled as solutions to the same class of SPDEs, but with different parameter values. Here, the FEM implementation is extended to allow for arbitrary smoothness as well as considering the spatial domain to be on the sphere instead of on the plane of longitude-latitude projections. That is, the mesh is created on a subset of the sphere.

Paper IV also introduces a bivariate random field model of $\log H_s$ and $\log T$ jointly. The model is constructed such that the marginal distribution of the two, univariate, random fields independently is equivalent to the model of Paper III. The dependency structure between the two random fields are introduced by a system of coupled SPDEs (Bolin and Wallin, 2018; Hu and Steinsland, 2016),

$$\begin{aligned}\sqrt{1+\rho^2}\mathcal{L}_X^{\alpha/2}X - \rho\mathcal{L}_Y^{\beta/2}Y &= \mathcal{W}\\ \mathcal{L}_Y^{\beta/2}Y &= \mathcal{V}.\end{aligned}$$

Here, $X(\boldsymbol{s}) = \log H_s(\boldsymbol{s})$ and $Y(\boldsymbol{s}) = \log T(\boldsymbol{s})$. The differential operators, $\mathcal{L}_X$ and $\mathcal{L}_Y$, are both of the class defined in Equation (7.1). The two spatial Wiener noise fields, $\mathcal{W}$ and $\mathcal{V}$, are identically distributed and independent. The parameter ρ explains the cross-correlation structure between the two random fields, X and Y. Since ρ is allowed to be spatially varying, it allows for a flexible bivariate random field model. It should be noted that the parameter ρ is not identical to the pointwise cross-correlation between the two fields. However, it is related and a negative ρ corresponds to a negative cross-correlation and vice versa.

Just as for the model of Paper III, the bivariate model can be approximated by the finite element method. This gives the same important beneficial properties as in the standard SPDE approach (Lindgren et al., 2011). Figure 7.6 shows a realization of such a bivariate random field with a negative cross-correlation, and with different anisotropy structures and smoothness in the two fields. As can be seen, even though the two realizations have quite different structure, for instance elongated in directions perpendicular to each other, the regions with high values in one of them tend to be regions with low values in the other one. This is an effect of $\rho < 0$.

The univariate models for H_s and T independently were shown to agree well with data of the north Atlantic during April month in the years 1979-2018. For the joint model, some degree of model-misspecification is present in the cross-correlation structure. This seems to be largely due to the dynamic nature of the sea states in space and time. The spatial points at which the maximum cross-correlation is reached between H_s and T are not aligned in space. Since the model of cross-correlation structure assumes an alignment between points of maximum cross-correlation, the ML estimates of the cross-correlation structure do not yield agreement with data. However, by fitting the cross-correlation structure as to explain the pointwise cross-correlation, the fitted model could explain the joint distribution relatively well.

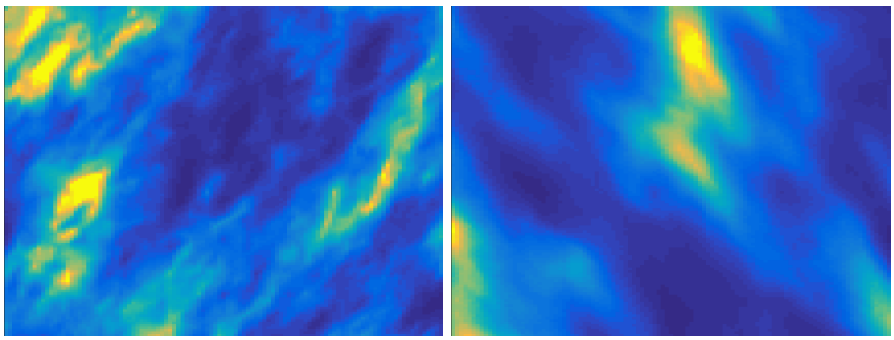


Figure 7.6: Realization of a bivariate, anisotropic and stationary Gaussian random field. The left field has a correlation range of 25 in the direction of the principal axis at 45° and a correlation range of 14 in the perpendicular direction. The right field has the principal direction at an angle of −45° with the correlation range 30, the perpendicular direction has a range of 15. The correlation between the fields are controlled by $\rho = -5$.

The fitted model was used to reevaluate the fatigue damage analysis of Paper III. In Paper III, data for T was not available, instead, the fatigue damage was computed using only H_s. By using a formula dependent on both H_s and T, a better approximation of true fatigue damage is acquired. In Paper IV, it is shown that the joint model explains the fatigue damage distribution better than just using H_s. However, similar results could be achieved by considering a univariate model of H_s and pointwise conditional means, $T|H_s$.

Also, a method for analyzing the risk of capsizing due to broaching-to was investigated. In the method, the intensity of encountering a "dangerous" wave is derived as a function of the sea state and the ship path. Furthermore, the risk of capsizing given a "dangerous" wave is evaluated as a log-linear function of H_s, T, and constants associated with the ship design. Putting these two components together, the risk of a capsizing event due to broaching-to for a ship traversing a route can be evaluated.

The risk of capsizing due to broaching-to is evaluated for a fictitious ship and the route of Figure 7.5. The risk is computed, both from the available data and from the spatial model. The distribution of capsize-risk using the bivariate spatial model compare well with the data.

Chapter 8

Future work

During research, some ideas turn out to be fruitful while others do not. It is common that new aspects of a problem are discovered while working on it. Therefore, after finishing a paper, the main problem is usually not completely solved. Instead, part of the problem has been solved and several new questions have arisen. Luckily, new ideas have often emerged as well. This section presents my view on possible paths of continuing the work presented in the appended papers.

8.1 Future work related to Paper I

The model of Paper I does only include spatial dependency in the classification field. However, in reality, the spatial dependency will also be present within the tissue types. Furthermore, it would be reasonable to model the substitute-CT as spatially continuous. That is, even though a CT image is spatially discrete, the attenuation it is measuring is rather spatially continuous. It would therefore be interesting to model the within-class random field as a spatially continuous Gaussian random field. The SPDE approach should probably be used since the spatial domain will be a complex three-dimensional region. This corresponds to the model used in (Bolin et al., 2019). Also, with the SPDE approach, the mesh could be refined in the face region while kept rather course in the larger brain region. This would make sense since more fine-scale variations are present in the face and nasal cavities.

Due to the spatial Markov random field, the parameter estimation and prediction relied on Monte Carlo simulations. Such methods will require some

computational overhead as well as lacking explicit expressions for important characteristics. It would be interesting to see if ideas from Paper II could be used in order to replace the Markov random field classification with a field based on level sets, such as in Paper II.

Since the level set model of Paper II requires an ordering between the classes, it would be interesting to see if such a constraint is reasonable for the CT/MR data. If not, would it be possible to define a level-set based classification that does not require an ordering? Such a model could be derived by considering a level set approach on a multivariate random field. However, this could possibly introduce issues with identifiability.

8.2 Future work related to Paper II

The level set Cox process model could benefit analysis in many areas of research and industry, e.g., biology, material science, epidemiology, and economic geography. In order to make it popular among such users, inference has to be fast and easy to perform without extended knowledge in mathematics .

The main drawback with the method of inference proposed in Paper II is that it is based on MCMC simulations of high dimensional spatial functions in two dimensions. This is computationally very costly—even when using the proposed Crank-Nicholson MALA algorithm. A popular software package for spatial statistical analysis is R-INLA (www.r-project.org). In this package, a LGCP model can be fitted and evaluated quickly using the SPDE approach and integrated nested Laplace approximations (INLA) (Rue et al., 2009). However, INLA relies on the latent field's link function to be linear. This is not the case for the LSCP—where a probit link occurs in the second layer of the latent Gaussian random fields.

Recently, the INLA-Bru package has been introduced (www.inlabru.org). With this, it is possible to approximate many non-linear link functions within the INLA framework. An approximation of the LSCP model could possibly be implemented using INLA-Bru. This would yield lightning-fast inference compared to the MCMC method used in Paper II. If such an implementation turned out to be successful, it will be both fast and easy for an analyst to use the LSCP model. Moreover, the user would not need to be an expert in neither mathematics nor computer science to perform the analysis and interpret the results.

8.3 Future work related to Paper III and IV

The model of Papers III and IV considers a diffeomorphism, F^{-1}, between the observed spatial domain, $\mathcal{G}$ and the deformed space, $\mathcal{D}$. We never parametrize F^{-1} directly—instead we parametrized a matrix-valued function of the Jacobian matrix of F^{-1}. Although we have yet to find a satisfiable parametrization of F^{-1}, doing so would open up new possibilities. For instance, mapping data into $\mathcal{D}$ in order to use methods applicable only to stationary processes.

A related issue is that of alignment between $\log H_s$ and $\log T$ in Paper IV. By considering the spatial domain of T to be deformed to a space where T is aligned with H_s—the suggested model of cross-correlation from Paper IV would explain the data better. This would probably be possible using the asymmetric shifted covariance model of Li and Zhang (2011) as suggested in Hu and Steinsland (2016). In fact, such a model would correspond to including yet another deformation (Sampson and Guttorp, 1992) into the model.

In Papers III and IV we used the ERA-Interim reanalysis dataset. This dataset was produced partly using atmospheric models. It would be interesting to compare the fitted model with real observed data, for instance from ships. Such real observations would be scattered irregularly in space. In this setting, it would make sense to evaluate the conditional predictive power of the model. That is, to use it for interpolation of measurements to continuous space. If shown to fit, such an application would be ideal since conditional distributions can be computed easily and with a low computational footprint using the proposed model.

The work of Papers III and IV concerns purely spatial models of sea state parameters. It is important to explain also the time evolution of the random fields. Hence, a future extension would be to consider a spatio-temporal model. Such a model would probably need to include the wave direction as well. Two important components for spatio-temporal models would be adding an advective term and to add time dependence into the SPDE. The spatio-temporal models of Krainski (2018) could possibly be used in such a model.

Bibliography

Åberg, S., I. Rychlik, and M. Leadbetter (2008). Palm distributions of wave characteristics in encountering seas. *Annals of applied probability 18*(3), 1059–1084.

Adler, R. and J. Taylor (2007). *Random fields and geometry*. Springer.

Baddeley, A. and R. Turner (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software 12*(6), 1–42.

Baxevani, A., K. Podgorski, and I. Rychlik (2003). Velocities for moving random surfaces. *Probabilistic Engineering Mechanics 18*, 251–271.

Bolin, D. and K. Kirchner (2018). The rational SPDE approach for Gaussian random fields with general smoothness. *arXiv preprint arXiv:1711.04333v3*.

Bolin, D. and J. Wallin (2018). Multivariate Type-G Matrn fields. *arXiv preprint arXiv:1606.08298v2*.

Bolin, D., J. Wallin, and F. Lindgren (2019). Latent Gaussian random field mixture models. *Computational Statistics and Data Analysis 130*, 80–93.

Brenner, S. and L. Scott (2008). *The mathematical theory of finite element methods*, Volume 3. Springer.

Bretschneider, C. (1959). Wave variability and wave spectra for wind generated gravity waves. *Technical memorandum* (118), 196.

Cotter, S., G. Roberts, A. Stuart, and D. White (2013). MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *Statistical Science 28*(3), 424–446.

Daley, D. and D. Vere-Jones (2003). *An introduction to the theory of point processes: Volume I: Elementary theory and methods*, Volume 2. Springer.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum Likelihood from incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Drinkwater, M., Q. Parker, D. Proust, E. Slezak, and H. Quintana (2004). The large scale distribution of galaxies in the Shapley Supercluster. *Publications of the Astronomical Society of Australia* (21), 89–96.

Dunlop, M., M. Iglesias, and A. Stuart (2016). Hierarchical Bayesian level set inversion. *Statistics and Computing*, 1–30.

Everitt, B. and D. Hand (1981). *Finite mixture distributions*. Chapman and Hall.

Farncombe, T. and K. Iniewski (2014). *Medical Imaging: Technology and Applications*, Volume First. CRC Press.

Gelfand, A., P. Diggle, M. Fuentes, and P. Guttorp (2010). *Handbook of spatial statistics*, Volume 2. Taylor and Francis.

Haidekker, M. (2013). *Medical Imaging Technology*, Volume First. Springer.

Hu, X. and I. Steinsland (2016). Spatal modeling with system of stochastic partial differential equations. *WIREs Computational Statistics 8*(2), 112–125.

Iglesias, M., Y. Lu, and A. Stuart (2016). A Bayesian level set method for geometric inverse problems. *Interfaces and free boundaries 18*(2), 181–217.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan (2008). *Statistical analysis and modelling of spatial point patterns*. Wiley.

Illian, J., S. Sørbye, and H. Rue (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation. *The annals of applied statistics 6*(4), 1499–1530.

Johansson, A., M. Karlsson, and T. Nyholm (2011). CT substitute derived from MRI sequences with ultrashort echo time. *Medical Physics 38*, 2708–2714.

Klebaner, F. (2012). *Introduction to Stochastic Calculus with Applications* (Third edition ed.), Volume 120. Imperial College Press.

Krainski, E. (2018). *Statistical Analysis of Space-Time Data: New Models and Applications*. Ph. D. thesis, Norwegian University of Science and Technology.

Lang, A. and J. Potthoff (2011). Fast simulation of Gaussian random fields. *Monte Carlo Methods and Applications 17*(3), 195–214.

Lange, K. (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(2), 425–437.

Li, B. and H. Zhang (2011). An approach to modeling asymmetric multivariate spatial covariance structures. *J. Multivar. Anal. 102*, 1445–1453.

Lindgren, F., H. Rue, and J. Lindstrøm (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the royal statistical society 73*(4), 423–498.

Longuet-Higgins, M. (1957). The statistical analysis of a random, moving surface. *Philosophical Transactions of the Royal Society, Series A 249*, 321–387.

Matérn, B. (1986). *Spatial Variations*, Volume 36. Springer-Verlag.

Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. *First edition*, 1104.

Ochi, M. (1998). *Ocean waves: The stochastic approach*, Volume First. Cambridge university press.

Olofsson, P. and M. Andersson (2012). *Probability, statistics, and stochastic processes*, Volume Second edition. Wiley.

Radon, J. (1986). On the determination of functions from their integral values along certain manigolds. *IEEE Transaction on medical imaging 5*(4), 170–176.

Ripley, B. (1977). Modelling spatial patterns. *Journal of the royal statistical society. Series B 39*(2), 172–212.

Roberts, G. and R. Tweedie (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli 2*(4), 341–363.

Rue, H. and L. Held (2005). *Gaussian Markov random fields*, Volume 104. Chapman and Hall.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: series B 71*(2), 319–392.

Sampson, P. and P. Guttorp (1992). Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association 87*(417), 108–119.

Simpson, D., H. Rue, A. Riebler, T. Martins, and S. S.H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science 32*(1), 1–28.

Stein, M. (1999). *Interpolation of spatial data*. Springer.

Stuart, A. (2010). Inverse problems: A Bayesian perspective. *Acta numerica 19*, 451–559.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography 46*, 234–240.

Whittle, P. (1954). On stationary Processes in the Plane. *Biometrika 41*(3/4), 434–449.

Winkler, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Springer.

Wu, C. (1983). On the convergence properties of the EM algorithm. *The annals of statistics 11*(1), 95–103.

Part II

Papers