Topological event history analysis for random fields with an application to global wind intensities

Hollie Johnson

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics, Statistics & Physics
Newcastle University
Newcastle upon Tyne
United Kingdom

September 2020

Acknowledgements

I wish to thank my supervisor Robin Henderson his unwavering support and enthusiasm in taking over my supervision. I have thoroughly enjoyed our meetings throughout my studies and will look back fondly. I am also forever grateful to Nathan for always being my biggest fan and believing in me when I didn't believe in myself. Finally, I would like to thank my parents Jennie and Roy for never letting me think I wasn't capable.

This work was supported by the Engineering and Physical Sciences Research Council, Centre for Doctoral Training in Cloud Computing for Big Data [grant number EP/L015358/1].

Abstract

Realisations of simulated climate variables from the CESM Large Ensemble (Kay et al., 2015) are frequently assumed to be independent and identically distributed random fields (Castruccio and Stein, 2013; Castruccio and Genton, 2014, 2016). Using concepts from the study of survival analysis and topological data analysis, we propose a methodology for the comparison of these realisations with specific application to global wind intensities.

Topological data analysis is becoming more widely used as the data available in many applications grows considerably, both in volume and complexity. Where computer science and machine learning often lean heavily on clustering techniques (Gan et al., 2007; Schaeffer, 2007), TDA, and more specifically persistent homology, allows a similar analysis with greater robustness to perturbations in data, for example. We extend ideas from topological data analysis using an event history approach. Survival analysis has wide-ranging applications, particularly in manufacturing and medical sciences where time-to-event data is common. We are interested in how event history methods can be used as tools for the comparison of topological features in random fields.

Drawing on work from these two areas of research, we consider specific topological features, connected components, on a random field and show that the number of these features differs between fields with different distributions or correlation structures. We use non-parametric survival models to model the rate of emergence of such features, achieving this through a reformulation of homological births and deaths as survival events.

We evaluate methods for modelling covariance of our wind intensities data on the surface of a sphere, comparing several common stationary models utilising a selection of distance measures. Our data is unusual in that we have multiple realisations of the same dataset, allowing us to examine the empirical correlation between each pair of points. We look at nonstationary approaches to modelling, including the incorporation of large-scale geographic descriptors, such as land, coast and ocean and consider the challenge of obtaining accurate covariance matrices on a single replicate.

We demonstrate how our proposed methods are informative for the assessment of Gaussianity in spatial data sets, comparing standard Gaussian data simulation packages. Finally, we apply our topological event history methods to multiple realisations from our large climate data set, identifying anomalous realisations.

Keywords: *survival analysis, topological data analysis, statistical topology, climate, spherical correlation, persistent homology*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The current climate crisis has lead to a significant increase in the urgency and volume of research into internal climate variability and future climate predictions. Assisted by the rapid advances in computational power available, climate scientists are now producing significant quantities of data available for research. Development of methods able to cut through masses of data to identify anomalous features or trends has become increasingly important. We propose methods based on theory from the fields of topological data analysis and event history analysis and demonstrate their application to global wind intensities data from the Community Earth System Model (CESM) Large Ensemble Project (LENS). We explore the challenges of covariance modelling for nonstationary data on the surface of a sphere and demonstrate how our methods have potential benefit as part of a suite of assessment metrics for covariance models.

It is widely accepted that the global climate is rapidly changing, with the rate of change becoming an increasing concern. Evidence of this change is present in a variety of contexts, from shrinking polar ice caps and the associated rise in sea levels to the increase in intensity and frequency of extreme weather events and changing ecological systems. Regardless of the increasingly challenging economic and political climate, high-quality research in this area is becoming more critical than ever, and hence our ability to effectively store and use appropriate data from a wide range of sources is vital.

The Intergovernmental Panel on Climate Change (IPCC) was founded in 1988 by the World Meteorological Organisation (WMO) and the United Nations Environment Programme (UNEP), to assist governments with climate policymaking through the provision of clear, comprehensive scientific information. The IPCC comprises 195 members, all governments who are members of the WMO or UN. The primary outputs of the IPCC include

assessment reports, comprehensive summaries of research into the drivers and impacts of climate change and how we can mitigate future risks.

The most recent assessment report from the IPCC, the 2014 AR5 Synthesis Report (Pachauri et al., 2014), provides details of the urgency of change required if we wish to succeed in adaptation and mitigation to the consequences faced by the global community. The primary conclusions of the report focused on evidence to support the findings of primary human cause alongside the message that as policymakers we *do* have the means to limit climate change if we are willing to make urgent and fundamental changes to policies and actions.

In this research, we work with data from the CESM Large Ensemble Project (see Chapter 2). The numerical climate model from which these data are simulated, is based on the RCP8.5 greenhouse gas concentration pathway, the most severe projection of greenhouse gas emissions from the IPCC AR5 report.

1.2 Purpose and aims

In this thesis, we present and demonstrate the applications of novel topological event history methods for data analysis, particularly in the context of global climate data, specifically wind intensities. We discuss the unrelated fields of topological data analysis and survival analysis and demonstrate how both can be useful for analysis of climate data on a grid. As a means of understanding the potential of different methods, we investigate the simulated data sets with different underlying distributions but the same marginal and correlation properties. Using ideas originating both from survival analysis and topological data analysis, we introduce new topological event history methods for data analysis, demonstrating their application both using simulation studies and wind intensities data.

Understanding and modelling correlation of random fields is a crucial element of the application of topological event history methods. We address the challenges of fitting correlation structures to highly nonstationary spatial data, assessing the difference between single and multi-replicate data for model fitting. In this thesis, we aim to show that topological event history methods can be valuable tools for climate data such as that from the CESM Large Ensemble, both for identification of non-Gaussian data and anomalous random fields, and to extract features of interest in the data. Further, we aim to demonstrate the use of these methods as an assessment metric for the local fit of a correlation matrix on Gaussian data.

1.3 Outline

In Chapter 2, we introduce wind intensities data from the CESM Large Ensemble project, alongside additional data sets used throughout this work and provide some exploratory data analysis. In Chapters 3 and 4, we explore topological data analysis for correlated random fields before demonstrating the use of topological data analysis (TDA) on global wind intensities data. We discuss event history analysis in Chapter 5, again considering the application of the methods to random fields, and we deal with the challenges of spatial correlation in data. In Chapter 6, we propose our method for the comparison of spatially correlated data, providing simulation studies to show its application. Chapters 7 and 8 consider the challenges of modelling covariance for data on a sphere, examining stationary and nonstationary methods respectively. We show how topological event history methods can be useful in assessing model fit and demonstrate the use of machine learning approaches for understanding nonstationarity during the modelling process. We test Gaussianity assumptions in Chapter 9 and we show how topological event history methods can be applied to identify certain non-Gaussian fields where existing methods can be insufficient. Finally, in Chapter 10 we explore a range of additional climate data sets using topological event history methods and present key findings before concluding in Chapter 11.

Chapter 2

The data set: global wind intensities

Global wind intensities on the surface of a sphere provide a valuable and topical application upon which to demonstrate the methods developed throughout this research. Further, the current climate crisis demonstrates the importance of developing techniques for working with global climate data. In this chapter, we introduce Earth System Models, before examining the primary data set used, defining conventions for the notation that will be used throughout the thesis and examining some exploratory summaries. While the primary data used in this research are wind intensities from the CESM Large Ensemble project, we also present auxiliary data sets, used mainly to enhance modelling approaches and provide insight into spatial variations in the wind data. We use a variety of such data sets, including the IPCC RCP8.5 pathway, the most severe greenhouse gas concentration trajectory presented in the IPCC 2014 assessment report (AR5), and geospatial data sets such as locations of land bodies, oceans, and continents.

2.1 Earth System Models

Numerical climate models are used by scientists to simulate the Earth's climate, through estimating solutions to differential equations. They exist in varying complexity from simple energy transfer models to fully coupled Earth System Models, in which model subsystems interact fully with each other in both directions. The Community Earth System Model (CESM) extends the latest version of the Community Climate System Model (Blackmon et al., 2001), CCSM4, released in 2010. The CCSM has seen several iterations, with simulations from the models contributing to wide-reaching research, including output from the IPCC and the US Global Change Research Program.

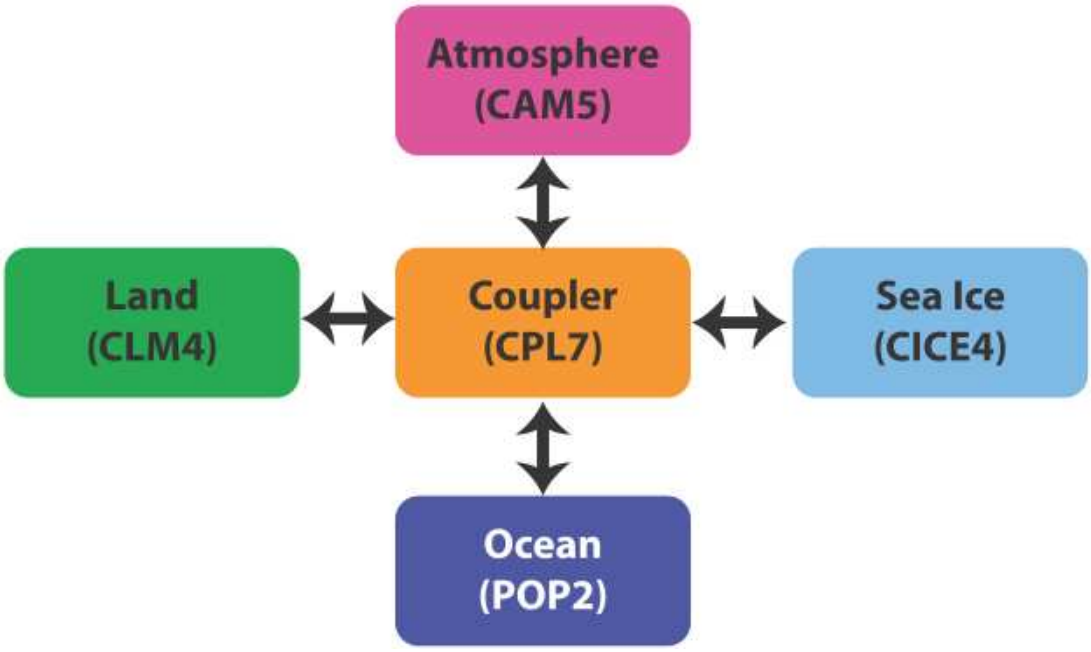

Figure 2.1: CESM1 is a fully coupled Earth System Model, in which elements including atmosphere, sea ice, land and ocean all interact.

Used to make predictions about future global climate, Earth System Models represent physical processes in the Earth's atmosphere, oceans, cryosphere and land surface and are able to simulate the response of global climate to changing greenhouse gas concentrations. More comprehensive than a climate system model, ESMs generally include a carbon cycle element which interacts fully with terrestrial and ocean ecosystems, as well as other model components (Hurrell et al., 2013). CESM1 is an example of a fully coupled (Figure 2.1) Earth System Model currently being used in climate and modelling research. Some elements of the CESM are shown in Figure 2.2.

2.1.1 Large space-time data sets from climate model output

ESMs produce large quantities of data which require considerable effort, expertise and cost to store and manage. The rapid developments in processor speeds over the last decade have hugely improved the accuracy of models through increased resolution, but improvements in storage capabilities have not occurred at the same rate. Data is often lost through the reduction of resolution or simulation length to meet storage requirements, negatively impacting scientific goals. This data loss could occur via a reduction in temporal or spatial resolution, simulation length or ensemble size. The CESM Large Ensemble Project is an essential example of this; its initial 30 ensemble members comprise over 300TB of data, yet less than 200TB was retained due to storage constraints (Kay et al., 2016). Extracting

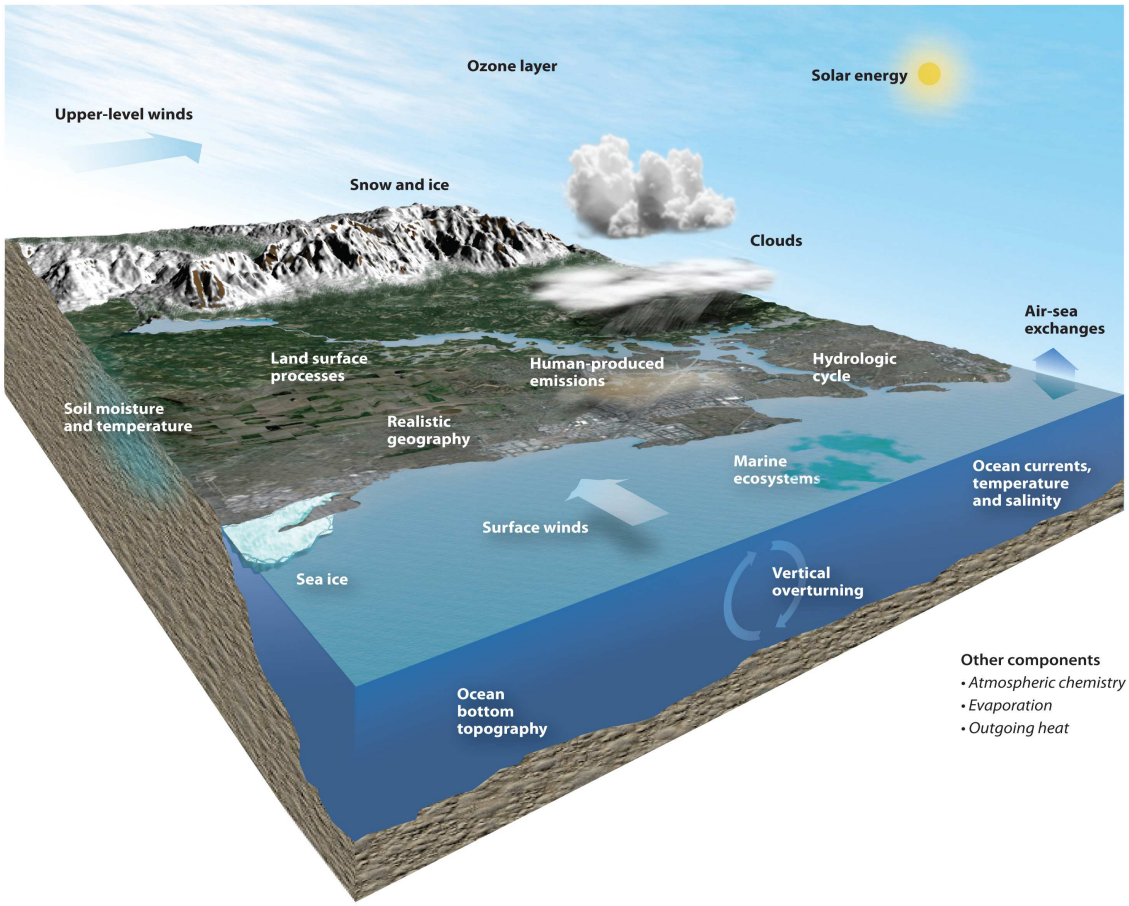

Figure 2.2: Elements of the NCAR-based Community Earth System Model (CESM). ©UCAR. https://www2.cisl.ucar.edu/software/community-models

value from the data relies on techniques to facilitate storage and management in addition to up to date computing methods.

2.2 Notation

The primary data set used in this research consists of $10m$ wind speeds measured in m/s for 192×288 sites spread across the surface of the Earth, at $0.9375°$ vertical resolution and $1.25°$ horizontal resolution. The data we are interested in span 1140 months between 2006 and 2100, and comprise 30 realisations (ensemble members), which can be considered independent due to the sensitivity of the model to initial conditions (Castruccio and Stein, 2013). From this set of monthly wind intensities, we aggregate the data by year to obtain 95×30 2-dimensional global subsets, one for each year-realisation pair. Denote by

$$\mathcal{W}_r(L, \ell, t)$$

the spatio-temporal near-surface wind intensity for realization r at the latitude L, longitude ℓ, and year t, where $r = 1, ..., 30$ and $t = 2006, ..., 2100$. Each subset corresponds to the set of points

$$\mathcal{W}_r^*(t) = \{\mathcal{W}_r(L, \ell, t); \forall L, \ell\},$$

for some realisation r and year t. Similarly, $\mathcal{W}_r^R(t)$ and $\mathcal{W}_r^S(t)$ correspond to the subset of residuals and standardised residuals, respectively, where residuals are calculated by subtracting the global mean over all realisations for a given year. Figure 2.3 shows these subsets (original data, residuals and standardised residuals) for $r = 1$ and $t = 2006$.

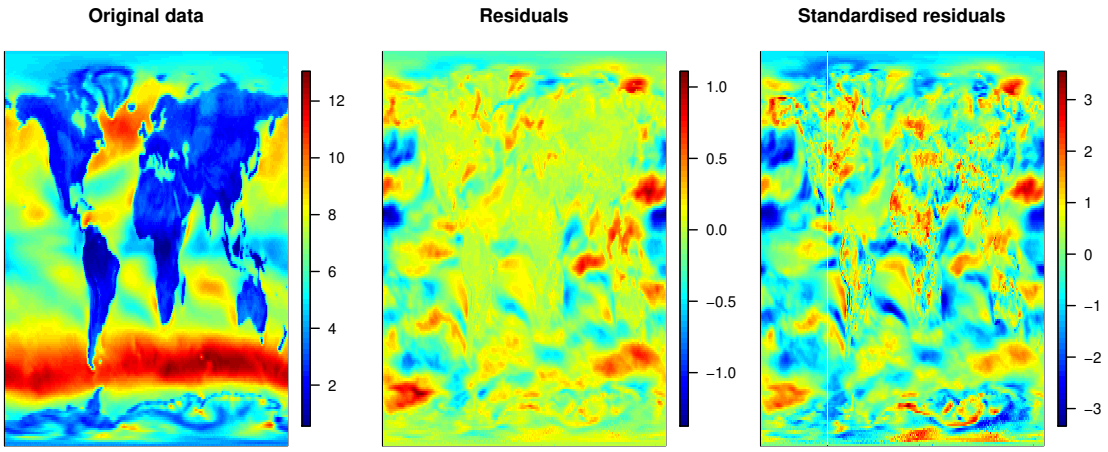

Figure 2.3: Processed data subsets for $r = 1$, $t = 2006$.

2.3 Year-realisation subsets

When working with large datasets, it can be challenging to extract value and to create meaningful visualisations. Methods for summarising the data are particularly useful in this regard, allowing us to see patterns and features within the data which would otherwise be difficult to expose. For our data set, which contains approximately 1.9×10^9 data points, development of interesting and informative summaries will help significantly in the search for patterns and anomalies and allow us to compare different subsets of the data.

In this research, we work primarily with site-by-site annual means and within-year variances for each of the 30 realisations, that is, we can consider our data as 30×95 sets of site-indexed mean values, and similarly for the within-year variances. Residuals and standardised-residuals are also calculated for both means and variances, as described in Section 2.2. We show a selection of summaries for these datasets, enabling comparison between realisations and across years. Here, we focus on the mean values, although the methods can be easily applied to the within-year variances and could contribute to future work.

2.3.1 Latitude band projections

Clear from Figure 2.3 above is the somewhat banded effect of wind intensities, particularly over oceanic regions of the globe. These are particularly noticeable in the Southern Ocean, where strong westerly winds circle Antarctica and can be seen clearly in red, in the left-hand plot of the figure. To better study and understand this banded effect, we also

calculate averages by latitude band, resulting in a length 192 vector for each subset, where each element in the vector is the average of all values in the corresponding latitude band. We also calculate this for residuals and standardised residuals. Calculating averages in this way enables us to see, via simple plots, how wind intensities change over time and realisation, and roughly where on the globe such changes are occurring. We denote by

$$\mathcal{W}_r^*(L,t),$$

the 192-vector as described above, corresponding to latitude L and year t. Similarly, we use $\mathcal{W}_r^R(L,t)$ and $\mathcal{W}_r^S(L,t)$ to denote the 192-vectors of residuals and standardised residuals.

Figure 2.4 shows these vectors plotted against latitude, for the original data, residuals and standardised residuals from realisation $r = 1$ at three years, $t = 2006$ (black), $t = 2047$ (red) and $t = 2100$ (green). We can see the average values diverging in time as the latitude increases towards the north pole. There is a less severe difference in time at approximately 50°S latitude, in the Southern Ocean region.

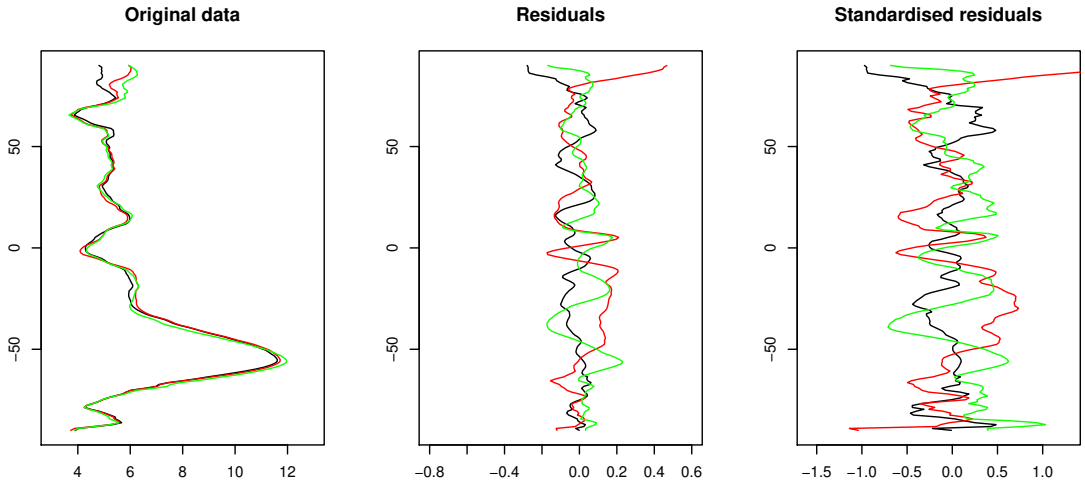

Figure 2.4: Average values by latitude, $t = 2006$ in black, $t = 2047$ in red, $t = 2100$ in green.

Comparison of the latitude effects is covered in further detail later. Methods discussed include one-dimensional topological analysis and functional data analysis. Since the impact of land and ocean on the wind intensities is clear, we can also investigate average latitude values corresponding to ocean-only or land-only sites. However, we encounter issues for latitudes that only comprise ocean sites or only land sites.

2.4 Exploratory data analysis

We show a selection of summaries to highlight patterns or changes in the distribution of the data over time and between realisations. Section 2.4.1 discusses simple summaries to describe the distribution of data points and Section 2.4.2 describes methods for summarising latitude vectors.

2.4.1 Distribution summaries

Figure 2.5 shows distribution summaries for the original data, with Figure 2.6 showing the top-left figure with detailed axes. In each plot, pixels represent individual year-realisation data sets, with realisations running from bottom to top, and years from left to right. The left column of plots show summaries calculated from pointwise yearly means, and the right column from pointwise within-year variances (calculated from monthly values). From top to bottom, the rows show minimum, lower quartile, mean, median, upper quartile and maximum of the distribution of these values.

Focusing on the mean value maps, shown in the left column, we can see some interesting patterns. As a general rule, as time increases, the minimum, mean and median increase, while the upper quartile decreases and the maximum and lower quartile show little change. Interpretation of these patterns is challenging, however plotting histograms of the data from a single realisation at years $t = 2006$, $t = 2047$ and $t = 2100$, shown in Figure 2.7, helps clarify the underlying effect.

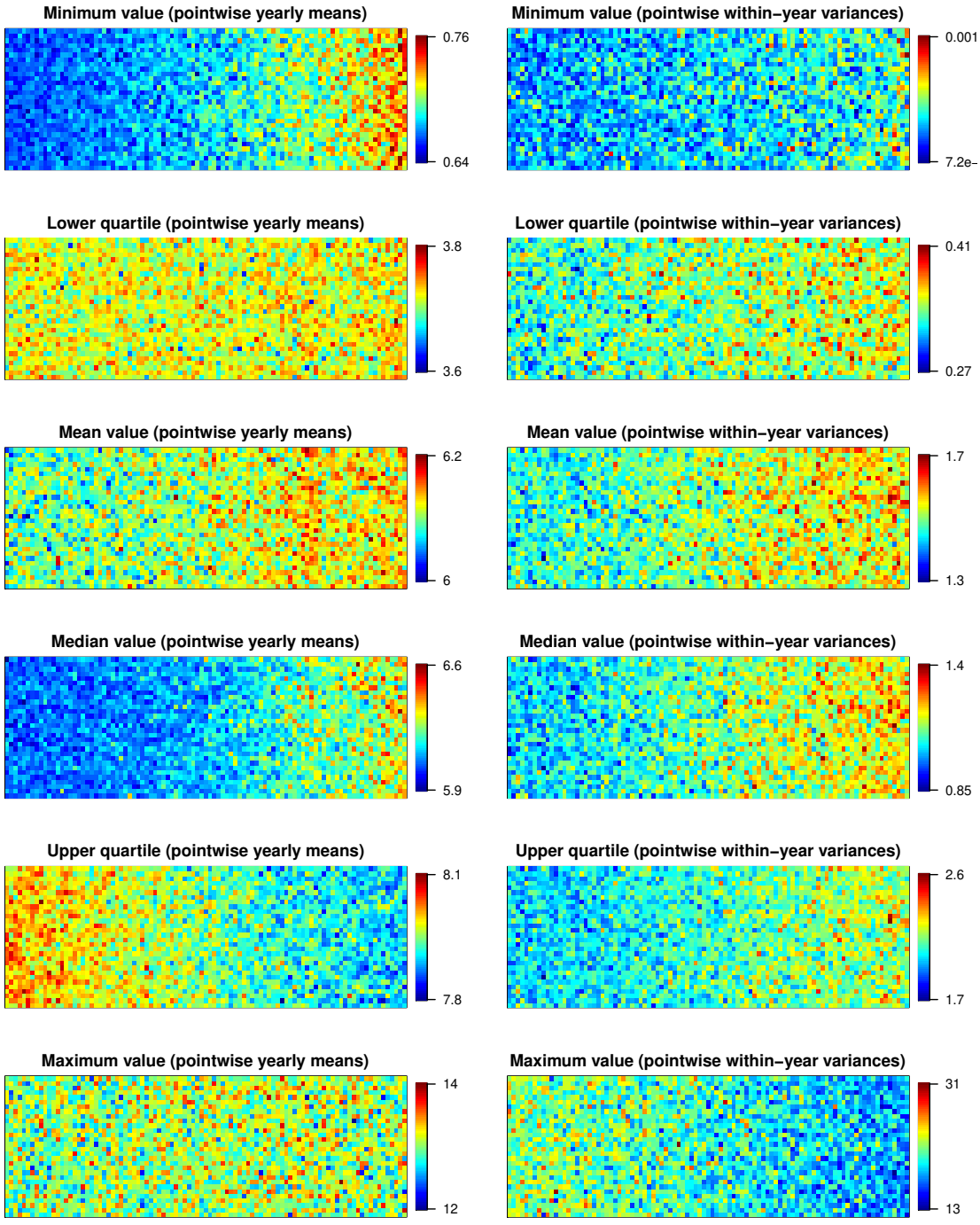

Figure 2.5: Distribution summaries. Left: summaries from pointwise yearly means, right: summaries from pointwise within-year variances. Top to bottom: minimum, lower quartile, mean, median, upper quartile and maximum.

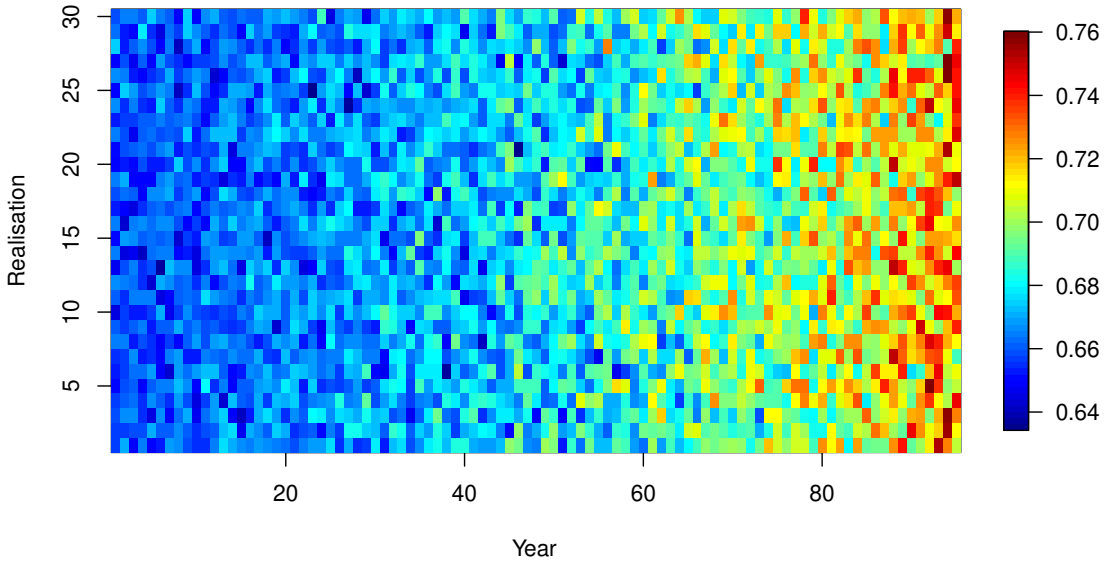

Figure 2.6: Minimum value across all sites by year and realisation

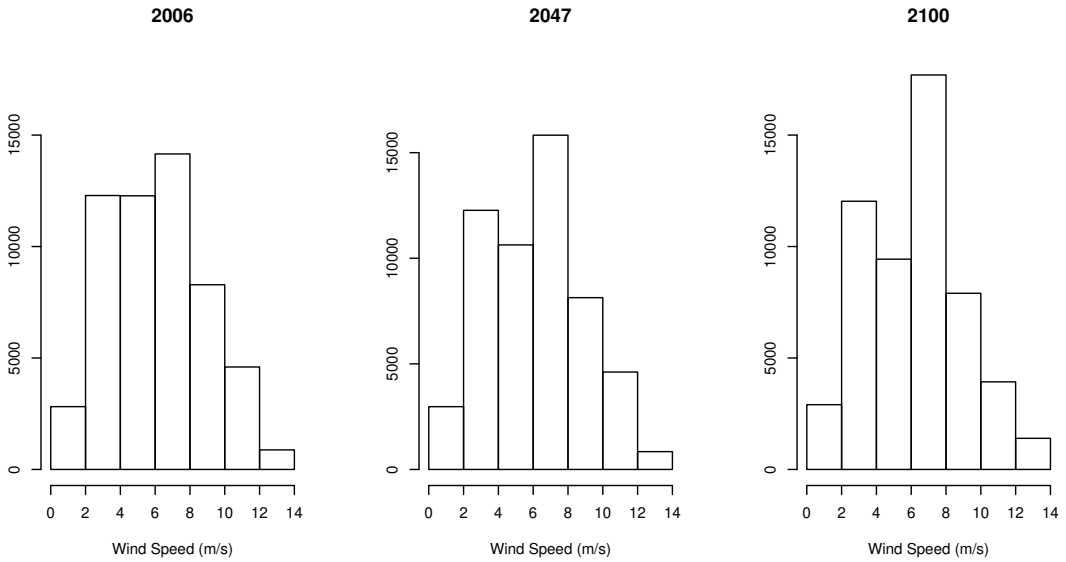

Figure 2.7: Changing distribution of wind speeds over time.

The histograms show frequencies of sites where wind speed falls between each $2m/s$ interval. What is immediately noticeable from these plots is the fall in frequency in the $4-6m/s$ range, and the corresponding increase in frequency in the $6-8m/s$ range. It appears that at least some of the sites that have average wind intensities in the $4-6m/s$ range at the beginning of the period have an increased average wind intensity in the $6-8m/s$ range at the end of the period. This effect can be seen more clearly in Figure 2.8, where sites in the Arctic region have a definite increase in average wind speed. In Figure 2.8, the top row plots show in red areas where wind intensities lie between $4m/s$ and $6m/s$. The bottom row show in red regions where wind intensities lie between $6m/s$ and $8m/s$. Figure 2.7 and 2.8 show only data from realisation $r = 1$ of the full data set; however, the result can be seen consistently across all realisations.

2.4.2 Summarising latitude vectors

Simple plots of the latitude vectors allow us to identify anomalous years or realisations. We also consider functional data analysis and use of functional box plots to identify subtle differences. Cloud plots of the vectors for each realisation over three years highlight realisations that are more or less 'normal' relative to the others. Figure 2.9 is an excellent example of this as we can see how the residual values from realisation six differ in some regions from other realisations, particularly in the year 2006 where we see more extreme differences than in other years.

Another way of visualising latitude vectors is through the use of functional boxplots, proposed by Sun and Genton (Sun and Genton, 2011). We treat each latitude vector as

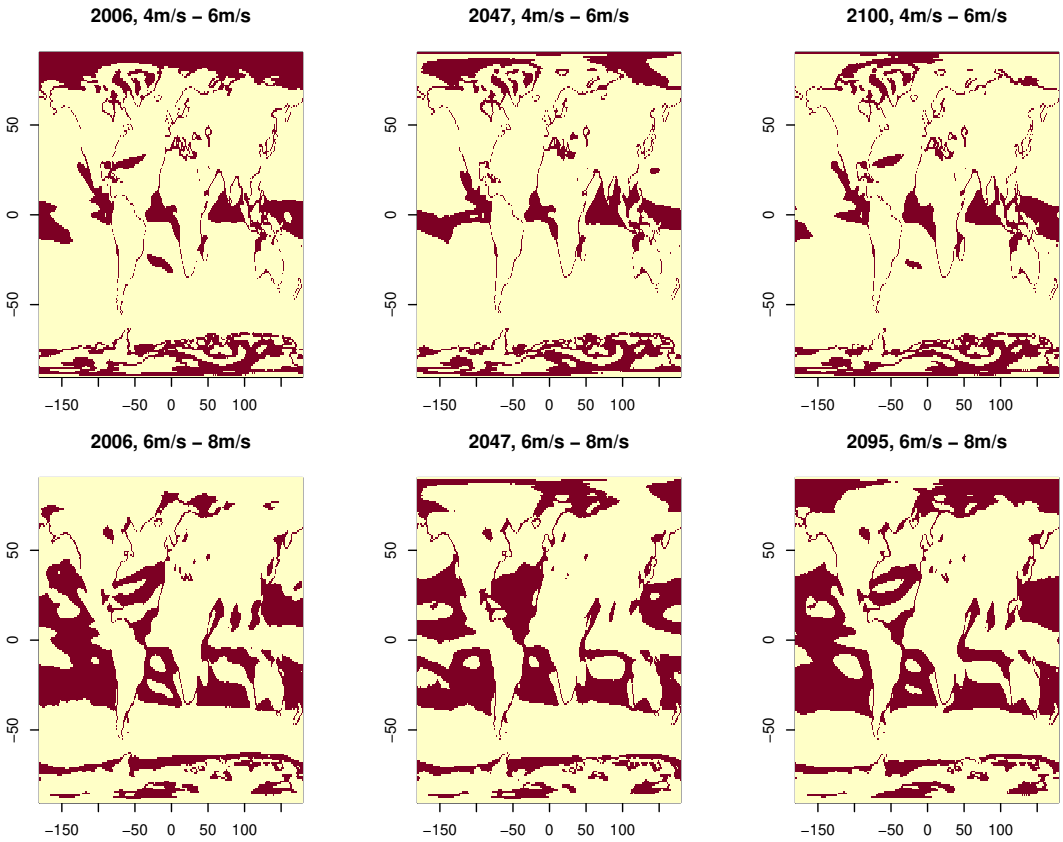

Figure 2.8: Changing wind intensities in the Arctic region. Site are shown in red where wind intensities lie between $4m/s$ and 6m/s (top row) and $6m/s$ and $8m/s$ (bottom row).

functional data, where the residual wind intensity is a function of latitude. We order the vectors using a notion called band depth, which allows ordering from the centre outwards, along with functional quartiles and the centrality, or 'outlyingess', of an observation. Functional boxplots for $t = 2006$, $t = 2047$ and $t = 2100$ can be seen in Figure 2.10.

2.5 Additional data processing

Our full data set consists of 95 years and 30 realisations, each containing data on 192 latitudes and 288 longitudes. A correlation matrix for a single year-realisation subset would be of size 55296×55296. Calculation and manipulation of such a large matrix are highly computationally intensive, and methods to reduce the size will be required to improve computation times. For this thesis, we use a spatial sampling approach (Bivand et al., 2013) selecting every nth latitude and every nth longitude to obtain a reduced data set. A value of $n = 3$ provides adequate computational speed-up for our requirements.

Further, due to the proximity of sites at the poles and statistically anomalous behaviour

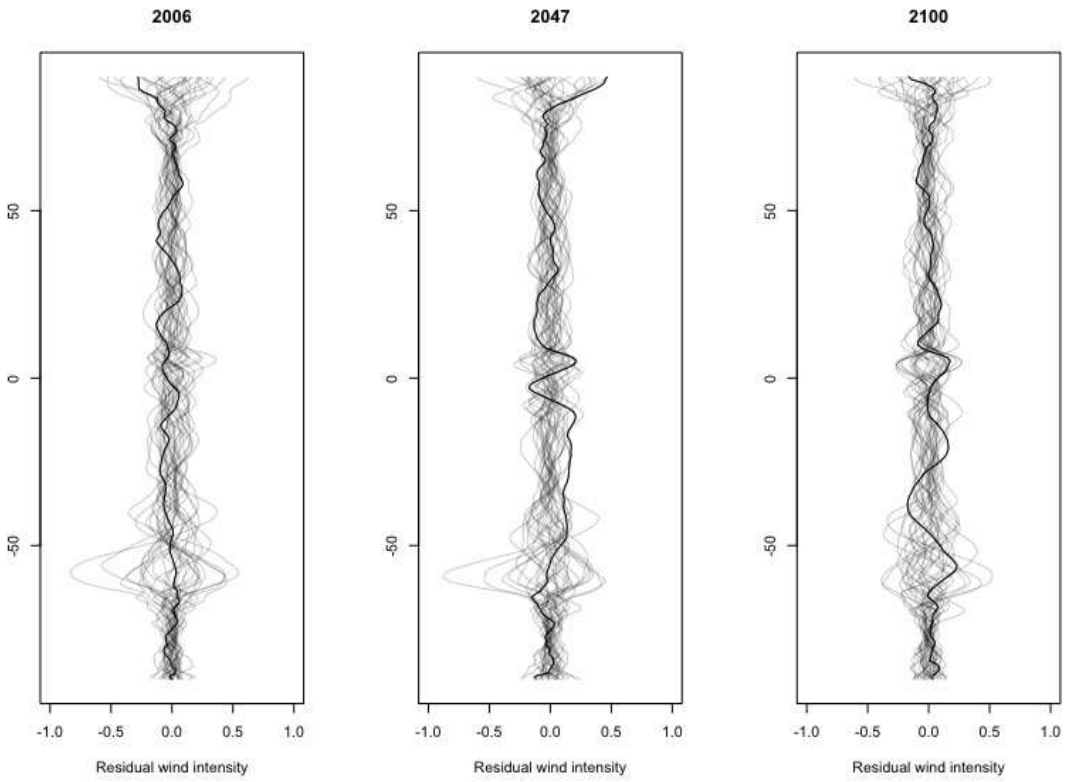

Figure 2.9: Latitude band cloud plots for residual wind intensities, years $t = 2006$, $t = 2047$ and $t = 2100$. Realisation $r = 6$ is shown with the darker black line and all other realisations with the paler lines.

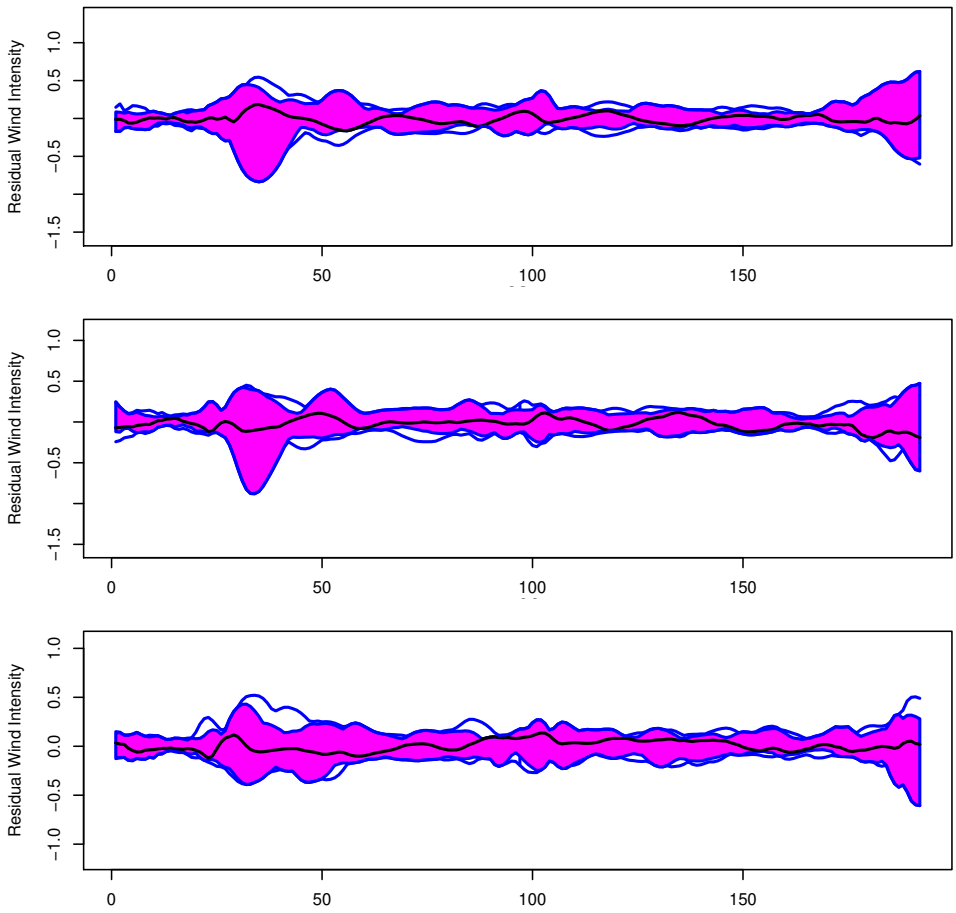

Figure 2.10: Functional box plots of all 30 realisations over three years, $t = 2006$ (top), $t = 2047$ (centre) and $t = 2100$ (bottom).

in the Antarctic region, we also remove latitudes north of 82° and south of −62°, as is common in work with this data (Castruccio and Genton, 2014, 2016). Castruccio and Genton (2014) cite the anomalous statistical behaviour in the Antarctic region as a reason for this restriction. We show the resulting data points on the sphere in Figure 2.11. Our final selection results in a 4896×4896 covariance matrix, for which many models are relatively fast to fit.

Throughout the following work, we will specify whether we are using the full or reduced data.

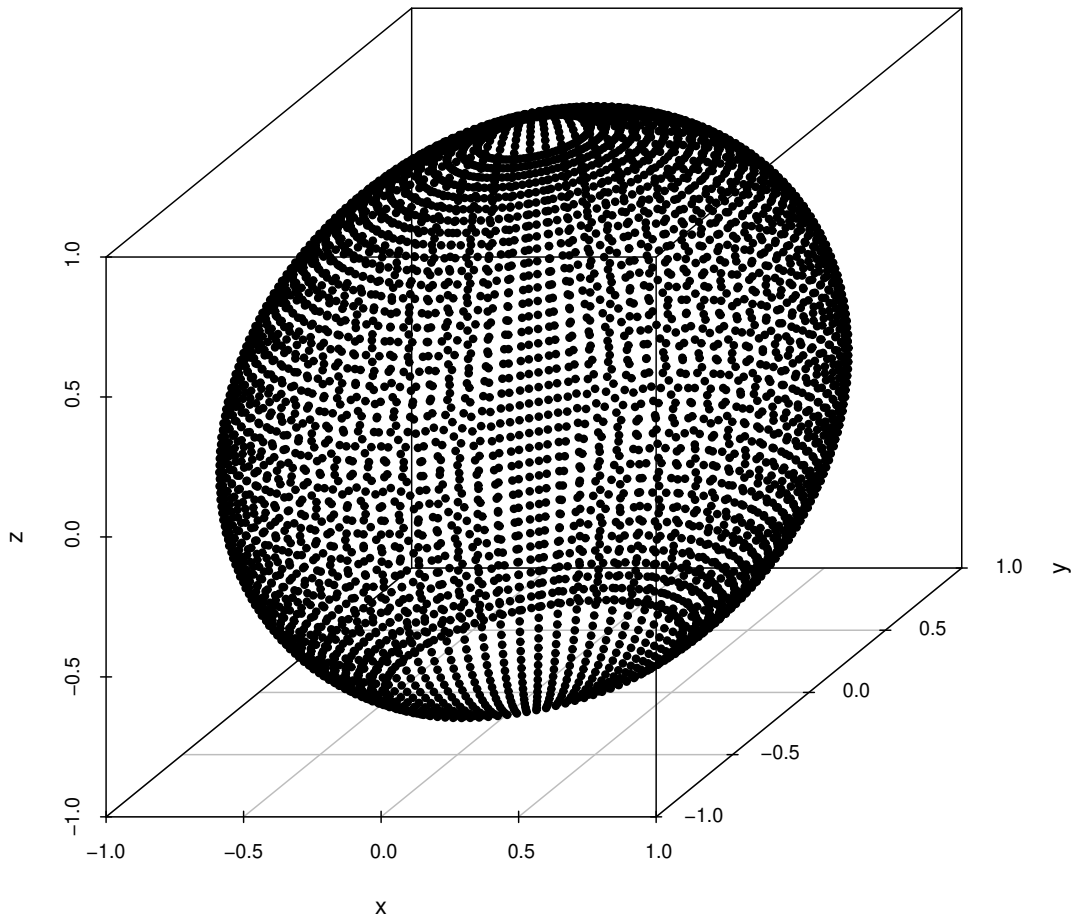

Figure 2.11: Reduced data locations (reduction by factor $n = 3$ and removal of polar latitudes).

Chapter 3

Topological data analysis on random fields

We are interested in topological data analysis as a tool for assessing model fit and verifying assumptions of Gaussianity in spatial data, particularly when applied to global wind intensities data. We believe that using statistical topology as a method for analysing model residuals may be more informative than standard, conventional methods alone and may allow us to discover otherwise unidentifiable differences between datasets. In this chapter, we look at one particular topological metric from persistent homology, connected components (the emergence of which can be equated to local maxima in 2-dimensional data, as will be further explained later) and investigate how and why their number differs between data sets with the same marginal distribution. We consider the impact of the covariance structure of the data on the results found.

We begin in Section 3.1 with a general discussion of topological data analysis and persistent homology. Section 3.2 introduces the technical details of statistical topology with some basic concepts. In Section 3.3, we look at theoretical results when we apply persistent homology to Gaussian random fields under a range of covariance structures and compare with empirical results from several simulations. Section 3.4 looks at the difference in results between distributions and examines the processes on a smaller scale. Finally, we look at how the differences emerge and the impact of different covariance structures in Section 3.5.

3.1 Topological data analysis: existing work

We use ideas from topological data analysis (TDA) for the analysis of our wind intensities data due to its flexibility and scope in assessing features in the data set not necessarily

apparent using conventional methods. Previous research into the applications of persistent homology has demonstrated the potential of applying these methods to our data. Henderson et al. (2020) used persistent homology to show differences between random fields that are otherwise assumed to be Gaussian and are indistinguishable using conventional marginal and correlation analyses. The findings provide evidence against the standard Gaussian model as well as showing topological differences at a local scale between different regions in the interstellar medium. In the context of global wind intensities, it is evident that these methods are of interest due to the assumed iid (independent and identically distributed) nature of the individual realisations (Castruccio and Stein, 2013).

Topological data analysis, sitting between data analysis, algebraic topology, computational geometry, computer science, statistics and other fields, is becoming more widely used as the data available in many applications grows considerably, both in volume and complexity. Where computer science and machine learning often lean heavily on clustering techniques (Goldenberg et al., 2010; Gan et al., 2007; Schaeffer, 2007), TDA, and more specifically persistent homology, allows a similar analysis with greater robustness to perturbations in data, for example. Many fields of research have already benefited from the flexibility of TDA. Some of these include but are certainly not limited to, analysis of sensor networks (de Silva and Ghrist, 2007), the study of proteins (Xia and Wei, 2014; Kovacev-Nikolic et al., 2016; Gameiro et al., 2015), robotics (Bhattacharya et al., 2015; Pokorny et al., 2016; Vasudevan et al., 2013), oncology (DeWoskin et al., 2010; Nicolau et al., 2011; Crawford et al., 2016; Singh et al., 2014), finance (Leibon et al., 2008; Gidea, 2017), contagion mapping (Lo and Park, 2018; Taylor et al., 2015) and neuroscience (Chung et al., 2009; Curto, 2016; Kanari et al., 2016; Lord et al., 2016; Dotko et al., 2017; Bendich et al., 2016; Yoo et al., 2016; Dabaghian et al., 2014). Some more novel applications of TDA include analysis of tennis players (Visser, 2018), music (Sethares and Budney, 2014) and Brexit (Stolz et al., 2016).

In the following survey of the field, we first take a broad view of statistical topology, summarising some of the key concepts and typical applications before focussing in on the area of persistent homology. Otter et al. (2017) provide a comprehensive roadmap of persistent homology which serves as an excellent guide to the field, as well as including an extensive list of applications of persistent homology. Here, we take particular interest in areas of research relating to filtrations and Betti numbers.

3.1.1 Motivations

Statistical topology, or topological data analysis as it is more commonly known, is a relatively new area in the much broader field of data analysis. The diverse nature of the applications to which it can be applied is partly responsible for the rapid growth in

the field. In the last few years data science has been applied to many different areas of research, but in some cases has been unable to keep up with the enormous growth in complexity and volume of data it attempts to deal with (Munch, 2017). TDA has the potential to fill these gaps through approaching the data differently, looking at shape and other qualitative features rather than focusing on its specific quantitative nature. In some applications, problems that can be solved by other methods are solvable more quickly, or with less restrictive assumptions using TDA methods. For example, in work by de Silva and Ghrist (2007), coverage of non-localised sensor networks is calculated using ideas from persistent homology. Existing methods to solve the problem used probabilistic methods and ideas from computational geometry, requiring assumptions about uniformity in the distribution of network nodes. The application of a TDA approach removes the requirement for this restrictive assumption, allowing the application to a broader class of similar problems. Kovacev-Nikolic et al. (2016) use persistence landscapes to enable the detection of conformational changes in forms of the maltose-binding protein. Persistence landscapes, one tool for interpretation and visualisation of results in persistent homology, are of particular use as it is simple to calculate statistical summaries upon them, such as means, variances and test statistics.

3.1.2 Algebraic and statistical topology

Algebraic topology, a somewhat niche field in its own right (described by Adler et al. (2010) as 'esoteric'), has historically been one in which research has been inward-looking, with less view to applications and more toward theoretical development. However, many researchers are now working on applications of the theory, from which topological data analysis has emerged. Traditional algebraic topology deals with the study of topological spaces through the application of concepts from group theory to topological spaces, such as homology, homotopy and cohomology. As we are interested in persistent homology, we bypass homotopy and cohomology here.

A topological space is defined as a set $\mathcal{X}$ with a class $\mathcal{T}$ of subsets of $\mathcal{X}$ upon which the following conditions are satisfied (Givant and Halmos, 2009):

- Both the empty set $\emptyset$ and $\mathcal{X}$ are in $\mathcal{T}$
- Any intersection of a finite number of members of $\mathcal{T}$ is also in $\mathcal{T}$
- Any union of an arbitrary family of sets in $\mathcal{T}$ is also in $\mathcal{T}$

In homology, the interest lies in the study of qualitative features in data, such as Betti numbers. For a topological space with N dimensions, there are $N + 1$ homology groups,

$H_i(\mathcal{X})$, where the rank β_k of $H_k(\mathcal{X})$ is the k-th Betti number (Adler et al., 2010). These algebraic features have 'homotopy invariance' (Otter et al., 2017), that is to say, they cannot be changed by deformations such as stretching and bending. We can also define the Euler characteristic as

$$\chi(\mathcal{X}) = \sum_{k=0}^{N} (-1)^k \beta_k$$

which is homotopy invariant.

The first Betti number β_0 counts the connected components, which form the zero-dimensional homology group, H_0. The homology group H_1 comprises holes enclosed by one-dimensional cycles and has rank β_1, and β_2 counts the number of three-dimensional voids (Edelsbrunner and Harer, 2008) enclosed by two-dimensional boundaries, forming homology group H_2.

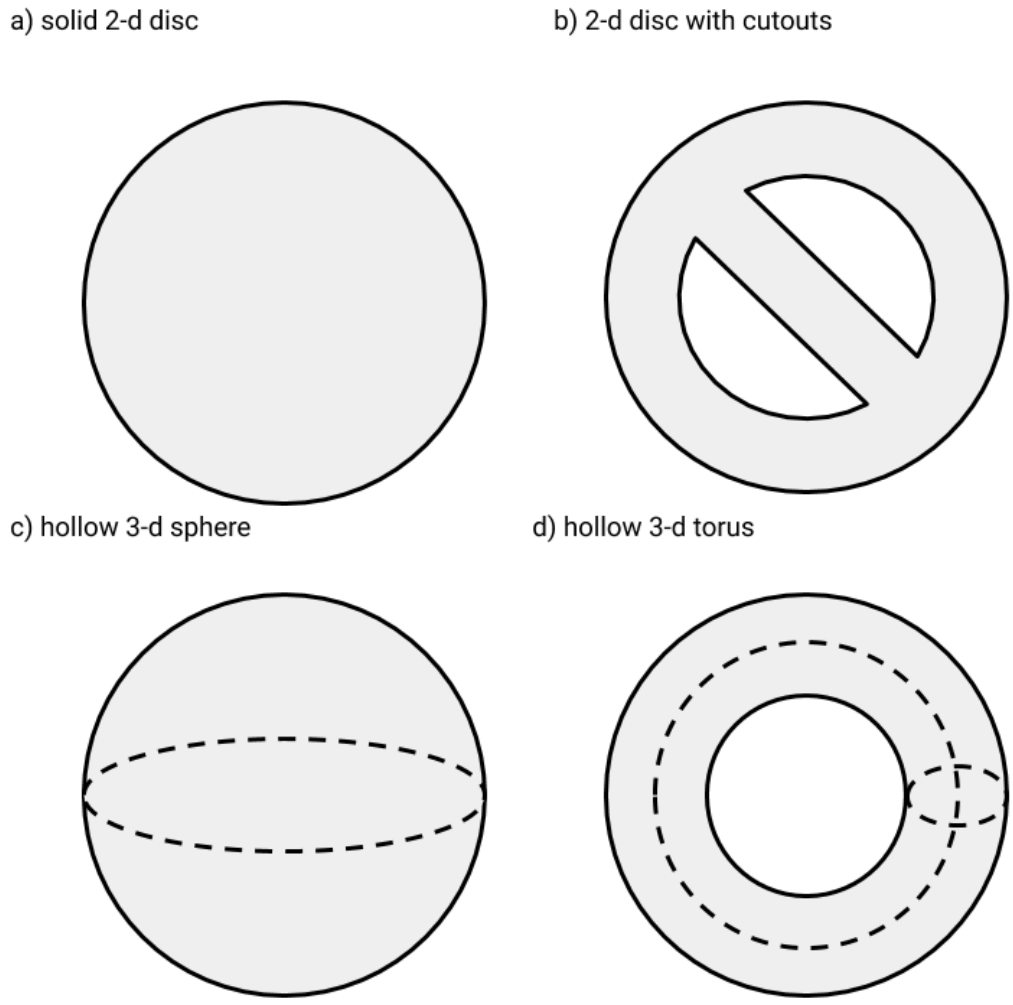

Figure 3.1: Examples of shapes with different topological features.

As a visual example, consider each of the shapes in Figure 3.1. We can see that each has the Betti numbers described in Table 3.1.

	β_0	β_1	β_2
a) solid 2-d disc	1	0	0
b) 2-d disc with cutouts	1	2	0
c) hollow 3-d sphere	1	0	1
d) hollow 3-d torus	1	2	1

Table 3.1: Betti numbers for shapes shown in Figure 3.1.

3.1.3 Persistent homology

Edelsbrunner et al. (2002) introduced the idea of persistence for Betti numbers, providing algorithms for computation and simplification of persistent homology. Persistent homology is nicely described by de Silva and Ghrist (2007) as 'homology classes which persist as one changes a parameter in the system'. For example, consider a point-cloud data set X in some Euclidean space. We can look at the point cloud at a selection of different resolutions, analysing the resulting shapes at each. For a set of points in $\mathbb{R}^2$, we define some distance parameter $e > 0$, and for varying values of e construct the following space, S_e:

1. An edge where the separation between two points is less than e
2. A triangle where all edges are in S_e

It is clear that for $e \leq e'$, S_e is contained in S'_e. S_e is then a 'simplicial complex' for the space X.

A simplicial complex K with a finite sequence of nested subcomplexes is called a filtered simplicial complex. Elements within each simplicial complex, or subcomplex, are described as faces (Guillemard and Iske, 2017) with faces of dimension zero and one corresponding to vertices and edges respectively. Homology can be applied to each of the subcomplexes, and one can examine the persistence of elements in K. An undirected network can be easily formulated as a 1-d simplicial complex, or a filtered 1-d simplicial complex if the network is weighted. This complex is constructed by filtering on increasing or decreasing weight. For example, Schauf et al. (2016) look at whether it is possible to obtain additional information about economies by looking at their shape or structure, as opposed to only considering their size. As in many applications, the problem is posed in the form of a weighted network, to which topological data analysis methods can be applied. A key concept in persistent homology is that there is some ordering of simplices within a simplicial complex. This

ordering, called a filtration, may come from a changing distance parameter in a point cloud data set, or a changing weight in a network (Edelsbrunner et al., 2002). We discuss these ideas in further detail in Section 3.2.

3.1.4 Computation of persistent homology

In addition to definitions of these homological features, we require the development of computational methods to calculate them in practice. Several open-source packages such as `TDAstats` in `R` have been developed to compute and provide visualisation of TDA methods (R. Wadhwa et al., 2018). Efficient computational methods are vital due to the often expensive nature of computing persistent homology (Patania et al., 2017), but emerging methods lack good documentation at present. The potential of TDA for data analysis is lessened by the temporal and spatial complexity, and hence expense, required in its computation. Various methods to reduce this computational complexity have been proposed, including cluster-based reduction (Moitra et al., 2018) in which the number of data points is reduced while minimising the impact on the important features in the data. This reduction in complexity is particularly valuable in dealing with computational challenges as in the majority of cases, the number of data points impacts the computation time exponentially. Computation of homology is often approximated using combinatorial structures known as simplicial complexes, to which computational algorithms can be applied.

3.1.5 Statistical interpretation and visualisation

Beyond theoretical definitions and computational challenges, we require some approaches for assessing the robustness and statistical interpretation of results. In persistent homology, commonly used statistical summaries include barcodes, persistence diagrams and persistence landscapes (examples can be seen in Section 3.2). A barcode is a selection of horizontal lines graphically representing the persistent homology of a filtered simplicial complex. Each line segment represents the interval (p, q) corresponding to the persistence of each element (Fugacci et al., 2016). An interesting question that frequently arises in the use of barcodes (and persistent homology more generally) is when we should consider elements as features and when as noise. Commonly, elements with short lifetimes are considered to be noise with longer lifetimes indicating signal. Fasy et al. (2014) developed confidence sets allowing the distinction of signal and noise. Persistence diagrams are a useful descriptor of topological approaches to data, due to their invariance to translation and rotations, as well as robustness to noisy data. However, a key limitation of topological data analysis is the difficulty in using common summaries such as barcodes and persistence diagrams with conventional statistics and machine learning methods. As a possible

solution to this problem, Bubenik (2015) proposed the persistence landscape, a topological summary that can be used alongside methods from machine learning and statistics and provided methods for their calculation (Bubenik and Dlotko, 2017). The persistence landscape can be thought of as something of a cross between the persistence diagram and the barcode. The primary convenience of the persistence landscape is that it can be viewed as a random variable with values in a function space. Further, it obeys the law of large numbers and has a central limit theorem. Kovacev-Nikolic et al. (2016) also argued that is is easier to 'do statistics' in the space of persistence landscapes as it is a vector space. Examples of barcodes and persistence landscapes can be seen in the following section.

Obayashi et al. (2018) used methods by which statistical features embedded in persistence diagrams can be extracted via machine learning, in cases where many data sets are available. They proposed a method for the unification of this and the inverse process, arguing that the inverse process is frequently more interesting and useful. That is, the question of what can be learned about the original data space using such statistical features obtained from persistence diagrams.

3.2 Application to random fields

As we wish here to utilise ideas from persistent homology for the analysis of random fields, we define some of the basic technical details in persistent homology as applied to such fields. Unlike many uses of persistent homology applied to point clouds in different spatial resolutions, we are interested in persistent homology on gridded data, in which we look for persistence across various 'temporal' filtrations of the data set. This strategy is in contrast to the standard approach of filtration by resolution. As an example, we look at a 2-d example data set of size 50×50. This data is shown in Figure 3.2.

Level sets and filtrations Let $z(x)$ be the value at site x on a 2-dimensional grid. For any real value t, the lower level set is the set of sites on the grid with values less than t, $\mathcal{F}_t = \{x : z(x) \leq t\}$. Increasing t from below defines a filtration of the grid, allowing us to describe changing topological features. A selection of level sets for our example data are shown in Figure 3.3. The top-left plot shows the level set at $t = 3$, which includes all sites with value less than or equal to t. We can see twelve unconnected regions in red, each of which can be described in topological terms as a connected component. These are described in the following sections. The bottom right plot shows the level set at $t = 9$, where we see only a single connected region.

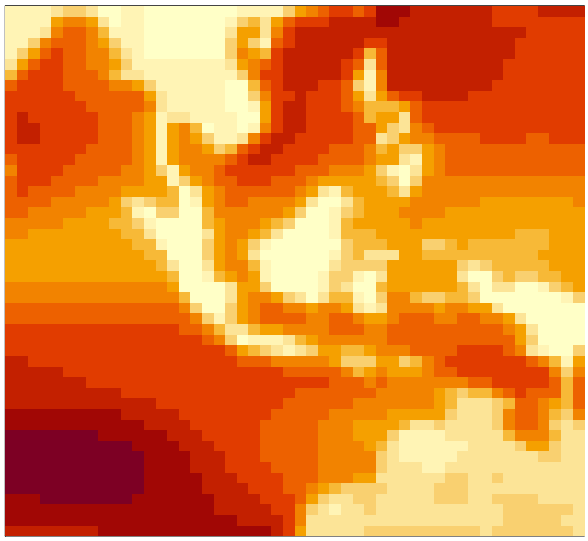

Figure 3.2: Example data set. In fact, this is a small section of our wind intensities data for year $t = 2006$ and realisation $r = 1$. The lighter regions show lower values and the darker regions show higher vales. Here we can see lower wind intensities over land for parts of South-East Asia.

Homology groups and Betti numbers In the 2-dimensional case, we can think about components and holes, corresponding to β_0 and β_1 respectively as described in Section 3.1.2. In the work that follows, we are primarily interested in components, a set of connected sites which are present in the level set, where we define neighbouring sites as being connected if they share a common edge, and define non-neighbouring sites as connected if there exists a path of connected sites between them. Counting the number of components in a level set provides the associated Betti number of order zero. As t increases, we describe the emergence of a new component as a birth. We describe the merger of two components as the continuation of the first to be born and the death of the other.

Euler characteristic The Euler characteristic

$$\chi(\mathcal{X}) = \sum_{k=0}^{N} (-1)^k \beta_k$$

is a topological invariant which describes the structure of a topological space, $\mathcal{X}$. In the 2-d case we simply have

$$\chi(\mathcal{X}) = \beta_0 - \beta_1.$$

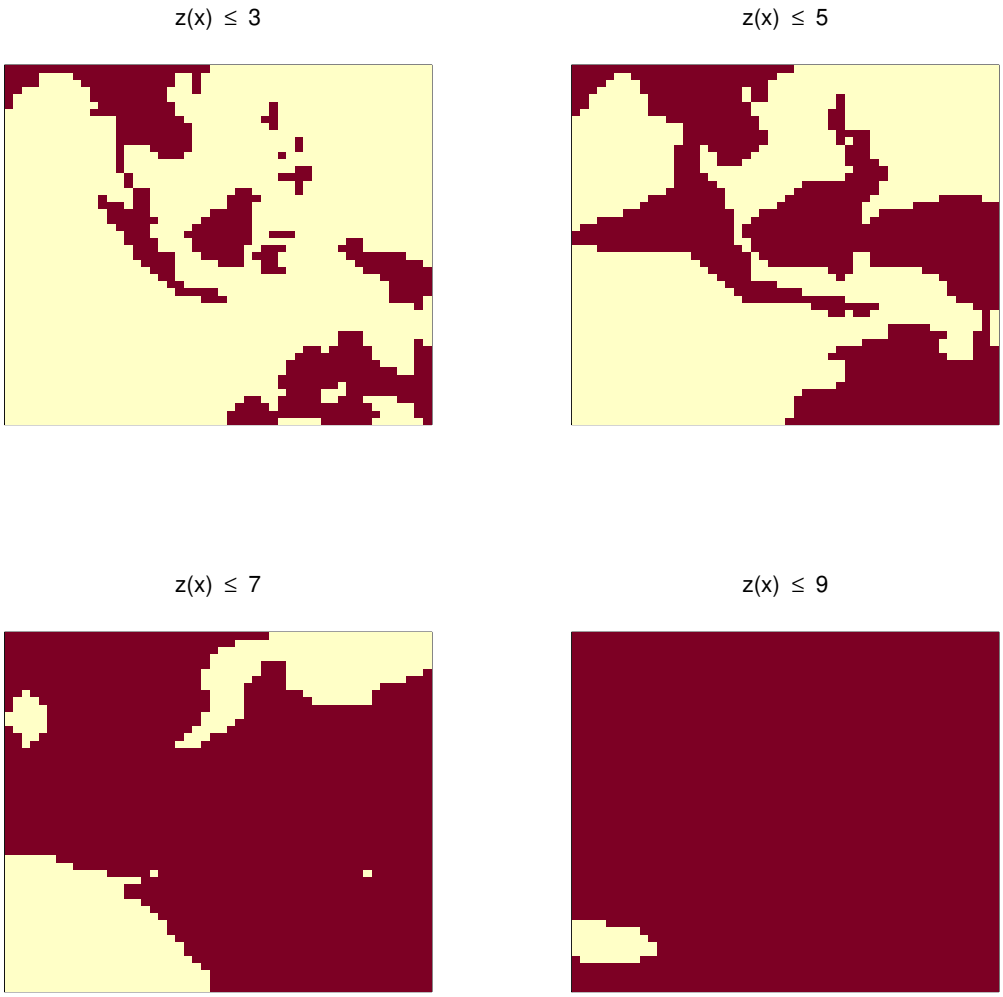

Figure 3.3: Four level sets for the example data set. Sites contained in the level set are shown in red.

Persistent homology Following the topology of level sets as a function of t is an example of persistent homology. As the level changes with t, there is no change in homology until a level is reached where there is a critical point of the field.

Barcodes A convenient way to visualise persistent homology is through barcodes, as can be seen in Figure 3.4 for our example data. Here, we also show lifetime of holes in red, for comparison. A barcode diagram for the level sets of an N-dimensional space, $\mathcal{X}$ is a collection of N graphs, each corresponding to a homology group. Here we have two, corresponding to connected components and holes. Generally, a bar in the kth graph shows the levels where an element of that group is born and dies. For the example data, we see that the connected components tend have a longer lifetime as well as an earlier birth time. Further, there are far fewer holes than connected components.

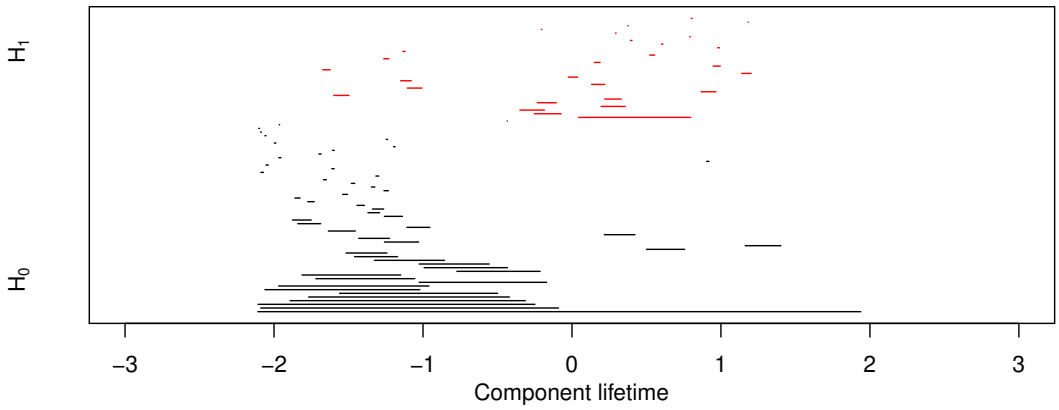

Figure 3.4: The barcode represents the persistent homology. Each black line represents a connected component (an element of homology group H_0), and each red line a hole (an element of homology group H_1), with the length and location of the line representing the lifetime of the element.

Persistence diagrams and landscapes A persistence diagram is a scatter plot of birth times against death times, for a particular homology group, as shown in Figure 3.5. Persistence diagrams allow us to see some of the structure in the data, depending on the distribution of points plotted. Points clustered along the diagonal indicate a field that is mostly noise, whereas points further from the diagonal suggest other, potentially more interesting features. In some contexts, it can be the noise that is of primary interest. For our data, the diagrams for components and holes are noticeably different, with the diagram for components containing many more points, with earlier birth times. This corresponds

directly to the barcode seen previously. The points in the persistence diagram for holes are all very close to the diagonal, indicating a short lifetime.

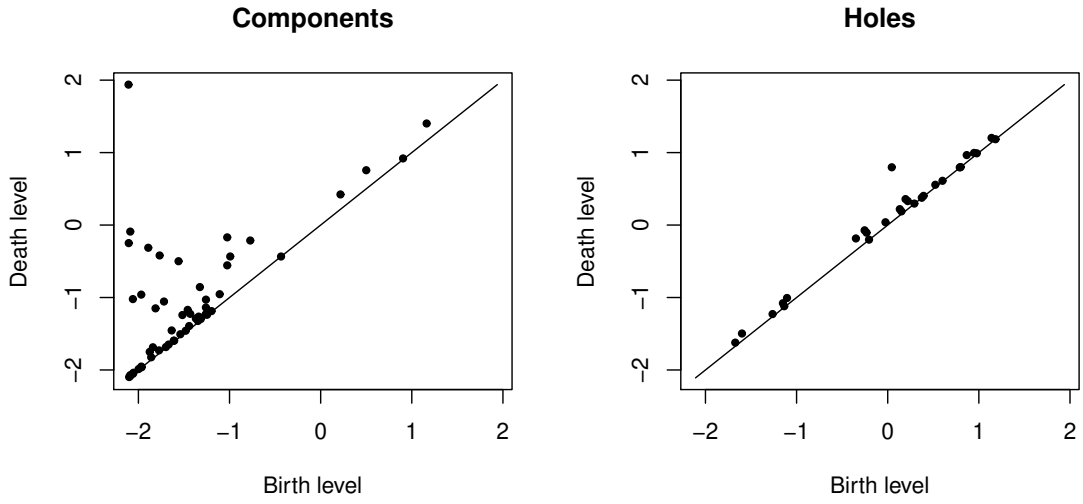

Figure 3.5: A persistence diagram for components and holes for our example data. In this example we have a relatively small number of components.

Finally, in Figure 3.6 we can see persistence landscapes for the same data. In practice, the persistence landscape can be understood as a rotation of the diagonal of persistence diagram to the x-axis. A persistence landscape is a mapping of a persistence diagram to a function space, on which a statistical and machine learning tools can be used (Bubenik, 2020). It is evident that the persistence landscape for components is dominated by a single connected component with a long lifespan, as can be seen in the persistence diagram in the top left corner. We can identify this component as the long bar in the barcode above.

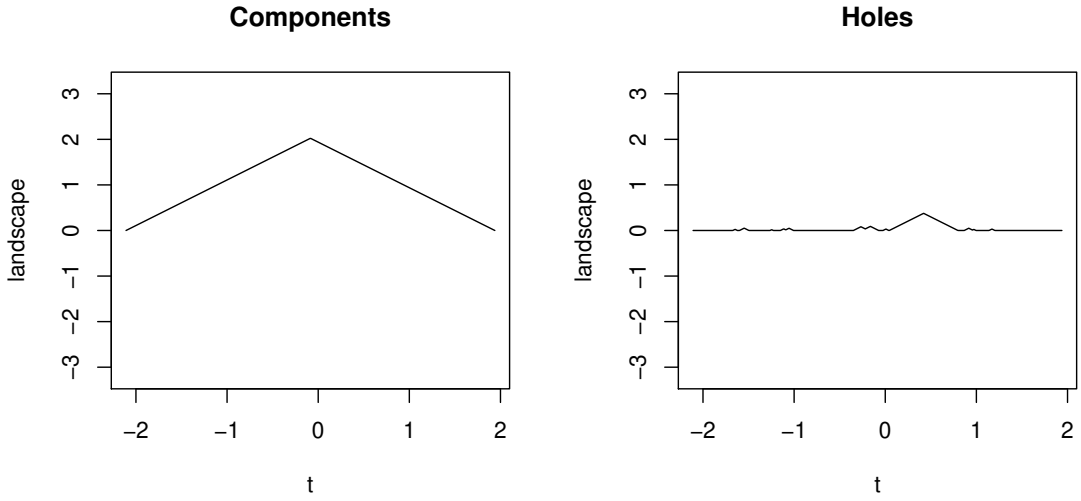

Figure 3.6: A persistence landscape for our example data.

3.2.1 Interpreting and summarising persistence diagrams

Interpretation of persistence diagrams can be challenging, so Henderson et al. (2020) propose the concept of peeling successive convex hulls until a proportion of points remain. This method allows us to see the general shape and properties of the persistence diagram without too much influence from points near the boundary and outliers. Additionally, the authors provide the following summaries for the convex hull:

1. Centroid coordinates, (C_b, C_d), the mean position of all points in the x and y directions
2. Perimeter, P
3. Area, A
4. Filamentarity $= \frac{P^2-4\pi A}{P^2+4\pi A}$

Application of these summaries is demonstrated in Chapter 4.

3.3 Local maxima on a 1-d Gaussian random field

We use some of these concepts from statistical topology to compare 1-dimensional random fields based on their topological properties, in particular, the number of connected components. Here, sites at which connected components are born are equivalent to local minima

when filtering from below as we have so far in this chapter. Alternatively, we could filter from above, taking level sets containing sites where $z(x) \geq t$. Filtering from above in this way, sites at which connected components are born would correspond instead to local maxima. We examine the differences between the numbers of these local maxima from a true Gaussian field and a marginally Gaussian field. In order to to understand the differences between true Gaussian data and data transformed to marginal Gaussianity, we first consider what we would expect to occur in the true Gaussian case. That is, we establish the expected number and variance of local maxima for a given correlation function.

3.3.1 Expected number of local maxima on a 1-d Gaussian field

To begin, we define some notation and calculate the expected number of local maxima over a discrete Gaussian random field (i.e. a 1-d grid), Z. We assume Z is marginally $N(0,1)$ with stationary and isotropic correlation. Let N_1 be the number of local maxima over a 1-d Gaussian field of length m, $z(x_i)$, where $i = 1, 2, 3, \ldots, m$. Here, each local maxima corresponds to a point in a persistence diagram. Let $z_1 = z(x_1)$ be the value at site x_1 and let $z_{(1)} = z(x_{(1)})$ be the 2-dimensional vector of values for the neighbours $x_{(1)}$ of x_1, where the neighbours of a location are the locations immediately adjacent. (For boundary points, there is only one neighbour, and results follow similarly to the non-boundary case discussed here.) Then, $(z_1, z_{(1)})$ is Gaussian with zero mean and variance matrix

$$\begin{pmatrix} 1 & r^T \\ r & R \end{pmatrix}$$

where r is the 2-vector $\mathrm{Cov}(z_1, z_{(1)})$ and $R = \mathrm{Cov}(z_{(1)}, z_{(1)})$. We have a local maximum at x_1 if $z_{(1)} < z_1 1_2$, where 1_2 is a 2-vector of ones, i.e. the value at a site is greater than that of its two neighbours. In order to calculate the probability of this occurring, we require the conditional distribution of $z_{(1)}$ given z_1 in addition to the marginal distribution of z_1. Let $\phi^{(p)}(x; \mu, \Sigma)$ be the Gaussian probability density in p dimensions with mean μ and covariance matrix Σ and let $\Phi^{(p)}(x; \mu, \Sigma)$ be the corresponding cumulative distribution function. Then z_1 has density $\phi^{(1)}(z_1; 0, 1)$, $z_{(1)}$ has density $\phi^{(2)}(z_{(1)}; 0_2, R)$, where 0_2 is the vector $(0,0)^T$ and the conditional density for $z_{(1)}$ given z_1 is

$$\phi^{(2)}(z_{(1)}; rz_1, R - rr^T).$$

Then

$$\begin{aligned} p_1 &= p(\text{local maximum at } x_1) \\ &= p(z_{(1)} < z_1 1_2), \end{aligned}$$

$$= \int_{z_1} \left\{ \int_{z_{(1)} < z_1 1_2} \phi^{(2)}(z_{(1)}; rz_1, R - rr^T) dz_2 \right\} \phi^{(1)}(z_1; 0, 1) dz_1$$

$$= \int_{z_1} \Phi^{(2)}(z_1 1_2; rz_1, R - rr^T) \phi^{(1)}(z_1; 0, 1) dz_1$$

$$= \int_{z_1} \Phi^{(2)}((1_2 - r)z_1; 0_2, R - rr^T) \phi^{(1)}(z_1; 0, 1) dz_1.$$

A general skew-Normal distribution can be used to show that (Henderson et al., 2020; Barrett et al., 2015; Arnold, 2009)

$$\Phi^{(q)}(D\mu; \nu, \Delta + \Delta\Sigma\Delta^T) = \int_y \Phi^{(q)}(Dy; \nu, \Delta) \phi^{(p)}(y; \mu, \Sigma) dy, \tag{3.1}$$

where μ is dimension p, ν is dimension q, Σ and Δ are $p \times p$ and $q \times q$ covariance matrices respectively, and D is an arbitrary $q \times p$ matrix. If we set $p = 1$, $q = 2$, $\mu = 0$, $\Sigma = 1$, $\nu = 0_2$, $\Delta = R - rr^T$, $y = z_1$ and $D = 1_2 - r$, then

$$p_1 = \Phi^{(2)}(0_2; 0_2, R - rr^T + (1_2 - r)(1_2 - r)^T)$$

$$= \Phi^{(2)}(0_2; 0_2, R + 1_2 1_2^T - 1_2 r^T - r 1_2^T).$$

In the case where the point x_1 is an endpoint, $x_{(1)}$ and hence $z_{(1)}$ are single values corresponding to the single neighbour of x_1. That is,

$$p_1 = \Phi^{(1)}(0; 0, R + 1 - r^T - r).$$

Summing over all locations we can obtain the expected number of local maxima,

$$E[N_1] = \sum_j p_j.$$

3.3.2 Variance of number of local maxima on a 1-d Gaussian field

Using a similar approach to that which we used for expectation above, we can estimate the variance of the number of local maxima (Henderson et al., 2020). We let I_j be an indicator of a local maximum at a point x_j. To determine the variance of $N_1 = \sum I_j$ we need $E[I_i I_j]$ for all pairs of locations x_i and x_j. In this section, as previously, we consider pairs of points of which neither are endpoints. In the situation where one or both are endpoints, the theory is similar, although endpoints have only one neighbour.

Let $z_{1,2} = (z(x_1), z(x_2))^T$ be the bivariate value at two locations x_1 and x_2. Let $z_{(1)} = z(x_{(1)})$ be the 2-vector of values over the neighbours $x_{(1)}$ of x_1, and let $z_{(2)} = z(x_{(2)})$

be the 2-vector of values over the two neighbours $x_{(2)}$ of x_2. Then $z_{(1,2)}$ is the 4-vector, $(z_{(1)}, z_{(2)})^T$.

There are three cases to consider. First, x_1 and x_2 have no common neighbours; second, where they have a shared neighbour and third, where they are neighbours of each other.

In the first case, x_1 and x_2 have distinct neighbours, and $(z_{1,2}, z_{(1,2)})$ is Gaussian with zero mean and covariance matrix

$$\begin{pmatrix} R_{11} & R_{12}^T \\ R_{12} & R_{22} \end{pmatrix}$$

where $R_{11} = \mathrm{Cov}(z_{1,2}, z_{1,2})$, $R_{12} = \mathrm{Cov}(z_{1,2}, z_{(1,2)})$ and $R_{22} = \mathrm{Cov}(z_{(1,2)}, z_{(1,2)})$.

We have $I_1 I_2 = 1$ if and only if we have local maxima at both sites, x_1 and x_2. That is, if $z_{(1)} < z_1 1_2$, and $z_{(2)} < z_2 1_2$.

We can see that $z_{1,2} = (z(x_1), z(x_2))$ has density

$$\phi^{(2)}(z_{1,2}; 0_2, R_{11}),$$

and $z_{(1,2)} = (z(x_{(1)}), z(x_{(2)}))$ has density

$$\phi^{(4)}(z_{(1,2)}; 0_4, R_{22}).$$

So the conditional density for $z_{(1,2)}$ given $z_{1,2}$ is

$$\phi^{(4)}(z_{(1,2)}; R_{12}R_{11}^{-1}z_{1,2}, R_{22} - R_{12}R_{11}^{-1}R_{12}^T).$$

Then

$$\begin{aligned} p_{12} &= p(\text{local maxima at both } x_1 \text{ and } x_2) \\ &= E[I_1 I_2]. \end{aligned}$$

Define the matrix

$$J = \begin{pmatrix} 1_2 & 0_2 \\ 0_2 & 1_2 \end{pmatrix}$$

where 0_2 and 1_2 are 2-vectors of 0s and 1s respectively. Then we have

$$p_{12} = p(z_{(1,2)} < J z_{1,2})$$

$$= \int_{z_{1,2}} \left\{ \int_{z_{(1,2)} < J z_{1,2}} \phi^{(4)}(z_{(1,2)}; R_{12}R_{11}^{-1}z_{1,2}, R_{22} - R_{12}R_{11}^{-1}R_{12}^T) dz_{(1,2)} \right\} \phi^{(2)}(z_{1,2}; 0_2, R_{11}) dz_{1,2}$$

$$= \int_{z_{1,2}} \Phi^{(4)}(Jz_{1,2}; R_{12}R_{11}^{-1}z_{1,2}, R_{22} - R_{12}R_{11}^{-1}R_{12}^T) \phi^{(2)}(z_{1,2}; 0_2, R_{11}) dz_{1,2}$$

$$= \int_{z_{1,2}} \Phi^{(4)}((J - R_{12}R_{11}^{-1})z_{1,2}; 0_4, R_{22} - R_{12}R_{11}^{-1}R_{12}^T) \phi^{(2)}(z_{1,2}; 0_2, R_{11}) dz_{1,2}.$$

By equation (3.1), if we set $p = 2$, $q = 4$, $\mu = 0$, $\Sigma = R_{11}$, $\nu = 0_2$, $\Delta = R_{22} - R_{12}R_{11}^{-1}R_{12}^T$, $y = z_{1,2}$ and $D = J - R_{12}R_{11}^{-1}$, then

$$p_{12} = \Phi^{(4)}(0_4; 0_4, R_{22} - R_{12}R_{11}^{-1}R_{12}^T + DR_{11}D^T).$$

Then for distinct, non-adjacent sites x_1 and x_2,

$$E[I_1 I_2] = \Phi^{(4)}(0_4; 0_4, R_{22} - R_{12}R_{11}^{-1}R_{12}^T + DR_{11}D^T),$$

where D is the 4×2 matrix described previously. Again, the results for edge points follow in the same way, but with only one neighbour for each edge point instead of two. In the second case, where x_1 and x_2 have a shared neighbour, the above result still holds. In the final case where x_1 and x_2 are neighbours of each other, there cannot be a local maximum at both points, and so $E[I_1 I_2] = 0$ and $\text{Cov}(I_1 I_2) = -E[I_1]E[I_2]$. This allows us to calculate the variance of the number of local maxima, N_1, using the following estimator,

$$\widehat{\text{Var}(N_1)} = \sum_i \left\{ E[I_i](1 - E[I_i]) + \sum_j \text{Cov}(I_i, I_j) \right\}.$$

Both the estimators for expectation and variance used above require the calculation of the distribution function of a multivariate normal distribution. This is achieved easily in `R` using the `pmvnorm` function within the `mvtnorm` package which uses a choice of algorithms, including GenzBretz (by default) and Miwa (Mi et al., 2009). In Appendix A we discuss issues found with the Miwa algorithm.

3.3.3 Simulation and estimation

Having established theoretical results for a 1-d Gaussian random field, we are in a position to compare these with empirical results. We simulate length-192 1-d Gaussian random fields, to correspond to the length-192 latitude band summaries from our original global

winds data set. We use a Matérn covariance function of the form

$$C(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{d\sqrt{2\nu}}{\eta}\right)^{\nu} K_{\nu}\left(\frac{d\sqrt{2\nu}}{\eta}\right),$$

where $K_{\nu}(.)$ is a modified Bessel function of the third kind and $\Gamma(.)$ is the Gamma function. This and other covariance functions are shown in Appendix B. We use six different values of the range parameter η and three different values of the smoothness parameter, ν. For each pair of parameters, we produce 1000 simulations in order to obtain empirical values for the mean and standard variation of the number of local maxima. We then calculate the expected mean and standard deviation using the estimators presented previously. Results are shown in Table 3.2. Running the procedure similarly with a selection of different spatial covariance structures from the `geoR` package provides results shown in Table 3.3.

η	ν	mean count	expectation	standard error	sd count	sd est
0	0.50	64.48	64.33	0.09	2.88	2.93
0	1	64.38	64.33	0.09	2.88	2.94
0	1.50	64.35	64.33	0.10	3.03	2.93
1	0.50	58.23	58.22	0.10	3.06	3.09
1	1	51.28	51.39	0.10	3.23	3.12
1	1.50	44.90	44.74	0.10	3.01	3.03
5	0.50	51.26	51.24	0.10	3.23	3.34
5	1	35.63	35.76	0.11	3.43	3.40
5	1.50	22.55	22.57	0.10	3.04	3.00
10	0.50	49.97	49.94	0.11	3.43	3.39
10	1	31.50	31.51	0.11	3.36	3.54
10	1.50	16.55	16.43	0.10	3.18	2.98
20	0.50	49.03	49.24	0.11	3.41	3.43
20	1	28.37	28.35	0.11	3.59	3.71
20	1.50	11.96	11.98	0.10	3.01	2.96
50	0.50	48.79	48.80	0.11	3.38	3.45
50	1	25.20	25.30	0.13	4.06	4.01
50	1.50	8.00	7.97	0.09	2.87	2.91

Table 3.2: Theoretical and simulated means and standard deviations, using a Matérn covariance function with different range (η) and smoothness (ν) parameters (1000 simulations per row).

It is clear that the theoretical approach and simulation results are very similar, and with a high number of simulations, we can produce results as expected for simulations of length 192.

model	η	ν	mean count	expectation	standard error	sd count	sd est
nugget	0		64.23	64.33	0.09	2.79	2.94
circular	20		48.48	48.54	0.12	3.66	3.51
cubic	20		16.63	16.74	0.09	2.79	2.77
exp	20		48.99	49.24	0.10	3.29	3.43
Matérn	20	1	28.34	28.35	0.12	3.73	3.71
spherical	20		48.69	48.58	0.11	3.43	3.49
powered exp	20	1	49.18	49.24	0.10	3.28	3.43

Table 3.3: Theoretical and simulated means and standard deviations, using a selection of covariance functions (1000 simulations per row).

3.4 Comparison with marginally Gaussian χ_1^2 fields

Having established an accurate simulation method we wish to compare the number of local maxima between Gaussian random fields and marginally Gaussian χ_1^2 random fields, referred from here on simply as χ_1^{2*} random fields. Henderson et al. showed that for the two-dimensional case, there could be significant differences between Gaussian and non-Gaussian random fields when counting the number of local maxima or minima. The authors observed these differences despite identical marginal distributions and indistinguishable correlation functions. We explore this further in the one-dimensional case by comparing Gaussian and χ_1^{2*} random fields as the latter showed the most significant difference from the Gaussian case, of a range of distributions considered.

3.4.1 Local maxima in the 1-d case

To begin, we simulate a 1-d Gaussian field containing elements z_i, with some chosen covariance structure and $N(0,1)$ marginals. We square each of the random variables to obtain z_i^2. Then for all i,

$$z_i^* = \Phi^{-1}(F_{\chi_1^2}(z_i^2))$$

is marginally $N(0,1)$, where $F_{\chi_1^2}$ is the χ_1^2 cumulative distribution function. We use asterisk notation to indicate marginal Gaussianity.

We simulated 10000 Gaussian random fields and 10000 χ_1^{2*} random fields, each of length 192. A single simulation for each of these is shown in Figure 3.7. There is no visually discernible difference between the two. For both, we used a Matérn covariance function with parameters as shown in Table 3.4. The χ_1^2 Matérn parameters are numerically estimated to maximise the likelihood of the parameters based on the covariance at various lags. This covariance matching procedure results in the covariance matrix of the two sim-

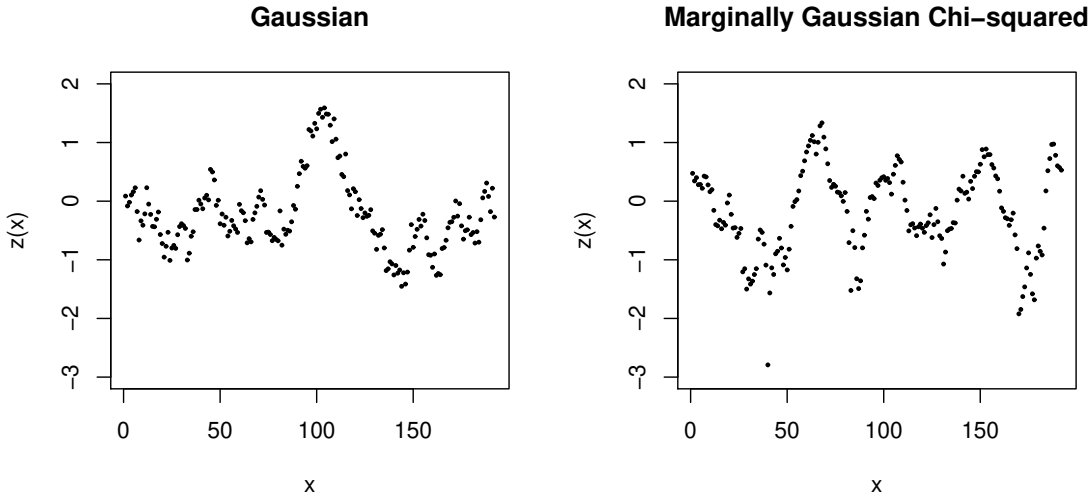

Figure 3.7: 1-d Gaussian and χ_1^{2*} random fields.

ulations being as close as possible. Table 3.4 shows the mean correlation at a selection of lags alongside the parameters used. We calculate these values using a subset of 100 fields for efficiency. We can see that the two distributions are relatively close at each of the lags considered.

	η	ν	$d=1$	$d=2$	$d=3$	$d=5$	$d=10$	$d=25$	$d=50$
Gaussian	20	0.500	0.963	0.928	0.897	0.840	0.721	0.538	0.423
			(0.017)	(0.032)	(0.045)	(0.068)	(0.104)	(0.129)	(0.120)
χ_1^{2*}	44	0.650	0.958	0.922	0.889	0.832	0.717	0.517	0.430
			(0.023)	(0.041)	(0.056)	(0.071)	(0.098)	(0.130)	(0.131)

Table 3.4: Parameters with mean and standard deviations of correlations for a selection of separation distances for a subset of 100 1-d Gaussian and χ_1^2 random fields of length 192.

For each of the 10000 replicates, we count the number of local maxima, calculating the mean and standard deviation for each distribution. Table 3.5 shows the parameters used, and these local maxima count means and standard deviations.

	η	ν	mean	sd
Gaussian	20	0.500	49.318	3.288
χ_1^{2*}	44	0.650	44.111	3.599

Table 3.5: Local maxima count means and standard deviations for Gaussian and χ_1^2 random fields, 10000 simulations per row.

As with work by Henderson et al. (2020) on a 256×256 grid, there are fewer local maxima evident in the χ_1^{2*} simulations than in the Gaussian case. We see about 10% fewer maxima lower for the χ_1^{2*} distribution than for the Gaussian, whereas Henderson shows a decrease of closer to 30%. Re-running the simulation procedure for all series lengths from $m = 3$ to $m = 1000$ shows that this ratio does not change as a function of series length. Figure 3.8 shows the ratio of the number of local maxima in χ_1^{2*} simulations to Gaussian simulations, with the mean calculated from ten simulations for each simulation length, m. Although the variation in the ratio decreases with increasing simulation length, there does not appear to be a trend in either direction from 0.88.

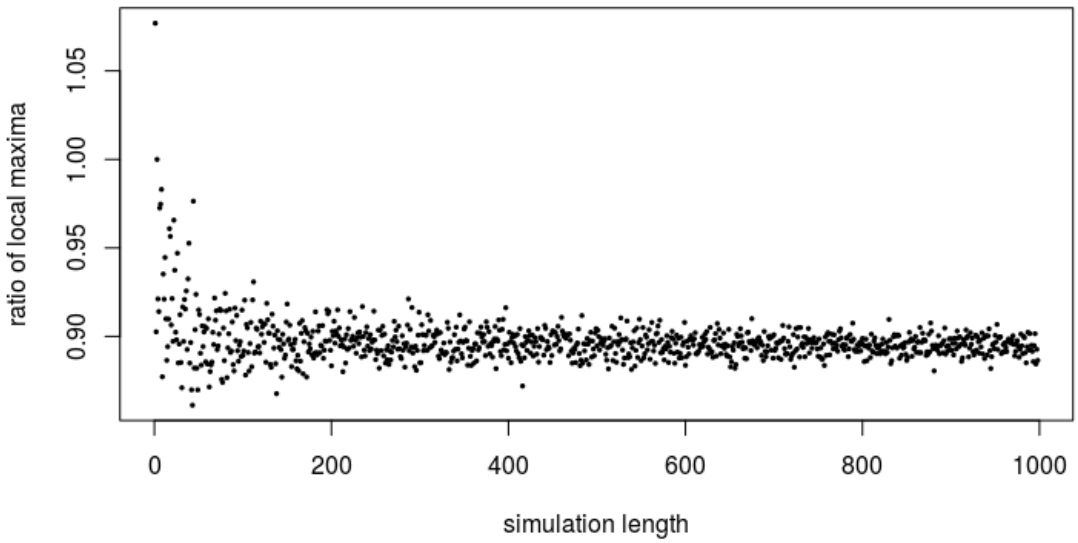

Figure 3.8: Ratio of mean number of local maxima in χ_1^{2*} to Gaussian random fields by simulation length, m.

3.4.2 Comparing simulations of length three

The marginally Gaussian chi-squared distribution provides a useful tool for assessing comparison methods, due to clear differences with a true Gaussian distribution. Thus far, although we have established a consistent difference between the number of local maxima occurring in one dimensional Gaussian random fields compared with χ_1^{2*} random fields, we have not established a reason why this might be. Both cases are matched for mean and standard deviation and have a close covariance, so intuitively one might expect the same proportion of local maxima. There is a clear difference between the two distributions, but the reasons why are less obvious. What are the underlying differences between our

Gaussian distribution and our χ_1^{2*} distribution? To answer this question, we note that understanding the behaviour of a distribution of length three is sufficient in understanding the behaviour of longer distributions, ignoring border cases for which we can easily adjust. We used this fact previously when calculating the expected number of local maxima over a length-192 random field. Hence, we must calculate the probability that the midpoint of a length-three random field is the maximum, under both the Gaussian and χ_1^{2*} distributions with a given covariance structure. From this, we will be able to calculate the expected number of local maxima in the χ_1^{2*} case. We first confirm expected results for the simpler cases, Gaussian under independence and χ_1^{2*} under independence, then investigate the χ_1^{2*} case under some covariance structure.

3.4.3 Gaussian under independence

We have a vector z of length three, containing independent random variables z_1, z_2 and $z_3 \sim N(0,1)$, where z_1 is the value at the midpoint and $z_{(1)} = (z_2, z_3)^T$ contains the values at the edgepoints. We know that the probability of any of the three points being the maximum is 1/3, by symmetry. That is

$$p(\text{midpoint is max}) = p(\text{midpoint is min}) = \frac{1}{3}.$$

The technical details are shown below for completeness. Define

$$I_Z = \begin{cases} 1, z_1 1_2 > z_{(1)} \\ 0, \text{ otherwise} \end{cases}$$

Then

$$\begin{aligned} p(I_Z = 1|z_1) &= \Phi^2(z_1) \\ \Rightarrow p(I_Z = 1) &= \int \Phi^2(z)\phi(z)dz \\ &= \left[\frac{\Phi^3(z)}{3}\right]_{-\infty}^{\infty} \\ &= \frac{1}{3} \end{aligned}$$

as expected.

3.4.4 χ_1^{2*} under independence

The above calculation for the χ_1^{2*} case is less trivial although again, for independence, we know the probability of any of the three points being the maximum is 1/3. We begin with a vector z of length three, containing independent random variables z_1, z_2 and $z_3 \sim N(0,1)$, where z_1 is the midpoint value and $z_{(1)} = (z_2, z_3)^T$ contains the edgepoint values, and transform to χ_1^{2*} variables, $z_1^*, z_{(1)}^*$ as described at the beginning of Section 3.4. Then we can define

$$I_{Z^*} = \begin{cases} 1, z_1^* 1_2 > z_{(1)}^* \\ 0, \text{ otherwise.} \end{cases}$$

We can see that

$$p(I_{Z^*} = 1) = p\big(z_1^* 1_2 > z_{(1)}^*\big)$$

$$= p\big(|z_1 1_2| > |z_{(1)}|\big).$$

If $z_1 > 0$,

$$p(-z_1 1_2 < z_{(1)} < z_1 1_2 | z_1) = \int_{-z_1 1_2}^{z_1 1_2} f(z_{(1)}|z_1) dz_{(1)}.$$

If $z_1 < 0$,

$$p(z_1 1_2 < z_{(1)} < -z_1 1_2 | z_1) = \int_{z_1 1_2}^{-z_1 1_2} f(z_{(1)}|z_1) dz_{(1)}.$$

Then for $z_1 > 0$,

$$p(I_{Z^*} = 1|z_1) = \{\Phi(z_1) - \Phi(-z_1)\}^2$$

and for $z_1 < 0$,

$$p(I_{Z^*} = 1|z_1) = \{\Phi(-z_1) - \Phi(z_1)\}^2.$$

So

$$p(I_{Z^*} = 1) = \int_{-\infty}^{0} \{\Phi(-z_1) - \Phi(z_1)\}^2 \phi(z_1) dy_1 + \int_{0}^{\infty} \{\Phi(z_1) - \Phi(-z_1)\}^2 \phi(z_1) dz_1$$

$$
\begin{aligned}
&= 2\int_0^\infty \{\Phi(z_1) - \Phi(-z_1)\}^2 \phi(z_1) dz_1 \\
&= 2\int_0^\infty \{2\Phi(z_1) - 1\}^2 \phi(z_1) dz_1 \\
&= 2\int_0^\infty \{4\Phi^2(z_1) - 4\Phi(z_1) + 1\} \phi(z_1) dz_1 \\
&= 8\left[\frac{\Phi^3(z_1)}{3}\right]_0^\infty - 8\left[\frac{\Phi^2(z_1)}{2}\right]_0^\infty + 2 \times \frac{1}{2} \\
&= \frac{8}{3} - \frac{8}{3 \times 8} - \frac{8}{2} + \frac{8}{2 \times 4} + 1 \\
&= \frac{1}{3}
\end{aligned}
$$

as expected.

3.4.5 χ_1^{2*} under a correlation structure

Having proved expected results in the independent case, we must find a way to understand behaviour when the random variables are not independent, but correlated according to some given correlation structure. In essence, this is similar to our previous approach, where we calculated the expected number of local maxima on a 1-d Gaussian field. It would be convenient if we could use the same methods for the χ_1^{2*} case, but we do not yet know the exact form of the covariance structure after the transformation. Again, we can begin working with fields of length three, as knowing the probability that the midpoint is the maximum is sufficient to calculate the expected number of maxima over a larger field.

Change in covariance under transformation

We can take two approaches to address this problem. Our first approach involves understanding how the covariance structure changes under the transformation of the Gaussian random variables. For a Gaussian 1-d field of length three, we have the following covariance matrix, R. We wish to understand the structure of the corresponding covariance matrix for the data after we transform it to χ_1^{2*}.

$$
R = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{12} \\ \rho_{13} & \rho_{12} & 1 \end{pmatrix}
$$

One method involves numerical calculation of the covariance after transformation. Consider two Gaussian random variables Z_1 and Z_2, with covariance ρ_{12}. We wish to know

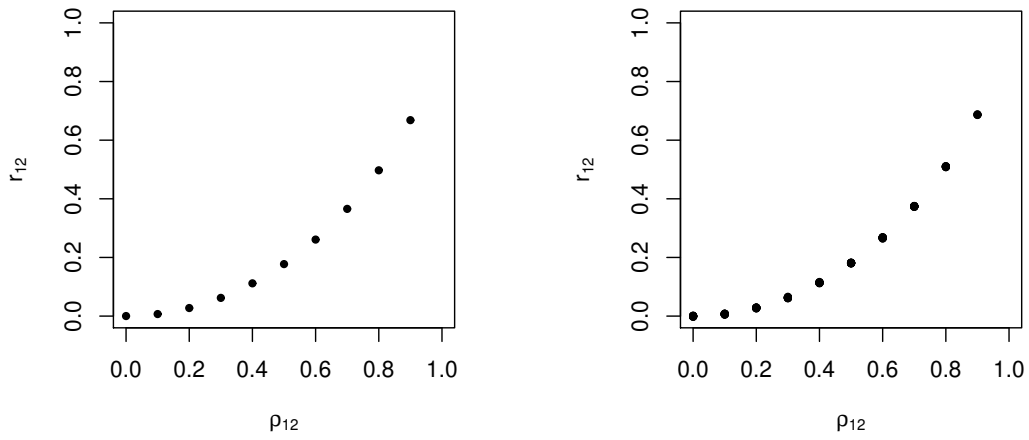

Figure 3.9: Relationship between initial and resulting covariances, ρ_{12} and r_{12}. The left hand plot shows results obtained using numerical integration. The right hand plot shows results obtained using simulation and estimation.

the covariance between the transformed χ^{2*}_1 random variables Z^*_1 and Z^*_2, $\text{Cov}[Z^*_1 Z^*_2]$. As described previously, $z^*_i = \Phi^{-1}\big(F_{\chi^2_1}(z^2_i)\big)$.

$$\begin{aligned}\text{Cov}[Z^*_1 Z^*_2] &= \text{E}[Z^*_1 Z^*_2] - \text{E}[Z^*_1]\text{E}[Z^*_2] \\ &= \text{E}[Z^*_1 Z^*_2],\end{aligned}$$

since Z^*_1, Z^*_2 are marginally Gaussian, so $\text{E}[Z^*_1] = \text{E}[Z^*_2] = 0$. Then

$$\text{E}[Z^*_1 Z^*_2] = \int\int \Phi^{-1}\big(F_{\chi^2_1}(z^2_1)\big)\Phi^{-1}\big(F_{\chi^2_1}(z^2_2)\big)\phi(z_1, z_2)dz_1 dz_2.$$

We calculate this integral numerically over a grid,

$$\text{E}[Z^*_1 Z^*_2] \approx \sum_{z_1, z_2} \Phi^{-1}\big(F_{\chi^2_1}(z^2_1)\big)\Phi^{-1}\big(F_{\chi^2_1}(z^2_2)\big)\phi(z_1, z_2)\delta^2$$

where $z_1, z_2 = -5, -5+\delta, -5+2\delta, \ldots, 5-\delta, 5$. Here we choose $\delta = 0.01$ to allow for manageable computation times. This gives us an approximate value for the covariance between the χ^{2*}_1 random variables, $\text{Cov}[Z^*_1, Z^*_2]$, r_{12}. The relationship between the initial ρ_{12} and subsequent r_{12} covariances can be seen in the left-hand plot of Figure 3.9.

As a test for this method, we simulate multiple 1-d random fields of length three, under different covariance matrices. We form 121 matrices, corresponding to 11 values for each of

ρ_{12} and ρ_{13} and take all valid covariance matrices. For each of these, we simulate 1000000 random fields of length three, transform to χ_1^{2*} as described previously and estimate the resulting covariance matrices. The relationship between the initial ρ_{12} and subsequent r_{12} covariances when calculated using this method can be seen in the right-hand plot of Figure 3.9.

This method has the disadvantage of requiring simulations, and hence includes a degree of uncertainty. We can minimise this uncertainty by increasing the number of simulations from which we calculate the covariance, but it cannot be eliminated.

As can be seen in Figure 3.9, simulation confirms results obtained using numerical integration. We can, therefore, estimate the covariance matrix of the transformed random variables to calculate the predicted number of local maxima.

3.5 Fundamental differences and impact of covariance structure

In this section we continue to consider the probability that the midpoint is the maximum for series of length $m = 3$, recalling that results correspond directly to the number of local maxima over a series of length $m \geq 3$. We are now able to calculate expected results for both the Gaussian and the χ_1^{2*} cases under any covariance structure to a reasonable degree of accuracy. From Section 3.4.1 it is clear that the number of local maxima (and hence the probability of the midpoint being a maximum) is different between the two distributions even when the covariance structures are close. We consider two main questions, firstly, how does the choice of covariances affect the differences between distributions and secondly, where do the differences come from when covariances are matched. We also look briefly at how the covariances differ when they are chosen to match the total probabilities between distributions.

3.5.1 Effect of relationship between ρ_{12} and ρ_{13}

We look first at how different choices of ρ_{12} and ρ_{13} impact $p(I_{Z^*} = 1)$, the probability that the midpoint is the maximum, in the χ_1^{2*} case. As established previously, when $\rho_{12} = \rho_{13} = 0$ (i.e. in the case of independence), the probability $p(I_{Z^*} = 1) = \frac{1}{3}$. In work on 2-d random fields Henderson et al. (2020) noted a significantly lower number of local maxima for χ_1^{2*} data when compared with true Gaussian data. The data used was all simulated using a Matérn covariance function, for which the covariance between two points decreases as the distance between them increases. In the 1-d length three case, we are interested in the how $p(I_{Z^*} = 1)$ changes with different values of ρ_{12} and ρ_{13}, where

ρ_{12} is not necessarily greater than ρ_{13}.

In order to asses this, we fix ρ_{12} at three values, 0.3, 0.5 and 0.7, and for each, we consider $\rho_{13} = 0.1, 0.2, \ldots, 0.9$ as shown in Figure 3.10. This shows that when $\rho_{13} < \rho_{12}$, $p(I_{Z^*} = 1) < \frac{1}{3}$, as was found in previous work, when using a Matérn covariance function. However, we see that when $\rho_{13} > \rho_{12}$, $p(I_{Z^*} = 1) > \frac{1}{3}$, with the case where when $\rho_{13} = \rho_{12}$ giving results very close to independence.

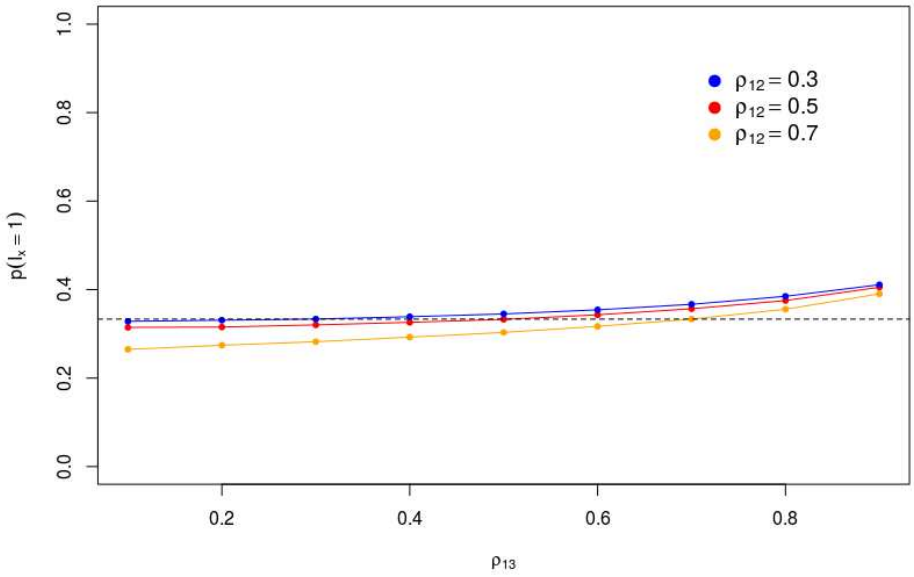

Figure 3.10: $p(I_{Z^*} = 1)$ for $\rho_{12} = 0.3, 0.7, 0.9$ and $\rho_{13} = 0.1, 0.2, \ldots, 0.9$. The black dashed line shows $p(I_{Z^*} = 1) = \frac{1}{3}$.

Intuitively, this is not a surprising result. The larger ρ_{13} is in relation to ρ_{12}, the more that the points x_1 and x_3 act as a single point. In the most extreme case, where $\rho_{13} = 1$, the two points act as one, and hence, $p(I_{Z^*} = 1) = \frac{1}{2}$.

It is worth noting that we have chosen values for ρ_{12} and ρ_{13} which result in a valid covariance matrix. The matrix is only positive-semi-definite and hence, only valid, when the following hold.

$$1 - 2\rho_{12}^2 + 2\rho_{12}^2\rho_{13} - \rho_{13}^2 \geq 0$$

and

$$0 \leq \rho_{12}, \rho_{13} < 1.$$

3.5.2 'Covariance matching' for Gaussian and χ_1^{2*} distributions

Having established the relationship between initial covariances and those which result after transformation of our Gaussian random field as seen in Section 3.4.5, we can compare our two distributions using these 'matched' covariances. That is, we can match the covariance values r_{12} and r_{13} of our χ_1^{2*} distribution, to the values ρ_{12} and ρ_{13} of our target standard Gaussian distribution. The process is as follows:

1. Select two root Gaussian covariances, ρ_{12}^* and ρ_{13}^*.
2. Calculate the corresponding target covariances, ρ_{12} and ρ_{13}, as discussed in Section 3.4.5. These will be the covariance values used in the Gaussian calculation, and are the corresponding values after transformation to χ_1^{2*}.
3. Using the root values (the covariances of the underlying Gaussian distribution), calculate $p(I_{Z^*} = 1)$ for the χ_1^{2*} distribution as discussed previously.
4. Use the target covariances to calculate $p(I_Z = 1)$ for the true Gaussian case.

Data simulated from χ_1^{2*} and Gaussian distributions in this way will have close to matching covariances. Table 3.6 shows the results of implementing the above procedure. We can see that regardless of the values when ρ_{12} and ρ_{13} are equal, the probabilities in both cases are equal to a third. When $\rho_{12} > \rho_{13}$, as is the case in the majority of standard models where covariance decreases with increasing distance, $p_{\chi_1^{2*}}(I_{Z^*} = 1) < p_{Gauss}(I_Z = 1)$. This result is in agreement with previous work on 2-d random fields. In the alternate case, where $\rho_{12} < \rho_{13}$, we see the opposite effect.

ρ_{12}^*	ρ_{13}^*	ρ_{12}	ρ_{13}	$p_{\chi_1^{2*}}(I_{Z^*} = 1)$	$p_{Gauss}(I_Z = 1)$
0.3	0.3	0.062	0.062	0.333	0.333
0.3	0.5	0.062	0.177	0.345	0.345
0.3	0.7	0.062	0.366	0.367	0.365
0.5	0.3	0.177	0.062	0.320	0.321
0.5	0.5	0.177	0.177	0.333	0.333
0.5	0.7	0.177	0.366	0.357	0.355
0.7	0.3	0.366	0.062	0.282	0.292
0.7	0.5	0.366	0.177	0.303	0.307
0.7	0.7	0.366	0.366	0.333	0.333

Table 3.6: Probability that the maximum occurs at the midpoint for χ_1^{2*} and Gaussian length three random fields.

We can see the underlying effect more clearly by looking at the plots of the conditional distributions, $p(x_1 > x_2 1_2 \mid x_1)$ and $p(x_1 > x_2 1_2 \mid |x_1|)$, shown in Figures 3.11 and 3.12,

for both true Gaussian and χ_1^2 distributions. (Generic notation is used here for both distributions.)

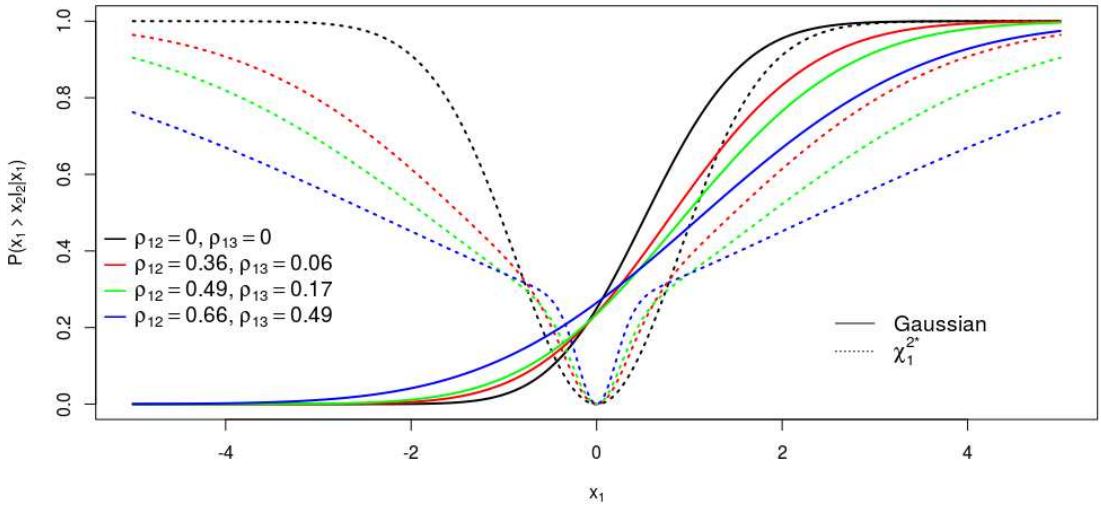

Figure 3.11: $p(x_1 > x_2 1_2 \mid x_1)$ for Gaussian and χ_1^{2*} distributions. Particularly around $x_1 = 0$, we can see the clear difference in probabilities that x_1 is a local maximum.

It is clear that there are significant differences between the distributions, particularly for smaller values of $|x_1|$. Given that smaller values of $|x_1|$ have a higher probability of occurring, the differences in this range will have a greater impact on the probability of x_1 being the maximum. This provides some explanation for the lower number of local maxima seen in the χ_1^{2*} case compared to the Gaussian (where $\rho_{12} > \rho_{13}$ as we have here).

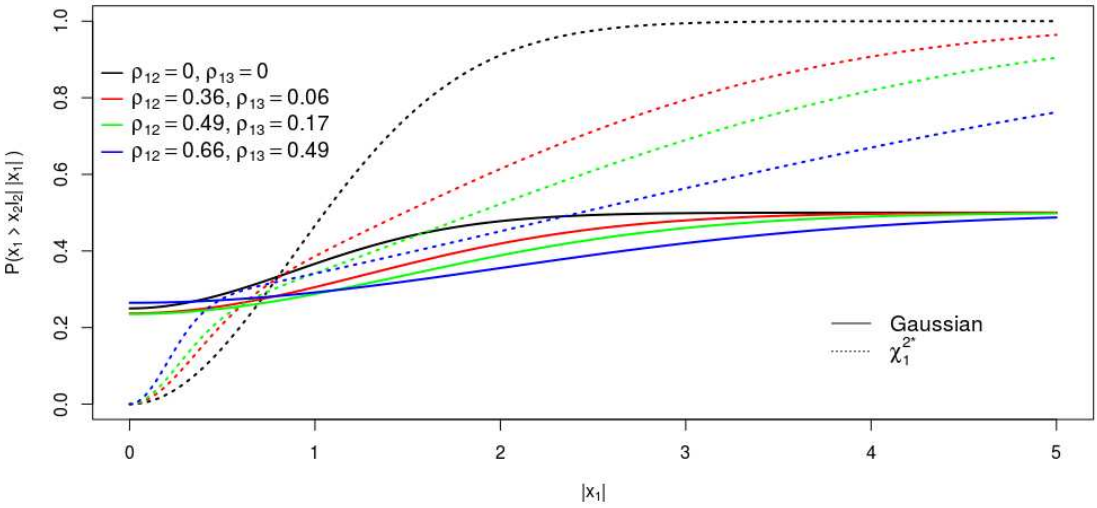

Figure 3.12: $p(x_1 > x_2 1_2 \mid |x_1|)$ for Gaussian and χ_1^{2*} distributions.

3.5.3 Matching total probabilities

The final question we wish to answer is what happens when we choose covariances such that the probabilities $p(z_1 1_2 > z_{(1)})$ in the Gaussian case and $p(z_1^* 1_2 > z_{(1)}^*)$ in the χ_1^{2*} case are as close as possible. After finding covariance values that equate the probabilities (or some approximation to this), we are able to plot the conditional distributions as in the previous section and compare the shape of the curves. Figure 3.13 shows this for covariances $\rho_{12} = 0.7$ $\rho_{13} = 0.3$, $r_{12} = 0.82$ and $r_{13} = 0.42$, resulting in $p(z_1 1_2 > z_{(1)}) \approx p(z_1^* 1_2 > z_{(1)}^* s) \approx 0.223$. As before we look at the probability conditional on the (general) midpoint x_1 but also at the probability conditional on the absolute value of the midpoint $|x_1|$.

We can see that where the midpoint x_1 has a positive value, the probability $p(x_1 > x_2 1_2 \mid x_1)$ for both Gaussian and χ_1^{2*} data have similar values, as does the probability conditional upon $|x_1|$ for all values of x_1. In contrast, as x_1 decreases from zero, the probability $p(x_1 > x_2 1_2 \mid x_1)$ for the two distributions diverges. However, considering the greater likelihood of values of x_1 close to zero (for marginally Gaussian data) the figure demonstrates how the likelihood of local maxima is related to the values of ρ_{12} and ρ_{13}.

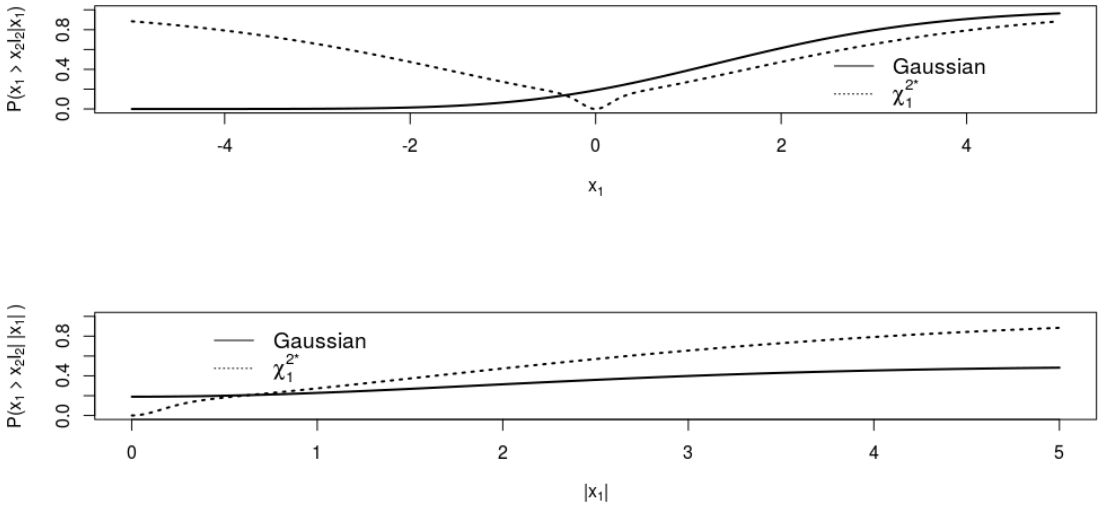

Figure 3.13: $p(x_1 > x_2 1_2 \mid x_1)$ and $p(x_1 > x_2 1_2 \mid |x_1|)$ for correlated Gaussian and correlated χ_1^{2*}, where both distributions have approximately equal total probability.

3.6 Conclusions

In this chapter we have introduced general ideas from topological data analysis, before showing how concepts could be applied to random fields on a discrete grid. We then investigated and compared numbers of local maxima on Gaussian and marginally Gaussian χ_1^2 both for spatially independent and correlated data. Finally, we considered the mechanism by which we see different results between distributions by looking into length-3 vectors with known correlations. Beyond introducing some of the topological theory upon which we will build in Chapter 4, we have shown how subtle differences in distributions have an impact on topological features, even when effort has been taken to account for differences in correlation structure. This forms a basis for the work to follow on distinguishing distributions using topological event history methods. In the following chapter, we return to some more standard topological data analysis methods and demonstrate their application to wind intensities data.

Chapter 4

Applications of TDA to wind data

To illustrate the topological data analysis approaches described in Chapter 3 we apply a selection of these approaches to the wind intensities data. This allows us to demonstrate some features of the data set as well as discuss some of the advantages and challenges of standard TDA methods. Since the data set consists of both multiple years and multiple realisations, where possible we look at comparisons across both variables. Prior to conducting the analysis, we may expect to observe few and/or relatively minor differences between realisations since they are generated to be statistically similar, although any such differences could be worth further investigation. This is particularly relevant in the context of climate variability, one of the main purposes for the development of the CESM LENS data set. Differences in analyses across years, however, may be less surprising, but potentially of interest from the perspective of investigating changes in climate over time. Throughout this chapter we use the full data set.

4.1 Local maxima and minima

We begin by looking at the numbers of local maxima and minima for each year-realisation data subset. Results can be seen in Figures 4.1 and 4.2, for the original data and standardised residuals respectively. Although there are no obvious trends across year or realisation, we see significantly more minima than maxima for the original data. Averaging the numbers of local maxima and minima by realisation for the standardised residuals we can see in Figure 4.3 that there is an increase in both maxima and minima as time increases, followed by a drop after peaking at around the 60th year (2067). This demonstrates how even a basic analysis approach based on TDA can highlight interesting trends across data sets. The remaining figures and analysis in this chapter all use standardised residuals.

In contrast with many standard analysis techniques, interpretation of results obtained via

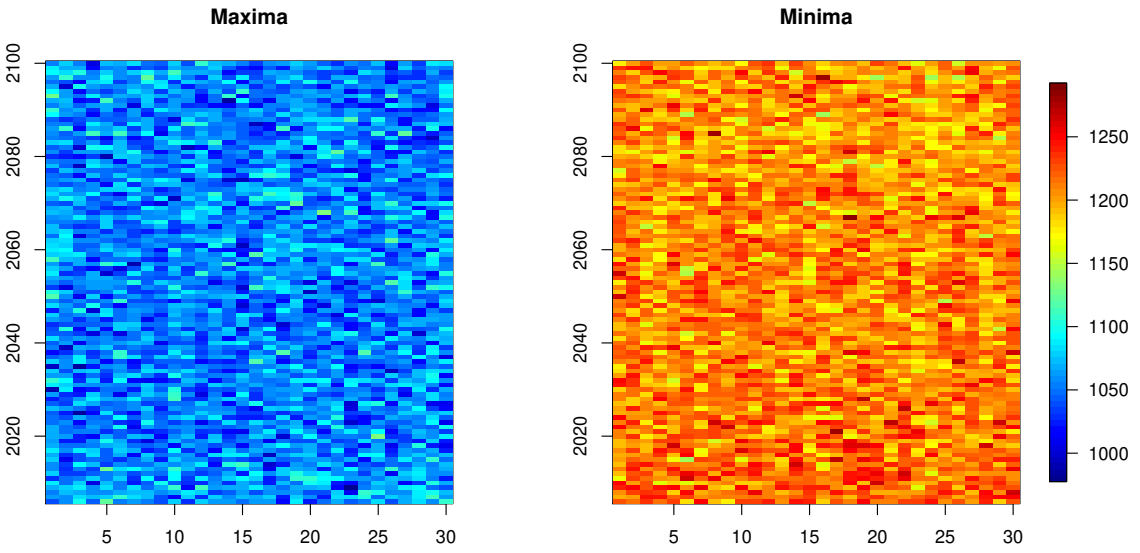

Figure 4.1: Number of local maxima (left) and minima (right) for all 30×95 year-realisation subsets (original data).

TDA methods can be challenging. We can generally say that higher numbers of local maxima and minima reflect greater noise in the random field, with less smoothness. However, the reasons for this occurrence could be a result of the methods used to simulate the original data or a feature of our changing climate under the specific conditions considered. Further understanding would require the knowledge of a climate scientist or researcher involved in the simulation of the data.

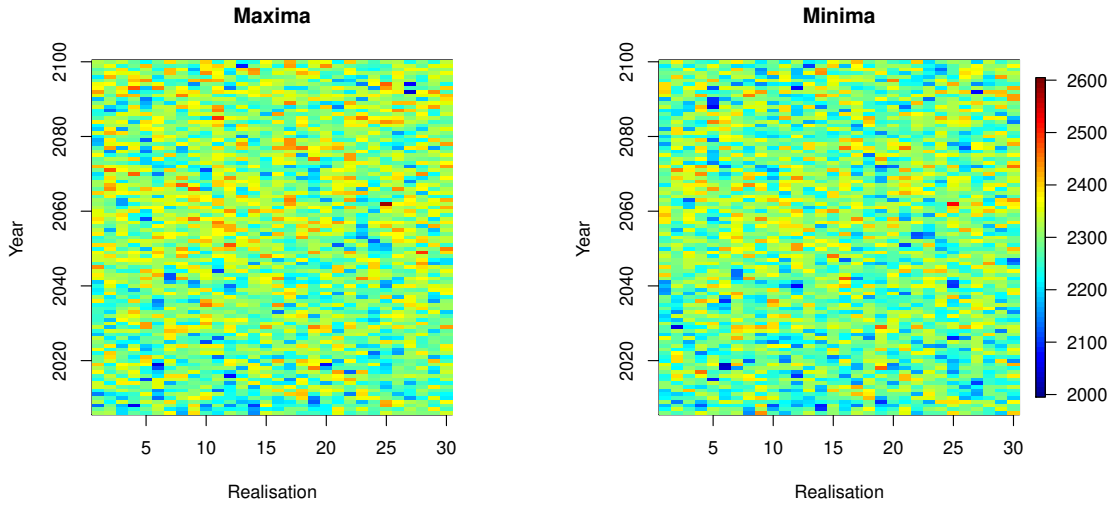

Figure 4.2: Number of local maxima (left) and minima (right) for all 30×95 year-realisation subsets (standardised residuals).

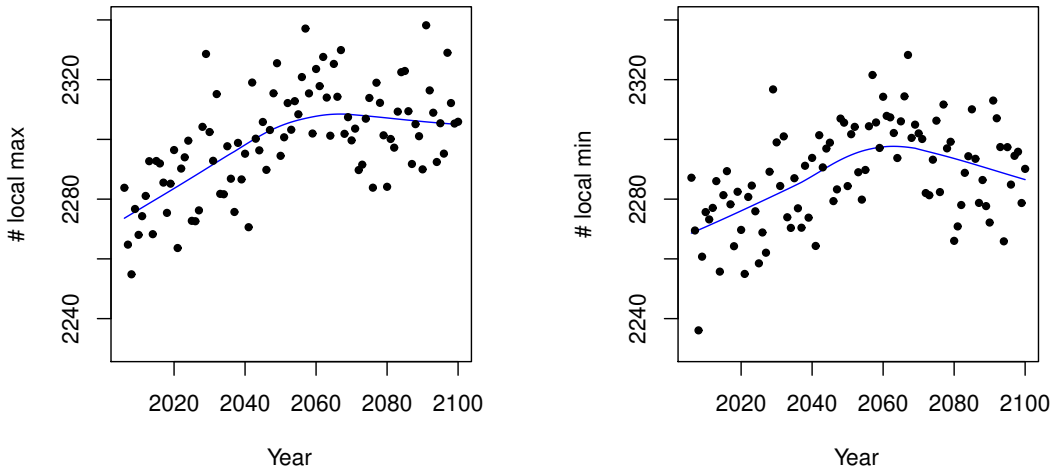

Figure 4.3: Average number of local maxima (left) and minima (right) for each year based on the standardised residuals.

4.2 Persistence

Recalling from the previous chapter, for a 2-d topological space there are two potentially non-zero Betti numbers, β_0 and β_1, corresponding to connected components and holes. For data on a grid, we define two pixels to be connected if they share an edge. We show connected component persistence diagrams for a selection of years and realisations, shown in Figure 4.4. Figures such as persistence diagrams for components and holes are less well suited for the comparison of numerous data sets; we are only showing $\sim 0.3\%$ of the full data set here.

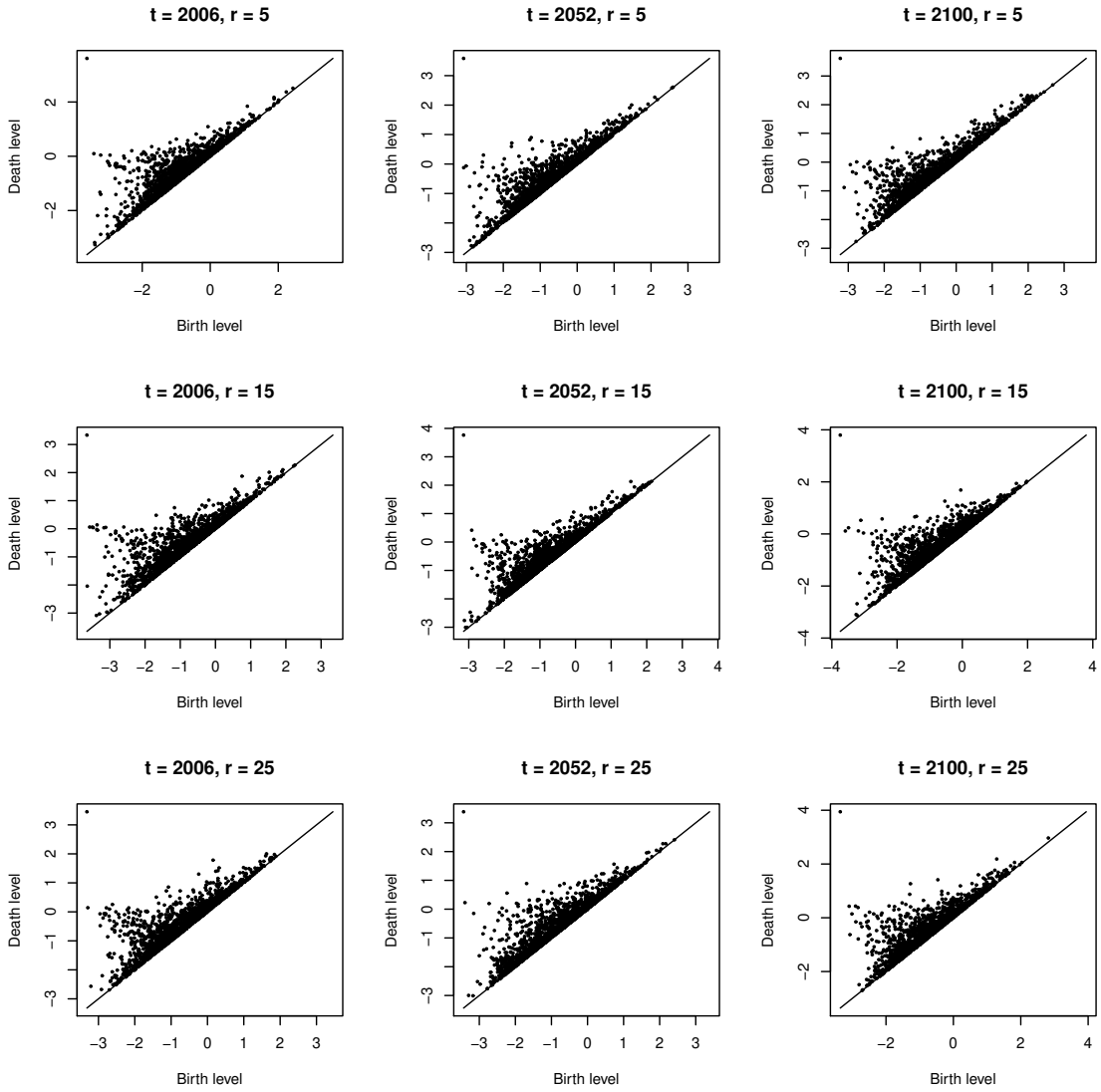

Figure 4.4: Persistence diagrams for connected components for a selection of years and realisations.

Although minor differences are apparent between each year-realisation pair, there is no

noticeable trend that can be seen from these figures. Persistence diagrams alone are poorly suited to data of this type; instead we require some way of summarising these figures.

Barcodes provide a means of visualising the same details in a different way. Figure 4.5 shows barcodes for three different years. The barcodes for each year are very similar, but each highlights the differences in persistence between components and holes. As would be expected with a filtration from below, we see generally earlier birth times for components as well as one component persisting to the maximum value.

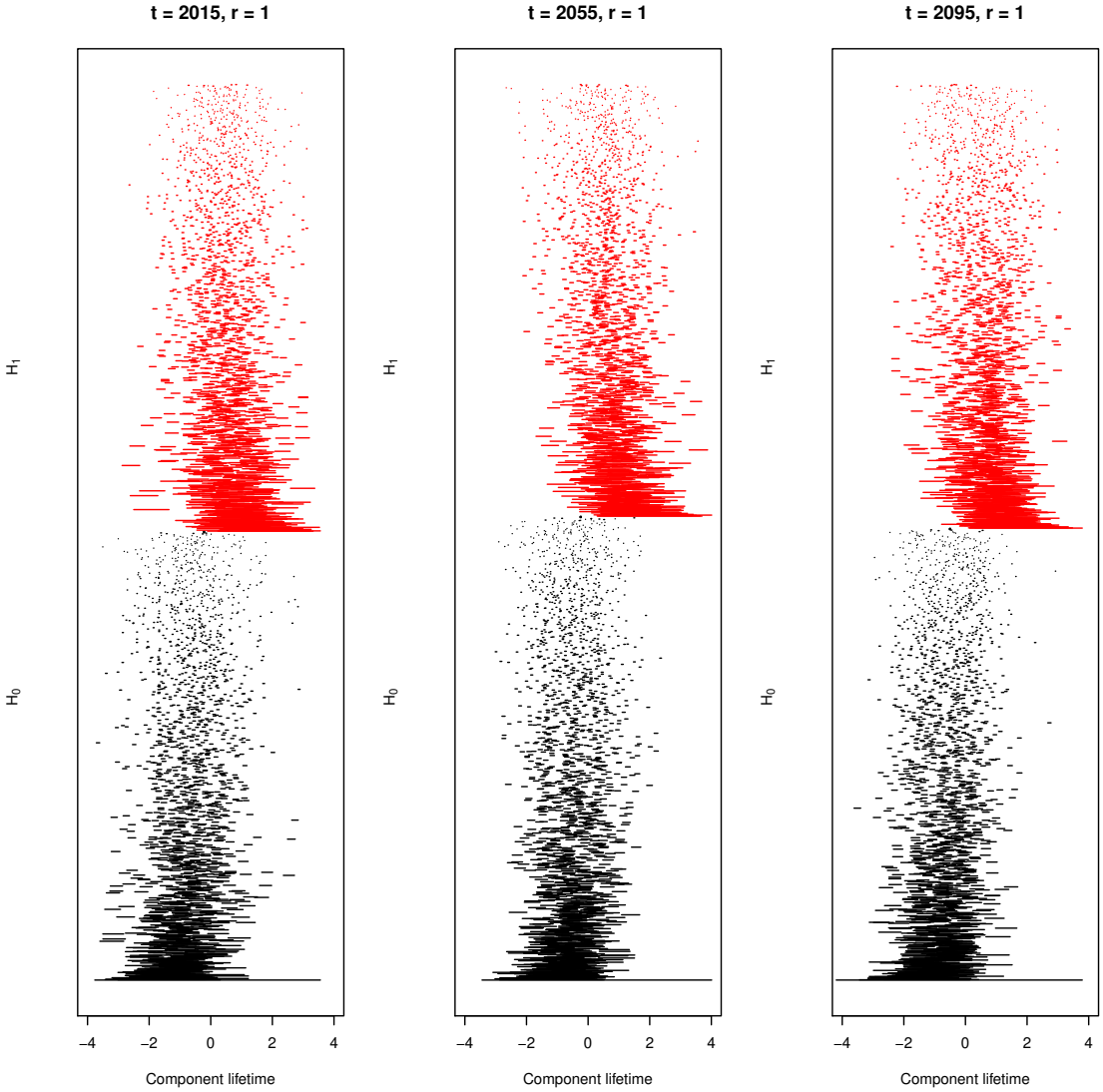

Figure 4.5: Barcodes for components (back) and holes (red), corresponding to elements from homology groups H_0 and H_1. $t = 2015, 2055, 2095$, $r = 1$.

4.2.1 Convex hull summaries

'Peeling' a persistence diagram to obtain successive convex hulls as proposed by Henderson et al. (2020), provides a means to uncover the general shape of a persistence diagram, without being obscured by the large number of points, particularly around the diagonal. Figure 4.6 illustrates.

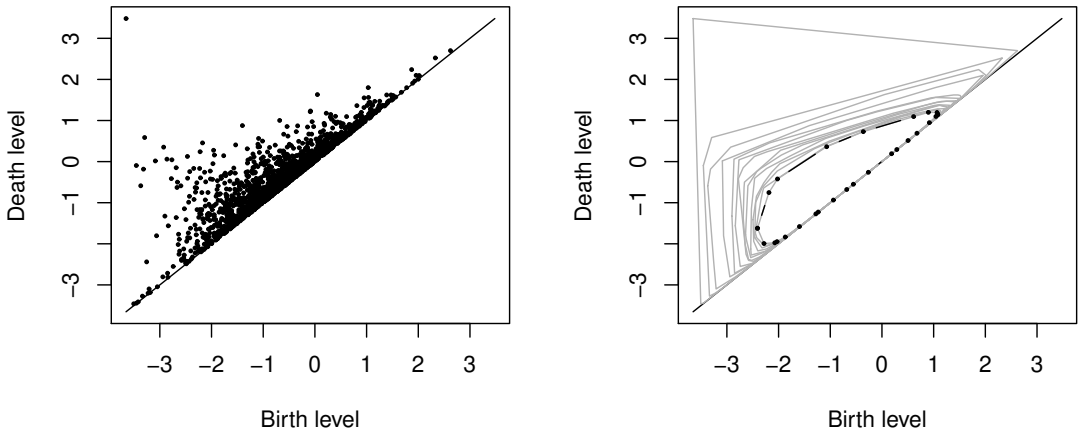

Figure 4.6: Persistence diagram and 85% convex hull for realisation $r = 10$ and year $t = 2015$. Grey lines show convex hulls that have been peeled.

We can use a convex hull peel to examine differences between persistence diagrams more clearly. As with persistence diagrams, we are able to compare a small number of data sets via their convex peels, and their component/hole counts. Figure 4.7 shows this for year one, realisations one, five, and ten, showing only minor differences between realisations. An improvement on the persistence diagram, this approach still limits us to a relatively small number of data sets.

Looking beyond simple descriptors of persistence, the convex hull summaries of persistence diagrams detailed in the Section 3.2.1 provide more valuable ways to compare persistence diagrams. We apply these to the persistence diagrams after convex peels such that 85% of points remain. Figure 4.8 looks at all years for realisations 10, 20 and 30, and Figure 4.9 looks at all realisations for years 2015, 2055 and 2095. Each of the years and realisations presented appear to have similar properties, with no clear outliers or trend.

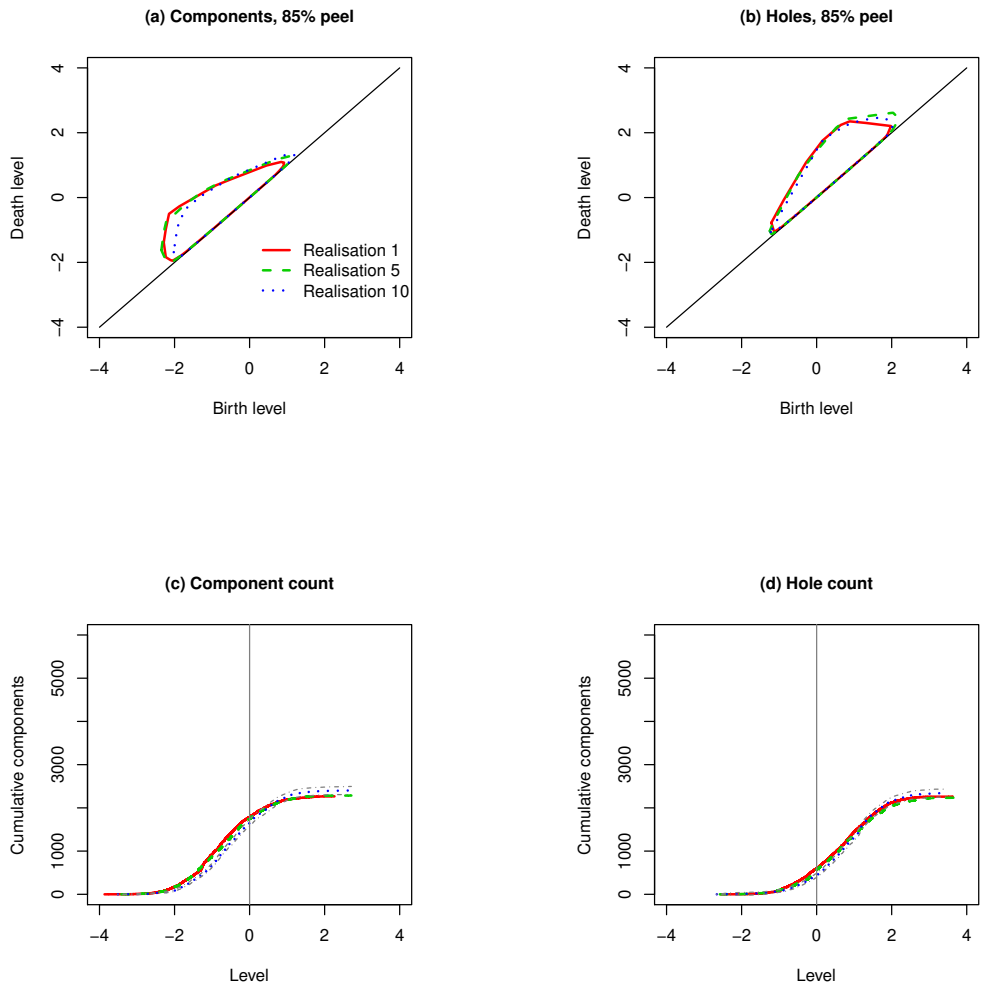

Figure 4.7: Convex peels for components and holes. $t = 2006$, $r = 1, 5, 10$, as well as a count of components and holes over increasing level sets.

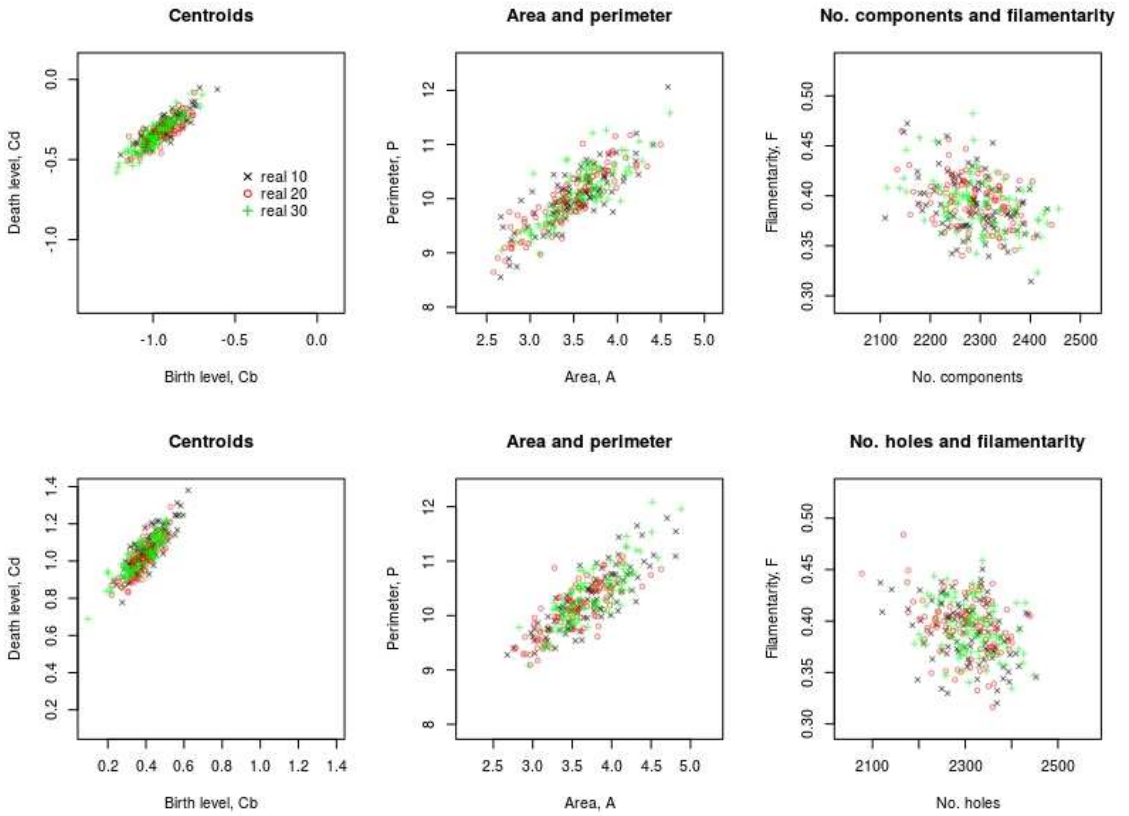

Figure 4.8: Convex hull summaries for all years, realisations $r = 10$, $r = 20$ and $r = 30$. Top row shows values for connected components, bottom row shows values for holes.

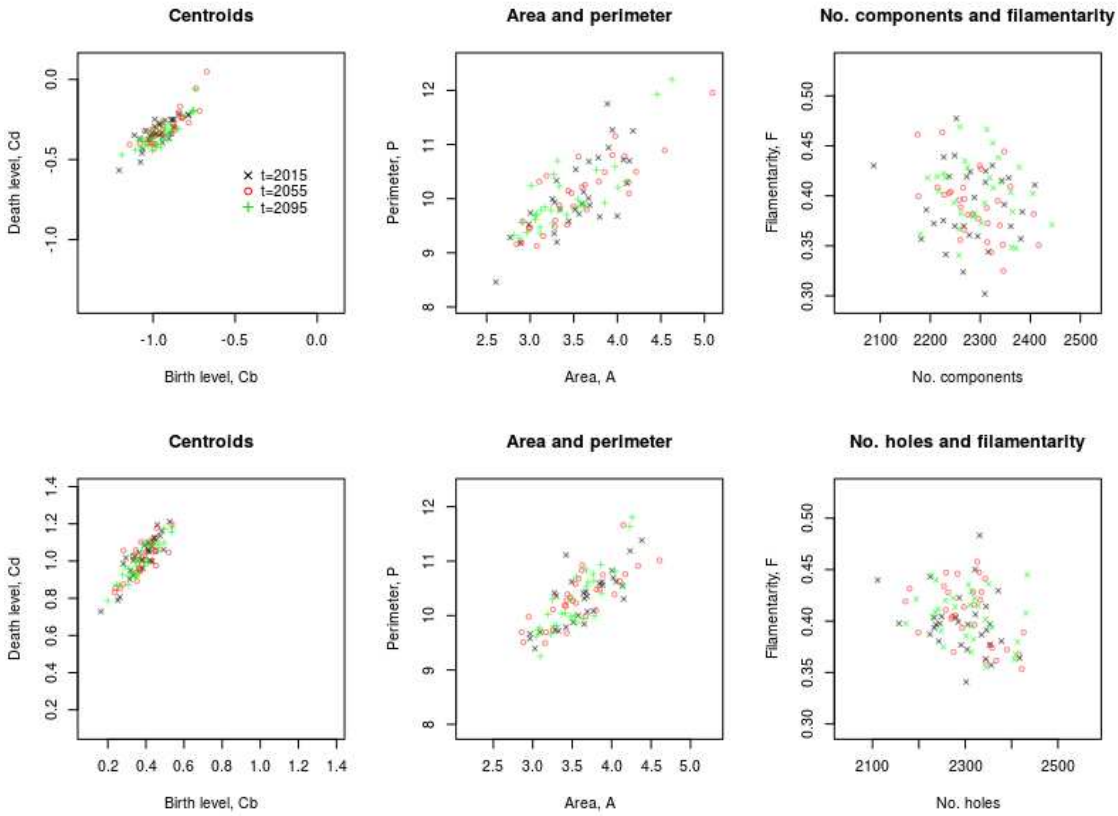

Figure 4.9: Convex hull summaries for all realisations, years $t = 2015$, $t = 2055$ and $t = 2095$. Top row shows values for connected components, bottom row shows values for holes.

4.3 Conclusions

In this chapter we looked at maxima and minima counts for our data sets over time, observing an interesting feature where counts of both maxima and minima appear to increase over time to a peak at around $t = 2067$ and decrease slightly after. We then looked at various visual representations of persistence, including persistence diagrams, barcodes and convex peels of persistence diagrams, before looking at values for several convex hull summaries. Beyond looking at the change in maxima and minima counts over times, these methods did not highlight any real differences between either years or realisations.

We can see how TDA is able to provide novel, visual methods for examining features of data. Calculation of local maxima and minima, and convex hull summaries allow us to compare between data and identify differences that may not be apparent by other means. However, it is clear that some of these methods, such as examination of persistence diagrams, are not well suited to comparing numerous data sets and instead allow a more detailed inspection of specific features. In the case of our wind intensities data, we are able to use these methods to gain insight into features of the data, however are restricted by the large quantity of year-realisation data sets we are working with.

Chapter 5

Event history analysis for spatially correlated data

As we described at the beginning of Chapter 3, existing work by Henderson et al. (2020) showed that methods based on persistent homology from the field of topological data analysis were able to distinguish between true Gaussian and marginally Gaussian χ^2_1 random fields. We are interested in extending these ideas using a survival analysis approach, with the eventual goal of combining methods from the two fields of research. In this chapter, we discuss ideas from survival analysis and investigate their application to spatially correlated data on a random field. We consider the problem of variance estimation and examine methods for adjusting existing variance estimators to work for correlated data.

The field of survival analysis has wide-ranging applications, particularly in manufacturing and medical sciences, where time-to-event data is common. For example, one might encounter data comprising time until failure of a part in a mechanical process, or until the recurrence of a disease in a medical trial. For this reason, survival analysis has been well researched and developed, with many publications describing approaches to the study of data sets where the variable of interest is time until the occurrence of some event. Owing to its high applicability in many areas of scientific research, much of the literature is concerned with the application of survival analysis to diverse scientific research areas.

In addition to comparing the outcomes of processes, analysis of time-to-event data is useful for understanding the underlying mechanisms. Hence, there is undoubtedly value in the interpretability of results to make inferences about each data set. Contrary to this idea of interpretation, in this chapter we are more interested in how concepts from survival analysis can be used as tools for the comparison of random fields. A primary difference between this and standard survival analysis is the spatial nature of the event times. In almost all scenarios we consider, there will be some degree of spatial correlation

in the data, a feature not present in the majority of time-to-event data sets. Further, the interpretability of results will not be of primary concern; instead, we wish to be able to distinguish between random fields with similar properties, where existing methods may be unable to identify differences.

Before continuing into a discussion of survival analysis for correlated data, we begin with a summary of standard survival analysis theory, including definitions of the survival function and cumulative hazard function and a review of concepts such as martingales, compensators and predictability. We describe two standard non-parametric estimators used in survival analysis and methods for assessing their variance. Since our work does not look at censored data (in a conventional sense), we omit most discussion of censoring from the following sections. We begin with a summary of basic survival analysis theory.

5.1 Basic concepts

Consider a set of n subjects, with survival times $T_1, T_2, \ldots, T_n$, observed without censoring. The T_i are continuously distributed from some survival function $S(\cdot)$, where

$$S(t) = P(T_i > t).$$

In many real applications, not all individuals will experience an event. For example, in a medical study of heart disease, it is almost certain that some patients will never experience the event of interest. In this case, the random variable T_i would be infinite and the survival function $S(t)$ will eventually decrease to a positive value as $t \to \infty$ (Aalen et al., 2009).

The distribution function and density function are then defined as $F(t) = 1 - S(t)$ and $f(t)$ respectively. The hazard function $\alpha(t)$ is defined as

$$\alpha(t) = \frac{f(t)}{(1 - F(t))}$$

and we can see that

$$\alpha(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t < T < t + \Delta t \mid T > t)}{\Delta t} \right\} = \frac{f(t)}{S(t)}. \tag{5.1}$$

The cumulative hazard function A is defined as the integral of the hazard function,

$$A(t) = \int_0^t \alpha(u) du. \tag{5.2}$$

In Section 5.2 we describe the non-parametric Kaplan-Meier and Nelson-Aalen estimators for the survival function and cumulative hazard function respectively.

The relationship between these two functions $A(t)$ and $S(t)$ can be shown simply. From 5.1 and 5.2 we can see that

$$A'(t) = \alpha(t) = \lim_{\Delta t \to 0} \left\{ \frac{S(t) - S(t + \Delta t)}{S(t)\Delta t} \right\} = -\frac{S'(t)}{S(t)}.$$

Since $S(0) = 1$, integrating obtains

$$-log\{S(t)\} = \int_0^t \alpha(s)ds$$

and hence

$$S(t) = exp\Big\{ - \int_0^t \alpha(s)ds \Big\}.$$

5.1.1 Counting processes

The theory behind counting processes provides many important ideas required to understand survival theory. For a subject i, T_i can be considered the survival time for that subject. Then we can define the counting process associated with T_i to be the indicator

$$N_i(t) = I(T_i \leq t).$$

The change dN_i can be written as

$$dN_i(t) = N_i(t) - N_i(t^-),$$

and takes value 1 at failure times and 0 everywhere else. Hence, over the whole sample, we can take

$$N(t) = \sum_{i=1}^{n} N_i(t)$$

where $N(t)$ is a step function which counts the number of observed events before time t.

Predictable processes A process is predictable if given the history $\mathcal{F}_{t\_}$ we know the value of the process at time t. $\mathcal{F}_{t\_}$ is everything known immediately prior to time t. Left-continuity is a sufficient (but not necessary) condition for predictability.

Risk sets A risk set at time t can be described as a set of subjects who, at time t, are at risk of failure. We can say

$$Y_i(t) = \begin{cases} 1, & \text{subject } i \text{ is at risk of failure at time } t \\ 0, & \text{otherwise.} \end{cases}$$

The total number of subjects at risk at time t is thus

$$Y(t) = \sum_{i=1}^{n} Y_i(t),$$

where $Y(t)$ and $Y_i(t)$ are both predictable processes.

Martingales

Martingales play an important role in survival analysis. First, suppose $M(t)$ is some stochastic, right-continuous process. We define the continuous time interval $\mathcal{T} = [0, \tau]$, where the terminal time τ is in the range $0 < \tau \leq \infty$ and $(\mathcal{F}_s : s \in \mathcal{T})$ of $M(t)$ is the history (or filtration) of the process up to time s.

Then $M(t)$ is a martingale if $M(t)$ is integrable, that is, if

$$E[\mid M(t) \mid] < \infty \qquad \forall t \in \mathcal{T}$$

and satisfies the martingale property

$$E[M(t) \mid \mathcal{F}_s] = M(s) \qquad \forall s < t$$

(Aalen et al., 2009).

Intensity The intensity of the counting process $N(t)$ can be defined as

$$\lambda(t) = Y(t)\alpha(t),$$

and we have

$$\mathrm{E}[dN(t) \mid \mathcal{F}_{t^-}] = \lambda(t)dt.$$

From this, we obtain the cumulative intensity,

$$\Lambda(t) = \int_0^t \lambda(s)ds \qquad t \geq 0,$$

so that $\mathrm{E}[N(t)] = \Lambda(t)$.

We can see now that $M(t) = N(t) - \Lambda(t)$ is a martingale since

$$\mathrm{E}[dM(t) \mid \mathcal{F}_{t^-}] = 0$$

and

$$\mathrm{E}[M(t) \mid \mathcal{F}_s] = M(s).$$

5.2 Non-parametric estimators

We look at two existing non-parametric estimators. First, briefly, we describe the Kaplan-Meier (KM) estimator $\hat{S}(t)$ for the survival function, followed by a more detailed presentation of the Nelson-Aalen estimator $\hat{A}(t)$ for the cumulative hazard function.

5.2.1 The Kaplan-Meier estimator for the survival function

It is likely that what is now usually known as the Kaplan-Meier estimator for the survival function was first proposed by Bohmer (1912) before being largely lost sight of until 1958, when it was presented as the Product-Limit estimator by Kaplan and Meier (1958). The estimator is defined as

$$\hat{S}(t) = \prod_{T_i \leq t} \left\{ 1 - \frac{1}{Y(T_i)} \right\}$$

where $Y(t)$ is the number of individuals contained in the risk set 'just before' time t (Aalen et al., 2008) and T_i are event times.

5.2.2 Variance of the Kaplan-Meier estimator

There are several variance estimators for the Kaplan-Meier estimator; the one provided by Kaplan and Meier (1958) is as follows:

$$Var\big(\hat{S}(t)\big) = \hat{S}(t)^2 \sum_{T_i \leq t} \frac{1}{Y(T_i)^2}$$

which is equal to $\hat{S}(t)^2$ multiplied by the Nelson-Aalen variance estimator (see Section 5.2.4).

In addition to the variance estimator provided above, Greenwood's formula (Greenwood and others, 1926) is also commonly used as a variance estimator for the Kaplan-Meier estimator:

$$Var\big(\hat{S}(t)\big) = \hat{S}(t)^2 \sum_{T_i \leq t} \frac{1}{Y(T_i)\big\{Y(T_i) - 1\big\}}. \tag{5.3}$$

Without censoring, Equation 5.3 is equivalent to standard binomial variance, $\hat{S}(t)\big(1 - S(t)\big)/n$. This can be shown as follows:

When no censoring occurs, the risk set at time T_{i+1} is simply the size of the risk set at time T_i minus the number of events that occurred at time T_i. Assuming no ties in the data, this means $Y(T_{i+1}) = Y(T_i) - 1$.

For $T_i \leq t < T_{i+1}$,

$$\begin{aligned}
\hat{S}(t)^2 \sum_{T_i \leq t} \frac{1}{Y(T_i)\{Y(T_i)-1\}} &= \hat{S}(t)^2 \sum_{T_i \leq t} \frac{Y(T_i) - Y(T_{i+1})}{Y(T_i)\{Y(T_i)-1\}} \\
&= \hat{S}(t)^2 \sum_{T_i \leq t} \left\{ \frac{1}{Y(T_{i+1})} - \frac{1}{Y(T_i)} \right\} \\
&= \hat{S}(t)^2 \left\{ \frac{1}{Y(T_{i+1})} - \frac{1}{Y(T_1)} \right\} \\
&= \hat{S}(t)^2 \left\{ \frac{Y(T_1) - Y(T_{i+1})}{Y(T_{i+1})Y(T_1)} \right\} \\
&= \hat{S}(t)^2 \left\{ \frac{1 - Y(T_{i+1})/Y(T_1)}{Y(T_1)Y(T_{i+1})/Y(T_1)} \right\} \\
&= \hat{S}(t)^2 \left\{ \frac{1 - \hat{S}(t)}{Y(T_1)\hat{S}(t)} \right\} \\
&= \frac{\hat{S}(t)(1 - \hat{S}(t))}{n},
\end{aligned}$$

the standard binomial variance.

5.2.3 The Nelson-Aalen estimator for cumulative hazard function

The Nelson-Aalen estimator for the cumulative hazard function $A(t)$, has the advantage of being a non-parametric estimator with no requirements for any assumptions on distribution (Borgan, 1997). This feature makes it suitable for a wide variety of cases where the distribution may not be known.

Given survival data without censoring, that is, n subjects or independent processes, with hazard function $\alpha(t)$ and cumulative hazard function

$$A(t) = \int_0^t \alpha(s)ds,$$

we can consider the occurrences of some event at times $T_1, T_2, \ldots, T_m$. We define

$$d_j = \#\text{events that occur at time } T_j$$

and

$$Y(T_j) = \#\text{subjects in the risk set immediately prior to } T_j.$$

The Nelson-Aalen estimator for the cumulative hazard function of the processes, is thus defined as

$$\hat{A}(t) = \sum_{T_j \leq t} \frac{d_j}{Y(T_j)}.$$

5.2.4 Variance of the Nelson-Aalen estimator

The variance of the estimator $\hat{A}(t)$ can itself be estimated by

$$\hat{\sigma}^2(t) = \sum_{T_j \leq t} \frac{(Y(T_j) - d_j)d_j}{(Y(T_j) - 1)Y(T_j)^2},$$

and both the Nelson-Aalen estimator and its variance estimator are close to unbiased (Borgan, 1997). We later call this standard variance estimator $V_S(t)$. Further, for large samples, the distribution of the Nelson-Aalen estimator at a time t is approximately normal. Hence we can define a standard $100(1-\alpha)\%$ confidence interval for $A(t)$ as

$$\hat{A}(t) \pm z_{1-\alpha/2}\hat{\sigma}(t)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ fractile of the standard normal distribution.

5.3 Comparing intensity and hazard rate models

The difference between the intensity function and the hazard function is subtle but critical. Here, we have mentioned the intensity function in the context of a counting process which produces multiple events, obtained by summing multiple survival processes. Similarly, the intensity function applies to Poisson processes, in which there are multiple arrivals over some period.

We can think of intensity as the number of events per unit time. This function may vary over time $\lambda(t)$ or may be a constant $\lambda > 0$, for example in a homogeneous Poisson process. We can then define the expected number of events or arrivals in some interval $(t_1, t_2]$ to be

$$E[N(t_1, t_2)] = \int_{t_1}^{t_2} \lambda(t)dt.$$

In contrast, in a survival process, we see only a single event before the system discontinues. The hazard rate function gives the conditional probability of an event occurring in the

time interval $(t, t + \delta t]$ given the that the event has not previously occurred,

$$\alpha(t) = \frac{f(t)}{1 - F(t)}$$

which integrates to the cumulative hazard function, $A(t)$, as previously discussed.

5.4 Nelson-Aalen for a Gaussian random field

To apply the Nelson-Aalen estimator to a Gaussian random field on a discrete space, we must consider sites on the field as individuals or subjects, each of which can experience an event. We then can consider the field value at each site as analogous to the event time in standard survival theory. In this scenario, an event will occur at every site at the field value for the site in question. For the following section, we consider this 'all events' data, where all sites experience an event. Later, we will consider an alternative definition of an event on a random field, building on the topological concepts discussed earlier.

5.4.1 Application

Given the relationship between event times in standard survival theory and the field values in our application, we could transform our Gaussian field values to log-normal survival times. Consider the random field $Z \sim N(0, \Sigma)$, where Σ is the covariance matrix. Survival times are then defined as $t_i = e^{z_i}$. However, for simplicity, in general we work with the original z_i which can have negative values. In the following analysis, we consider the untransformed data values, z_i. The same methodology applies, and this more closely reflects the type of 'real world' data with which we will be working in future.

5.4.2 Expectation

For spatially independent survival times $Z \sim N(0, \Sigma)$, where $\Sigma = I$ is the identity matrix of appropriate size, calculating the expectation of the Nelson-Aalen estimate (the theoretical cumulative hazard function) is straightforward.

We simply have

$$E[\hat{A}(t)] = A(t) = \int_{-\infty}^{t} \frac{\phi(u)}{1 - \Phi(u)} du$$

where ϕ and Φ are standard Normal density and distribution functions. An example of this is shown in the top-left plot of Figure 5.1. The result does not hold when there is spatial correlation in the random field, i.e. when $\Sigma \neq I$, corresponding to correlated event times.

This can be seen clearly in Figure 5.1, where each of the four figures shows data simulated using an exponential correlation function with different correlation length parameters, $exp(\nu = 0.5, \eta)$. Each figure shows 1000 1-d simulations each of length 192. As the correlation length parameters increase, we see that the mean of the Nelson-Aalen estimates (green dots) gets further from the expected cumulative hazard curve (red line). This is not a surprising result, as we make no assumption of spatial correlation in the calculation of the expected cumulative hazard curve. In this application where all sites experience an event, the finishing point of each Nelson-Aalen estimate is always at $\hat{A}(t_{max}) = 5.84$, regardless of the maximum field value t_{max}. This is due to the calculation of the Nelson-Aalen estimator as the cumulative sum of the inverse of the risk sets. That is, for this 1-d data of length 192 (still assuming no ties),

$$\hat{A}(t_{max}) = \sum_{u=1}^{192} \frac{1}{u} = 5.84.$$

5.4.3 Performance of the naive variance estimator

Further, in Figure 5.1, we show as green bands the observed mean plus and minus two (observed) standard deviations. The blue bands show the observed mean plus and minus two standard deviations as calculated by the naive Nelson-Aalen variance estimator. We can see that the standard Nelson-Aalen variance estimator does not perform well for correlated data, and we see significant under-coverage. Further, as the correlation length parameter increases, we see an increasing empirical variance.

In order for the Nelson-Aalen to provide value as a method for events on a correlated random field, we require an adjustment for correlation.

5.5 Dealing with spatial correlation

The evolution of survival analysis as a field from one considering single event processes to multi-event processes has required the development of new techniques and methods. Similarly, where conventional survival theory concerns itself with (spatially) independent processes, our interest in spatially correlated events requires further examination of existing methods.

Some work has touched upon the idea of correlated survival times. For example, Williams (1995) describes a novel approach to estimating the variance of correlated product-limit survival times, where Greenwood's formula is not appropriate due to the requirement for independent survival times. However, Williams develops a robust variance estimator using

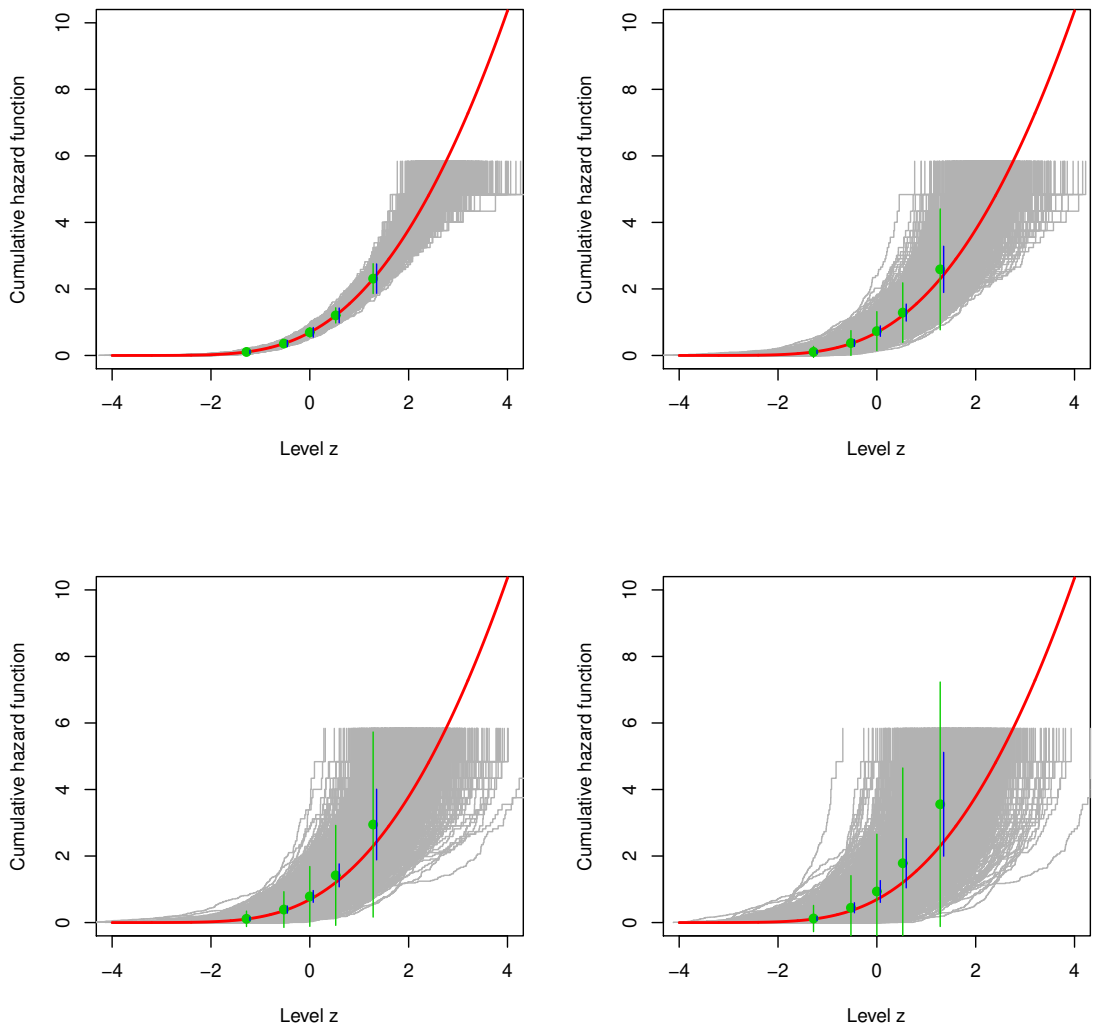

Figure 5.1: Expected cumulative hazard function and Nelson Aalen estimates for 1000 simulations of length 192 under different correlation parameters. Top-left: independent, top-right: $exp(\nu = 0.5, \eta = 10)$, bottom-left: $exp(\nu = 0.5, \eta = 20)$, bottom-right: $exp(\nu = 0.5, \eta = 50)$. The red lines show the expected cumulative hazard function under independence, and the green dots show five pointwise averages for the calculated Nelson-Aalen estimates, calculated at $\Phi^{-1}(0.1)$, $\Phi^{-1}(0.3)$, $\Phi^{-1}(0.5)$, $\Phi^{-1}(0.7)$ and $\Phi^{-1}(0.9)$. The green bands represent the observed mean plus and minus two (observed) standard deviations. The blue bands represent the observed mean plus and minus two standard deviations as calculated by the naive Nelson-Aalen variance estimator.

Taylor series linearised values and a commonly used between-cluster variance estimator often seen in multi-stage surveys. The proposed methods are useful for cluster-correlated survival times but do not allow for spatial correlation between individual observations.

Other work by Li and Ryan (2002) uses frailty models in which there is spatial correlation of the random effects. Random effects with spatial correlations are multiplicatively incorporated into the baseline hazard, without any requirement for parametric forms to be assumed. The research by Li and Ryan includes simulations with a range of different spatial correlation structures. However, less well researched is data where there is a spatial relationship between all individuals, i.e., scenarios where correlation exists between all survival times and is related to the physical distance between individuals.

5.5.1 Mean adjustment

Before working with the data, we correct for the mean of the event times. We have $Z_1, \ldots, Z_n$ marginally $N(0,1)$ with $n \times n$ correlation matrix, R. Let $W_i = Z_i - \bar{Z}$, $Z = \{Z_i\}$ and $W = \{W_i\}$, $i = 1, \ldots, n$. Then with I_n being the identity matrix and J_n being an $n \times n$ matrix of ones, we have $W = (I_n - J_n/n)Z$ and so

$$W \sim N(0, C), \qquad C = (I_n - J_n/n)R(I_n - J_n/n)^T.$$

5.5.2 Expectation for correlated data

For correlated data after mean-adjustment, we have a marginal distribution $W_i \sim N(0, c_i)$, where c_i is the ith diagonal element of C. Hence

$$E[\hat{A}(t)] = A(t) = \int_{-\infty}^{t} \frac{\phi(u; 0, \sqrt{c_i})}{1 - \Phi(u; 0, \sqrt{c_i})} du.$$

5.5.3 Adjusted variance estimator

In order to estimate variance where individuals are spatially correlated, we look at a fixed time, t. For an individual j, we have

$$N_j(t) = I(W_j < t),$$

equal to one if an event has occurred and zero if not. Given $W \sim N(0, C)$, we can calculate

$$e_j = P(W_j < t) \text{ and } e_{jk} = P(W_j < t, W_k < t).$$

Since

$$N(t) = \sum_j N_j(t),$$

we have

$$\begin{aligned} Var(N(t)) &= \sum_j Var(N_j(t)) + \sum_{j,k} Cov(N_j(t), N_k(t)) \\ &= \sum_j e_j(1-e_j) + 2\sum_{k>j}(e_{jk} - e_j e_k). \end{aligned}$$

By definition,

$$\hat{S}(t) = 1 - \frac{N(t)}{n}$$

and so

$$\begin{aligned} Var(\hat{S}(t)) &= Var\left(1 - \frac{N(t)}{n}\right) \\ &= \frac{Var(N(t))}{n^2}. \end{aligned}$$

The Nelson-Aalen estimate, by definition is

$$\hat{A}(t) = -log(\hat{S}(t))$$

and using a delta-method approximation (see Appendix C) we can obtain the adjusted variance estimator

$$\mathrm{V}_{ADJ}(\hat{A}(t)) = \frac{Var(N(t))}{n^2(1-\frac{N(t)}{n})^2}.$$

5.5.4 Performance

Known correlation

We run initial tests using the known correlation structure of the data to obtain the variance estimate described above and assess how the coverage of plus or minus two standard deviations based on our two variance estimators compares to that of intervals based on plus or minus two empirical standard deviations. Tables 5.1 and 5.2 show the results for a selection of correlation parameters and grid sizes of $n = 200$ and $n = 500$ respectively. We compare results at five levels as before, $\Phi^{-1}(0.1)$, $\Phi^{-1}(0.3)$, $\Phi^{-1}(0.5)$, $\Phi^{-1}(0.7)$ and $\Phi^{-1}(0.9)$.

As we would expect, V_E has good coverage for all levels and parameter sets, for both grid sizes. We see significant undercoverage when using V_S particularly further from

Level	$\Phi^{-1}(0.1)$	$\Phi^{-1}(0.3)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.7)$	$\Phi^{-1}(0.9)$
			$\nu = 1$, $\eta = 5$		
V_E	95.8	95.5	94.3	95.9	96.1
V_S	77.0	91.2	95.7	92.3	80.0
V_{ADJ}	95.6	95.7	94.8	95.6	94.7
			$\nu = 1$, $\eta = 10$		
V_E	97.3	95.0	94.8	94.8	95.5
V_S	61.4	79.6	87.7	78.9	64.6
V_{ADJ}	98.1	96.1	96.4	96.2	93.2
			$\nu = 2$, $\eta = 5$		
V_E	94.9	96.2	95.1	95.1	94.0
V_S	74.4	87.2	91.5	87.9	74.5
V_{ADJ}	95.7	95.9	94.1	95.7	93.7

Table 5.1: Coverage of two standard deviations using empirical (V_E), standard Nelson-Aalen (V_S) and adjusted (V_{ADJ}) variances for a selection of correlation parameters and grid size $n = 200$. Here we use the Matérn correlation function.

Level	$\Phi^{-1}(0.1)$	$\Phi^{-1}(0.3)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.7)$	$\Phi^{-1}(0.9)$
			$\nu = 1$, $\eta = 5$		
V_E	96.6	95.8	95.0	95.2	95.8
V_S	77.4	90.0	95.5	91.8	77.1
V_{ADJ}	96.4	94.5	95.5	95.8	95.5
			$\nu = 1$, $\eta = 10$		
V_E	94.5	95.1	95.4	95.1	94.0
V_S	61.9	78.0	85.5	76.4	64.0
V_{ADJ}	95.2	94.4	95.1	93.8	95.6
			$\nu = 2$, $\eta = 5$		
V_E	94.6	95.8	95.2	95.7	94.5
V_S	75.8	87.4	92.7	86.5	74.7
V_{ADJ}	95.1	96.1	96.0	94.3	94.3

Table 5.2: Coverage of two standard deviations using empirical (V_E), standard Nelson-Aalen (V_S) and adjusted (V_{ADJ}) variances for a selection of correlation parameters and grid size $n = 500$. Here we use the Matérn correlation function.

the mid-level, again seeing similar results for both $n = 200$ and $n = 500$ grid sizes. Our adjusted variance estimator V_{ADJ} gives good coverage generally, although we see slight undercoverage at the last shown level and slight overcoverage for for first level and parameters $\eta = 1, \nu = 10$ for grid size $n = 200$. For grid size $n = 500$ coverage is much improved in both of these cases.

Estimated correlation

We have seen that performance of the adjusted variance estimator is generally good, given known correlation. The following results measure coverage equivalently to the results above; however, we estimate the correlation parameters from the mean-corrected data, here assuming a Matérn correlation function. Tables 5.3 and 5.4 show the results for the selection of parameters used above and grid sizes of $n = 200$ and $n = 500$ respectively, each based on 1000 simulations.

Level	$\Phi^{-1}(0.1)$	$\Phi^{-1}(0.3)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.7)$	$\Phi^{-1}(0.9)$
			$\nu = 1$, $\eta = 5$		
V_E	95.5	95.9	95.2	95.3	94.3
V_S	77.3	90.4	96.6	91.8	78.9
V_{ADJ}	96.8	96.3	96.3	96.1	93.5
			$\nu = 1$, $\eta = 10$		
V_E	95.6	95.0	94.9	95.7	95.2
V_S	60.3	80.5	89.0	79.7	63.0
V_{ADJ}	96.3	96.7	97.0	95.5	91.4
			$\nu = 2$, $\eta = 5$		
V_E	95.3	94.9	95.1	95.6	94.9
V_S	73.4	87.8	93.1	88.5	75.3
V_{ADJ}	95.9	95.9	95.1	95.8	93.9

Table 5.3: Coverage of two standard deviations using empirical (V_E), standard Nelson-Aalen (V_S) and adjusted (V_{ADJ}) variances over 1000 simulations for a selection of correlation parameters and grid size $n = 200$, with correlation estimated from the data. Here we use the Matérn correlation function.

For grid size $n = 500$, we see similarly good results using estimated correlation as we saw for known correlation. However for $n = 200$ we see more noticeable undercoverage for the adjusted variance estimator in the last shown level as well as slightly more overcoverage at lower levels. Here, we are estimating correlation parameters based on a known Matérn correlation structure. For data with unknown correlation function we might expect to see a worse performance of the adjusted variance estimator when assuming a Matérn correlation

Level	$\Phi^{-1}(0.1)$	$\Phi^{-1}(0.3)$	$\Phi^{-1}(0.5)$	$\Phi^{-1}(0.7)$	$\Phi^{-1}(0.9)$
			$\nu = 1$, $\eta = 5$		
V_E	95.3	95.3	95.1	94.9	95.5
V_S	77.1	92.0	96.7	89.5	80.2
V_{ADJ}	95.0	96.0	96.5	95.0	95.3
			$\nu = 1$, $\eta = 10$		
V_E	95.1	94.7	94.9	95.3	95.6
V_S	61.2	77.0	86.3	76.1	62.3
V_{ADJ}	95.5	95.5	95.2	95.3	94.1
			$\nu = 2$, $\eta = 5$		
V_E	95.3	95.3	95.6	95.0	95.0
V_S	74.0	88.4	92.1	88.1	73.5
V_{ADJ}	95.5	95.5	94.7	95.0	94.3

Table 5.4: Coverage of two standard deviations using empirical (V_E), standard Nelson-Aalen (V_S) and adjusted (V_{ADJ}) variances over 1000 simulations for a selection of correlation parameters and grid size $n = 500$, with correlation estimated from the data. Here we use the Matérn correlation function.

function.

5.6 Random fields

To investigate the potential of the Nelson-Aalen estimator as a means for comparing random fields, we simulate five fields for illustration, each from a different distribution. A marginally Gaussian version of F2 was used previously in Section 3.4, calculated as

$$z^* = \Phi^{-1}(z),$$

where $\Phi()$ is the $N(0,1)$ distribution function. This will also be used in the following sections.

- F1: GRF $z \sim N(0, \Sigma)$ where Σ is Matérn with parameters ν, η.
- F2: χ^2_1, constructed from a single GRF z_1 as

  $$z = F_{\chi^2_1}(z_1^2),$$

  where $F_{\chi^2_1}$ is the χ^2_1 distribution function.
- F3: χ^2_3, constructed from three three independent GRFs, z_1, z_2, z_3 as

  $$z = F_{\chi^2_3}\left(\sum_{i=1}^{3} z_i^2\right),$$

  where $F_{\chi^2_3}$ is the χ^2_3 distribution function.
- F4: T_3, constructed from four independent GRFs, as

  $$z = F_{T_3}\left(\frac{z_1}{\left(\sum_{i=2}^{4} z_i^2/3\right)^{1/2}}\right),$$

  where F_{T_3} is the T_3 distribution function.
- F5: $F_{3,3}$, constructed from six independent GRFs, as

  $$z = F_{F_{3,3}}\left(\frac{\sum_{i=1}^{3} z_i^2/3}{\sum_{i=4}^{6} z_i^2/3}\right),$$

  where $F_{F_{3,3}}$ is the $F_{3,3}$ distribution function.

5.7 Power

To assess the effectiveness of the Nelson-Aalen estimator as a means of comparing the rate of occurrence of 'events' for sites on random fields (i.e. field values), we observe how the estimates differ for a range of distributions. Figure 5.2 shows the differences between Nelson-Aalen estimates for random fields over the five distributions outlined above. Each plot shows Nelson-Aalen curves for 100 2-d random fields each of size 200×200. Here we only consider random fields with no spatial correlation and at this point without transformation to marginal Gaussianity. It is clear that by this method the Nelson-Aalen estimator distinguishes between spatially independent Gaussian, χ^2_1, χ^2_3, T_3 and $F_{3,3}$ random fields.

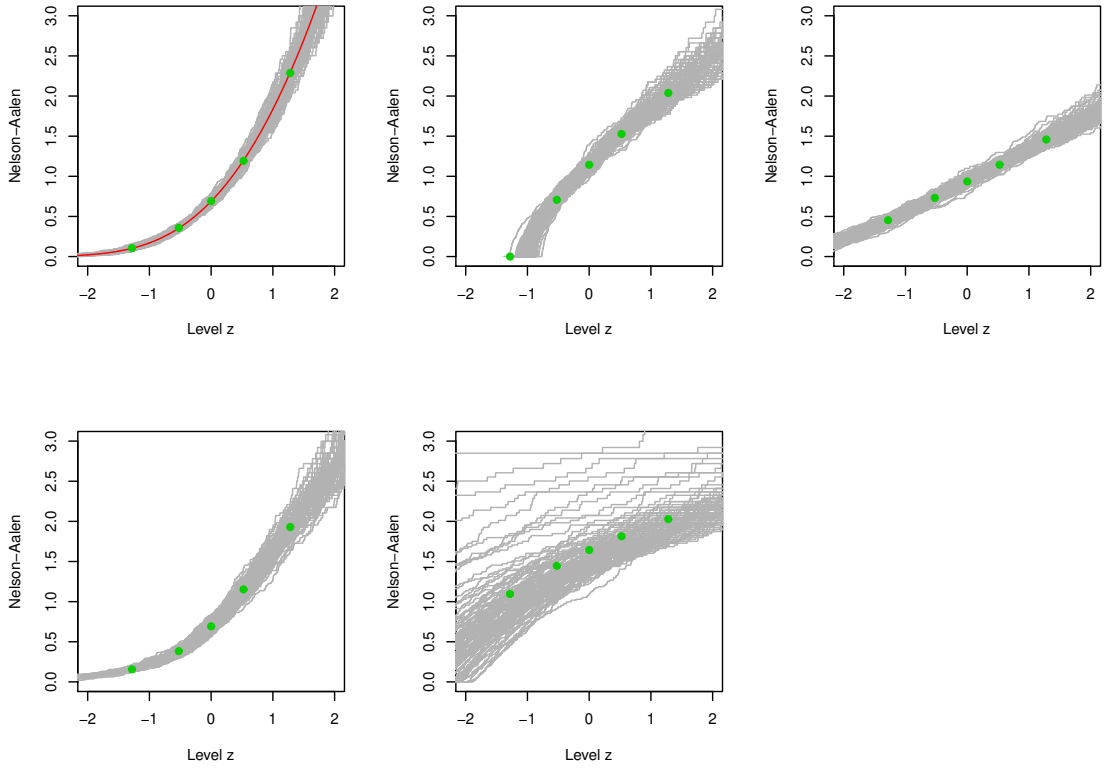

Figure 5.2: Nelson-Aalen estimates for independent random fields. Left-right, top-bottom: Gaussian, χ^2_1, χ^2_3, T_3, $F_{3,3}$. Green dots show the average of the Nelson-Aalen estimates at $\Phi^{-1}(0.1)$, $\Phi^{-1}(0.3)$, $\Phi^{-1}(0.5)$, $\Phi^{-1}(0.7)$ and $\Phi^{-1}(0.9)$.

However, we are primarily interested in the performance of the Nelson-Aalen estimator as a means to distinguish between fields when they are marginally Gaussian. Next, we look at marginally Gaussian random fields with matched correlation.

5.7.1 Comparing Gaussian and marginally Gaussian χ^2_1 random fields with matched correlation

We look at correlated random fields, for which the Matérn covariance parameters have been chosen such that the correlation for the Gaussian and χ^2_1 fields are close. We use the parameters described previously in Section 3.4. Without marginal transformation, there is a significant difference between Nelson-Aalen estimates as can be seen in the two left-hand plots of Figure 5.3. However, following marginal transformation of χ^2_1 data with matched correlations, there is not a clear difference between Nelson-Aalen estimates.

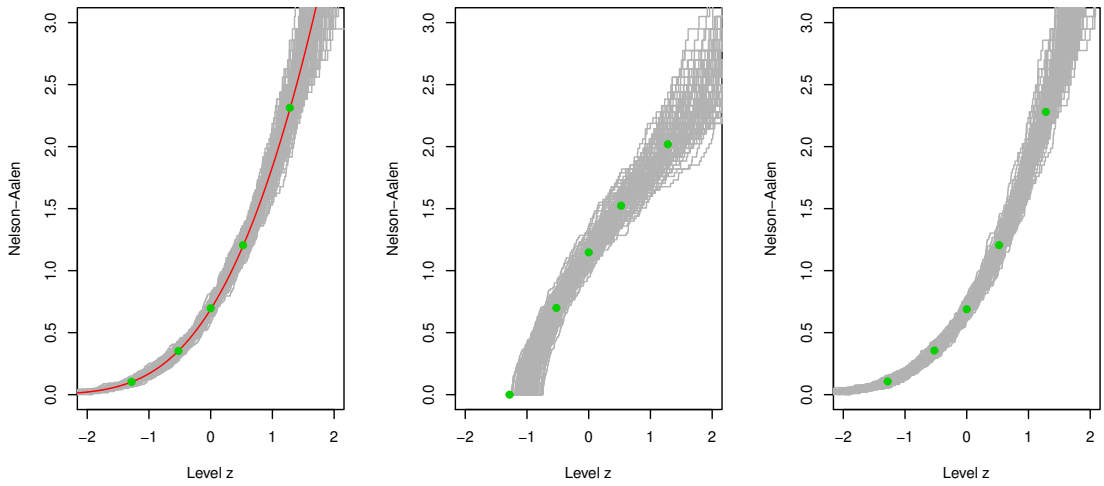

Figure 5.3: Left to right: Gaussian, χ^2_1 with matched correlation structure, marginally Gaussian χ^2_1 with matched correlation structure. Green dots show the average of the Nelson-Aalen estimates at $\Phi^{-1}(0.1)$, $\Phi^{-1}(0.3)$, $\Phi^{-1}(0.5)$, $\Phi^{-1}(0.7)$ and $\Phi^{-1}(0.9)$.

Evidently the ability of the Nelson-Aalen estimator to distinguish between random fields from different distributions is limited to cases in which the non-Gaussian distributions have not been transformed to marginal Gaussianity. Ideally, we would like a method that is able to distinguish these fields where correlation is matched and where both fields have Gaussian marginals.

5.8 Conclusions

In this chapter we introduced ideas from survival analysis such as counting processes and martingales. We discussed non-parametric estimators for the survival function and the cumulative hazard function of survival data, in particular the Nelson-Aalen estimator for the cumulative hazard function. We considered how data on a discrete random field could

be thought of in a survival context and looked at the application of both the Nelson-Aalen estimator and the Nelson-Aalen variance estimator in this scenario. Introduction of spatial correlation to the random fields illustrated the need for adjustments to both these estimators in order to be appropriate for spatially correlated data. Finally we demonstrated the effectiveness of these adjusted estimators for a selection of random fields from different distributions. Comparison of Gaussian and marginally Gaussian χ_1^2 fields with matched correlation demonstrated that applying the Nelson-Aalen to all values on a random field as events in a survival context was not sufficient to distinguish between these cases. In the following chapter we introduce methods to address this, returning to ideas from topological data analysis.

Chapter 6

Topological event history analysis

Although topological data analysis has wide-ranging applications, it is yet to be seen frequently in statistical literature. We show how standard methods from event history analysis can be applied to topological features, developing methods with relatively low computational overheads for the comparison of random fields. In this chapter, we demonstrate the potential of the method via a simulation study and apply it to our wind intensities data. We further show how topological event history methods can be used to assess the local-level fit of correlation models to data. Work in this chapter is included in our paper, 'Event History and Topological Data Analysis' (Garside et al., 2020).

6.1 Nelson-Aalen estimates for births of connected components

Throughout Chapter 3, we were interested in the persistence of topological features, defined by the difference between their time of birth and time of death. In Chapter 5, we looked at standard event history analysis as well as applications to random fields to understand the emergence of events, particularly in the context of spatially correlated data. Here, we consider how the application of event history methods to birth levels of topological features in data can be used for comparison of random fields.

In contrast to the previous chapter, in which we considered every site on the discrete grid to experience an event at its field value, we can define events as the emergence of local minima or maxima in the data. That is, an event only occurs when the field value is a local maximum or minimum. On a 1-d random field, this is clear to picture; a site is a local minimum if its value is lower than that of its two neighbours. In 2-d this definition is less clear. How do we define the neighbourhood of a site? We could define as its neighbours

all sites with a shared edge, or all sites diagonally adjacent (in both cases the number of neighbours is four). Alternatively, we could include both adjacent and diagonally adjacent sites in this definition, resulting in a neighbourhood of size eight.

In this chapter, we treat random fields on a grid as we did in Chapter 3, with sites connected by shared edges such that each site has four neighbours, excluding edge cases. This choice of neighbourhood allows a constant distance (if we assume constant spacing between rows and columns of the grid) between the site x_i of interest and all its neighbours, $x_{(i)}$, as would not be the case with all choices. The events of interest in this work are the birth times of connected components, although the ideas could be applied to other features and death times instead of births. We assume continuous values on the field so will not encounter tied event times.

Given a filtration of a random field Z from below, as described in Section 3.2, the birth of a connected component at site x_i and level t corresponds to a local minima of the field with value $z_i = z(x_i)$, where a site being a local minima can be understood as the site having lower value that those four sites $x_{(i)}$, with which it shares an edge ($x_{(i)}$ would comprise fewer than four sites for boundary cases).

Figure 6.1 shows, for an example 10×10 random field, a sequence of level sets, at each of which a new connected component is born. The top left image shows data on a discrete random field, where darker pixels represent higher values and lighter pixels, lower values. We can see the three lowest values which correspond to the first three connected components being born in each of the first three level sets shown. In the forth level set (top-right figure) we see not only the birth of an additional connected component but the growth of the first to be born. After this we see additional growth of existing connected components as well as new components being born (by design, a new birth in each level set shown). By the final level set shown, we see one large connected component of 28 sites, along with eight additional smaller connected components. In these level sets, we see no examples of holes, areas fully surrounded by connected components.

We wish to apply the standard Nelson-Aalen estimator as in Chapter 5, however counting only births of connected components instead of values at all sites. In Chapter 5 we applied survival techniques to spatial data on a discrete grid by treating every site as experiencing an event, with time corresponding to field value. Here, we treat a site as experiencing an event, only when a connected component is born at that site. For example, one event would occur at each of the level sets shown in Figure 6.1 corresponding to the birth of each new connected component. As before, the event time correspond to the field value.

A key difference from the previous approach is the number of sites that will experience an event. Whereas in the 'all events' case we knew that the total number of events is equal to the number of sites on the grid, here the number of events will vary depending on the

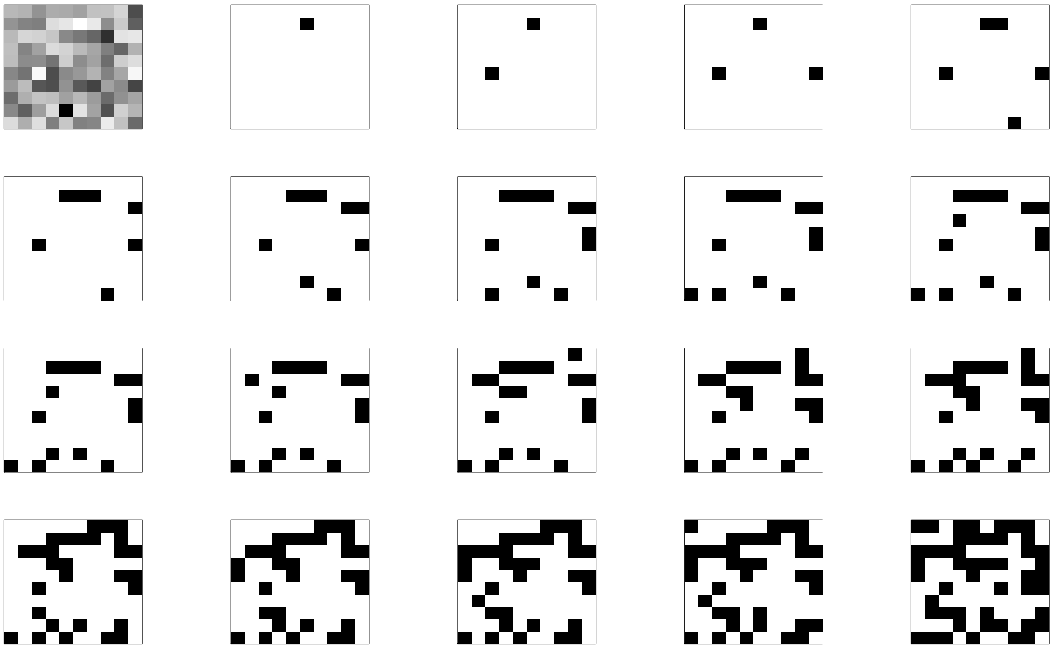

Figure 6.1: The top left image shows an example of a random field. Each subsequent image shows a level set of the field and the birth of a new connected component.

structure of the data. This is related to the number of local minima observed as discussed in Chapter 3 and will differ with different distributions and correlation structures.

Consider the predictable at-risk indicator function of site x_i at level t, i.e. whether at time t is is possible for site x_i to be a local minimum,

$$Y_i(t) = \begin{cases} 1 & z_i \geq t, \quad z_{(i)} \geq t1_i \\ 0 & \text{otherwise} \end{cases}$$

where 1_i is a unit vector the same length as $z_{(i)}$.

Then

$$N_i(t) = \begin{cases} 1 & z_i \leq t, \quad z_{(i)} > z_i 1_i, \\ 0 & \text{otherwise} \end{cases}$$

is the number of connected components born at site x_i up to and including level t.

Summing over all sites on the field, we have

$$Y(t) = \sum_i Y_i(t)$$

and

$$N(t) = \sum_i N_i(t).$$

Then as seen for discrete values in Section 5.2.3,

$$\hat{A}(t) = \int_{-\infty}^{t} \frac{dN(u)}{Y(u)}$$

is the Nelson-Aalen estimator for the cumulative hazard function, in this case for the birth times of connected components when filtering from below (corresponding to local minima).

6.2 Obtaining variance estimates

Obtaining accurate variance estimates for the Nelson-Aalen estimator is non-trivial. As we saw when applied to all sites on a random field in Chapter 5, the standard Nelson-Aalen variance estimator provides significant under coverage for the variance of the Nelson-Aalen estimator where data are spatially correlated. Since we are looking at events on a correlated random field, we can assume non-independence between the locations of births.

Further, the adjusted variance estimator presented in Section 5.5.3 underestimates the variance (results not shown), likely as a result of temporal non-independence, resulting in informative 'censoring'. That is, for some site x, the knowledge that x is no longer in the risk set at time t tells us something about other locations. Although this is not censoring the data directly, we observe a 'knock-out' effect, whereby neighbours of a site where a birth event occurs can no longer be in the risk set. Hence the rate at which births occur significantly affects the size of the risk set. This interesting feature cannot be treated as censoring as is commonly studied in survival analysis, partly due to the correlation between 'censored' sites and sites at which events occur, but invalidates our existing variance estimators and hence has implications when attempting to obtain confidence intervals for our methods.

6.2.1 Pointwise confidence intervals

Multiple replicates

Consider N iid replicates of a random field, $Z_1, \ldots, Z_N$ on a common space $\mathcal{X}$, with corresponding Nelson-Aalen estimators $\hat{A}_1(t), \ldots, \hat{A}_N(t)$. Then each $\hat{A}_i(t)$ is non-negative and bounded above and for all t has finite variance. Hence, the standard central limit theorem can be applied, and pointwise inference for large N is straightforward using the

sample mean $\bar{A}(t)$ and the sample standard deviation.

Single replicate

There are two ways in which we obtain pointwise estimates of the variance from a single replicate, depending on our knowledge of a parametric model for the data. The assumption of a parametric model for the underlying data field allows us to use a parametric bootstrap in order to obtain a variance estimate. Assuming some known parametric model F_θ for a random field Z, we can use a parametric bootstrap approach to obtain pointwise confidence intervals. Given a consistent estimator $\hat{\theta}$ for θ, we can simulate N realisations of Z from F_θ and obtain corresponding bootstrap Nelson-Aalen estimators, $\hat{A}_i(t)$, $i = 1, \ldots, N$. We can then calculate pointwise confidence intervals using bootstrap quantiles, which have asymptotically correct coverage in N.

Alternatively, in cases where we have no parametric model and only a single replication, we might assume some parametric model and account for local model uncertainly via doubling the variance as proposed by Copas and Eguchi (2005). We would expect this approach to provide a conservative variance estimate.

6.2.2 Simultaneous confidence bands

In addition to estimating pointwise confidence intervals, we consider the estimation of simultaneous confidence bands. Discontinuities in $\hat{A}(t)$ require us to be more careful, since there is no guarantee of a central limit theorem for uniformly bounded cadlag processes, dependent on the nature of the discontinuities (Hahn, 1977). Given certain sufficient conditions provided by Bloznelis and Paulauskas (1994), $\sqrt{N}\bar{A}(t)$ does converge weakly to a Gaussian process with finite variance and smooth sample paths in $(0, \tau)$ for any arbitrarily large τ for which $\text{pr}\{(Y(\tau) > 0\} > 0$. We verify these conditions in Appendix D. Thus, a functional central limit theorem applies.

Additionally, for a fixed space $\mathcal{X}$ we can show that $\bar{A}(t)$ is consistent in N for

$$A(t) = \int_{-\infty}^{t} J(u) \frac{\sum_x P\left(z_{(x)} > u1_x \mid z_x = u\right) f_x(u)}{\sum_x P\left(z_x > u,\ z_{(x)} > u1_x\right)} du,$$

where $f_x(.)$ is the marginal density of z_x and $J(u)$ is the indicator $J(u) = I(Y(u) > 0)$. If there is only a single realisation, the estimator $\hat{A}(t)$ is also consistent for $A(t)$, but this time as the cardinality of $\mathcal{X}$ increases, provided $\text{cor}(z_x, z_{x'})$ approaches zero as the distance between x and x' increases.

Again we consider two cases, firstly where we assume the existence of multiple replicates,

and secondly where we have access to only a single replicate.

Multiple replicates

Firstly, we assume the availability of multiple independent replicates, and hence multiple Nelson-Aalen estimates $\hat{A}_i(t)$, $i = 1 \dots B$ corresponding to fields Z_i. The Nelson-Aalen estimators have discontinuities at values corresponding to local minima, so considering the estimators $\hat{A}_i(t)$ as functional data, we discretise the level t of the estimator into M distinct levels, $t_1, \dots, t_M$, and let $\tau_1 = t_1$ and $\tau_2 = t_M$. Since we assume a continuous field, we can be certain that these discontinuities will not occur at any of the M levels. A common approach in functional data analysis, this enables us to handle the unavoidable presence discontinuities in the estimator. Whilst many techniques are available for the analysis of functional data (Wang et al., 2016; Chiou and Muller, 2009), we apply a method for the construction of simultaneous confidence bands as described by Degras (2011) and Cao et al. (2012).

Let $\hat{A}_{ij}$ be the Nelson-Aalen estimate for field Z_i and level t_j, with $\bar{A}_j$ the sample mean and $\hat{\sigma}_j^2$ the sample variance of the estimates at level t_j. Additionally, let $\hat{\rho}_{jk}$ be the estimated correlation between estimates $\hat{A}_{\cdot j}$ and $\hat{A}_{\cdot k}$. As shown by Degras (2011), when τ_1 and τ_2 are fixed and B and M are increasing,

$$\bar{A}_j \pm c_\alpha \frac{\hat{\sigma}_j}{\sqrt{B}}$$

has asymptotic coverage $1 - \alpha$, where c_α is the upper α-quantile of the maximum absolute value over $(\tau_2 - \tau_1)/\tau_2$ of a Gaussian process with standard margins and with correlation function equal to an appropriately scaled limit of $\hat{\rho}_{jk}$.

Parametric bootstrap for a single replicate

For a single replicate, as with the pointwise confidence intervals, we assume availability of a parametric model F_θ and a consistent estimator $\hat{\theta}$ for θ. We then take a bootstrap approach, generating B realisations of Z from F_θ and obtaining corresponding Nelson-Aalen estimates for each Z_i, $i = 1 \dots B$. Taking a Monte Carlo approach (Crainiceanu et al., 2012) avoids the need to explicitly estimate the covariance function. As with multiple replicates, we discretise the levels of the estimators to M distinct levels, $t_1, \dots, t_M$ where $\tau_1 = t_1$ and $\tau_2 = t_M$. Let $\hat{A}_j$ be the original Nelson-Aalen estimate at level t_j, and $\hat{A}_{ij}$ the Nelson-Aalen estimate at level t_j for bootstrap replicate i. Then let $\bar{A}_j$ and $\hat{\sigma}_j$ be the bootstrap mean and standard deviation at t_j.

We define

$$G_i = \max_j \left\{ \frac{|\hat{A}_{ij} - \hat{A}_j|}{\hat{\sigma}_j} \right\},$$

taking $\hat{d}_\alpha$ to be the upper α-quartile of the empirical distribution of $G_1, \ldots G_B$. Then for large M an approximate simultaneous confidence band for $A(t)$ over the interval (τ_1, τ_2) is

$$\hat{A}_j \pm \hat{d}_\alpha \hat{\sigma}_j \qquad j = 1, \ldots, M.$$

6.3 Application to random fields

We illustrate the use of topological event history analysis via a selection of simulated, marginally $N(0,1)$ random fields. Each is stationary and isotropic, on a discrete 60×60 grid and their correlation functions are indistinguishable by design. We additionally conducted the following simulation study on a 50×50 grid and obtained similar results. We simulate data from three models. Model 1 is a stationary and isotropic Gaussian random field with standard $N(0,1)$ marginals and Matérn correlation function with $\theta_1 = (\eta, \nu)$ Model 2 is a marginally Gaussian χ^2_1 random field and Model 3 is a marginally Gaussian $F_{3,3}$ random field. The fields are simulated from the three models as described in Section 5.6 and the underlying Gaussian fields used for Model 2 and Model 3 have Matérn correlation with parameters θ_2 and θ_3 respectively. We consider three sets of parameters, as shown in Table 6.1 where for each θ_1, θ_2 and θ_3 are chosen such that the correlation of the resulting Model 2 and Model 3 fields is as close as possible to that of the Model 1 field. This choice of correlation allows us to evaluate the topological event history methods beyond simply comparing marginal distributions, and show that differences in topological results are not due to differences in correlations.

Figure 6.2 shows examples of these random fields. In panel (a) are data simulated from Model 1, in (b) and (c) are data independently simulated from Model 3 and panel (d) shows data simulated from Model 2, using the first set of parameters shown in Table 6.1,$\theta_1 = (5.0, 1.0), \theta_2 = (8, 4, 2.2), \theta_3 = (9, 1.2)$. Panel (d) is immediately distinguishable from the others with the appearance of filaments in the image, however there is no obvious difference between panels (b) and (c), and the Gaussian random field in panel (a).

Correlation parameters

The simulations in this section use three choices of parameter θ_1 for the Gaussian random field Model 1. The values of θ_2 and θ_3 that we used in generating Model 2 and Model 3 are given in Table 6.1. We chose these to match the final correlation structure of Model 2 and Model 3 simulations to that of Model 1, as described in Section 3.5.2. Table 6.2 confirms

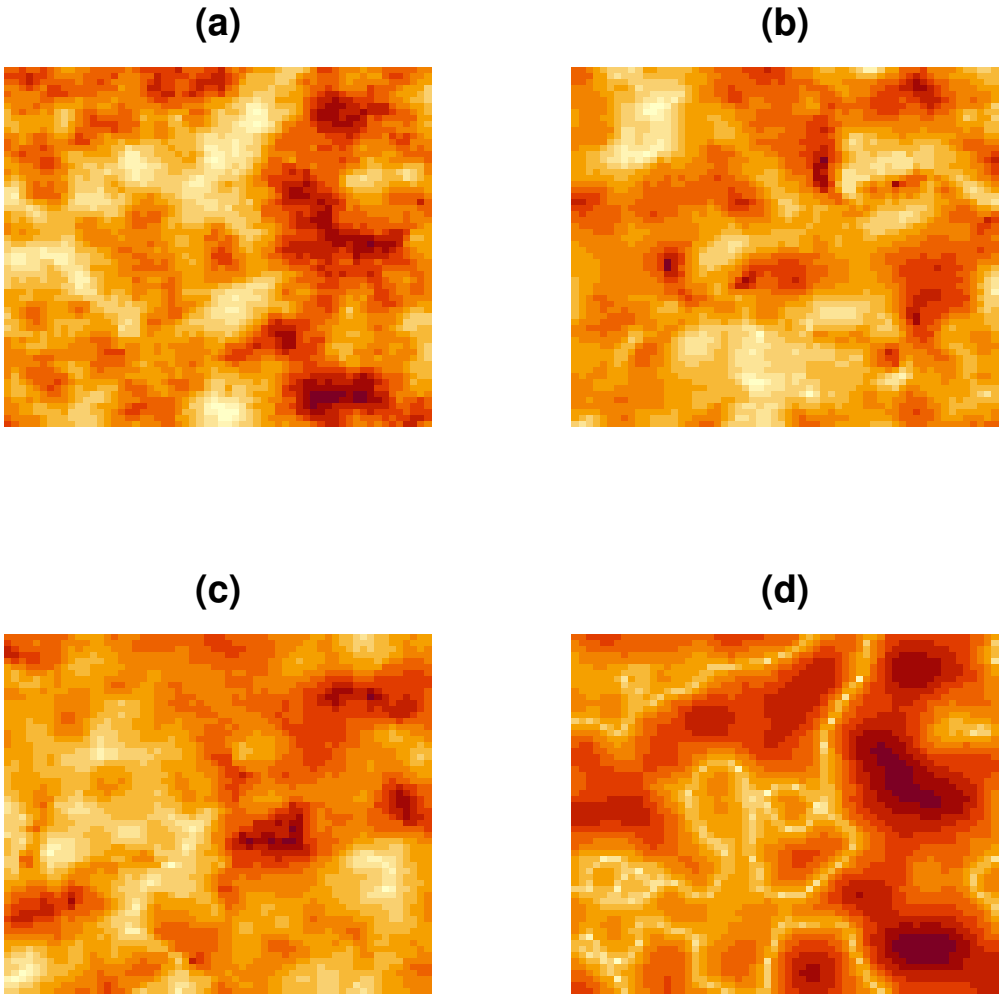

Figure 6.2: Four simulated random fields with matched correlations, each on a 60×60 lattice. Panel a: Model 1, b and c: Model 3, d: Model 2.

that there are no important systematic differences between the correlation functions of simulations from Model 2 and Model 3 and that of the Gaussian random field simulated from Model 1.

We saw in the previous chapter how choice of distribution and correlation structure impacts the expected cumulative hazard functions when all sites experience an event. Figure 6.3 shows for Gaussian random fields with Matérn correlation functions, the impact of correlation parameter choices on the expected cumulative hazard curves for event history data. We can see the difference between the curves for uncorrelated and spatially correlated data, with each parameter set resulting in a curve further from the uncorrelated case.

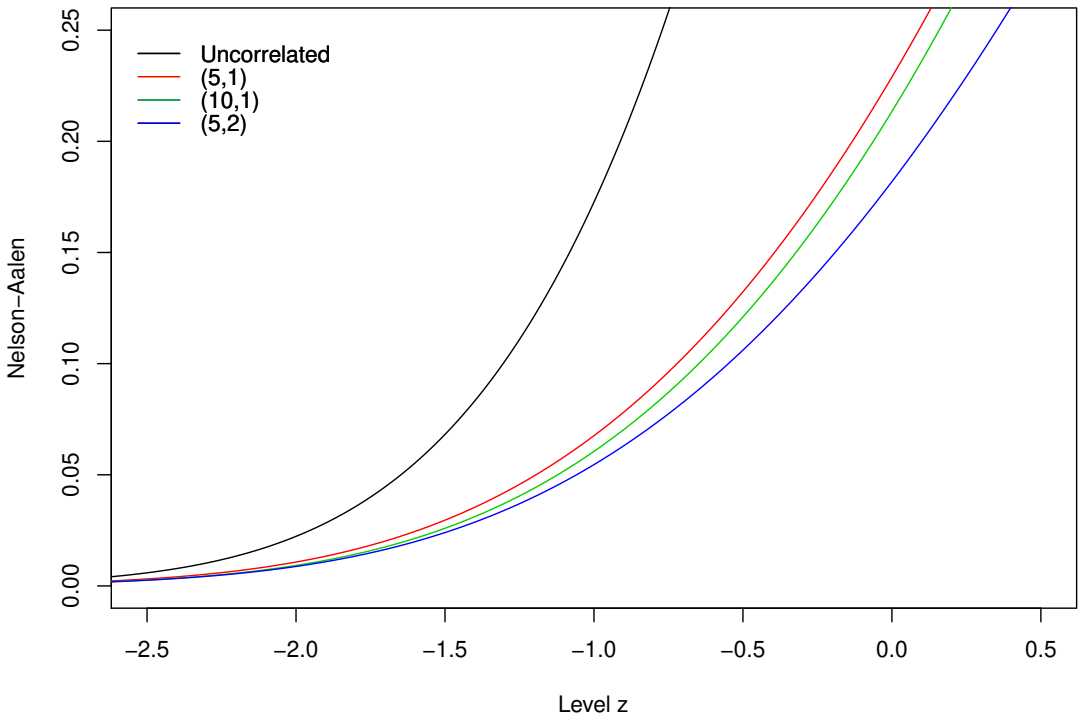

Figure 6.3: Excepted cumulative hazard curves for Gaussian random fields with Matérn correlation structure using the parameters shown.

Table 6.1: Parameter values for simulations in Section 6.3. For each model M1, M2, and M3 with corresponding parameters θ_1, θ_2, θ_3, we consider three choices of parameters $\theta = (\eta, \nu)$.

Model parameters	(η, ν)	(η, ν)	(η, ν)
θ_1	(5.0,1.0)	(10.0,1.0)	(5.0,2.0)
θ_2	(8.4,2.2)	(16.5,2.4)	(8.4,5.1)
θ_3	(9.0,1.2)	(18.7,1.1)	(8.0,3.6)

Table 6.2: True and empirical correlations for the simulation models of Section 6.3. The empirical values are the means of $100,000$ simulations

	Distance						
	1	2	3	4	5	10	20
True M1	0.924	0.798	0.668	0.548	0.444	0.140	0.011
Empirical M1	0.924	0.797	0.666	0.546	0.441	0.137	0.010
Empirical M2	0.925	0.810	0.685	0.563	0.455	0.123	0.000
Empirical M3	0.930	0.801	0.665	0.541	0.431	0.127	0.004
True M1	0.974	0.924	0.863	0.798	0.732	0.444	0.140
Empirical M1	0.974	0.924	0.863	0.797	0.731	0.444	0.138
Empirical M2	0.971	0.926	0.872	0.812	0.750	0.455	0.123
Empirical M3	0.976	0.924	0.861	0.794	0.727	0.435	0.140
True M1	0.963	0.870	0.751	0.626	0.508	0.139	0.006
Empirical M1	0.963	0.870	0.751	0.626	0.507	0.140	0.006
Empirical M2	0.941	0.848	0.741	0.629	0.521	0.144	0.005
Empirical M3	0.965	0.875	0.756	0.629	0.508	0.129	0.003

Results

Models 2 and 3, although of no direct interest to this work, provide a means to assess the power of topological event history methods due to their Gaussian marginal distributions. Figure 6.4 shows the Nelson-Aalen plots for each of the fields shown in Figure 6.2. The solid line shows the expected Nelson-Aalen estimate for a Gaussian random field with the correlation structure of the data in panel (a) of Figure 6.2.

We can immediately see the difference between the Nelson-Aalen curves for each model. In order to assess the significance of these differences, we obtain confidence intervals for each. We investigate the coverage of 95% confidence intervals for Gaussian random fields only, looking at confidence intervals obtained via three methods. First, we consider the standard (naive) Nelson-Aalen variance estimator. The second approach involves a parametric bootstrap. Here, for each sample, we estimate Matérn parameters using maximum likelihood and use these estimates to generate an additional 100 random fields. As with

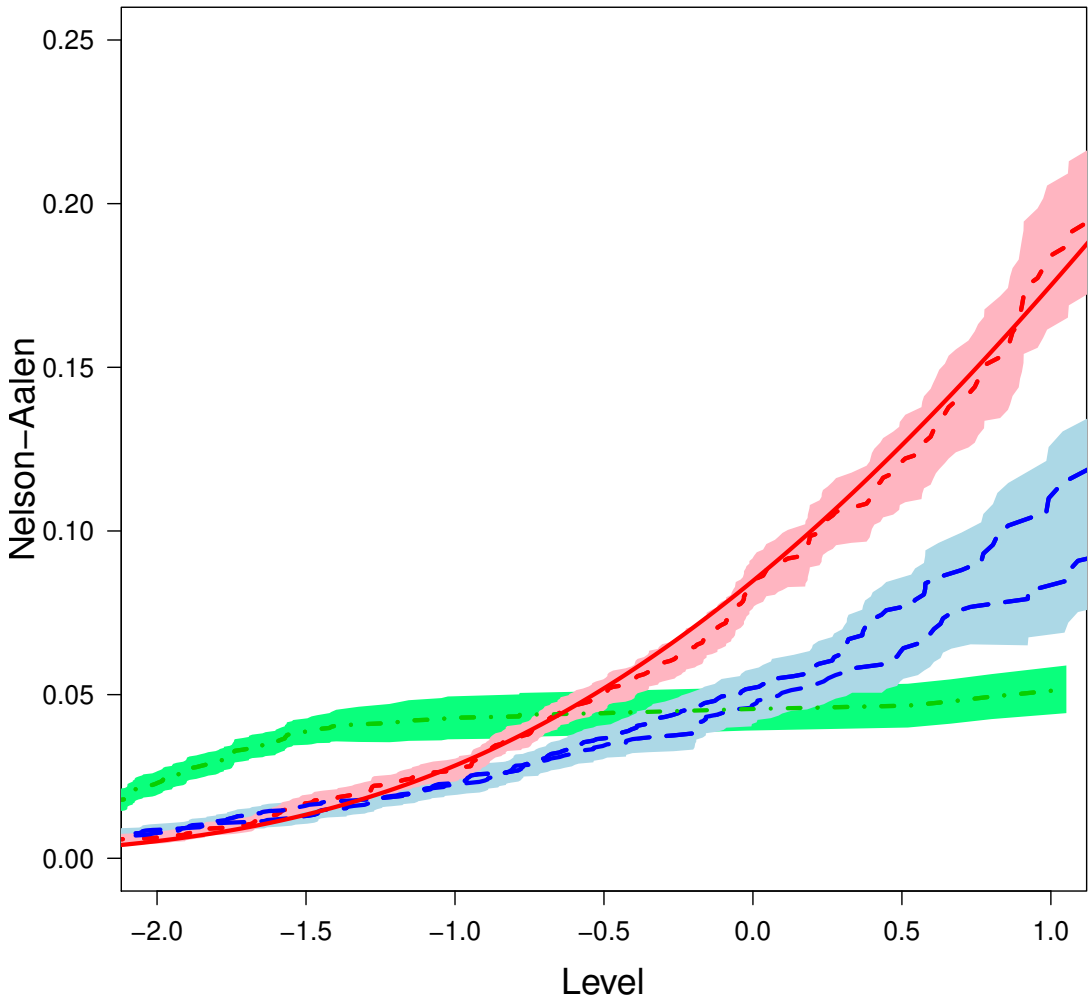

Figure 6.4: Nelson-Aalen plots for connected components for the random fields of Figure 6.2. Short dashed lines correspond to panel (a), long dashed lines to panels (b) and (c), and dot-dash lines to panel (d). The shaded regions indicate ± one standard deviation, obtained numerically from 1000 simulations,. The solid line shows the expected value for a Gaussian random field with the same correlation structure as the data in Figure 6.2.

when we use multiple replicates, this allows us to obtain standard errors. Finally, we assume the existence of multiple replicates, from which empirical standard deviations can be calculated. For these three approaches, we calculate both pointwise estimates and simultaneous confidence bands. Each method produces good results in comparison to the standard Nelson-Aalen variance estimator, which results in significant under coverage, as seen in Table 6.3.

For the non-Gaussian random fields, we compare the standard Nelson-Aalen variance

Table 6.3: Coverage of nominal 95% pointwise confidence intervals and 95% simultaneous confidence band (SCB) around Nelson-Aalen component plots for Gaussian random fields. Results from 1000 simulations on 60×60 lattices, with Matérn correlation function with parameters η and ν

		Percentile					
(η, ν)	Method	0.9	0.7	0.5	0.3	0.1	SCB
(5,1)	Standard	84.0	92.2	93.5	88.9	82.2	73.9
	Par. bootstrap	95.3	94.4	95.8	95.3	96.2	94.1
	Replications	94.2	94.7	95.7	93.9	94.7	94.4
(10,1)	Standard	67.8	82.7	83.3	73.7	65.5	39.1
	Par. bootstrap	95.3	95.1	95.6	95.7	94.7	93.6
	Replications	95.1	94.0	94.9	95.9	94.0	96.4
(5,2)	Standard	93.4	94.6	95.0	93.2	90.8	82.9
	Par. bootstrap	95.4	95.2	95.7	95.5	95.8	95.9
	Replications	94.0	94.6	95.0	94.3	93.7	93.4

estimator to a parametric bootstrap based on a Gaussian random field, with the variance adjustment proposed by Copas and Eguchi (2005). Results can be seen in Table 6.4. Again, the use of the standard estimator results in under coverage. The parametric bootstrap performs well generally, although it can be conservative as expected.

We look at the results as pointwise effect sizes and powers as percentages at $10, 30, 50, 70$ and 90% of the at-risk distribution as can be seen in Table 6.5. To calculate these values we assumed single realisations and used maximum likelihood for parameter estimation under an assumed Gaussian random field. Size is good when the Gaussian random field assumptions are correct (Model 1) and power is excellent for Model 2 at the central and later percentiles. Power can be low at some percentiles, since the expected curve can cross that of a Gaussian random field. This demonstrates the advantages of using a simultaneous confidence band where appropriate. Power is lower for Model 3, which is closer to a Gaussian random field, although still very good at $\theta = (5, 2)$, which has the highest local correlations.

Table 6.4: Coverage of nominal 95% pointwise confidence intervals and 95% simultaneous confidence band (SCB) around Nelson-Aalen component plots for non-Gaussian random fields. Results from 1000 simulations on 60×60 lattices. The parameters given are the targets θ_1 as described in the text

			Percentile					
Model	(η, ν)	Method	0.9	0.7	0.5	0.3	0.1	SCB
M2	(5,1)	Standard	89.4	71.2	71.8	75.7	81.0	65.8
		Adjusted par. bootstrap	97.0	91.9	98.6	100.0	100.0	98.2
		Replications	95.0	93.6	93.3	93.4	94.4	94.1
	(10,1)	Standard	65.3	60.6	61.0	64.4	70.3	58.4
		Adjusted par. bootstrap	95.8	97.3	97.8	98.0	98.9	97.2
		Replications	94.7	95.4	95.5	95.9	95.7	93.6
	(5,2)	Standard	86.6	73.5	74.4	76.2	78.3	72.5
		Adjusted par. bootstrap	97.6	98.2	99.8	100.0	100.0	97.8
		Replications	94.2	94.8	94.7	94.5	94.9	95.3
M3	(5,1)	Standard	85.1	92.3	91.6	91.8	86.7	75.4
		Adjusted par. bootstrap	98.6	98.9	99.7	99.9	99.6	99.0
		Replications	94.4	96.1	95.8	92.0	95.6	94.6
	(10,1)	Standard	71.4	79.3	84.2	75.1	68.4	42.6
		Adjusted par. bootstrap	98.6	99.3	99.3	99.5	98.9	98.0
		Replications	93.5	94.7	95.6	94.0	94.4	94.5
	(5,2)	Standard	91.5	94.5	95.8	95.5	94.1	87.7
		Adjusted par. bootstrap	98.1	99.9	100.0	100.0	100.0	98.9
		Replications	95.2	94.5	95.6	95.9	95.2	96.8

Table 6.5: Pointwise and simultaneous size and power, as percentages, for testing for a Gaussian random field. Results from 1000 simulations on 60×60 lattices. The parameters given are the targets θ_1 as described in the text.

		Percentile					
Model	(η, ν)	0.9	0.7	0.5	0.3	0.1	SCB
M1	(5,1)	5.1	5.3	5.6	5.0	3.6	5.1
	(10,1)	5.9	5.2	4.7	4.6	7.3	8.2
	(5,2)	4.5	5.0	5.0	4.6	3.3	4.7
M2	(5,1)	100.0	27.8	95.6	100.0	100.0	100.0
	(10,1)	40.6	74.1	98.1	99.1	99.3	99.9
	(5,2)	99.6	21.6	99.7	100.0	100.0	100.0
M3	(5,1)	11.2	19.0	47.9	57.2	41.4	61.7
	(10,1)	6.0	18.2	24.7	21.5	11.1	31.9
	(5,2)	9.8	59.9	92.5	96.3	98.1	98.4

6.4 TEH for global wind intensities data

We return briefly to wind intensities data from the Community Earth System Model Large Ensemble project (Kay et al., 2015). Although the Earth System Models from which our data are generated are deterministic, spatial and spatio-temporal statistical modelling is often applied as a means to both explore and concisely summarise the output (Castruccio, 2016; Edwards et al., 2019). Such models often assume that the modelled residuals form a Gaussian random field. We can use the topological event history approach to assess this assumption, at least in part, and as a quick and easy method to compare ensemble members and to examine the suitability of assumed correlation structures.

We show topological Nelson-Aalen plots for connected components in Figure 6.5 for year one ($t = 2006$). The data used is reduced from the original data set, as described in Chapter 2. Each of the 30 realisations thus consists of values on a 2-d grid comprising 96 longitudes and 51 latitudes for a total of 4896 sites. In red we show the expected cumulative hazard curve, calculated from the empirically obtained covariance matrix and realisation one is shown in black. No ensemble member stands out as being particularly unusual relative to the set of 30. Further, we see consistency between our data and a Gaussian random field with our empirical correlation. Later, in Chapter 10 we investigate this further, applying topological event history methods to a wide range of data from the LENS data set including different years and variables other than wind intensities.

These expected Nelson-Aalen plots provide a simple means for the assessment of a correlation model and the assumption of Gaussianity. It is important to note that a 'good' fit of an expected Nelson-Aalen curve under that model only tells us about the local level correlation and nothing about large-scale fit, but a poor fit tells us that our model is wrong.

6.5 Conclusions

In this chapter, we introduced topological event history analysis, through the application of the non-parametric Nelson-Aalen estimator for the cumulative hazard function to emergence of topological features. We demonstrated different methods for obtaining both pointwise confidence intervals and simultaneous confidence bands and compared these in a small simulation study. Finally we showed a simple example of the application of TEH to our wind intensities data. In the following chapter, we use this method in combination with more general assessment methods to investigate correlation models for our wind intensities data. In Chapter 9, we explore the use of this method in testing for Gaussianity in data.

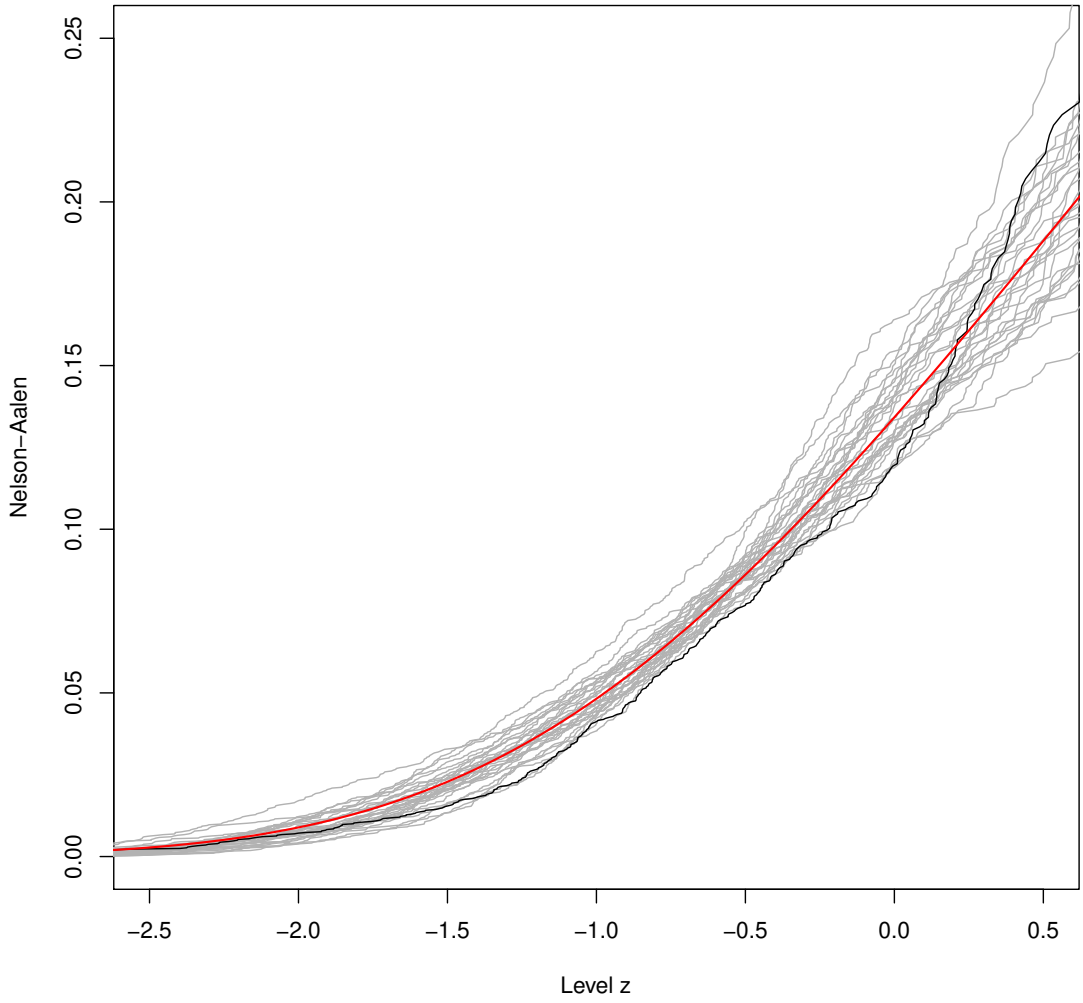

Figure 6.5: Nelson-Aalen plots for each of 30 realisations. The expected cumulative hazard function under the empirical covariance is shown in red. Realisation one is shown in black to illustrate an individual NA plot.

Chapter 7

Stationary covariance modelling for geospatial data

Many of the methods presented in previous chapters rely on accurate knowledge of the covariance structure of gridded data. In the case of wind intensities from the CESM Large Ensemble Project, we can take advantage of multiple realisations to obtain a highly accurate covariance structure empirically. However, in other scenarios, it is more common only to have a single data set, requiring us to obtain the covariance structure through the fitting of some model. We use the wind intensities data to investigate and assess several modelling approaches for geospatial data on the surface of a sphere. The cumulative hazard curves based on the empirical and modelled covariance matrices can serve as a measure of the accuracy of the model fit, specifically at a local scale which is of primary interest for topological event history analysis. Further, we use unsupervised clustering algorithms to investigate different spatial behaviours in the covariance, an approach which can provide a simple means for compression of a large covariance matrix.

The use of covariance structures for modelling spatial relationships in data has been widely studied. Early research included that of Fisher (1937), who described, concerning the study of crop yields, that 'patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are farther apart'. Observing that this phenomenon could not be explained by a naive model in which the data depend on some underlying variables (in addition to some zero-mean, independent stochasticity), he introduced a technique known as blocking. This primitive form of covariate adjustment allowed the incorporation of some spatial structure into the process, subsequently improving model fit. Since Fisher's work, the field has developed significantly, in part due to the range of applications requiring the description of some form of spatial variation. In fact, the emergence of the covariance function as a means to model the covariance between data points as a function of their

separation emerged almost simultaneously in two very different areas, namely forestry and mining, by Matérn (1960) and Krige (1951) respectively. The commonly used Matérn family of covariance functions was proposed and is still widely used today.

As early as 1963, spherically and axially symmetric stochastic fields on a sphere were studied, being represented as spectral processes by Jones (1963). In the present day, there has developed a growing interest in the changing global climate (Pachauri et al., 2014) as well as a significant progression in technologies allowing us to capture data from a wide range of locations, on an equally wide range of variables. This progression has led to the capture of numerous spatial data sets on the globe, resulting in a significant volume of research into the validity of existing covariance functions on the surface of a sphere, as well as the development of new, more flexible functions. The requirement for development and understanding of correlation functions valid on the surface of a sphere is indisputable, thanks to the ever-increasing number of large global data sets emerging for research into climate change and renewable energy. The CESM Large Ensemble from which we obtain our wind intensities data (Kay et al., 2015) contains many examples of data sets on spheres.

In the remainder of this chapter, we evaluate methods for modelling covariance on the surface of a sphere, when applied to our wind intensities data. We begin with a summary of key concepts and a discussion of distance measures, in addition to mentioning some of the computational challenges that are commonly faced when working with large data sets (and hence large covariance matrices). After introducing informal diagnostic simulation plots, which we frequently use throughout the work, we compare several standard stationary models using a selection of distance measures. Our data is unusual in that we have multiple realisations of the same dataset, allowing us to examine the empirical correlation between each pair of points. Following this analysis, we look at nonstationary approaches to modelling, including the incorporation of large scale geographic descriptors, such as land, coast and ocean. Finally, we use unsupervised clustering to define regions with similar local scale correlations.

7.1 Modelling covariance on the surface of a sphere

With the emergence of many global climate data sets comes the need to model covariance accurately on the surface of a sphere. Here, there are new challenges, frequently not present with many commonly used spatial data sets. For instance, the choice of distance measure on a sphere and the effects of geophysical descriptors on the covariance function can both influence the choice of model. In this section, we look at some fundamental concepts of covariance modelling, Euclidean and geodesic distance measures and discuss some of the computational challenges that may arise, with specific reference to our global

winds data. For all work in this chapter, we approximate the Earth as a perfect sphere. As we see later, for our data in which sites close to the poles are removed, the curvature of the Earth does not impact the modelling in any significant way. Hence, we argue that minor departures from a perfect sphere are trivial in this particular application.

7.1.1 Some preliminaries

Here we provide a brief introduction to some of the basic concepts required for modelling covariance on a random field, including isotropy and stationarity. For a zero-mean random field $z(x)$ we can describe the following properties, where x is some site on our random field and $z(x)$ the value of the field at that site. Frequently, in the work that follows, we assume that $z(x)$ is a Gaussian random field.

Covariance function

A covariance function C can be used to model the covariance between two variables. In the context of spatial data, $\mathrm{C}(z(x_1), z(x_2))$ defines the covariance between values at sites x_1 and x_2 on a Gaussian random field. Commonly used covariance functions include the Matérn and exponential functions. Full definitions of these and other common covariance functions can be seen in Appendix B.

Positive-definiteness

For any finite collection of sites $(x_1, x_2, \ldots, x_N)$, the covariance matrix R of

$$(z(x_1), z(x_2), \ldots, z(x_N))$$

is positive definite if and only if

$$u^T R u > 0$$

for all non-zero u. Further, a covariance function $C(z(x_1), z(x_2))$ is positive definite if and only if the corresponding covariance matrix is strictly positive definite.

Stationarity

A Gaussian random process can be described as stationary when

$$\mathrm{E}[z(x)] = \mathrm{E}[z(x+h)]$$

for all x and any separation vector h and

$$\text{Cov}(z(x), z(x+h)) = Cov(z(0), z(h)) = C(h).$$

Standard covariance functions are only valid for stationary data.

Isotropy

A stationary process and corresponding covariance function can be considered isotropic when the covariance function depends on the spatial separation vector only through its scalar length (Gelfand et al., 2010). That is,

$$C(h) = C(\|h\|).$$

For a process in $\mathbb{R}^d$ we can view isotropy as an invariance property under translation and rotation (Stein, 2012).

7.1.2 Distance measures on the surface of a sphere

The choice of distance measure on the surface of a sphere is essential in the modelling of spatial covariance, as it can have implications on whether or not a proposed covariance function is guaranteed to be positive-definite. We consider three measures with the global wind intensities data in mind. First, we use a simple projection of the data points to a regular 2-d grid in order to simplify the problem and establish a baseline. Second, we look at the chordal or Euclidean distance and third, the geodesic or great-circle distance.

Projection to 2D: a regular grid

Here, we consider the data as existing on a regular 2-d grid, where there is a unit distance between each latitude band, as well as between each longitude band. This projection is unrealistic, effectively stretching apart sites near the poles and pushing together those near the equator. However, it allows the comparison of a naive approach to methods using the distance metrics described below. Considering the data in this manner is highly convenient as the data sets are made available in this gridded latitude-longitude format.

Chordal (Euclidean) distance

The chordal distance, or Euclidean distance, is the distance measured in Euclidean space. To use chordal distance for spherical data, we must restrict the set of points between which

it can be calculated to those on the surface of the sphere. The Euclidean distance between points $x_1 = (L_1, l_1)$ and $x_2 = (L_2, l_2)$ can be calculated as

$$d(x_1, x_2) = 2r\left[\sin^2\left(\frac{L_1 - L_2)}{2}\right) + \cos L_1 \cos L_2 \sin^2\left(\frac{l_1 - l_2}{2}\right)\right]^{\frac{1}{2}}$$

where r is the radius of the sphere, L is the latitude and ℓ is the longitude (Jeong and Jun, 2015b).

Geodesic (great-circle) distance

The great-circle distance or geodesic distance is the shortest distance between two points on the surface of a sphere, measured along the surface. A great-circle is any circle on the sphere whose midpoint coincides with that of the sphere. Through any two points which are not directly opposite each other, there is a unique great-circle, but through any two antipodal points, there are an infinite number. As the Earth is almost spherical, great-circle distance formulas are accurate to within about 0.5%.

Given two points x_1, x_2 with latitude and longitude (L_1, l_1) and (L_2, l_2) respectively, we can define the central angle between the points using the spherical law of cosines (Weisstein, 2020)

$$\Delta\sigma = \arccos\big(\sin L_1 \sin L_2 + \cos L_1 \cos L_2 \cdot \cos(l_1 - l_2)\big). \tag{7.1}$$

Then the great-circle distance is the arc length

$$d = r\Delta\sigma$$

where r is the radius of the sphere, as can bee seen in Figure 7.1.

7.1.3 Choice of distance measure and model

For small distances, the spherical law of cosines (Equation 7.1) can have significant rounding errors. For 64-bit floating point numbers, this is not serious for distances over a few metres. If we were interested in much smaller distances, the Haversine formula is designed to improve accuracy (Robusto, 1957). We assume the radius of the Earth to be $r = 1$ unit distance, simplifying our calculations to only depend on the latitude and longitude points.

There are two common approaches taken when fitting a correlation structure on a 2-sphere (i.e. a 'standard' sphere in 3-d space). The first is to restrict a function valid on $\mathbb{R}^3$ to the surface of the 2-sphere using the chordal distance. This restriction maintains the validity of isotropic positive-definite covariance functions on the surface of a sphere

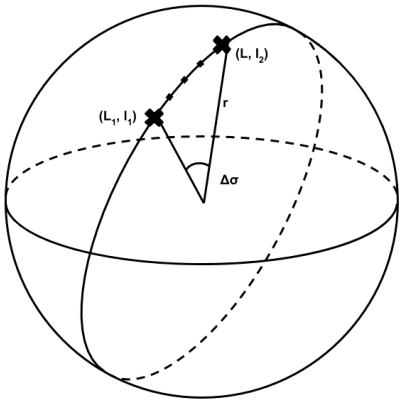

Figure 7.1: The Great-circle distance is measured between two points along a 'great-circle' on the surface of a sphere.

but can result in significant spatial distortions, particularly at greater distances (Jeong and Jun, 2015b). The second approach commonly taken is to use the geodesic distance directly in place of the chordal distance. The geodesic distance has many advantages; notably, it is a more intuitive way of measuring distance on a sphere. Indeed for many covariance functions, this is a valid approach, subject to some parameter restrictions. As shown by Gneiting (2013), this has negative consequences for the flexibility of the popular Matérn function, requiring a significant restriction of the smoothness parameter for the function to be guaranteed positive-definite. Hence, we encounter a significant impact on the usefulness of the Matérn function, as its popularity derives mainly from its ability to control local behaviour of a process through the smoothness parameter.

Jeong and Jun (2015a) presented a method to overcome the parameter limitations of fitting the Matérn correlation function with the geodesic distance. In this, the proposed covariance functions were able to guarantee positive-definiteness and model smoothness on a sphere where the restricted Matérn function failed, as well as being easily extensible to nonstationary cases. However, they found that the chordal Matérn function was at least equivalent in performance and often outperformed models specifically developed for use on the surface of a sphere. This result was also argued by Guinness and Fuentes (2015). In a later paper, Jeong and Jun (2015b) investigated a selection of parametric covariance functions using both Euclidean and geodesic distance metrics and compared spatial prediction. When the true spatial range is large, they found that functions defined with the geodesic distance do outperform functions defined (and restricted) in Euclidean space, with the result amplified in the presence of significant negative spatial correlations at large lags.

In addition to restrictions of chordal models as described above, various other models have been proposed to be more effective for data on a sphere. Guinness and Fuentes (2015) presented the 'circular Matérn', which they described as an 'analogue to the Matérn covariance function', valid on spheres using the geodesic distance. The circular Matérn (using the geodesic distance) is always valid on spheres in up to three dimensions and in common with the standard version, has parameters to control the range and smoothness of the process. Further, it has a closed form when the smoothness parameter ν is equal to a half-integer, but the authors presented methods for efficient computation when the parameter takes on other values. Several further flexible covariance functions valid on spheres were proposed by Guinness and Fuentes (2016), but these did not outperform the chordal Matérn function when applied to satellite and climate model data.

7.2 Estimating correlation structure directly from the data

The data we have is unusual, in that we have 30 realisations, assumed to be independent and identically distributed. We consider these as independent, due to the chaotic nature of global climate data (Castruccio et al., 2014) and rapid diversion from initial conditions. Hence, we can obtain 30 distinct sets of standardised residuals and estimate the correlations between sites directly from the data. We do not require our previous assumptions of isotropy and stationarity, allowing individual correlations to be defined entirely using the values of the corresponding 30 pairs of sites across the realisations.

The effect of correlation structure at the local level is of most interest. Due to this, we restrict the calculation to a specific neighbourhood of 12 sites around each site, as shown in Figure 7.2. The central site is marked 'X'. Symmetry allows us to calculate correlations for only half of the neighbouring sites those marked 'O'.

In Figure 7.3, we can see the empirical correlations for all sites, for the set of lags and directions shown in Figure 7.2. What is clear is that, particularly in the East-West direction, both the lag-one and lag-two correlations are noticeably different over land than over ocean regions, indicating the presence of nonstationarity. Nonstationarity is also evident in the diagonal directions, but less evident in the North-South direction. These differences imply that the land-ocean effect is more significant with changing longitudes. We can see that the assumption of stationarity may be invalid. This effect has been the basis for a widely used class of models known as axially symmetric models, described in more detail in Section 8.2.1.

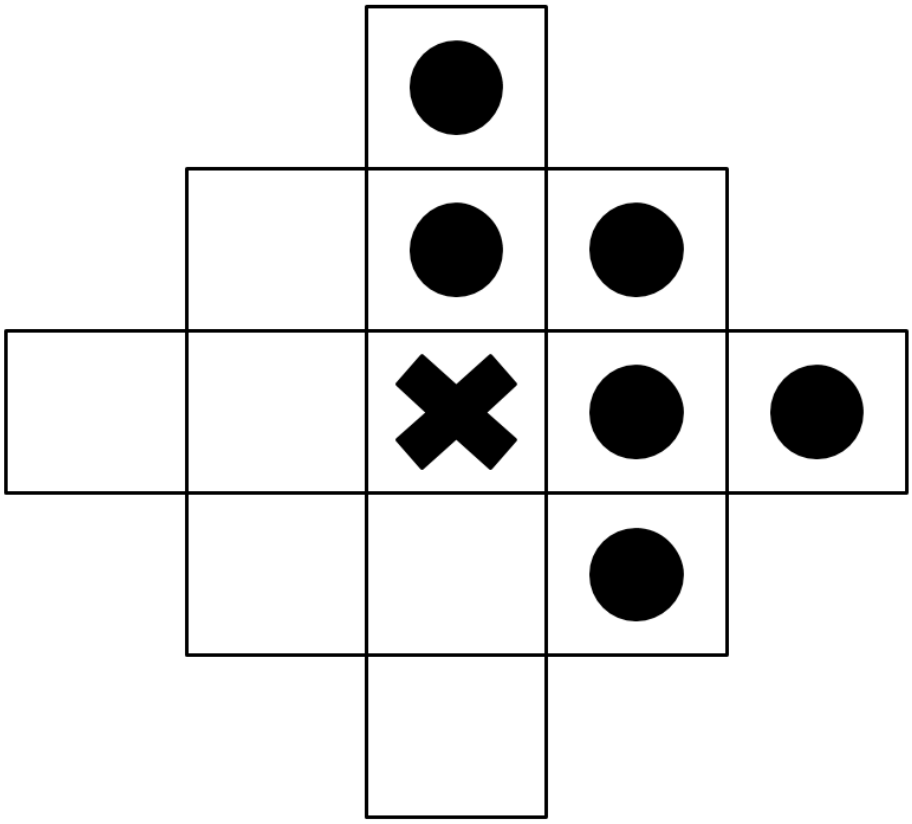

Figure 7.2: Neighbourhood of interest around a site.

7.3 Assessing model fit

The empirical covariance matrix provides us with knowledge of a 'truth' about the covariance structure of the data. This knowledge helps allow us to assess the fit of various models. In this chapter, we consider several assessment measures, including informal diagnostic simulation plots, the correlation matrix distance, standard AIC and expected cumulative hazard plots for local-scale assessment.

7.3.1 Informal diagnostic simulation plots

In the following section, we fit several covariance models using different distance measures. We use simulation plots as an informal diagnostic tool, which show an example of a data set simulated from a Gaussian random field using the relevant correlation structure and estimated parameters. While not providing a formal assessment of the performance of each model, these plots provide a quick way to determine the presence of any numerical or computational errors visually. Figure 7.4 shows the original data, and Figure 7.5 shows

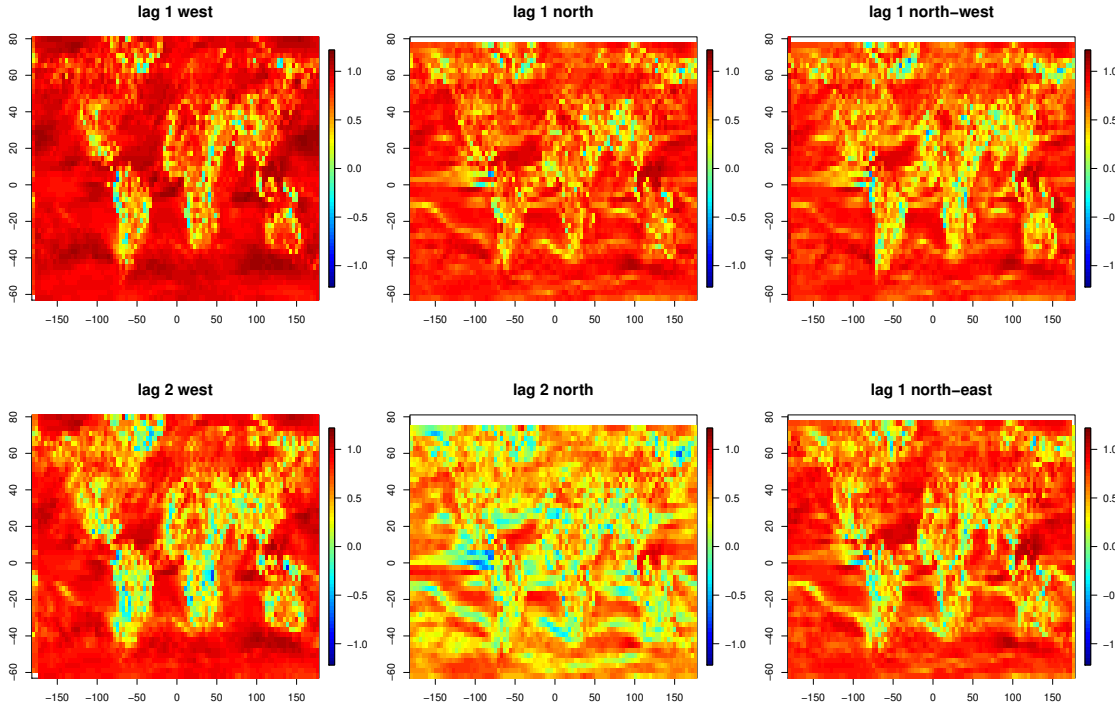

Figure 7.3: Empirical correlations calculated over 30 realisations.

two examples of simulations using poorly fitted Matérn correlation models for the original data.

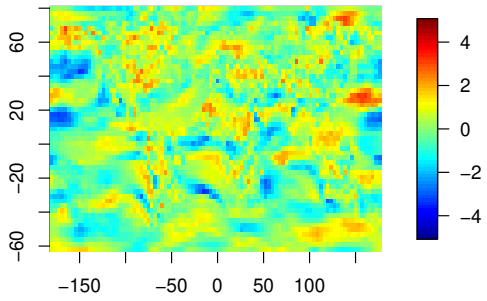

Figure 7.4: Standardised residuals for year $t = 2006$ and realisation one.

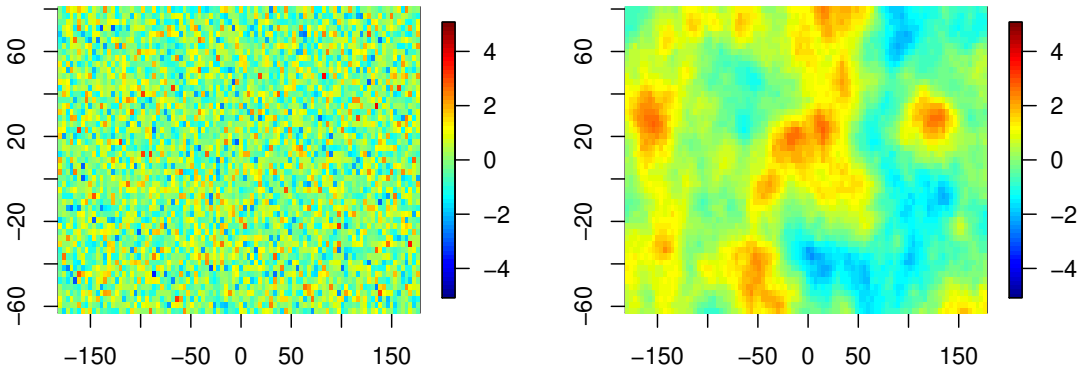

Figure 7.5: Informal diagnostic simulation plots showing data simulated from Matérn models with poor parameter estimates. On the left ($\eta = 0.1$, $\nu = 1$), the simulated data is hard to distinguish from uncorrelated noise; on the right ($\eta = 10$, $\nu = 1$), the simulated field is significantly smoother than the original data. In both cases, the fitted models are inappropriate.

7.3.2 The 'Correlation Matrix Distance'

Our research will include the use of informal diagnostic simulation plots, as described above, to compare original data to that simulated using an estimated covariance matrix. It would be useful to also include some more formal, or quantitative method in our comparisons. As we do not know the 'true' covariance matrix, with which to compare our estimates, we can calculate an empirical matrix using our 30 realisations of the data set and consider this as a baseline of truth, against which we can measure our models. A challenge arising from this method is that the empirical covariance matrix is not guaranteed to be positive-definite. Hence, we favour formal comparison methods which do not require the inversion of our empirical matrix.

Many methods exist for the comparison of covariance matrices, the simplest of which are based on the pairwise comparison of matrix elements. We consider a simple distance measure called the Correlation Matrix Distance (CMD) proposed by Herdin and Bonek (2004) and further developed by Herdin et al. (2005), primarily for application in MIMO (multiple-input, multiple-output) data.

The CMD between two matrices of equal size R_1 and R_2 is calculated as follows:

$$d_{corr}(R_1, R_2) = 1 - \frac{\text{trace}(R_1, R_2)}{||R_1|| \cdot ||R_2||}$$

and has the property that $d_{corr}(R_1, R_2) \in [0, 1]$, where $d_{corr}(R_1, R_2) = 0$ implies equal correlation matrices (up to some scale factor) and $d_{corr}(R_1, R_2) = 1$ implies maximally different correlation matrices.

This method has many advantages in our application. First, the measure is somewhat interpretable, as described above. Second, it is quick and easy to calculate over relatively large matrices (we have size 4896×4896 covariance matrices to compare). Thirdly, and perhaps most importantly in this context, we are not required to invert or decompose the covariance matrices in question, which is vital given the non-positive-definite nature of our empirical covariance matrix. However, although a more formal and quantitative measure of similarity, the CMD is arguably less informative than our diagnostic plots, providing only a single value as a measure.

7.3.3 Expected cumulative hazard plots for local-scale assessment (using topological event history methods)

Previous work on topological event history analysis (see Chapter 6) has demonstrated the potential of expected cumulative hazard function plots as a way to compare local performance of correlation models when fitted to random fields. These allow us to compare various covariance matrices for a data set via the resulting cumulative hazard functions, specifically assessing the model fit at a local scale. We use this method throughout this and the following chapter as an assessment method for correlations.

Since we are interested in correlation modelling in cases where we have access to only a single realisation of the data, all of the following models are fitted to a single realisation and time point, year $t = 2006$ from realisation one. In the expected cumulative hazard plots, we show the Nelson-Aalen estimates for year $t = 2006$ from all realisations, with realisation one shown in a darker grey. We can use this to visually compare the model to the empirical covariance obtained from all realisations, but also to the individual data set to which we have fit the model in question.

7.4 Fitting covariance functions to global wind residuals

We use standardised residuals from global wind intensities, reduced as described in Section 2.5, as a sample data set on which to fit a variety of different correlation models, described below and summarised in Appendix B. In this section, we assume both isotropy and stationarity, fitting standard Matérn and other common functions. For all cases below, we use the reduced data set for computational purposes. Table 7.1 shows a summary of the functions fitted and the maximum-likelihood (ML) parameter estimates we obtain in the

following section.

Table 7.1: Correlation functions fitted.

Covariance function	Distance metric	Parameters	ML estimates
Matérn	Regular grid	$\eta \geq 0$, $\nu \geq 0$	$\eta = 2.140$, $\nu = 0.884$
Powered exponential	Regular grid	$\eta \geq 0$, $\nu \in (0, 2]$	$\eta = 2.386$, $\nu = 1.339$
Matérn	Chordal	$\eta \geq$, $\nu \geq 0$	$\eta = 0.133$, $\nu = 0.598$
Powered exponential	Chordal	$\eta \geq 0$, $\nu \in (0, 2]$	$\eta = 0.131$, $\nu = 1.152$
(unrestricted) Matérn	Geodesic	$\eta \geq 0$, $\nu \in (0, 1]$	$\eta = 8.087$, $\nu = 0.822$
Powered exponential	Geodesic	$\eta \geq 0$, $\nu \in (0, 1]$	$\eta = 7.416$, $\nu = 0.592$
$\mathcal{F}$-family	Geodesic	$\tau \geq 0$, $\alpha \geq 0$, $\nu \geq 0$	$\tau = 1.514$, $\alpha = 19.932$, $\nu = 0.591$

Table 7.2 shows the Correlation Matrix Distance when we compare each of the models in the following section to the empirical matrix, considered to be our 'truth'. It is evident that the CMD values are fairly close and do not provide major insight into the fit of the matrices. The TEH curves can be seen at the end of this Chapter in Figure 7.16.

Table 7.2: Correlation matrix distance for each model compared to the empirical matrix.

Covariance function	Distance metric	d_{corr}
Matérn	Regular grid	0.794
Powered exponential	Regular grid	0.796
Matérn	Chordal	0.790
Powered exponential	Chordal	0.790
(unrestricted) Matérn	Geodesic	0.806
Powered exponential	Geodesic	0.825
$\mathcal{F}$-family	Geodesic	0.845

7.4.1 Isotropic and stationary covariance functions

The following models all assume isotropic and stationary covariance structure in the data. We make this assumption for convenience, although it is not necessarily appropriate for all data sets, as seen earlier. Most common models assume these properties and allow us to fit covariance functions to the data easily. Several methods exist for testing for isotropy and stationarity; a common practice when testing for isotropy is to assess sample variograms in multiple directions. Figure 7.6 shows an assessment of the (an)isotropy in our data using this method. With each panel corresponding to a different direction (North, North-East, East, South-East), we plot the semivariance, measuring spatial correlation between points

at different separation distances. There is a difference between the North-South direction and the East-West direction, however putting a quantitative measure on the significance of this is less straightforward. Guan et al. (2004) argued that although useful, this test is subjective and open to interpretation. They proposed a non-parametric approach which compares sample variograms in different directions based on their asymptotic joint normality. Fuentes (2005) developed ideas presented by Priestley (1965) testing for stationarity and isotropy of processes by assessing the homogeneity of spectral functions at different points in space. We take a related approach in Chapter 8 to test for non-stationarity between different geophysical regions.

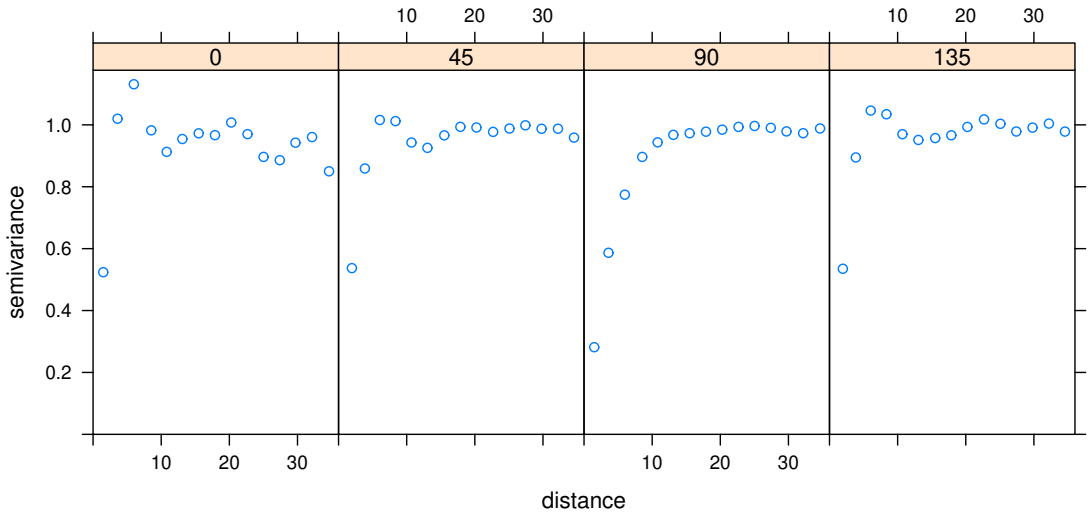

Figure 7.6: Semivariograms for the following directions (left-to-right): Northerly, North-Easterly, Easterly, South-Easterly.

7.4.2 Modelling on a projection to a regular 2-d grid

Matérn on a regular grid

We begin by fitting the Matérn correlation function,

$$C(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{d\sqrt{2\nu}}{\eta}\right)^{\nu} K_{\nu}\left(\frac{d\sqrt{2\nu}}{\eta}\right)$$

considering data points to occur on a regularly spaced 2-d grid. Maximum likelihood estimation results in parameter estimates of $\eta = 2.140$ and $\nu = 0.884$.

Figure 7.7 shows that data simulated using the fitted matrix does not appear significantly different in structure and smoothness to the original data. Although the process requires

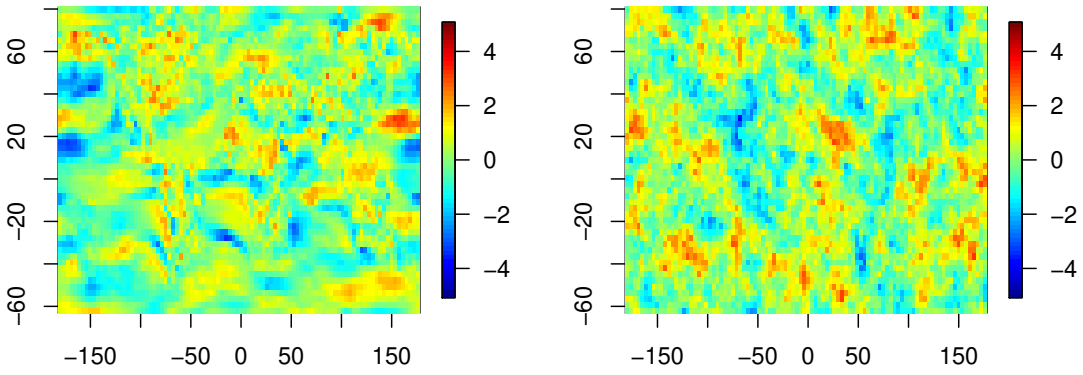

Figure 7.7: Original (left) and simulated (right) data using Matérn parameter estimates on a regular grid.

some assumptions and simplifications, it provides a baseline against which to compare the chordal Matérn, in addition to other functions and distance measures.

Powered exponential on a regular grid

As with the Matérn previously, we fit the powered exponential correlation function

$$C(d) = \exp\left\{-(\frac{d}{\eta})^{\nu}\right\}, \quad \eta > 0, 0 < \nu \leq 2$$

again assuming all data is distributed on a regular grid. Using maximum likelihood estimation we obtain parameter estimates of $\eta = 2.386$ and $\nu = 1.339$. Figure 7.8 shows original and simulated data using this method. Both the TEH figure and the CMD suggest that there is no significant difference between the Matérn function on a regular grid and on the sphere. The correlation matrix distance for this matrix is marginally lower than that for the gridded Matérn function, and the expected cumulative hazard curves are almost indistinguishable.

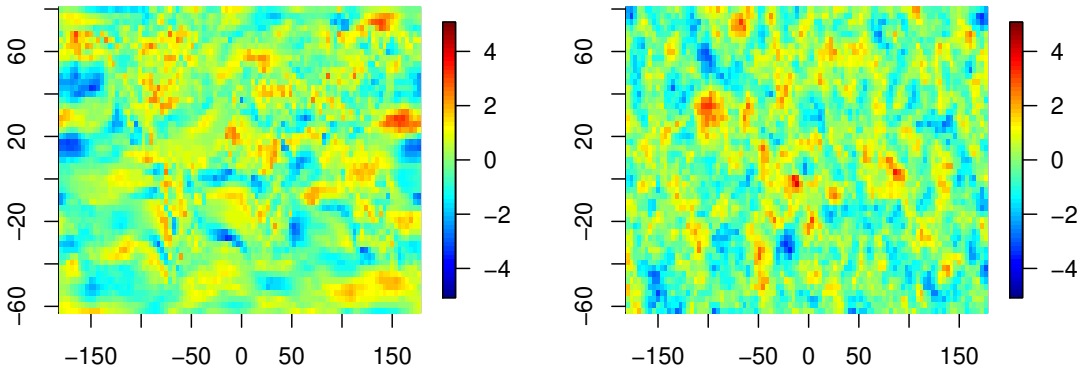

Figure 7.8: Original (left) and simulated (right) data using powered exponential parameter estimates on a regular grid.

7.4.3 Modelling with the chordal distance

Chordal Matérn

Next, we fit the Matérn correlation function in 3-d using chordal distance between sites on the surface of the sphere. We define the origin as being the centre of the Earth, (assuming the Earth is a true sphere). We convert latitude-longitude values to points in Euclidean space as follows:

Given latitude L_1, longitude l_1 and Earth's radius r, the Euclidean coordinate is

$$s_{euc} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\cos L_1 \cos \ell_1 \\ r\cos L_1 \sin \ell_1 \\ r \sin L_1 \end{pmatrix}.$$

We now fit the Matérn as before. Here we obtain maximum likelihood parameter estimates of $\eta = 0.133$ and $\nu = 0.598$. The range parameter η is much smaller than for the Matérn function on a regular grid. Figure 7.9 shows original and simulated data using this method. Relative to the grid Matérn function, we see a slightly lower CMD, but a poorer fit based on the TEH curves.

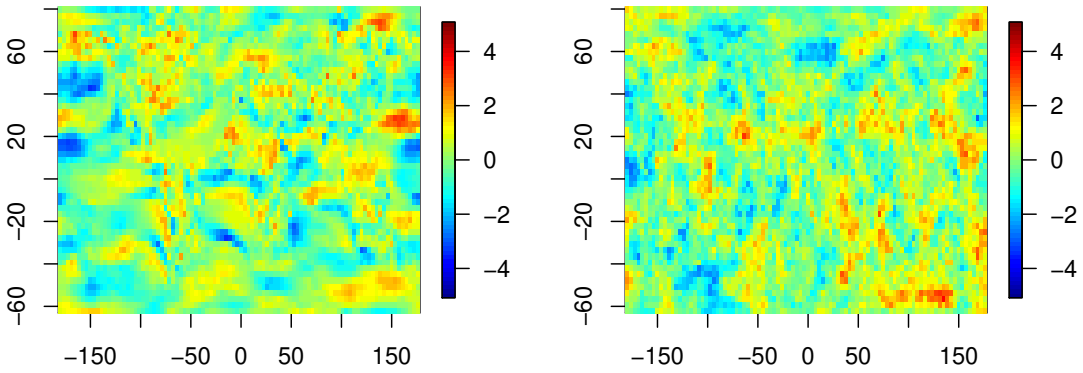

Figure 7.9: Original (left) and simulated (right) data using chordal Matérn parameter estimates.

Chordal powered exponential

We fit the powered exponential covariance function using chordal distance as we did for the Matérn. We obtain ML parameter estimates of $\eta = 0.131$ and $\nu = 1.152$, with simulation results shown in Figure 7.10. As seen with the Matérn model, the range parameter η for the chordal powered exponential is much smaller than for the powered exponential on a regular grid.

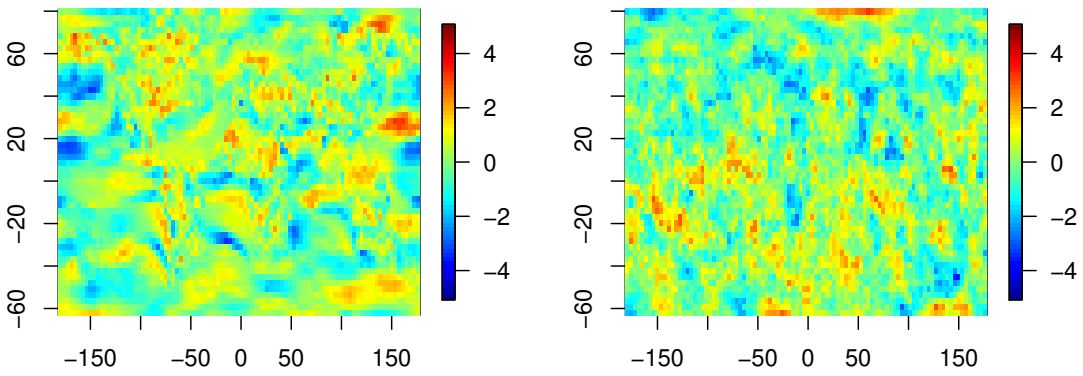

Figure 7.10: Original (left) and simulated (right) data using powered exponential parameter estimates and chordal distance.

There is not a noticeable difference between the IDSPs for the gridded and chordal versions of this model and again, similar to in the Matérn case, the chordal model has a marginally lower CMD, but a poorer fit when assessing based on TEH curves.

7.4.4 Modelling with the geodesic distance

Geodesic Matérn

For completeness, we fit the Matérn correlation function with the geodesic distance without putting any restrictions on the smoothness parameter. Maximum likelihood estimation gives parameter estimates of $\eta = 8.087$ and $\nu = 0.822$, both much higher than the parameter estimates for our previous Matérn models. Here, the smoothness parameter ν, is outside the restriction of $\nu \in (0, \frac{1}{2}]$ that Gneiting (2013) described as being required for guaranteed validity on the surface of a 2-dimensional sphere. Figure 7.11 shows simulation results obtained with this model.

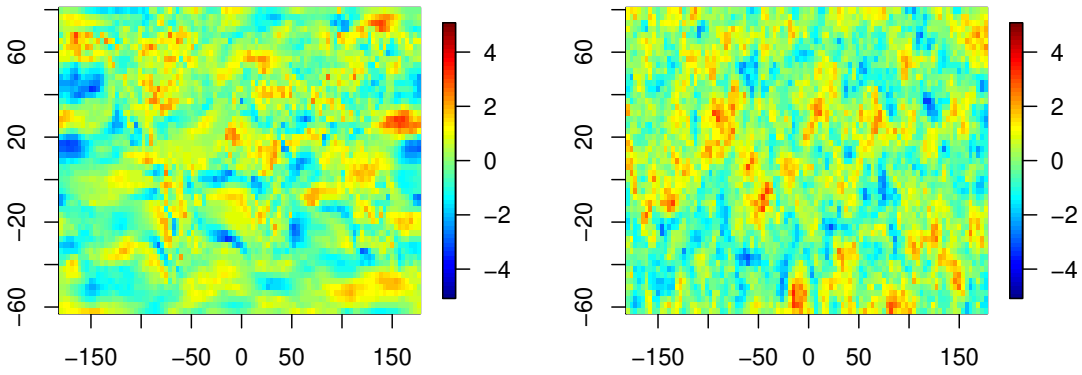

Figure 7.11: Original (left) and simulated (right) data using Matérn parameter estimates and geodesic distance.

For this model, we see a higher CMD value, indicating less similarity with our empirical covariance matrix than previous models and a cumulative hazard curve very close to those obtained with the chordal models.

Geodesic powered exponential

Fitting the powered exponential with the geodesic distance results in parameter estimates of $\eta = 7.416$ and $\nu = 0.592$. Like the geodesic Matérn function, we see higher parameter estimates than for the alternative distance metrics. We present simulated data from this model in Figure 7.12.

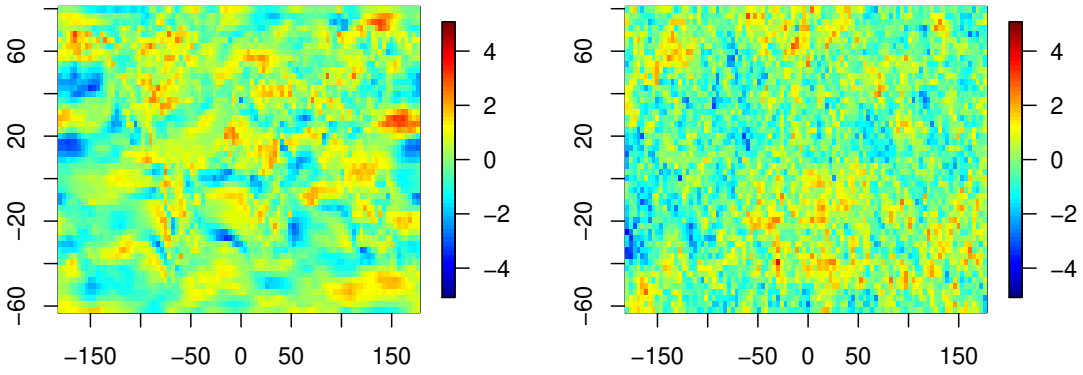

Figure 7.12: Original (left) and simulated (right) data using geodesic distance for powered exponential correlation function.

Visual inspection of the diagnostic plot for this model shows a noticeable difference between data simulated with this model and the original data when compared to previous models. Additionally, the CMD, in this case, is the highest of those measured and the expected cumulative hazard curve furthest from our empirical baseline. Arguably, the geodesic powered exponential covariance function has the weakest performance of those tested when applied to our data.

$\mathcal{F}$-family

As discussed in detail in Section 7.1, the geodesic or great-circle distance is considered to be a more natural distance measure on a sphere than the chordal distance. Although this has its advantages, the availability of correlation functions valid with the chordal distance is limited. Those that have been developed often lack closed forms, thus requiring a degree of approximation. One correlation function that does not require this, is the $\mathcal{F}$-family correlation function proposed by Alegria et al. (2018) and based on the Gauss Hypergeometric function (Beukers, 2007).

The value in this new family of functions lies in its analogousness to the common and popular Matérn function. As discussed previously, the Matérn covariance function has proved popular in many applications due to its flexibility and the inclusion of the smoothing parameter, which allows control over the mean-square differentiability of the process. This property is highly challenging to estimate empirically, hence the ability to control it via a parameter in the covariance function contributes to the appeal of the Matérn function. For both spatial prediction under infill asymptotics (an increase in sample size over a fixed domain) and parameter estimation, accurate knowledge of the mean-square differentiability of the process is important. The $\mathcal{F}$-family of covariance functions shares this property but has the convenience of being valid for random fields on the surface of a sphere. In particular, one of the parameters in the $\mathcal{F}$-family of functions allows control of the mean-square differentiability of the field. Analogous to the Matérn smoothing parameter, this parameter ν controls the mean-square differentiability of the corresponding random field and closed-form expressions are available for the function whenever $\nu = \frac{1}{2} + k$ for $k \in \mathbb{Z}_+$.

We fit this function using a maximum likelihood approach, using software kindly provided by Alegria et al. (2018), to obtain parameter estimates of $\tau = 1.514$, $\alpha = 19.932$ and $\nu = 0.591$.

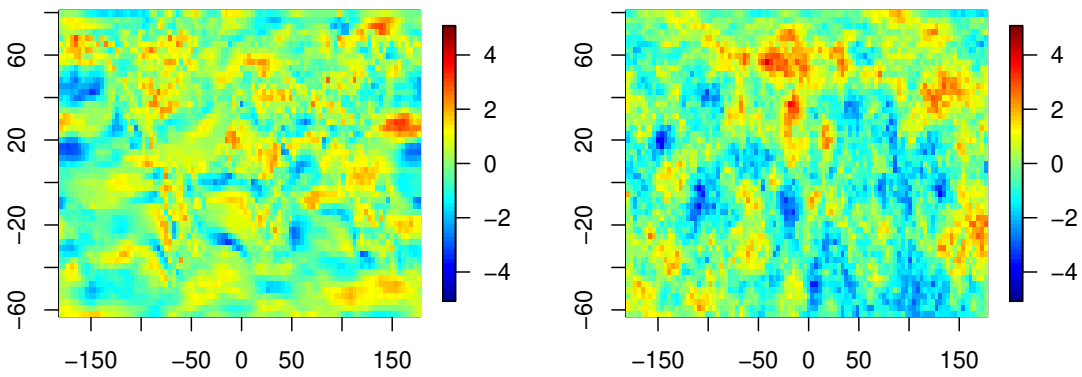

Figure 7.13: Original (left) and simulated (right) data using geodesic distance for $\mathcal{F}$-family correlation function.

Data simulated using the fitted matrix appears noticeably different to the original data, producing larger areas at the extreme values, as can be seen in Figure 7.13. Further, the CMD is one of the highest of those measured, yet despite this, the expected cumulative hazard curve is closest to our empirical curve of all models fitted so far. This is likely

a result of the fact that the cumulative hazard function is influenced only by local scale correlation. This model is attractive due to its flexibility and validity on the sphere.

7.4.5 Some comparisons

We can make some simple comparisons between the different distance measures. The first point of note is that the functions we fit on a regular grid treat the Earth as a finite plane, such that the points which appear to be furthest apart on a given latitude are in fact neighbouring. Figure 7.14 shows the covariance by separation for three different latitudes, chosen to reflect behaviour where spatial lag is equivalent to significantly different geodesic distances. We see the chordal and grid Matérn matrices have very similar values at all sites shown, with the $\mathcal{F}$-family showing much higher covariance values.

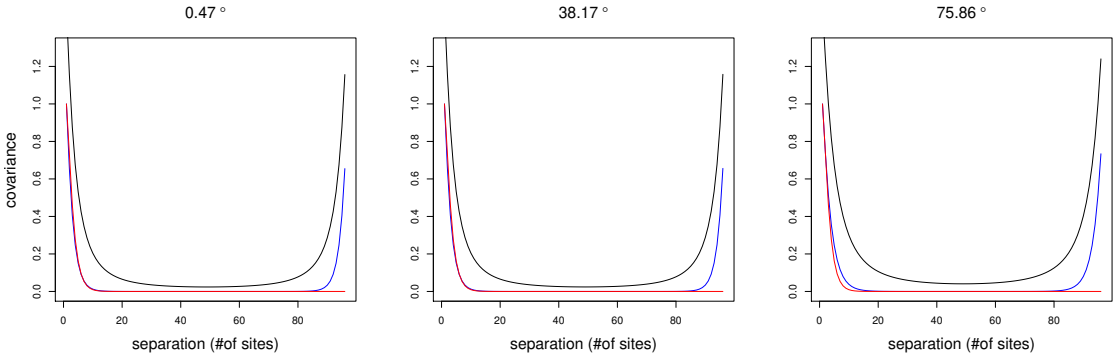

Figure 7.14: Covariance by separation at three different latitudes. $\mathcal{F}$-family (black), Chordal Matérn (blue) and Matérn on a regular grid (red).

Measuring by Correlation Matrix Distance, all models perform very similarly, with the chordal powered exponential being closest to the empirical matrix and the geodesic models, in general, performing the least well. We argue that the closeness in value of these different models suggest that the CMD is a poor measure of correlation matrix similarity, particularly for this application.

In Figure 7.15, we can see the mean empirical covariance for each latitude up to a spatial lag of five sites in the East-West direction. This mean empirical covariance is the average value across the latitude to allow for comparison with our model-based methods - since stationarity is not assumed, the true value changes with longitude. As would be expected, the covariance decreases as distance increases. It is important to note that for all the plots in Figure 7.15, the physical distances between lags one to five will differ depending on latitude. At equatorial latitudes, the distance between each site is considerably larger than at polar latitudes. As a result of this, we might expect plot one to show higher correlations for the polar latitudes, relative to the equatorial latitudes. While somewhat

evident, we see a noticeable peak in covariances just north of the equator.

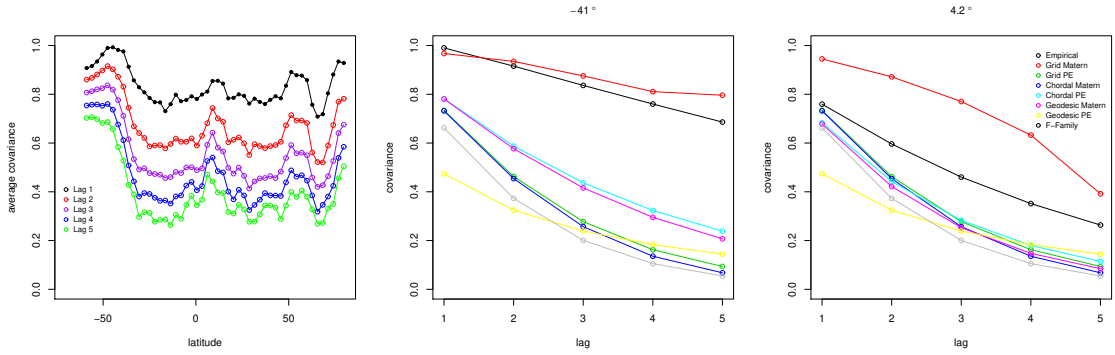

Figure 7.15: Left-hand plot shows directly estimated covariances at lags one to five for all latitudes. Remaining plots show lags one to five for all models for 41°S and 4.2°N.

For the two latitude bands shown, we can see little in the way of consistency between the best and worst-performing models. These latitudes are representative of the variation in model performance across all latitudes. The gridded Matérn curve, which sits very close to the empirical values for 41°S, is arguably the furthest away for 4.2°N.

7.4.6 Comparison to empirical covariance

For use in future work on topological data analysis and spatially correlated survival times, we are most interested in the covariance structure at the local level. Figure 7.16 shows the expected cumulative hazard functions under each of our fitted stationary models, alongside the empirical curve and the Nelson-Aalen estimates from the data. We see that at the local level, the models fitted on a grid perform equally, and better than the remainder of the models, although neither are particularly close to matching the curve achieved with the empirical covariance matrix.

One observation that we make based on this work is that identifying the optimal covariance model is not straightforward. One's choice of comparison method can give noticeably differing results. Further, as we are most interested in covariance at a local level, we note that some model with a good local fit may not be optimal for other purposes and equally, a model with an overall 'good' fit, may perform relatively poorly at a local level. This observation highlights the importance of identifying what is meant by an optimal covariance model in any particular application and choosing a comparison metric accordingly. By its definition, our topological event history approach is ideal for identifying models with a good local fit, in comparison to measures such as the Correlation Matrix Distance, which does not treat local or large-scale covariance with differing importance.

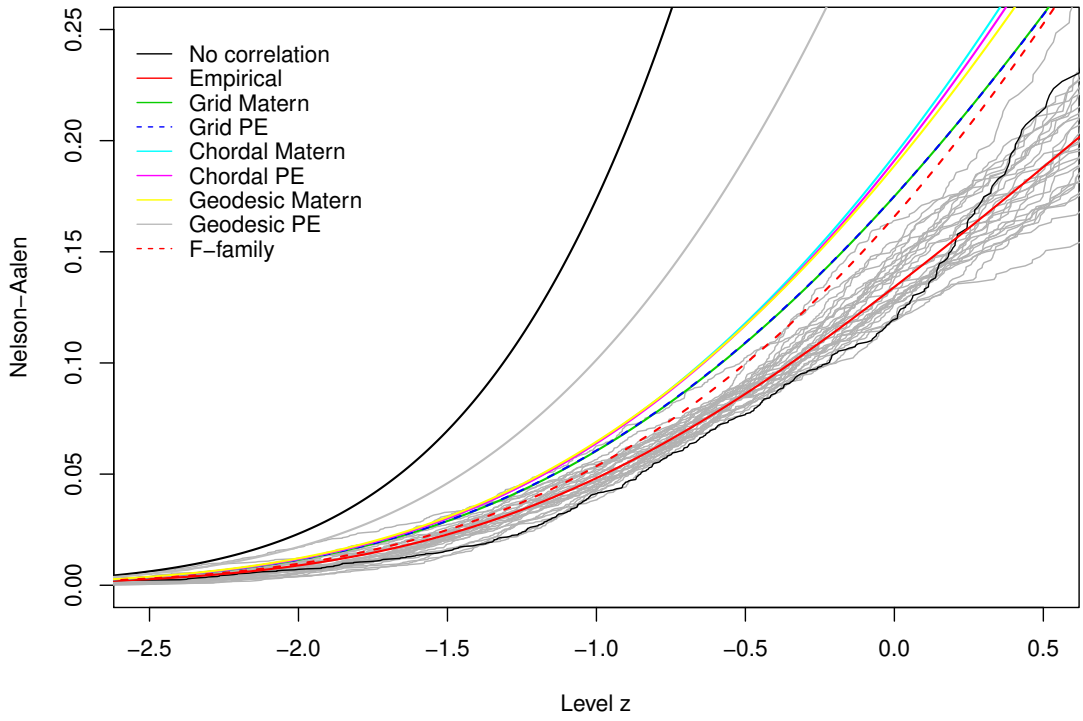

Figure 7.16: Expected cumulative hazard functions under a range of fitted covariance models. The grid Power Exponential curve is shown as a dashed line due to its similarity to the grid Matérn curve. We highlight the Nelson-Aalen curve for realisation one in black.

7.5 Conclusions

The purpose of this chapter was to investigate the performance of stationary covariance models for nonstationary data. We discussed various distance measures for data on the surface of a sphere and presented a number of assessment metrics for the fit of a correlation model. Comparing expected cumulative hazard curves for our fitted models to one using the empirical covariance matrix showed the $\mathcal{F}$-family model to perform the best, closely followed by the Matérn and powered exponential models fitted on a grid. We saw no improvement in general from the use of the chordal or geodesic distance, relative to assuming data to be on a regular grid, likely due to removal of data near the poles where difference in distance between neighbouring sites is more pronounced. In the following chapter we assess the degree of nonstationarity in the data and fit a number of nonstationary covariance models, assuming data to be on regular grid throughout.

Chapter 8

Modelling nonstationarity: challenges and approaches

In the following section, we look at more flexible correlation structures, for which we do not require the assumption of stationarity or isotropy in the data. We first consider the challenge of modelling covariance in the presence of significant land-ocean nonstationarity as observed in the previous chapter, confirming the results before examining solutions. We discuss a selection of common approaches to modelling covariance for spatially correlated data, before comparing model fits, including one taking a weighted combination of stationary models for each location, another in which we fit independently to different regions of the globe and an approach using radial basis functions. Throughout, we evaluate local model fit via expected cumulative hazard function curves, as seen in previous chapters. We examine the extent to which the curves have value for a more general assessment of covariance models. To conclude this chapter we explore site clustering to understand further the nature of the nonstationarity in the data and briefly discuss the handling of large covariance matrices in this context.

8.1 Land-ocean nonstationarity

The feature of land-ocean nonstationarity described above is not unique to this application. It has been observed in work on global temperatures, where Castruccio and Guinness (2017) identified similar dependence of the covariance structure on large-scale geographic descriptors. Although sometimes apparent from a plot of the global data, nonstationarity can be shown more formally by looking at a periodogram of the data at different latitudes for the regions of interest, in this case land, ocean and coastal regions. To assign site classifications, we first overlay a world map to obtain a binary land-ocean classification. We

class a site as coastal (on the original, unreduced data) if any of the direct neighbourhood of four sites with which it shares an edge differs from it.

As described by Castruccio and Guinness (2017), we calculate the values

$$|\hat{f}^j_{L_m}(c)|^2 = \frac{1}{N}|\sum_{n=1}^{N} h^j(\ell_n)Z(L_m,\ell_n)e^{-i\ell_n c}|$$

where $Z(L_m,\ell_n)$ is the standardised residual value at latitude L_m and longitude ℓ_n and $j = 1,2,3$ for ocean, land and coast respectively, h^i is a smooth function equal to zero when $\ell_n \notin$ region i and c is the wavelength. Thus $|\hat{f}^i_{L_m}|^2$ is a periodogram for region i at latitude L_m. We average across all latitudes to obtain Figure 8.1. Parallel periodograms would indicate similar correlation structure (Castruccio and Guinness, 2017), however the periodograms for land, ocean and coast are not parallel. Further, it is evident that the correlation structure over the ocean is significantly smoother than over land or coast, with correlation for land being slightly smoother than for coastal regions. This verifies the results seen in the empirical covariance lag plots in Chapter 7.

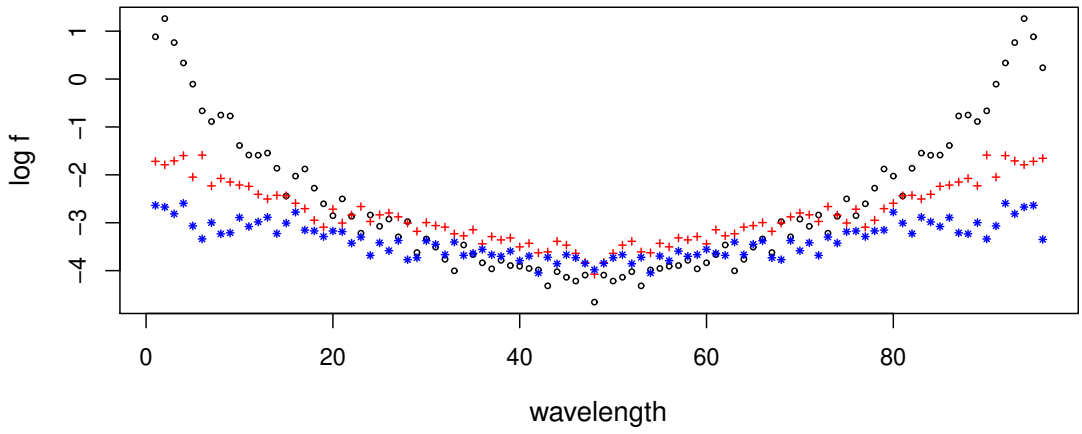

Figure 8.1: Periodograms of ocean (black), land (red) and coast (blue), averaged across all valid latitudes. The differing curvatures of the periodograms indicates different correlation structures.

8.2 Nonstationary modelling

There are many ways to approach the presence of nonstationarity in data. Here, we examine a few of these, in the context of global climate data.

8.2.1 Axially symmetric models

One observation that we made from the empirical covariance plots in the previous chapter was the clear difference between correlations in a longitudinal direction, compared to those in a latitudinal direction. This difference is unsurprising given the global nature of the data. Axially symmetric models have been widely used (Jun and Stein, 2008; Castruccio and Stein, 2013) to capture correlation without requiring an assumption of stationarity in the longitudinal direction. These models involve fitting Matérn (or Matérn-like) processes separately across latitude bands, before estimating the multi-band dependence. These allow more flexibility than a fully stationary model, since stationarity is assumed in longitude but not latitude, and are better suited to global climate data, where the impact of latitude frequently implies nonstationarity. However, it has been acknowledged by Castruccio and Guinness (2017) that this type of model is too restrictive in certain cases, specifically that of surface temperature data, upon which presence of land or ocean has a significant impact. As we have seen, for global winds the impact of land and ocean is noticeable from figures of the data alone and has been confirmed with periodograms.

Castruccio and Guinness (2017) proposed an extension of the axially symmetric model to incorporate an evolutionary spectrum approach, which they fitted using a multi-step conditional likelihood approach. The increased flexibility of this model involves the incorporation of different regimes (in this case, land/ocean). Jun (2014), on the other hand, approached the problem with a modified Matérn model, in which the smoothness of the process changes over the different regimes. This approach is highly suited to our problem, but it has computational limitations. Computational restrictions required the authors to sample only 1000 locations; to achieve this data size, we would have to reduce our data by at least a factor of four, in addition to our existing reduction of the original data set.

8.2.2 Moving window approaches

Moving window approaches can provide a means to deal with nonstationarity by fitting stationary models to local scale correlation over small windows. These methods are based on ordinary kriging and are prevalent in statistical climatology (Kuusela and Stein, 2018; Haas, 1995, 1990; Ver Hoef et al., 2004). Generally used for estimation of data, moving window approaches do not provide the user with a complete covariance matrix. However, they remove large-scale dependencies and allow variation of local properties over the full space. The estimation process takes advantage of quasi-stationarity, introduced by Journel and Huijbregts (1978), where a large nonstationary field will contain smaller areas of more homogeneous covariance. This has the effect of allowing the use of methods suited to stationary covariance structures, within some limited region. The proposers of quasi-

stationarity argue that the size of the window should depend on the degree of accuracy required in the estimation. A window should be large enough such that the data contained is sufficient for this level of accuracy, and hence the size of the window should, in theory, vary based on each site to be estimated.

The moving window methods have several advantages, particularly in statistical climatology applications. The flexibility in window size allows the user to gain control over the accuracy of estimates, without a requirement for either fixed window size or fixed accuracy. The method, therefore, can cope with fields in which specific areas contain a higher concentration of sites. Further, the method makes no assumptions about the relationship between covariance structure in different windows, allowing flexibility for complex nonstationary fields. We later use models that build on many of the ideas in moving window approaches.

8.2.3 Modelling with covariates

For scenarios where physical properties such as land masses impact the covariance properties, it would be valuable to take an integrated approach where we can incorporate knowledge of large-scale geophysical descriptors into the modelling process itself. Several approaches have been taken to incorporate these effects on the correlation structure of geophysical data. Pourahmadi (1999) used a Cholesky decomposition of the inverse of the covariance matrix to obtain a unique lower triangular matrix and a unique diagonal matrix. We can then interpret entries of these matrices as regression coefficients. This approach was generalised by Pan (2003) for growth curve, longitudinal and multi-level data, to include a polynomial representation of the main structure and the fit of a joint mean-covariance model.

An alternative approach by Castruccio and Guinness (2017) looked to incorporate large-scale geophysical descriptors through modelling an evolutionary spectrum across different regions. Evolutionary spectra for nonstationary time series data were initially introduced by Priestley (1965). More significant opportunities for application became available due to the development of computationally efficient fitting methods provided by Guinness and Stein (2013). An extension of previous work on spectral modelling across latitudes by Castruccio and Genton (2014), the evolutionary spectrum model allows for the incorporation of different regions as well as capturing longitudinal patterns. The spectral domain is fit across longitude, assuming independence between latitude bands, before modelling the between-latitude dependence. The evolutionary spectrum model is fit using a multi-step conditional likelihood method, which preserves nonstationarity features in the data. The work was further developed by Jeong et al. (2017) in which the set of geophysical descriptors is extended to include high mountainous areas in addition to land and ocean.

Finally, Jun (2014) proposed a Matérn based model for nonstationary cross-covariance processes in which smoothness is allowed to vary over different geophysical regions. Although able to capture nonstationarity, the model parameters are difficult to interpret, and the fitting process is not computationally feasible for large data sets.

8.2.4 Deformation

Another method commonly used for dealing with nonstationarity in spatial data uses a deformation of the spatial coordinate system to one in which stationarity is a valid assumption. Sampson and Guttorp (1992) proposed an approach for observations on a spatiotemporal random field, in which stationarity could be assumed temporally but not spatially. The approach models spatial dispersion (the variance of the difference between observations) as a smooth function of the geographic coordinates of the observation sites. Here, the term 'variogram' is not used since there is no assumption of isotropy. The spatial coordinates are transformed using nonmetric multidimensional scaling (MDS) such that the covariance structure of the observations as a function of the spatial dispersions is both stationary and isotropic. Finally, the geographic representation of the sites is mapped to the MDS representation using thin-plate splines. Thus, a nonparametric estimator for the covariance can be obtained through a composition of the mapping and a monotone function belonging to a class of conditionally non-positive definite variogram functions. This ensures that the estimator is a valid non-negative covariance model. The work is developed further by Perrin and Senoussi (2000) who extend the approach to a wider class of correlation functions.

The frequently unrealistic assumption of temporal stationarity encountered in spatial deformation research is challenged by Castro Morales et al. (2013). The authors use state-space models to model the temporal trend to obtain a more complete spatiotemporal model. Meiring et al. (1998) outline some of the challenges of this approach. The authors argue that estimation of the transformation and model parameters is a difficult multidimensional problem, with the dimension increasing with the number of sites. Despite the improvements in computational processing power, this is a limitation for large spatial datasets, which since the time of publication have become a common application for nonstationary models. Examples provided in this work use a set of 20 observation sites, significantly fewer than even our reduced data subset of 4896 sites.

In methods proposed by Fouedjio et al. (2015), the spatial deformation solution is extended from requiring multiple observations at each spatial location (whether at multiple points in time, or otherwise), to only requiring a single realisation of the data. This development is particularly valuable for use of the method on larger data sets.

8.3 Comparison of nonstationary models using TEH Nelson-Aalen plots (single replicate)

We use expected cumulative hazard function curves to compare the local fit of a selection of nonstationary covariance models. All of the following are fit to a single replicate of the data set. We focus on the single replicate case as we know that using the full set 30 replicates we can achieve excellent results simply using the empirical covariance matrix. We look at a method proposed by Reich et al. (2011) in which a weighted combination of stationary models is used to capture nonstationarity and a simple block approach whereby we model regions both separately and pairwise with stationary models before adjustment to obtain a valid covariance matrix. We then investigate the `LatticeKrig` package from Nychka et al. (2016) and finally look at work on a cyclic covariance function using B-spine basis functions from Konzen et al. (2019). Figure 8.2 shows the expected cumulative hazard function for a Gaussian random field, assuming independence between sites (black), the empirical covariance calculated from all 30 replicates (red) and the fitted grid Matérn model from the previous chapter for reference (green). Throughout this chapter we will add to this figure in order to compare fitted models.

8.3.1 Weighted combination of stationary models

Reich et al. (2011) propose a nonstationary covariate-dependent covariance function for modelling spatial correlation in different regions, in which a weighted combination of covariance functions enable environmental conditions to influence the prevalence of specific functions in different spatial regions. This is built on work by Schmidt and Rodrguez (2011) and Schmidt et al. (2011), who investigated how covariates could influence the covariance structure of spatial processes.

The motivation for the development of the model was the study of ozone in the troposphere, a secondary pollutant formed from the photochemical reactions and whose levels on a given day and in a given location are linked closely with meteorological variables, including cloud cover and wind intensities and direction. The authors aimed to show if and how the changing meteorology affects not only the mean concentration of ozone but also the variance and spatio-temporal correlation.

The model can be described formally as follows. Given

$$z_i = s(x_i, t_i)'\beta + \mu(x_i, t_i) + \epsilon_i$$

where z_i is the response, $s(x_i, t_i)$ is the vector of covariates at location $x_i \in \mathcal{R}^2$ and time $t_i \in \mathcal{R}$, β is a vector of regression coefficients and μ is a spatio-temporal process, with iid

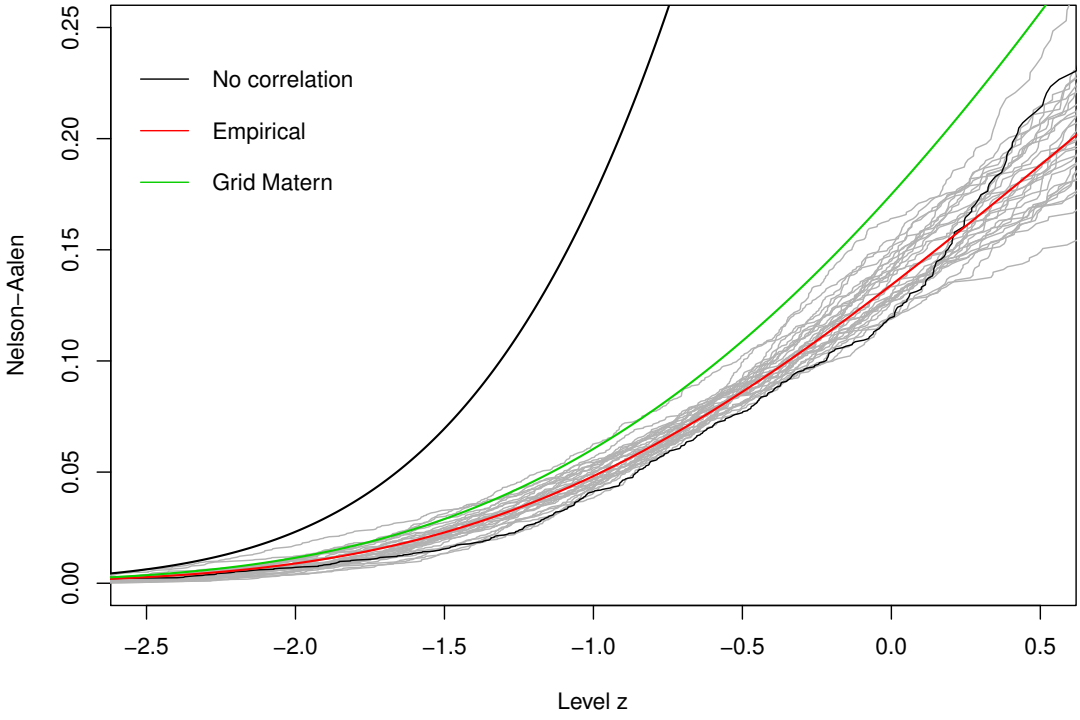

Figure 8.2: Expected cumulative hazard functions. The black curve shows the expected cumulative hazard assuming independence between sites. The stationary grid Matérn from the previous chapter is included for reference in green and the expected curve using the empirical covariance from all 30 realisations is shown in red. The Nelson-Aalen plots for the thirty realisations are shown in grey, with the single realisation used throughout this chapter in black.

error ϵ.

Then the spatiotemporal effect $\mu(x,t)$ is modelled as a linear combination of stationary processes,

$$\mu(x,t) = \sum_{j=1}^{M} w_j[s(x,t)]\theta_j(x,t)$$

where θ_j are zero mean Gaussian processes, with covariance K_j. The weighting w_j allows the resulting covariance structure to vary over the domain.

By integrating over θ_j, we obtain a model for the covariance as

$$\text{Cov}(\mu(x,t),\mu(x',t')) = \sum_{j=1}^{M} w_j[s(x,t)]w_j[s(x',t')]K_j(x-x',t-t').$$

The primary advantage of this approach is the relative simplicity and interpretability of the model, which limits the number of parameters to be estimated. It uses several possibly separable, stationary and isotropic functions to build a (generally) nonseparable, nonstationary and anisotropic model. It is important to note that there are cases where the environmental conditions (or known covariates, more generally) do not wholly explain nonstationarity, and hence this approach would be inefficient, or suboptimal. Further, effective use of the model requires detailed knowledge of all covariates that may significantly impact the covariance.

Fitting the full model is very computationally expensive. Hence we choose to fit a simplified version of the model proposed by Reich et al. (2011). In this, a weighted combination of stationary covariance models is fit according to underlying covariates, which in our case is the geophysical region only, i.e. land, ocean or coast. Further, we use a version of our data set with a reduction factor of four, in contrast to all other work in this chapter, in order to make this method computationally feasible. This gives us a data set of size 72×38.

We model the spatial covariance between responses at sites x and x' as

$$\text{Cov}(\mu(x),\mu(x')) = \sum_{j=1}^{M} w_j[s(x)]w_j[s(x')]K_j(x-x')$$

where K_j are individual stationary covariance functions and w_j weights each function according to spatial covariates $s(x)$. The squared weights for each observation sum to one, and each has multinomial logistic form

$$w_j[s(x)]^2 = \frac{exp(s(x)^T\alpha_j)}{\sum_{l=1}^{M} exp(s(x)^T\alpha_l)}$$

where $\alpha_1, \ldots, \alpha_M$ are coefficients controlling how the covariates affect the covariance. We select $M = 3$, and fit this model based on three stationary chordal Matérn functions, each with fixed ν and η estimated by maximum likelihood. Setting $\alpha_1 = 0$, we then estimate α_2 and α_3 using a maximum likelihood approach to obtain $\alpha = (0, -84.3, -124.8)$. This method allows us to incorporate land-ocean-coast structure into our modelling process.

Our implementation of the model could be extended to better capture the effects of the geophysical descriptors on the covariance structure, through fine-tuning the impact of the covariates on the weightings or using a higher number of underlying stationary models. However, these adjustments require an increase in the number of parameters, with a significant adverse effect on computation time. Figure 8.2 shows expected cumulative hazard curves for all the models fitted in this chapter, including the expected cumulative hazard function curve under this model. It does not perform better than the stationary Matérn model, indicating a poor local fit.

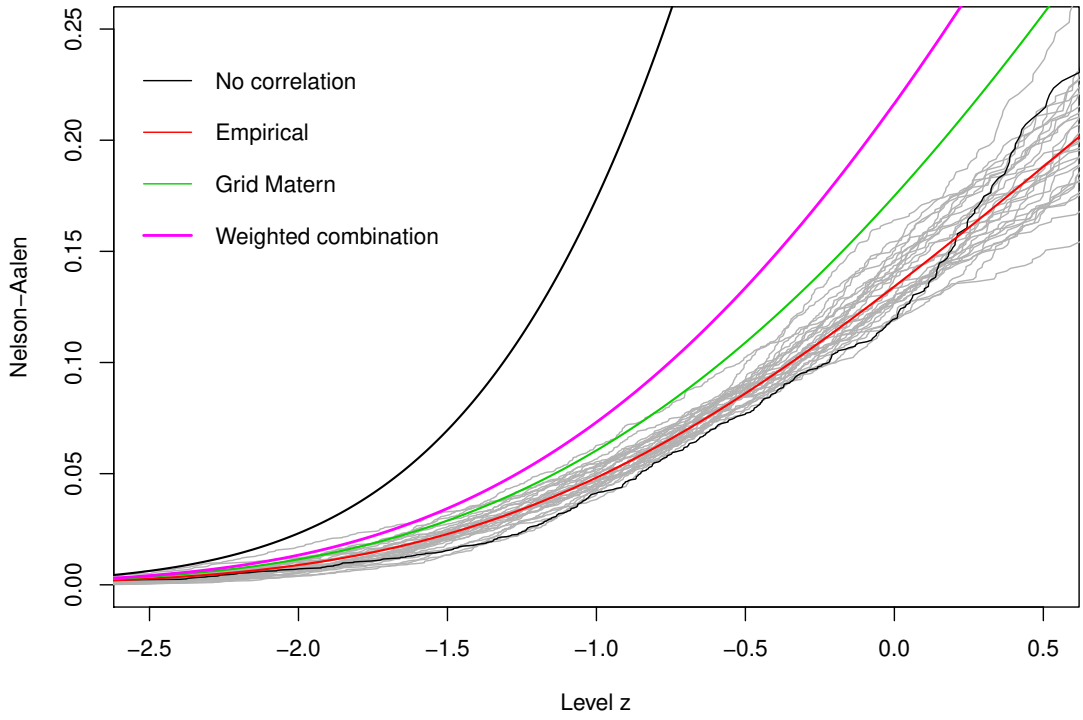

Figure 8.3: Expected cumulative hazard functions. This figure shows in pink the expected cumulative hazard curve assuming the fitted weighted combination model.

8.3.2 Regional block model

The nonstationarity results justify the incorporation of different regions into our modelling. We propose a regional block model, which allows the covariance parameters of some underlying stationary model to vary between different regions and pairs of regions. Initially, we divide sites into three regions: land, ocean and coast.

We can describe the model as follows. For each site x_i, with value $z_i = z(x_i)$, we define some region $k_i = k(x_i)$. Then with N being the number of regions,

$$\text{Cov}(x_i, x_j) = \sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_{i,j} K_{m,n}(x_i, x_j)$$

where $K_{m,n}$ is the covariance model for regions m and n and $\alpha_{i,j} = 1$ when $k_i = m$, $k_j = n$ and is zero otherwise.

In practice, we run the following multi-step process, described here for the three-region model with underlying Matérn covariance but easily extendable to a higher number of regions.

1. Fit a Matérn correlation model for land-land pairs only.
2. Fit a Matérn correlation model for ocean-ocean pairs only.
3. Fit a Matérn correlation model for coast-coast pairs only.
4. Use a maximum likelihood approach to fit a Matérn correlation model for all land-ocean sites, fixing known (previously estimated) correlations for land-land and ocean-ocean pairs in the correlation matrix.
5. Repeat step 4 for all ocean-coast sites.
6. Repeat step 4 for all land-coast sites.
7. Populate all land-ocean, ocean-coast and land coast sites in the correlation matrix using calculated values.
8. Populate land-land, ocean-ocean and coast-coast sites in the correlation matrix with calculated values.

We describe results obtained at various steps of fitting the three-region model. Standard maximum likelihood estimation produces parameters for each region, as shown in Table 8.1. These estimates further highlight the differences in smoothness between regions and verifies the results found when examining the periodogram in Figure 8.1. We see similar

results for the range parameter over all three regions but a much higher shape parameter for the ocean than for land or coast.

Table 8.1: Number of sites and Matérn parameter estimates for land, ocean and coast.

Region	#sites	η	ν
land	1191	0.0970	0.4075
ocean	3188	0.1211	1.3531
coast	517	0.0948	0.4442

We then fit the model for each pair of regions. Using this multi-step approach, we obtain the additional set of parameter estimates shown in Table 8.2.

Table 8.2: Matérn parameter estimates for each pair of regions.

Region pair	η	ν
land-ocean	2.1718	1.3056
ocean-coast	1.1582	1.0212
land-coast	2.1652	1.3192

Methods to ensure positive-definiteness

The pairwise block approach, while interpretable and fast to fit, can often (and does, in this case) result in a covariance matrix which is not positive-definite. This causes problems when we wish to calculate the inverse of the matrix, for example, when using it for simulation or emulation. A problem frequently encountered in finance, this is simple to adjust for, and it is possible to obtain good results after such adjustment.

A valid correlation matrix is symmetric, with ones along the primary diagonal and elements in $(-1, 1)$ off the primary diagonal. In addition, the matrix must be composed with some degree of consistency. For example, if we know the relationship between elements x_1 and x_2, and the relationship between x_2 and x_3, we should be able to say something about the relationship between x_1 and x_3. As a trivial example, if we know $\text{cor}(x_1, x_2) = \text{cor}(x_2, x_3) = 1$, we know that $\text{cor}(x_1, x_3) \neq 0$. More formally, we can see that a matrix that satisfies these requirements has a Cholesky decomposition and is positive definite.

Issues arise when the correlations are calculated, or as is common in financial applications, estimated inconsistently (Joubert and Langdell, 2013). In our case, for each region pair we consider, we have disregarded the existence of the remaining region or regions, thereby potentially introducing inconsistencies into the matrix. In these cases, the matrix may

have both positive and negative eigenvalues. A simple repair mechanism can be utilised, taking advantage of the spectral decomposition of the correlation matrix. We replace negative eigenvalues by zeros (or a very small positive number), and the matrix is rescaled to obtain ones along the primary diagonal.

`repairMatrix`

A simple method (Gilli et al., 2011; Rebonato and Jaeckel, 2011) for the repair or adjustment of such matrices involves exploiting the nature of a valid, positive-definite matrix, in particular, the fact that it always has positive eigenvalues. One can decompose some matrix R into $R = Q\Lambda Q^T$ where Q contains the eigenvectors of R as its columns and Λ is a diagonal matrix containing the corresponding eigenvalues of R. Replacing negative elements of Λ with zero (or a small positive number, one can obtain Λ', and reconstruct the original matrix to obtain $R' = Q\Lambda' Q^T$, before rescaling to obtain ones on the primary diagonal. This is implemented as the `repairMatrix` function in the `NMOF` R package (Schumann, 2020). Although effective, the method is limited in that it is difficult to assess the relationship between R and R'.

`nearPD`

In an alternative approach, Higham (2002) proposes the closest correlation matrix method in which the distance between the two entries is minimised. The distance used in the calculation is the Frobenius distance, defined as

$$||R, R'|| = \sqrt{\sum_i \sum_j (R_{i,j} - R'_{i,j})^2}.$$

This method is implemented in R as the `nearPD` function in the `Matrix` package.

Comparisons

On examination of both methods, we find no notable differences between the two. Figure 8.4 shows the differences between the empirical positive semi-definite matrix and the two matrices obtained using the repair methods, for a selection of rows in the matrices. Neither method performs consistently better than the other, depending on the row considered. However, the computation time of the two methods differ, and the algorithm used in `nearPD` is not guaranteed to converge in a small number of steps. For a smaller version of the data set (used in Section 8.3.1), `repairMatrix` takes just over one minute to run and `nearPD` approximately 14 minutes. For our full 4896×4896 matrix, `repairMatrix` takes approximately 6 minutes to run and `nearPD` does not converge in under 100 iterations. For these reasons, we proceed using the `repairMatrix` function from the `NMOF` package

when adjusting our nonstationarity covariance matrix.

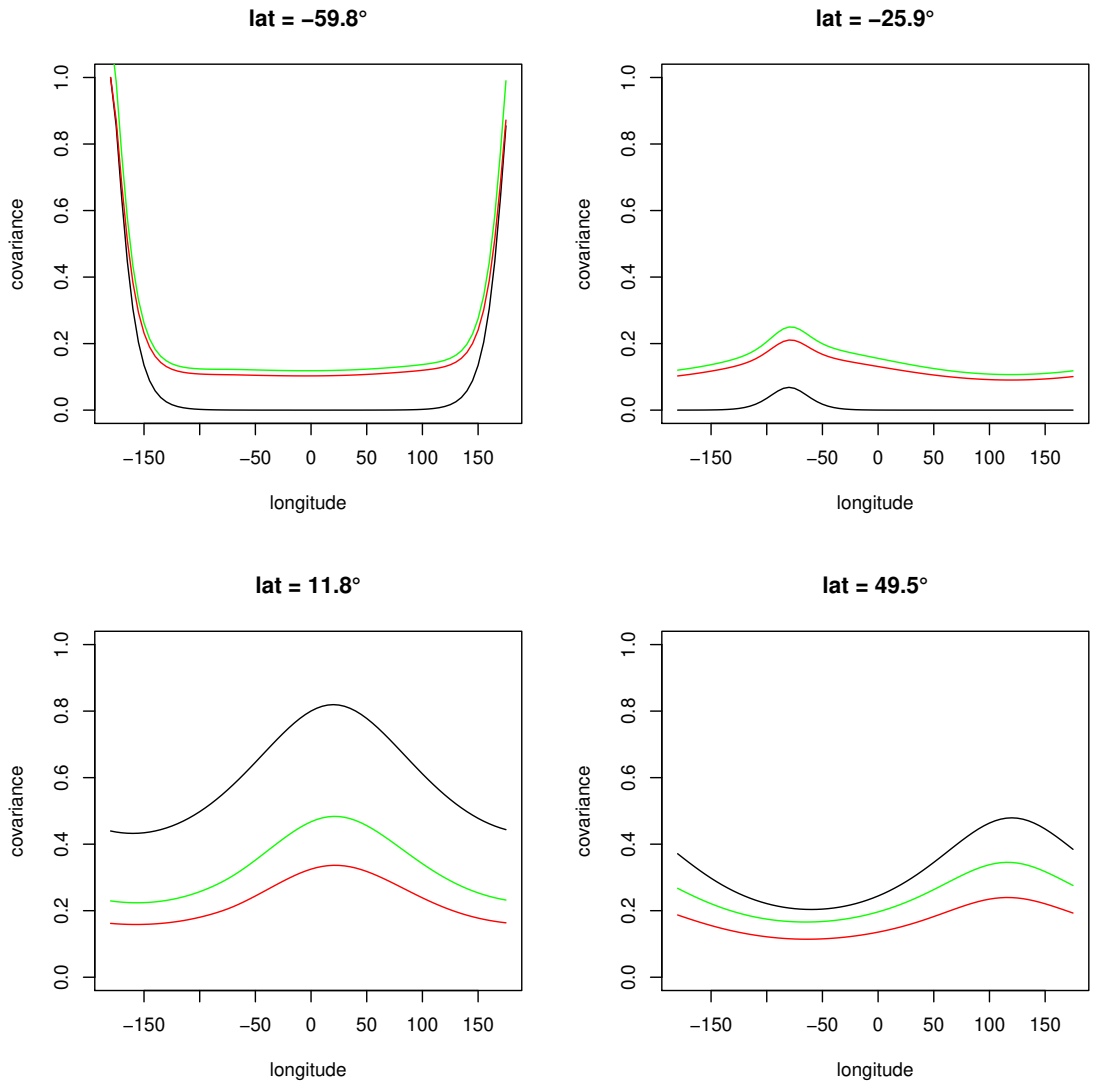

Figure 8.4: Values of the unrepaired (black) `repairMatrix` (red) and `nearPD` (green) correlation matrices for a selection of latitudes. Here we use the smaller data set for which we have valid repaired matrices using both methods.

8.3.3 Results

The expected cumulative hazard function curve for a Gaussian field with this correlation matrix can be seen in Figure 8.2. Although an improvement on the grid Matérn and weighted combination models, the local fit of this model indicated by the curve remains poor, likely due to the assumption of local stationarity over large areas of the data.

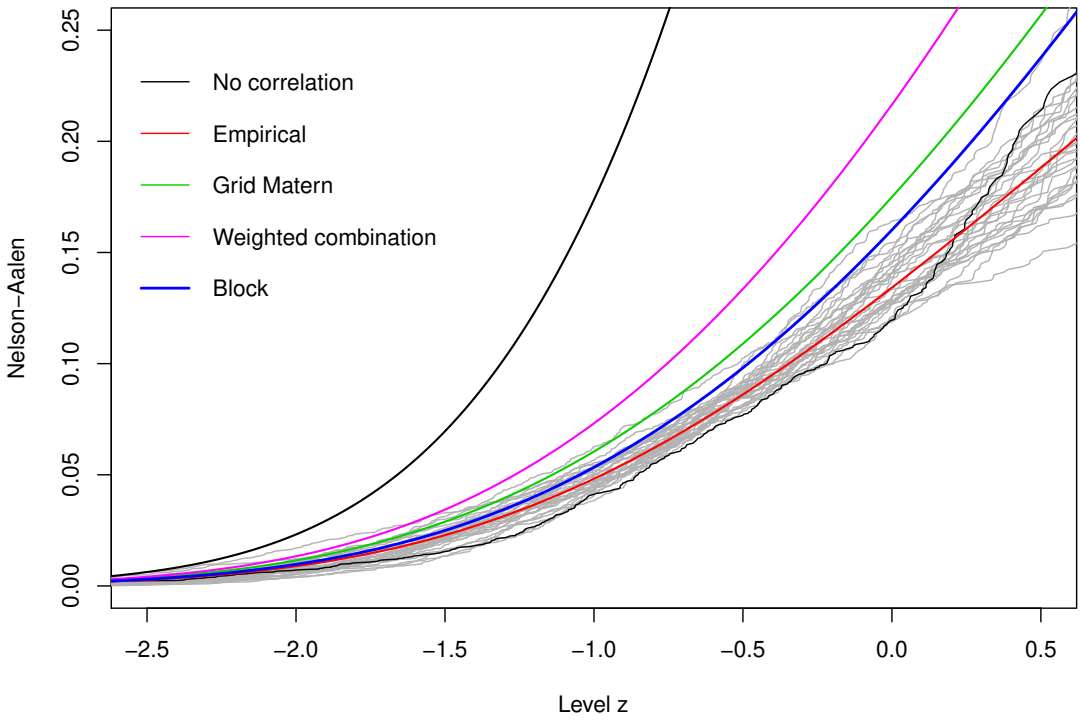

Figure 8.5: Expected cumulative hazard functions. This figure shows in blue the expected cumulative hazard curve assuming the fitted block model.

15-region model

We investigate whether division into 15 geographical regions improves the results, as it allows for a greater degree of nonstationarity. The 15-region model takes advantage of the natural division of sites between continents and oceans, although arguably these divisions are not based on geophysical properties. We continue to include coastal sites as a single unique region due to the notable differences between the coastal region and the remainder of the globe. We classify non-coastal sites as one of the following continents:

- Asia
- Africa
- Australia
- Europe
- North America
- South America

or oceans:

- Arctic
- Indian
- North Atlantic
- North Pacific
- South Atlantic
- South Pacific
- Southern
- other seas.

After usual reduction of the data set and removal of polar latitudes, there are no remaining sites in Antarctica, hence its omission from the classifications. Table 8.3 shows the parameter estimates for each region. Pairwise parameter estimates are not provided here for succinctness.

Despite the increased number of regions, the 15-region model provides only a marginal improvement over the original three-region approach. The pattern of nonstationarity

Table 8.3: Number of sites and Matérn parameter estimates for each region.

Region	#sites	η	ν
coast	517	0.0947	0.4442
Asia	237	0.0674	0.4670
Africa	216	0.1018	0.8169
Australia	54	0.0693	1.1759
Europe	338	0.0827	0.4548
North America	206	0.1896	0.2007
South America	140	0.0585	1.4703
Arctic	75	0.0800	1.7972
Indian	570	0.1107	2.1122
North Atlantic	284	0.1004	1.8062
North Pacific	562	0.1058	2.1178
South Atlantic	376	0.0936	2.2666
South Pacific	667	0.1010	2.3218
Southern	94	0.1623	1.5049
other seas	560	0.1093	0.9943

does not strictly correspond to the oceanic or continental divisions, resulting in a final fit indistinguishable from the initial three region approach, as can be seen in Figure 8.6. Later, in Section 8.5 we look in more detail at how we might best cluster the sites, and assess the validity of choosing land, ocean and coast as site classifications.

8.3.4 LatticeKrig - a convolution-based model for nonstationary Gaussian processes

We consider the LatticeKrig (LK) convolution-based model, both since its design was motivated by climate data sets similar to our wind intensities (in fact from the same CESM large ensemble project) and due to its suitability for large nonstationary data. The LK model allows for a spatially varying correlation range and variance, assuming local stationarity. This work by Nychka et al. (2018) builds on previous approaches for the design of convolution-based models and has a particular focus on highly efficient simulation of nonstationary fields. To be precise, the use of locally stationary models allows efficient computation, whilst incorporating these into a global model allows simulation of random fields. Further, computations of local covariances can be run in parallel, allowing the use of multiple processors for reduction of computation time. As with our previous work, we are using a reduced data set, however the efficiency of this method would make it a suitable candidate for model fitting on the full data. For this reduced data set the full process takes just over one minute, compared to over an hour for some of the methods

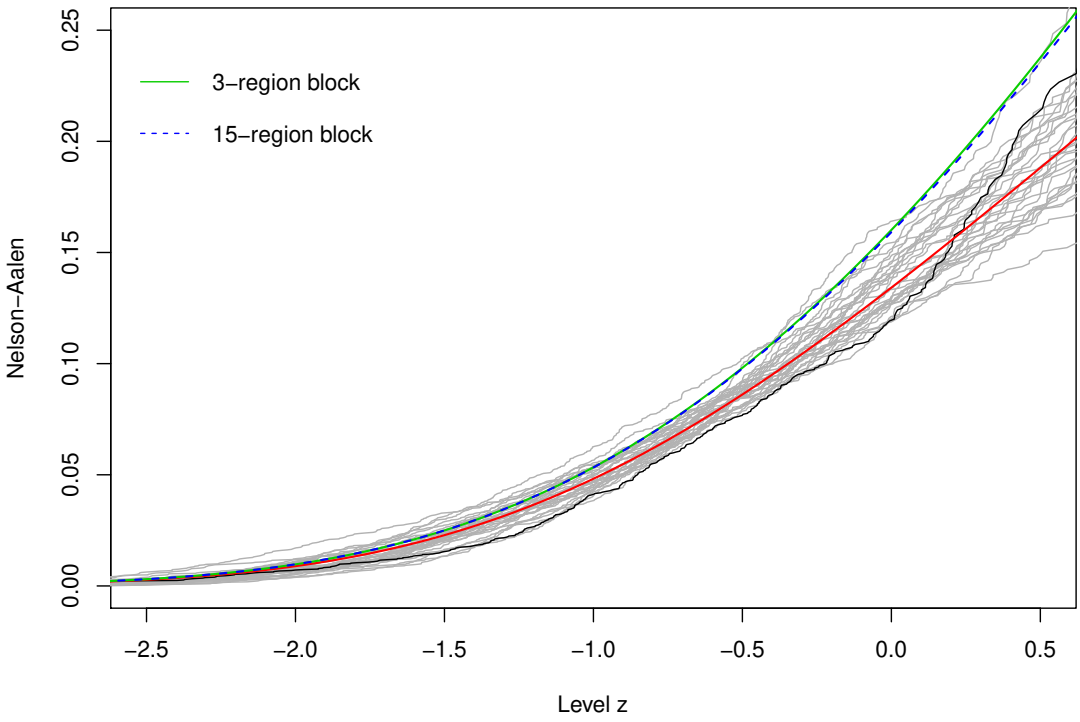

Figure 8.6: Expected cumulative hazard functions. The three region model is shown as a solid green line and the 15 region model as a dashed blue line.

previously described.

Our approach is as follows. We simulate 29 additional replicates using an LK model fit on a single initial realisation (year $t = 2006$, realisation one). Whilst in practice, we are able to calculate a covariance matrix directly from our original 30 realisations, we investigate whether a similar result can be achieved starting with a single realisation and simulating an additional 29 using the LK model. We use the `LatticeKrig` package in `R` (Nychka et al., 2016) to fit a nonstationary model for the same data subset as we have used throughout this and the previous chapter.

The LatticeKrig model

The LK model builds on fixed-rank kriging but models the precision matrix of the coefficients as a sparse matrix, allowing the incorporation of a greater number of basis functions relative to the size of the data. The multi-resolution basis allows enhanced flexibility and superposition of latent processes at different spatial scales forms the basis for the nonstationary version of the model.

For a Gaussian random field $z(x)$,

$$z(x) = s(x)^T\beta + g(x) + \epsilon(x),$$

where $s(x)^T\beta$ is the vector of covariates with corresponding linear fixed-effect parameters, $g(x)$ is a mean-zero, smooth Gaussian process and $\epsilon(x)$ is an independent Gaussian white noise process. We are interested in modelling the stochastic process $g(x)$.

The LK model can be described as follows. Let $g_l(x)$ corresponding to the l^{th} level have mean zero and marginal variance $\sigma_l(x)^2$. Then $g(x)$ is the sum of L independent processes

$$g(x) = \sum_{l=1}^{L} g_l(x)$$

with marginal variance

$$\sum_{l=1}^{L} \sigma_l(x)^2.$$

Then each latent process $g_l(x)$ can be defined as

$$g_l(x) = \sum_{k=1}^{m(l)} c_k^l \phi_{k,l}(x).$$

Here, $\phi_{k,l}$ is a sequence of fixed basis functions where $1 \leq k \leq m(l)$ and c^l is a vector

of mean-zero multivariate normal coefficients independent between levels, with covariance matrix Q_l^{-1}. Then let

$$\Phi_{i,j}^l = \phi_{j,l}(x_i)$$

with columns corresponding to the $m(l)$ basis functions for level l. Hence, the observation model for the random field can be written as

$$z = s\beta + \Phi c + \epsilon,$$

where $c = \{c_1, \ldots, c_L\}$ and $\Phi = [\Phi^1, \ldots, \Phi^L]$.

For details on the formulation of the independent basis functions see Nychka et al. (2018).

Each coefficient $c_{k,l}$ is associated with a site $v_{k,l}$, and since the data is defined on a grid, each site will have a neighbourhood of up to four additional sites (i.e. its nearest neighbours), $\mathcal{N}_k$.

Assume that

$$a_k c_{k,l} - \sum_{k* \in \mathcal{N}_k} c_{k*,l} = v_{k,l},$$

where $a_k < 4$ is a sequence of parameters and $v_{k,l}$ are iid $N(0,1)$ random variables. Then let B_t be a square spatial autoregression matrix, with diagonal elements a_k.

Using this formulation we have $B_l \mathbf{c}_l = \mathbf{v}_l$ and $Q_l = (B_l^T B_L)$ is the precision matrix for $\mathbf{c}_l$. Construction of Q in this way constrains it to be positive definite.

Modelling nonstationarity

Nonstationarity in the LatticeKrig model can be achieved through spatial variation in the weighting across scales through σ_l, the overall correlation range, and/or through spatial variation in a, which control the dependence of the random field and the correlation range within a level l of g_l.

Results

We use the `R LatticeKrig` package to fit a nonstationary LK model on the data subset for year $t = 2006$, and realisation one as has been used previously. For the specification of a nonstationary model, we allow the `a.wght` parameter to vary between lattice points. We choose not to vary the parameter between levels. The efficiency of this approach is a significant improvement on previous methods. The model itself allows simulation of additional random fields, but not an explicit covariance matrix, hence we decide to take a bootstrapping approach. After fitting the model, we simulate an additional 29 realisations

to obtain an artificial version of our original 30 data subsets. Using these additional 29 replicates, we obtain a covariance matrix directly as in Section 7.2. Figure 8.2 shows the cumulative hazard function for expected local maxima under a Gaussian model with this covariance matrix, a noticeable improvement on previous models at higher levels, but performing poorly at lower levels.

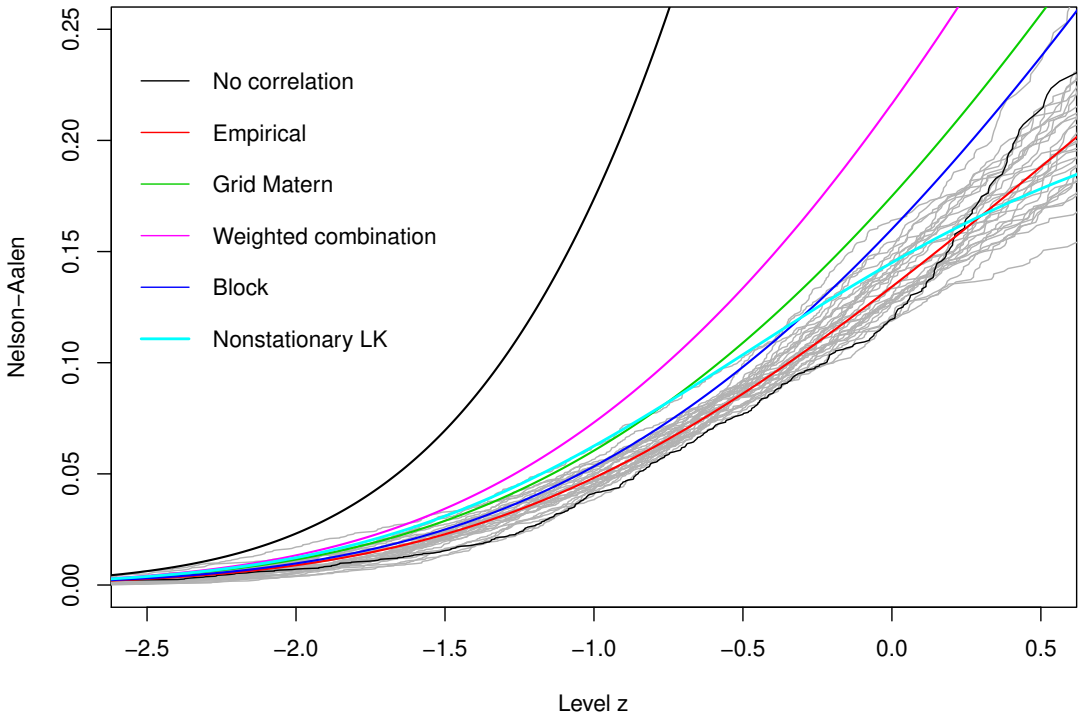

Figure 8.7: Expected cumulative hazard functions. This figure shows in cyan the expected cumulative hazard curve assuming the fitted LatticeKrig model.

8.3.5 Nonstationary Gaussian process model

To conclude this section we look at a Bayesian approach to nonstationary process modelling. In this, a convolution-based spatially varying kernel is used in which parameters can be modelled as a nonparametric function of spatial location, in order to handle nonstationary and/or nonseparable covariance structures. We describe this approach proposed by Konzen et al. (2019) before demonstrating its application to a single wind intensities data subset. We also illustrate the improvement in the model fit when we use the full 30 replicates.

We are interested in, but also most challenged by the accurate estimation of the covariance

function. In Q dimensions, as Q increases, the 'curse of dimensionality' is often dealt with by assuming separability of the covariance function. That is, the covariance function is separable if it can be factorised into Q covariance functions, one for each dimension. This approach has many advantages, namely it improves computational efficiency and guarantees the covariance function to be positive definite. In present research, nonseparable covariance functions are largely restricted to stationary fields.

Gaussian process regression model

The method considers observations as realisations of an underlying random process $z(x)$, with mean function $\mu(x)$ and covariance $C(x, x') = \text{Cov}[z(x), z(x')]$.

The Bayesian process model

$$z(x) = f(x) + \epsilon(x), \quad \epsilon(x) \sim N(0, \sigma^2)$$

treats f as an unknown process, about which inferences can be made via a prior distribution over the random function f, such as a Gaussian process prior.

Parameterised by its mean and covariance functions, $\mu(\cdot)$ and $C(\cdot, \cdot)$ respectively, the Gaussian process prior can be written as

$$f(\cdot) \sim \text{GP}(\mu(\cdot), C(\cdot, \cdot)).$$

When used with the GP prior, the Bayesian process model above is known as the Gaussian process regression model (GPR).

Appropriate choice of C allows numerous forms of f to be modelled, contributing to the popularity of GPRs. In one or two dimensions, both parametric and nonparametric functions are available. For example, given some valid correlation function g, we could use a stationary parametric covariance function of the form

$$\text{Cov}[z(x), z(x+h)] = \sigma^2 g(\sqrt{h^T B h}) + \sigma_\epsilon^2 \delta_h,$$

where B is the anisotropy matrix. Here, $\delta_h = 1$ if $h = 0$ and 0 otherwise. The diagonal elements of B, $\text{diag}(b_1, \ldots, b_Q)$ are known as decay parameters, and $1/b_q$ can be interpreted as the range parameter. Each controls how fast the function f varies in each direction. If the off-diagonal elements of B, b_{pq} are non-zero, there is interaction between the covariance in different dimensions and thus the covariance function is nonseparable. It is clear that in order to fix properties such as stationarity and separability for f, the covariance function must be chosen with some care. To allow for nonstationarity, the authors define a convolution-based approach using a spatially varying kernel, with nonparametric

modelling of the parameters.

A convolution-based approach

The convolution-based approach can be defined as follows (Higdon et al., 1999). A spatial process $f(\cdot)$ can be written as the convolution

$$f(x) = \int_{\mathbb{R}^Q} C_x(u)s(u)du,$$

where $s(\cdot)$ is a Gaussian white noise process and C_x is a spatially varying kernel. Thus the covariance function for f is

$$\text{Cov}[f(x), f(x')] = \int_{\mathbb{R}^Q} C_x(u)C_{x'}(u)du$$

which is positive definite when

$$\sup \int_{\mathbb{R}^Q} C_x(u)^2 du < \infty.$$

The ease of finding a kernel that satisfies the above, relative to direct specification of a covariance function, has contributed to the popularity of the convolution approach. The covariance function for f above is valid in every Euclidean space (Paciorek and Schervish, 2006).

Assuming a Gaussian kernel, $C_x(u)$, the covariance of f is

$$\text{Cov}[f(x), f(x')] = \sigma^2|\Sigma(x)|^{1/4}|\Sigma(x')|^{1/4}\left|\frac{\Sigma(x)+\Sigma(x')}{2}\right|^{-1/2} exp\{-Q_{xx'}\}.$$

Here,

$$Q_{xx'} = (x-x')^T\left(\frac{\Sigma^{-1}(x)+\Sigma^{-1}(x')}{2}\right)^{-1}(x-x').$$

A more general class for nonstationary covariance functions is

$$\text{Cov}[f(x), f(x')] = \sigma(x)\sigma(x')|\Sigma(x)|^{1/4}|\Sigma(x')|^{1/4}\left|\frac{\Sigma(x)+\Sigma(x')}{2}\right|^{-1/2} g(\sqrt{Q_{xx'}}),$$

for some valid isotropic correlation function $g(\cdot)$ (Paciorek and Schervish, 2006).

The anisotropy matrix $\Sigma(x)$, which measures how quickly the fluctuation of random processes vary over x, must be positive definite, a restriction which can be ensured via parameterisation. Konzen et al. (2019) chose a spherical parameterisation due to its inter-

pretability, although many other parameterisations exist and would be suitable.

Results

To fit this model we use software provided by Konzen et al. (2019). With a single replication, we are unable to obtain a good local fit using this model as can be seen in Figure 8.8. Using the expected cumulative hazard curve as an assessment of model fit, we do not see an improved performance when compared with other nonstationary models. However, as we show in the next section, we are able to closely match the empirical curve using all 30 replicates to fit the model. It is clear that modelling highly nonstationary correlation using a single replication is challenging and all of the methods discussed in this chapter could produce improved fits provided with more data.

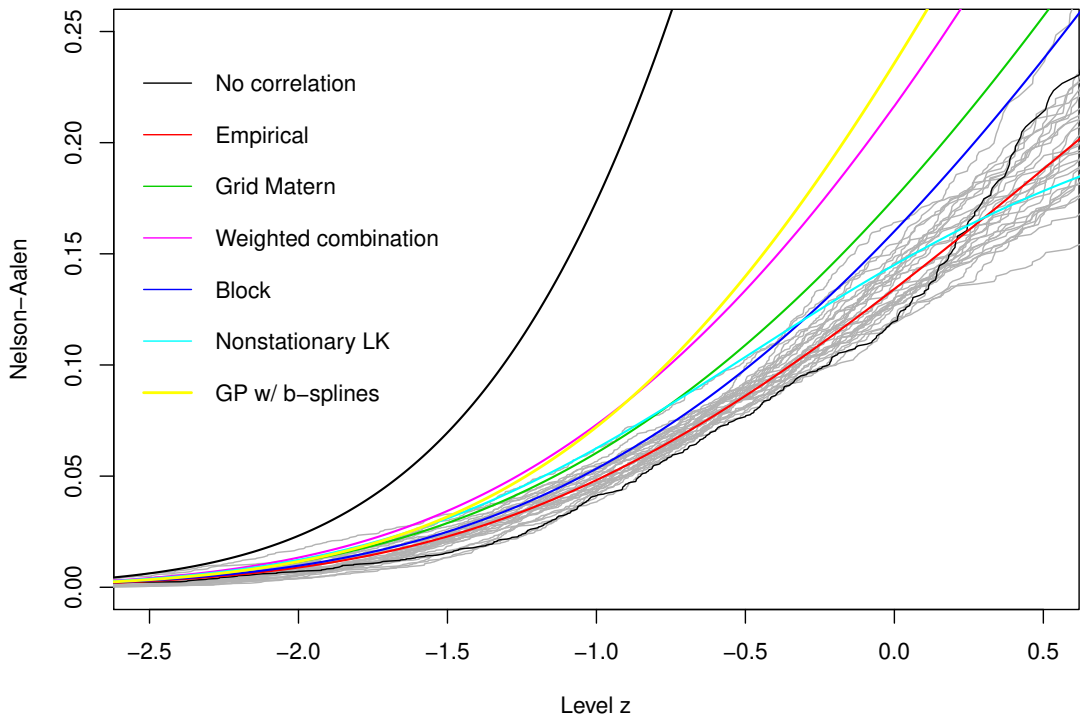

Figure 8.8: Expected cumulative hazard functions. This figure shows in blue the expected cumulative hazard curve assuming the fitted Gaussian Process model.

8.4 Comparison of nonstationary models using TEH Nelson-Aalen plots (multiple replicates)

Finally, after examining the extent to which nonstationary models can be accurately fit using single replicates, we return to the final Gaussian Process model when fit on all 30 replicates. Since our primary interest in this and the previous chapter is understanding the use of TEH methods for assessing model fit, we do not investigate additional models when fit on all 30 replicates. Clearly, as shown in many of our TEH assessment figures, use of the empirical covariance matrix provides excellent results. However, we briefly consider fitting the GP model on 30 replicates as an example of the improvements available with additional data. For the Gaussian Process model, although fitting on a set of replicates takes several hours (~ 10), it is relatively straightforward and produces an excellent fit as can be seen in Figure 8.9. The estimated covariance matrix is almost indistinguishable from that calculated directly from the data. As previously discussed, the expected Nelson-Aalen curves do not give a full picture, rather they illustrate the fit on a local level. To compliment the TEH figures, we also look back at an alternative method of covariance matrix comparisons for this method.

Figure 8.10 shows the covariance values at various lags and directions for both our empirical matrix and that from the nonstationary GP model. The empirical data figures were also shown previously in Section 7.3. We can immediately see the nonstationarity of the Gaussian Process matrix, with correlations in the Northerly direction standing out relative to other directions. The difference between the two matrices is clear from these figures, despite being almost indistinguishable from the topological event history plot. The smoothness of the GP matrix is forced by the parameterisation of the anisotropy matrix used and this smoothness is noticeably present in the GP matrix in contrast to the empirical matrix. Data simulated using the fitted matrix (Figure 8.11) does not stand out as being particularly different in structure from the original data, although we can see more smaller areas at the extremes.

We can see that whether we choose to fit a model (for inference about the covariance structure) or calculate a covariance matrix from the replicates directly, the results are a significant improvement on any of the models fit to a single replicate.

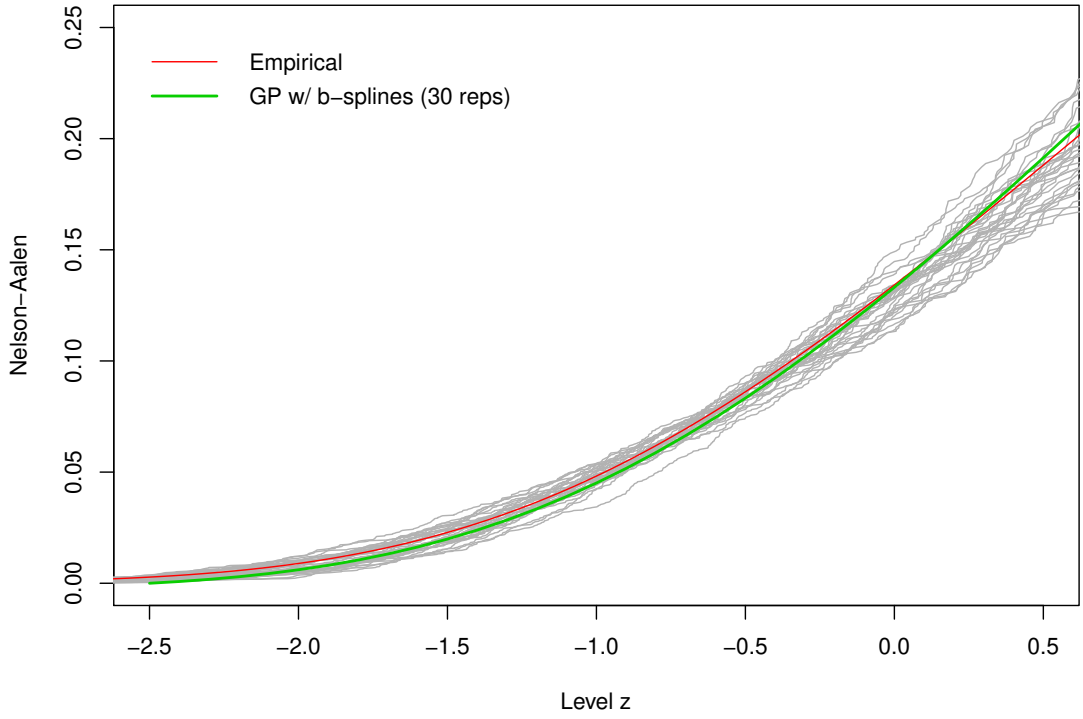

Figure 8.9: Expected cumulative hazard functions. This figure shows in green the expected cumulative hazard curve assuming the fitted Gaussian Process model fitted on all 30 replicates. When fitting with this full set of replicates we see an excellent fit.

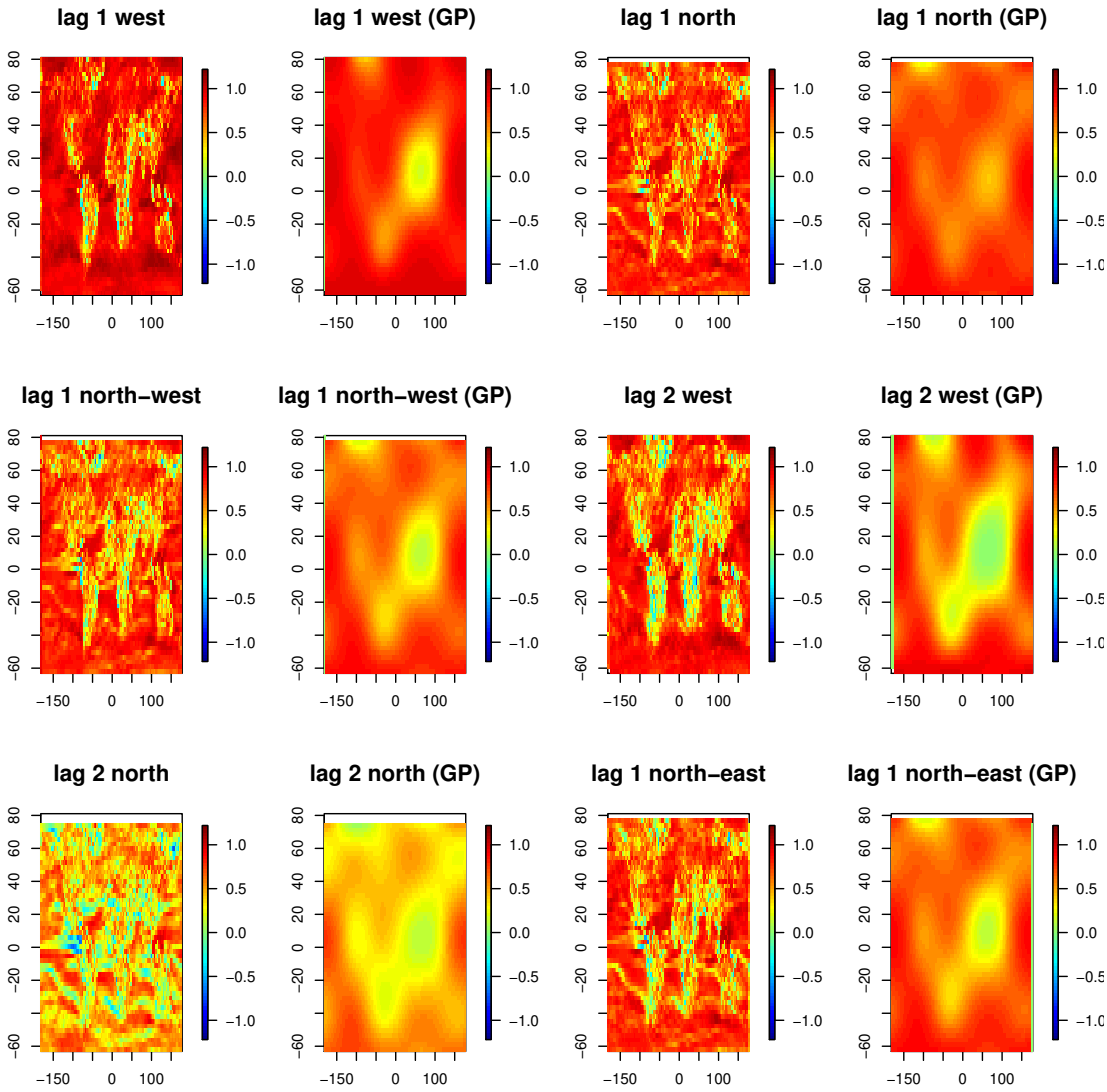

Figure 8.10: Correlations from the empirical and nonstationary GP matrices for a selection of lags and directions. We see significantly more smoothness over a given direction and lag for the nonstationary GP matrix.

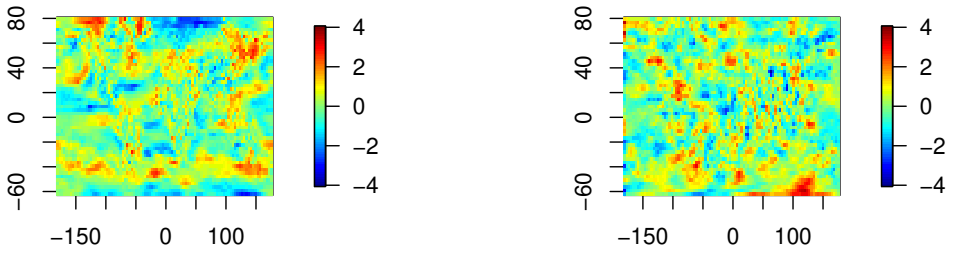

Figure 8.11: Original (left) and simulated (right) data using the nonstationary Gaussian Process model fitted on all 30 replicates.

8.5 Site clustering

We established in Section 8.3.2 that dividing the sites into geographical regions is an effective way to capture some of the nonstationarity in geospatial data. Previously, our choice of regions has been based solely on geophysical descriptors or political boundaries. Given the empirical covariance matrix, obtained from the 30 data realisations, we cluster sites into regions using an unsupervised clustering algorithm and compare the results to our earlier land-ocean-coast approach. We investigate how unsupervised clustering into regions has the potential to reduce storage requirements for the large empirical covariance matrix, by reducing it to a vector of site classifications and several model parameters.

For each of the 4896 sites we have four neighbours of interest and hence can extract a 5×5 covariance matrix for each site. Taking advantage of symmetry, we reduce this to a 10-vector for each site. In edge cases, we have missing data due to a reduced set of neighbours. There are many methods for dealing with this, either pre or post-clustering (Graham, 2012; Little, 1987; Friedman et al., 2001). Methods used before clustering include complete case analysis (ignoring sites with missing data completely) and various imputation approaches. We chose to take a two-step approach running a complete case analysis, followed by nearest-neighbour assignment for sites with missing data. Since the proportion of sites with missing data is relatively low, this is not problematic for the initial clustering, and visual inspection does not show anomalous results.

For the clustering itself, we use a simple k-means algorithm, investigating $k = 3$, 5, 10 and 20 clusters. Figure 8.12 shows the distribution of sites into clusters for each k. For the $k = 3$ case, we note the clear distinction between land (primarily in class three) and ocean (primarily in class two), justifying the positive results in our land-ocean-coast block model. The remaining cluster (class one) is more complex, covering some coastal regions as well as much of North America and striations through the oceans. In the following section, we focus on the $k = 3$ case.

Returning to our Nelson-Aalen curves, we take the average local correlation for sites in each class. Producing separate expected NA curves for each shows clearly the effect of local correlation on the position of the curves as can be seen in Figure 8.13. Class 2, the largest class (containing 2355 of the 4896 sites), is the only one to fall below the empirical curve, with the two smaller classes sitting above the empirical.

We can see the $k = 3$ results more clearly in Figure 8.14, in which the expected NA curves for each site are plotted individually, coloured according to class. We can see the differences not just in position but also in shape and spread of the curves, suggesting different distributions for the different regions.

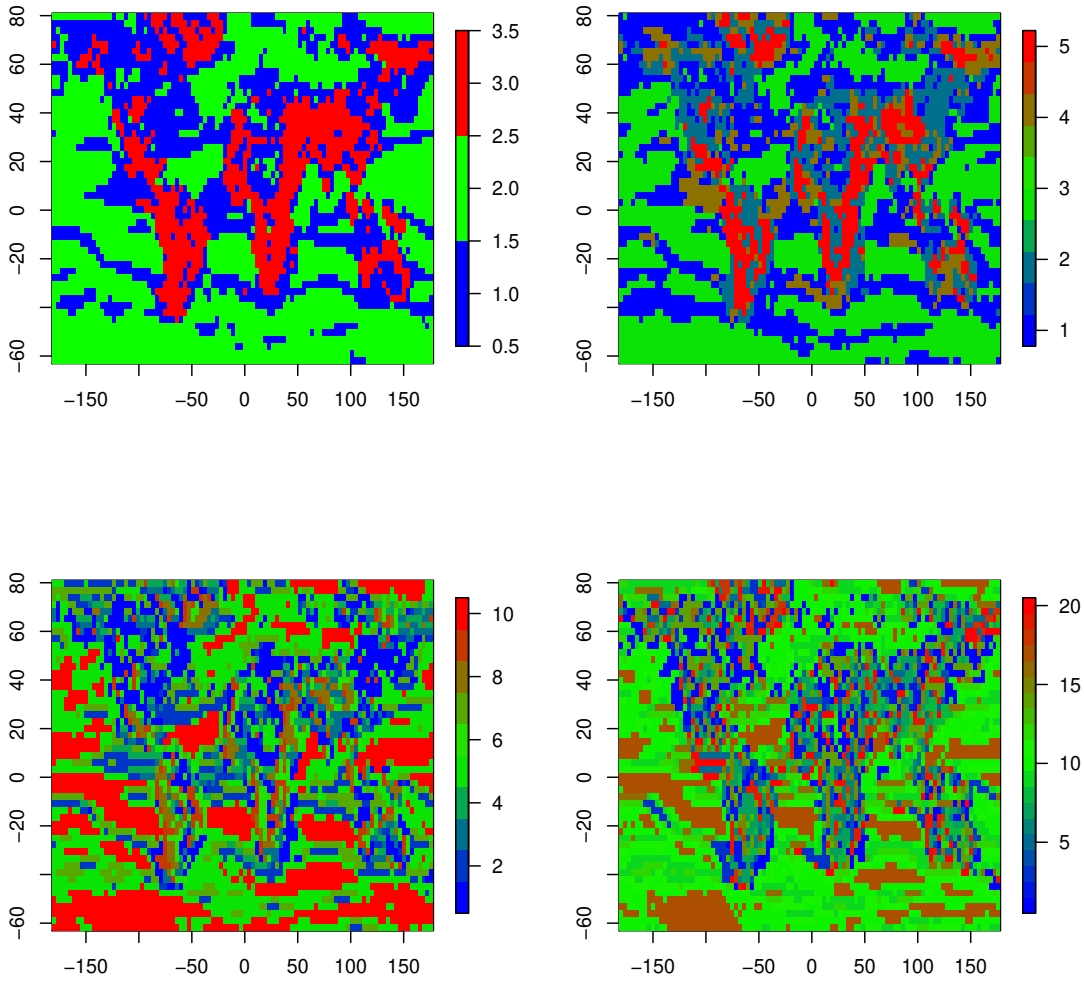

Figure 8.12: Classification of sites into $k = 3$, 5, 10 and 20 clusters using the `kmeans` algorithm.

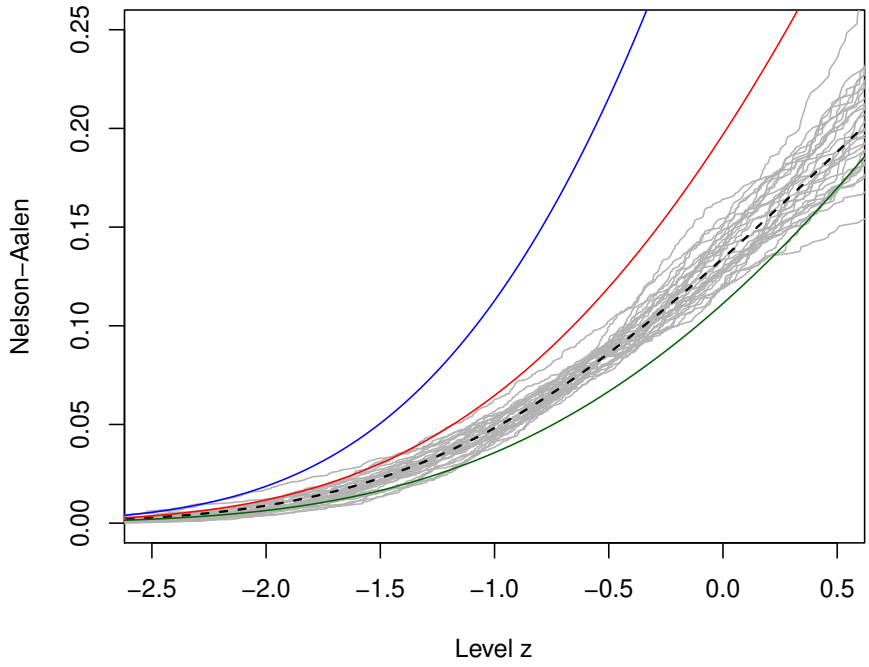

Figure 8.13: Expected NA curves based on average local correlation for $k = 3$. Class one: red, class two: green, class three: blue. The black dashed line shows the expected NA based on empirical covariance.

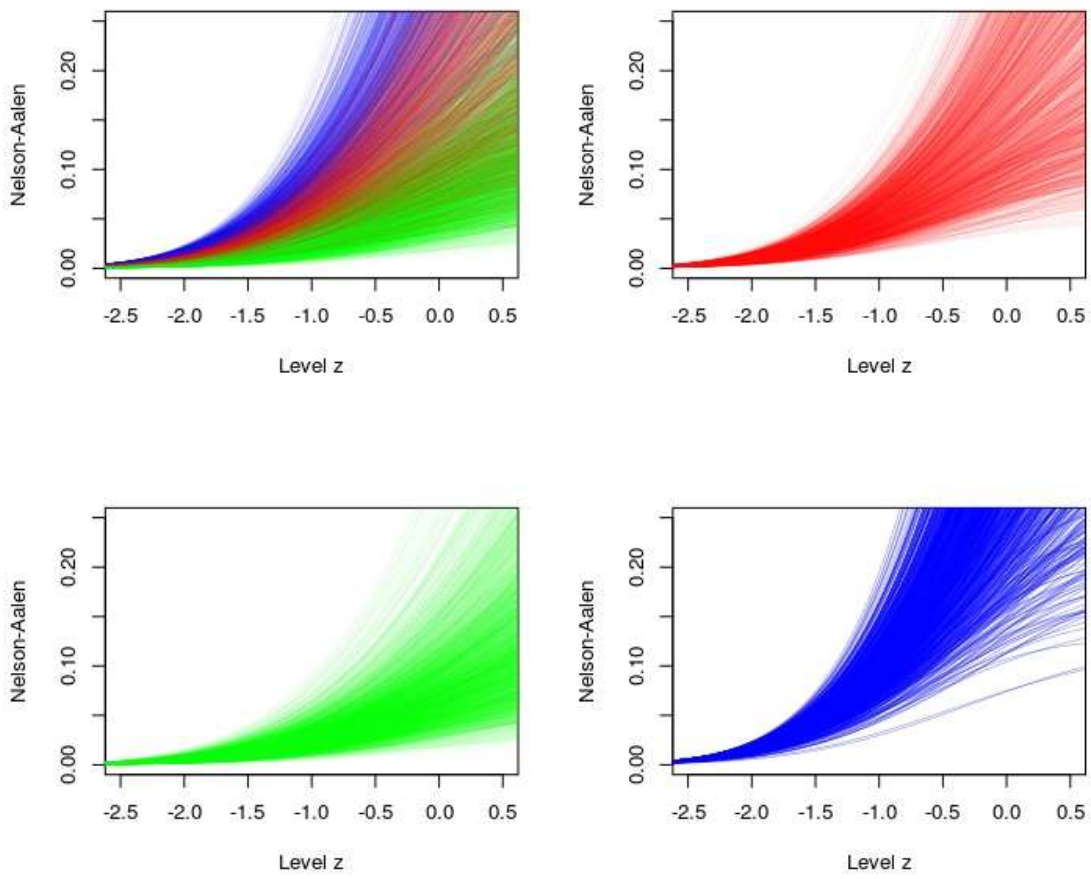

Figure 8.14: Individual NA curves by site for $k = 3$ clustering. Class one: red, class two: green, class three: blue.

We can further see the differences between classes via the actual values of the average local correlations, shown in Figures 8.15, 8.16 and 8.17 for classes one, two and three respectively. The figures show the 10 local correlations within the neighbourhood of a site. Most striking here are the noticeably lower values between all sites in class three, relative to classes one and two. The corresponding curve, therefore, sits closer to where we would see a curve calculated from uncorrelated sites. We also see for classes one and two the lower correlations in the North-South direction than in the East-West direction, as we saw previously for the data set as a whole.

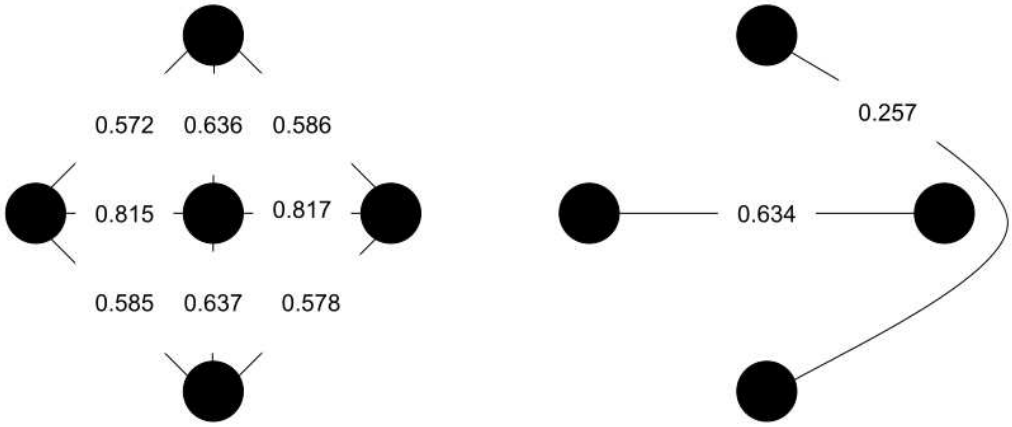

Figure 8.15: Average local correlation for class one. In the left figure we see lag one correlations in all directions. In the right figure we see lag two correlations in the North-South and East-West directions.

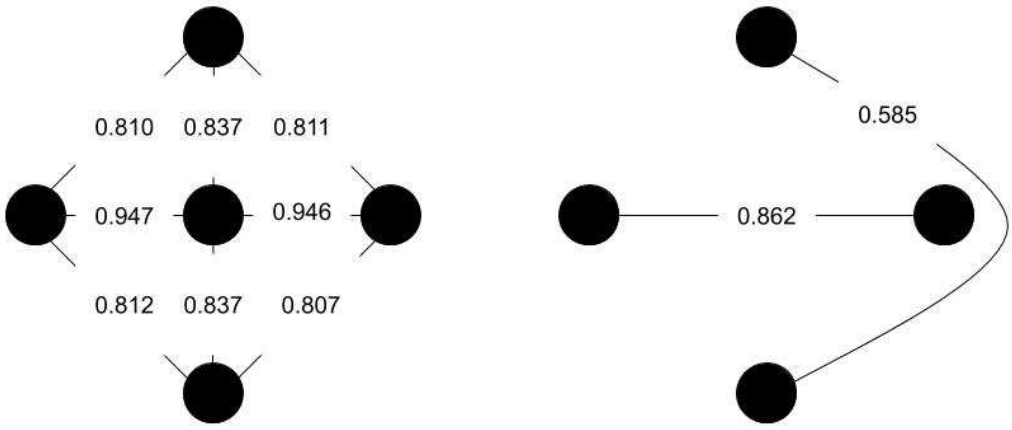

Figure 8.16: Average local correlation for class two.

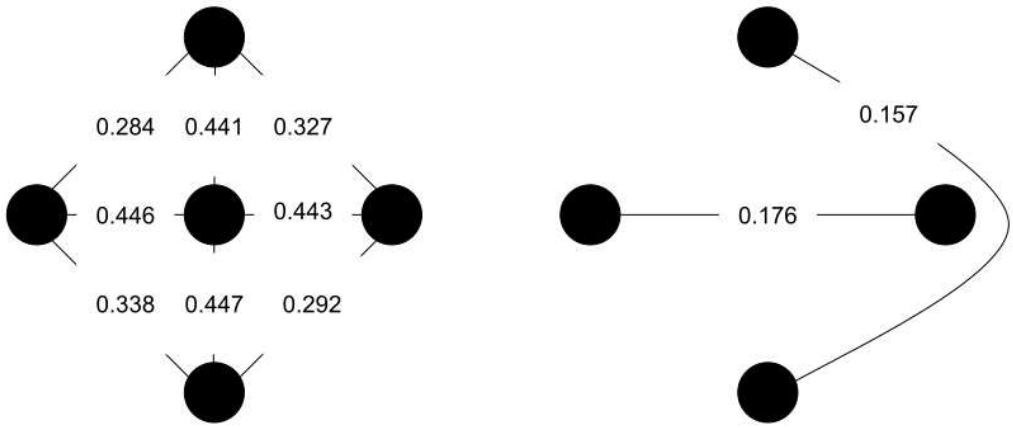

Figure 8.17: Average local correlation for class three.

8.6 Conclusions

The exercise of modelling correlation structure for a highly nonstationary, spatial data set on the surface of a sphere is complex and comes with many challenges, several of which we addressed in this chapter. We investigated stationarity in the data and considered methods to overcome this. We fit a selection of correlation models and demonstrated how expected cumulative hazard function curves can be used to compare and assess the local fit. Our results showed that when only a single realisation is available, it is difficult to produce a fit close to the empirical correlation. In scenarios such as ours, where we have multiple realisations of a data set, it makes sense to take advantage of this and use the empirical covariance structure, however adjustment to obtain a positive-definite matrix may be required. Having established an effective and computationally efficient method, this will not be a significant undertaking. As we have mentioned throughout the chapter, numerous methods exist for the modelling of nonstationary data on the surface of a sphere, largely from research into similar types of geospatial data. A valuable extension of this work would include looking more closely at these, in particular, evolutionary spectrum models (Castruccio and Guinness, 2017) and covariate-dependent models (Reich et al., 2011). Further, one of the primary limitations to this work is the size of the data; a single realisation of our reduced data set contains 4896 values, compared with 55296 in a realisation of the full data set. Fitting models to a data set of this size comes with additional challenges, exploration of which would be an interesting extension to this work.

Chapter 9

TEH as a test for Gaussianity

Perhaps one of the more valuable applications of the topological event history methods proposed in this work is the ability to identify non-Gaussian random fields that are not distinguishable by other means, for example by standard marginal analysis. We showed via simulation studies in Chapter 6 how this method was able to distinguish between true Gaussian random fields and other random fields that had been transformed to have marginal Gaussianity and indistinguishable correlation structures. In this chapter, we look at the alternative methods available for assessing Gaussianity and investigate their effectiveness in comparison to the topological event history approach. Further, we use TEH to verify Gaussian data produced using a selection of packages. Our methods show the extent to which the output from standard simulation packages produce Nelson-Aalen curves corresponding to true Gaussian data.

9.1 Methods for testing Gaussianity

When modelling data or doing statistical tests, knowing or being able to make an informed assumption of the distribution of the data is important. In many scientific applications, data is considered to be Gaussian, usually following some test to validate this assumption. A Gaussian assumption allows the application of useful large sample properties and common statistical tests. Although there exist a number of tests for Gaussianity, most are able to test only for the absence of marginal Gaussianity. Beginning with some of the simpler tests and moving onto the more advanced, we look at a selection of methods commonly used.

9.1.1 Marginal Gaussianity

We are particularly interested in distinguishing between Gaussian random fields and those which show only marginal Gaussianity. To understand the different tests available, we consider nine data sets, each on a 100×100 grid. Three are Gaussian random fields with Matérn covariance and three are transformations of Gaussian random fields (with underlying Matérn covariance) to be χ^2_1 fields, with additional transformation to marginal Gaussianity. The final three are transformations of Gaussian random fields (with underlying Matérn covariance) to be χ^2_1 fields, without marginal Gaussianity. Each of the random fields are not matched for resulting correlation structure, since here we are not interested specifically in comparing them, rather we aim to identify those from a true Gaussian distribution only.

Table 9.1: Gaussian and marginally Gaussian transformed data for assessment of test methods.

Data	Distribution	Matérn parameters
rf1	Gaussian	$\eta = 5$, $\nu = 1$
rf2	Gaussian	$\eta = 10$, $\nu = 1$
rf3	Gaussian	$\eta = 5$, $\nu = 2$
rf4	Marginally Gaussian χ^2_1	$\eta = 5$, $\nu = 1$
rf5	Marginally Gaussian χ^2_1	$\eta = 10$, $\nu = 1$
rf6	Marginally Gaussian χ^2_1	$\eta = 5$, $\nu = 2$
rf7	χ^2_1	$\eta = 5$, $\nu = 1$
rf8	χ^2_1	$\eta = 10$, $\nu = 1$
rf9	χ^2_1	$\eta = 5$, $\nu = 2$

Table 9.1 and Figure 9.1 show these nine different data sets. Visually we can distinguish between the three models, although the differences between true Gaussian and marginally Gaussian χ^2_1 are less clear and could be attributed to the difference in correlation structure. The untransformed χ^2_1 data sets on the bottom row of Figure 9.1 are immediately differentiable from the others.

Methods which test for marginal Gaussianity should not be able to distinguish between the true Gaussian and marginally Gaussian χ^2_1 data sets. This includes the visual tests shown below, as well as the tests which look for skewness in the data. Regardless, we give a brief outline of these tests and the results obtained.

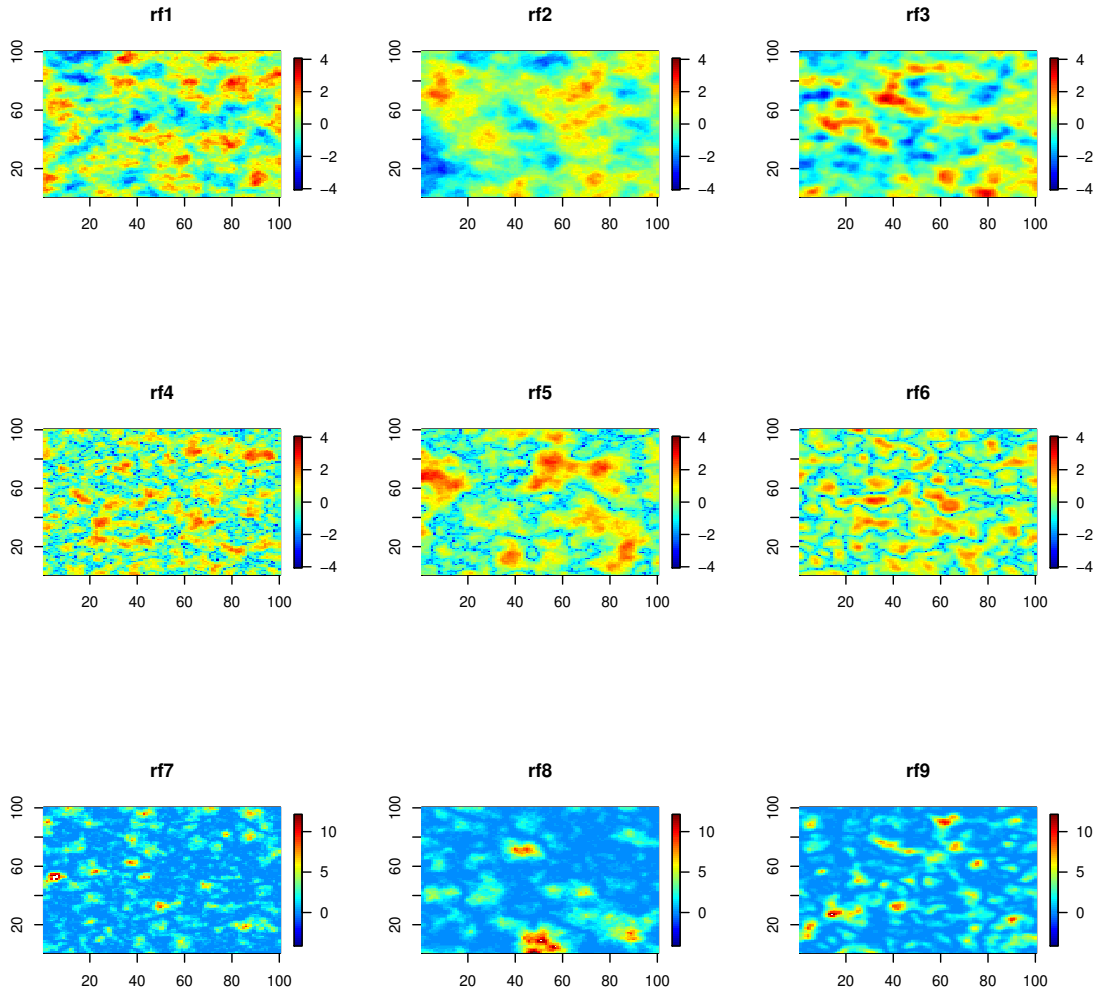

Figure 9.1: Data sets for testing. Top row: Gaussian, middle row: marginally Gaussian χ^2_1, bottom row: untransformed χ^2_1. Note the different scale on the bottom row plots.

9.1.2 Visual tests

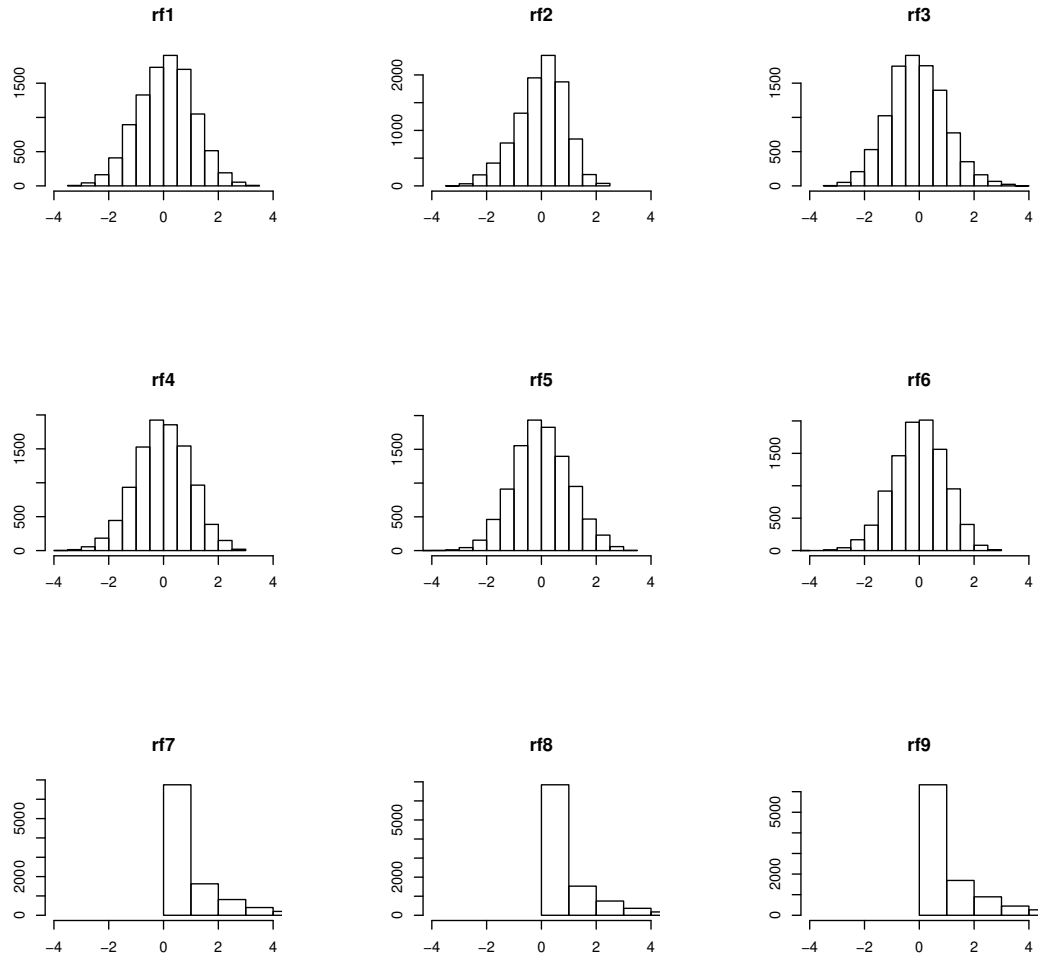

Figure 9.2: Histograms for each of our nine data sets. As expected, these do not distinguish between the marginally Gaussian χ^2_1 and true Gaussian data sets. The untransformed χ^2_1 data sets are clear.

Visual tests provide a quick, clear method for one to assess normality in a data set. A plot indicating normality provides no guarantee, but a poor looking plot is a simple and accessible way to identify data with non-Gaussian marginals. The following are examples of visual aids commonly used to verify assumptions of normality.

Figures 9.2 and 9.3 demonstrate this. Visually, we can not distinguish between the top and middle rows of each and these figures do not inform us beyond reasonable verification of marginal Gaussianity. The bottom row, without marginal Gaussianity are immediately

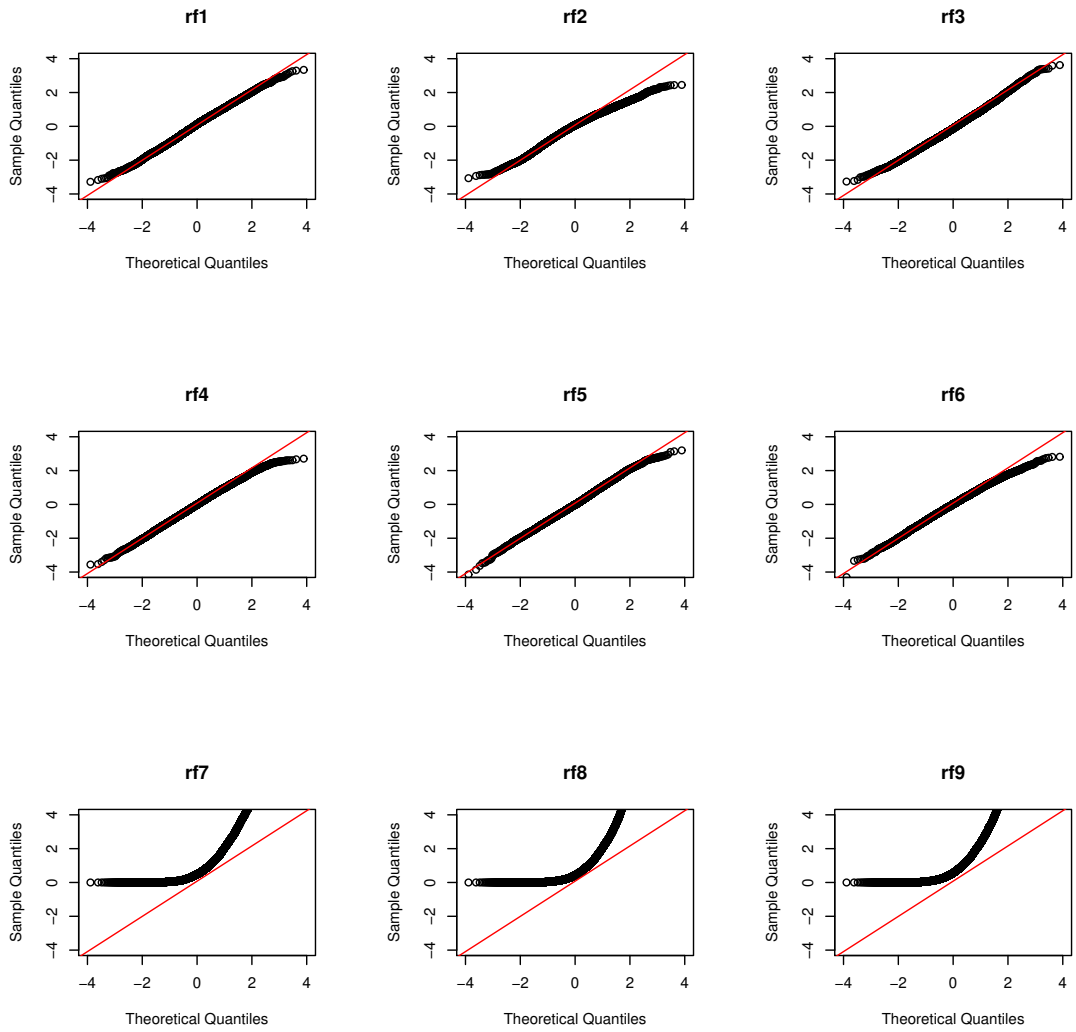

Figure 9.3: QQ plots for each of our nine data sets. Again, we cannot distinguish between the marginally Gaussian χ^2_1 and true Gaussian data sets. QQ plots for the untransformed χ^2_1 data sets are clearly different.

evident from both the histograms and the QQ plots.

9.1.3 Numerical tests

In addition to visual tests, we consider a selection of numerical tests for Gaussianity. Designed for testing uncorrelated data, these tests are insufficient for testing spatially correlated data, as can be seen in the following sections. In Table 9.2 we include results for a Gaussian $N(0,1)$ random field of the same size with no spatial correlation, rf0 for comparison.

Table 9.2: Standard numerical tests for Gaussianity applied to random fields with different distributions and spatial correlation.

Data	p-values				
	D'Agostino's	Jarque -Bera	Kolmogorov -Smirnov	Anderson -Darling	Cramer von Mises
rf0	0.563	0.6821	0.2695	0.545	0.549
rf1	< 0.001	< 0.001	0.001	< 0.001	0.009
rf2	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
rf3	< 0.001	< 0.001	0.016	< 0.001	0.023
rf4	< 0.001	< 0.001	0.061	< 0.001	0.107
rf5	0.132	0.036	0.055	< 0.001	0.060
rf6	< 0.001	< 0.001	0.056	< 0.001	0.036
rf7	< 0.001	< 0.001	< 0.001	< 0.001	0.018
rf8	< 0.001	< 0.001	< 0.001	< 0.001	0.030
rf9	< 0.001	< 0.001	< 0.001	< 0.001	0.018

D'Agostino's K-squared test The D'Agostino K-squared test (D'Agostino, 1970) is a goodness-of-fit measure, assessing the likelihood of data having originated from a Gaussian distribution. This is achieved via transformations of the skewness and kurtosis of the data and thus is only able to distinguish Gaussian data from skewed alternatives. Hence, we would not expect it to identify our transformed data, and this can be seen in Table 9.2.

Jarque-Bera test The Jarque-Bera test (Jarque and Bera, 1987; Bera and Jarque, 1981) is another goodness-of-fit test, again testing sample data for Gaussian skewness and kurtosis. We use the implementation `jarque.bera.test` from the `R tseries` package. As with the D'Agostino test, it is unable to identify our transformed data (see Table 9.2) as we would expect.

Kolmogorov-Smirnov test The Kolmogorov-Smirnov test (Justel et al., 1997; Marsaglia et al., 2003; Lopes et al., 2007) tests for difference between two sets of sample data, or between a sample data set and a reference distribution. For the latter, it tests by measuring the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. We use the implementation `ks.test` from the `R stats` package.

Anderson-Darling test The Anderson-Darling test (Anderson and Darling, 1954) is able to test whether a data sample comes from a specific distribution. It is a more sensitive modification of the Kolmogorov-Smirnov test and utilises the specific distribution being tested against to calculate critical values. The test belongs to a class of tests based on the empirical distribution function. We use the implementation `ad.test` from the `R nortest` package. We do not see a difference in results for the different data sets.

Cramer-von Mises criterion The Cramer-von Mises criterion (Cramer, 1928; Anderson, 1962) compares the empirical distribution function of a sample of data to the cumulative distribution function of a reference distribution. We use the implementation `cvm.test` from the `R goftest` package.

Conclusions

It is clear from the results shown that as expected, none of the methods are able to identify data from our marginally Gaussian χ^2_1 distribution as distinct from data from a 'true' Gaussian distribution. This is not surprising, however, the tests are further unable to identify 'true' Gaussian data with spatial correlation structure, as can be seen by comparing results for rf1, rf2 and rf3 (Gaussian fields with spatial correlation) to rf0 (Gaussian field with no spatial correlation). As expected, tests designed for univariate sample data are not helpful for spatially correlated data. This confirms the value of methods that are able to distinguish between these types of data sets with a spatial correlation structure.

9.1.4 Topological event history

Figures 9.4, 9.5 and 9.6 show the Nelson-Aalen plots for our nine random fields. For each, we show the Nelson-Aalen estimator and the expected cumulative hazard function under an assumption of Gaussianity. The differences between the three types of random field are immediately evident on inspection and the Nelson-Aalen curves for the marginally Gaussian χ^2_1 fields clearly fall outside what would be expected for a true Gaussian random

field. This clearly demonstrates the potential of topological event history methods for the purpose of identifying non-Gaussianity in spatial data, and at the time of writing we are not aware of any other methods able to distinguish these data sets as clearly.

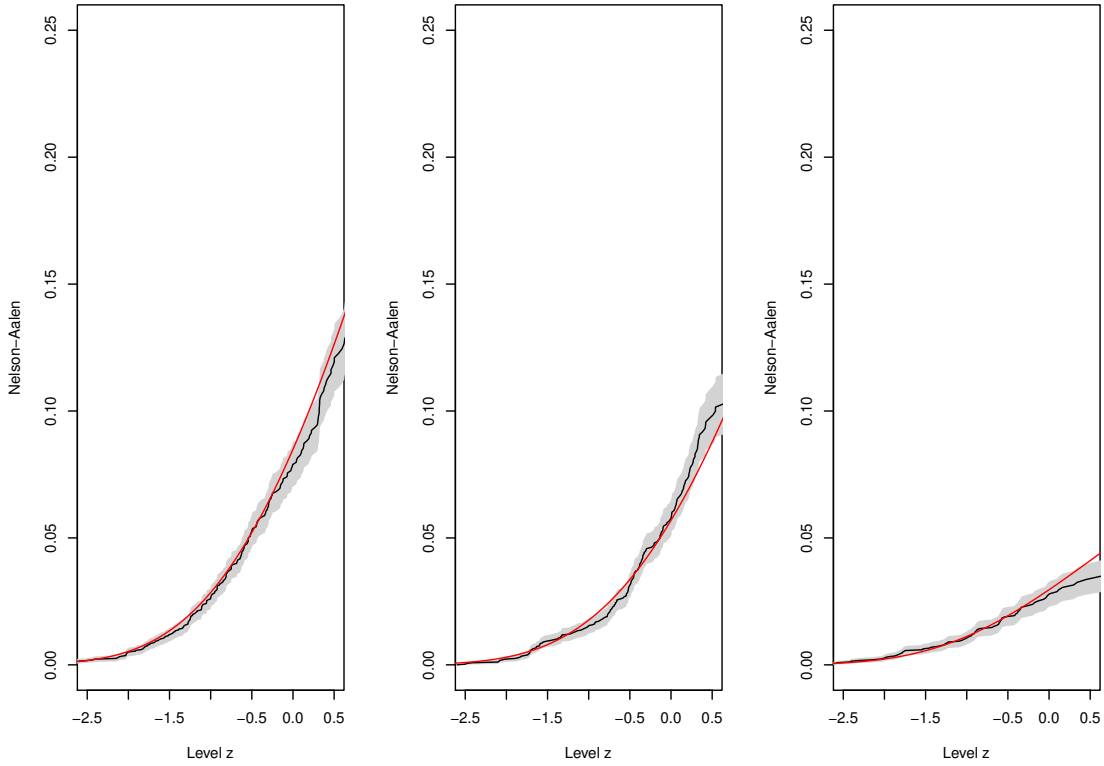

Figure 9.4: Nelson-Aalen plots for the three Gaussian random fields. The Nelson-Aalen estimator is shown in black, and the expected cumulative hazard function under an assumption of Gaussianity is shown in red. Grey bounds show plus and minus two standard deviations.

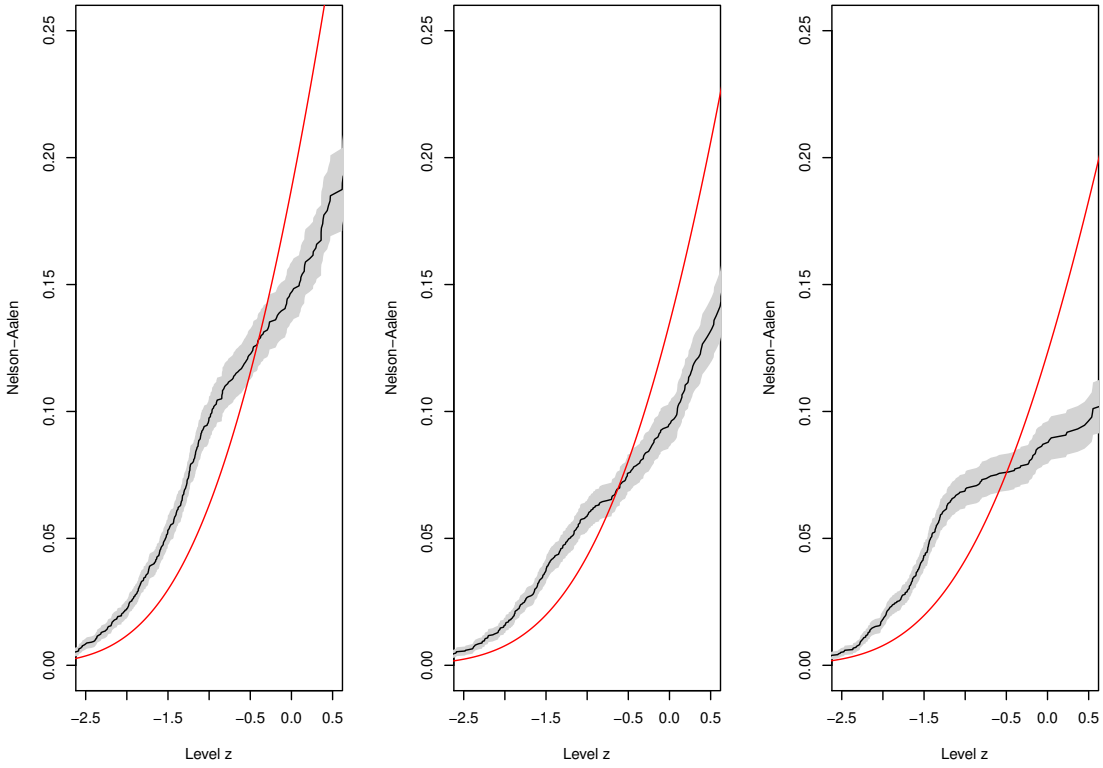

Figure 9.5: Nelson-Aalen plots for the three marginally Gaussian χ^2_1 random fields. The Nelson-Aalen estimator is shown in black, and the expected cumulative hazard function under an assumption of Gaussianity is shown in red. Grey bounds show plus and minus two standard deviations.

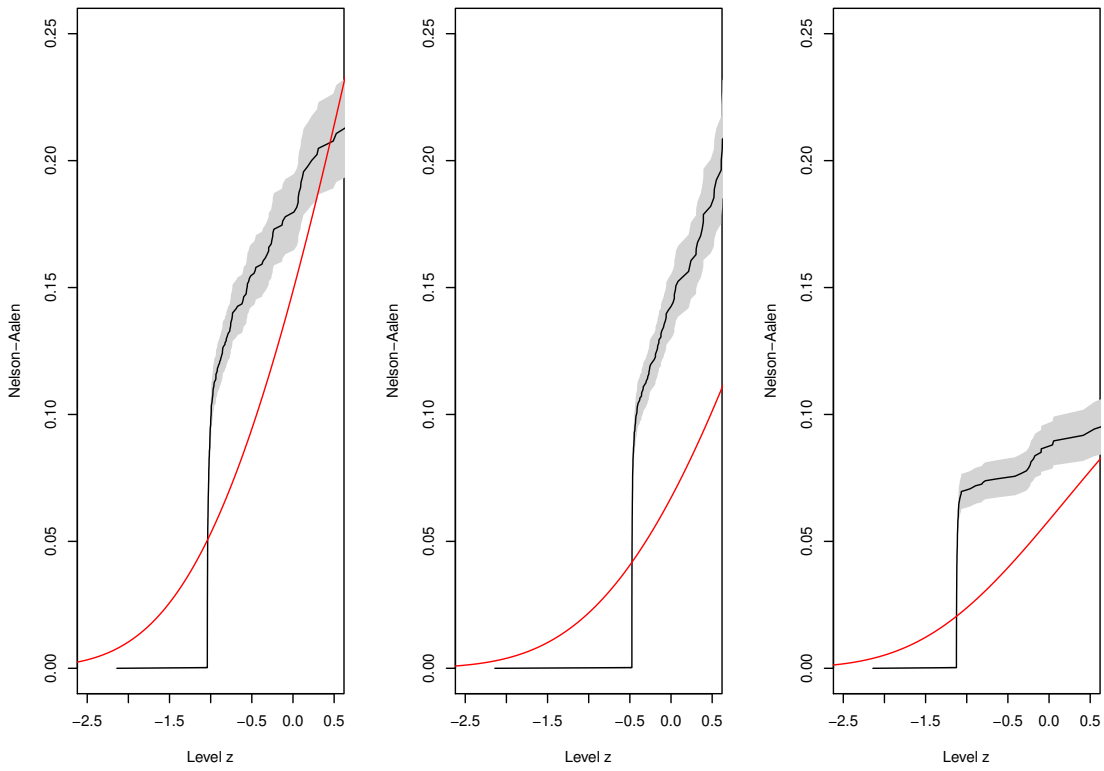

Figure 9.6: Nelson-Aalen plots for the three untransformed χ^2_1 random fields. The Nelson-Aalen estimator is shown in black, and the expected cumulative hazard function under an assumption of Gaussianity is shown in red. Grey bounds show plus and minus two standard deviations.

9.2 Testing alternative Gaussian data

From previous work we are able to establish a theoretical baseline for a Gaussian random field with a given correlation structure. That is, we have an expected cumulative intensity curve for local maxima on such a field. This allows us to look at fields simulated using different `R` packages and compare to theoretical results.

We consider the three following sets of Matérn correlation parameters.

Table 9.3: Matérn covariance parameters

	η	ν
cor1	5	1
cor2	10	1
cor3	5	2

For each set of parameters and each of the simulation methods below, we simulate 10 Gaussian random fields of size 100×100.

`rnorm` We first generate 10000 independent Gaussian random variables using the `R` `rnorm` function and obtain our correlated random fields via covariance matrix decomposition. Given independent Gaussian random variables X with mean zero and unit variance, and covariance matrix R, our correlated fields can be calculated as

$$Y = CX,$$

where $CC^T = R$ is the Cholesky decomposition of R. Figures can be seen in the left-hand column of Figure 9.7.

`RandomFields - RFsimulate` The `RFsimulate` function from the `RandomFields` package (Schlather et al., 2020) in `R` allows simulation of unconditional random fields from a variety of distributions and conditional Gaussian random fields. We simulate 10 correlated Gaussian random fields of size 100×100 for each Matérn parameter set. The Nelson-Aalen curves for these fields are shown in the central column of Figure 9.7 along with the expected cumulative hazard function under Gaussianity with each covariance structure.

`fields - sim.rf` We finally look at another `R` function, `sim.rf` from the `fields` package (Furrer et al., 2012). This generates Gaussian simulations using a given covariance

object. Again, we simulate 10 correlated 100×100 random fields for each set of Matérn parameters. The results can be seen in the right-hand column of Figure 9.7 with the expected curve under each correlation.

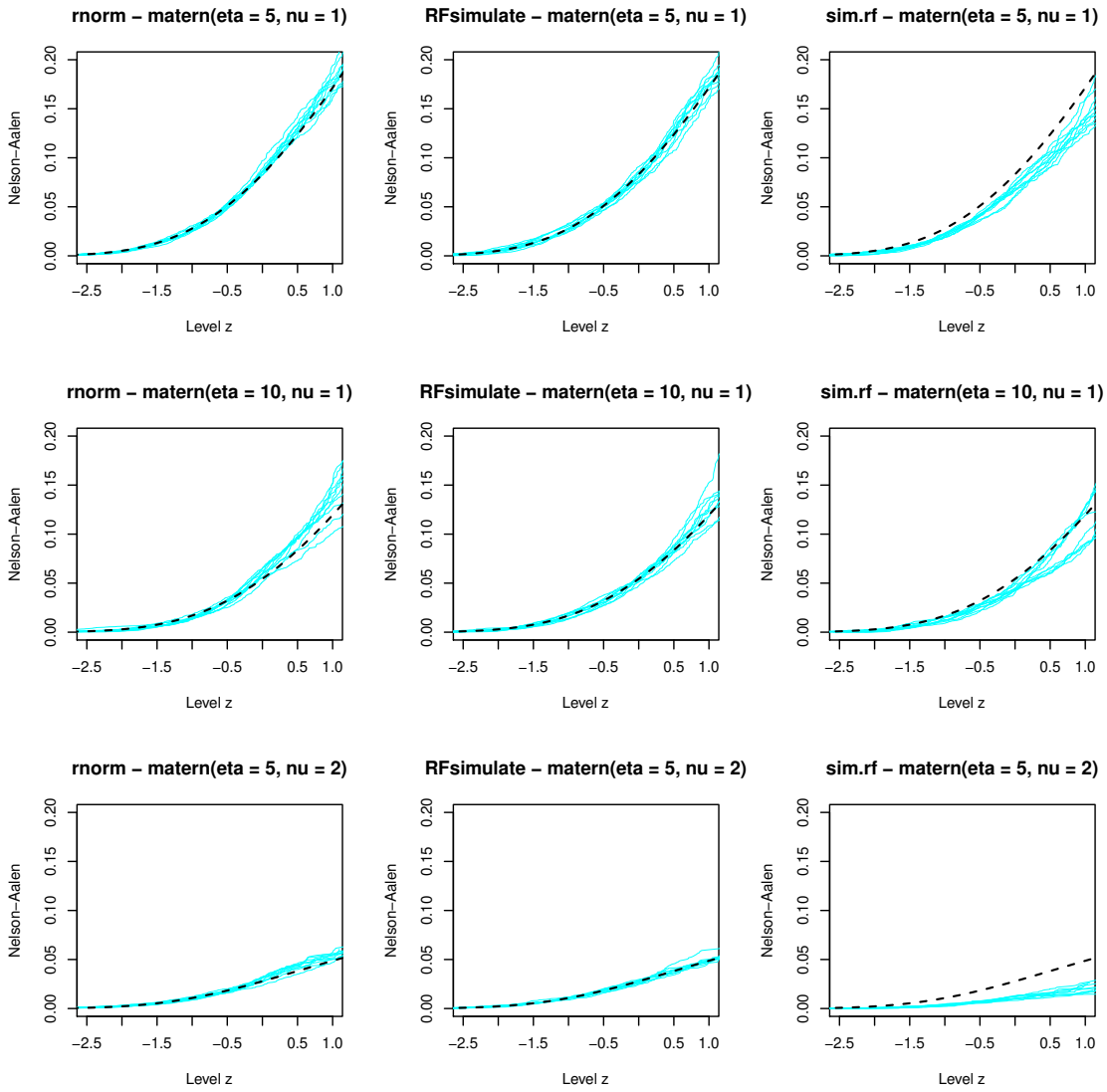

Figure 9.7: Nelson-Aalen curves and expected cumulative hazard function for correlated Gaussian random fields, simulated via three different methods. In blue are Nelson-Aalen curves for the simulated data, the dashed black line shows the expected cumulative hazard function for that correlation structure.

Figure 9.7 shows that all methods simulate data with topological event history features close to what we would expect to see. However, the variance within methods varies, with NA curves for simulations using 'cor3' parameters having lower variance than those with 'cor1' and 'cor2' parameters. Further, fields simulated with `RFsimulate` have Nelson-Aalen estimates for the cumulative hazard function closest to what we could expect to see, with

data simulated with `rnorm` and `sim.rf` slightly higher and lower than expected, respectively. The `sim.rf` plots in particular show Nelson-Aalen curves that are consistently lower than the expected cumulative hazard curve for all three sets of correlation parameters, especially visible for the 'cor3' parameter set. This might indicate some systematic issue and should be questioned.

9.3 Conclusions

In this chapter we applied a selection of common tests for Gaussianity to true Gaussian, marginally Gaussian χ^2_1 and untransformed χ^2_1 sets. We demonstrated that as one would expect, these tests are unable to identify non-Gaussian data with Gaussian marginals. We showed how use of our topological event history method effectively identifies non-Gaussianity, simply and in a visually clear way. We believe that this method has significant value when deciding whether to assume Gaussianity in spatial data. Finally we applied the method to Gaussian data simulated using different `R` packages and showed the difference in output from different simulation methods. Investigation of these and other simulation methods using TEH methods would make an interesting avenue for further work.

Chapter 10

Applications to climate data

Throughout earlier chapters, we have demonstrated the application of topological event history methods for testing for Gaussianity and assessing correlation models. We propose that topological event history is also a valuable comparison tool, allowing us to distinguish between seemingly indistinguishable data, or to identify more subtle differences between data that existing methods may be unable to detect. In this chapter we take advantage of the large quantity of available global climate data, looking at additional years, monthly output, new realisations and other climate variables, all of which are available in the CESM LENS project. Finally, we verify topological event history analysis as a method for identifying anomalous data sets from an ensemble.

10.1 TEH for data investigation

As a basis for comparisons in work to follow, we use average wind intensities for $t = 2006$ from the original 30 ensemble members, as we have throughout the thesis so far. We have already established that these 30 realisations for $t = 2006$ are consistent with data from a Gaussian distribution and no single realisation stands out as notable. To these, we compare several different data examples. Primarily we have considered $t = 2006$ only, although we have data available from 2006 to 2100. Here, we apply TEH methods to a selection of additional years from those available, with the primary aim to identify any temporal trend, if present. Secondly, in previous work, we have considered annual averages, eliminating any seasonal components. Here, we investigate the difference between months in a year with interest in identifying the potential differences between seasons. Since the commencement of this work, additional ensemble members have been added to the LENS project; we apply our methods to these as we did for the original ensemble. Finally, we investigate some of the numerous additional variables available in the LENS project,

including precipitation, temperature and pressure.

10.1.1 Trend over time

During the introduction to the data and exploratory analysis in Chapter 2, we presented examples of the data over a range of years, but through the majority of this work we focused only on $t = 2006$. From a climate perspective, we are particularly interested in effects over time, so we investigate whether there are any differences between each of our 95 years of data, or any noticeable temporal trend. We have seen previously that the wind intensity values do change over the 95 years, and we examine to what extent this is evident for topological features. Figure 10.1 shows yearly Nelson-Aalen curves for all realisations. With earlier years towards the red end of the colour gradient, and later years towards the blue end, this figure would allow us to see clearly if there was a temporal effect in the data. We do not see this, and instead can conclude that the change in actual wind intensity vales over time does not impact the topological features studied here of the standardised residuals.

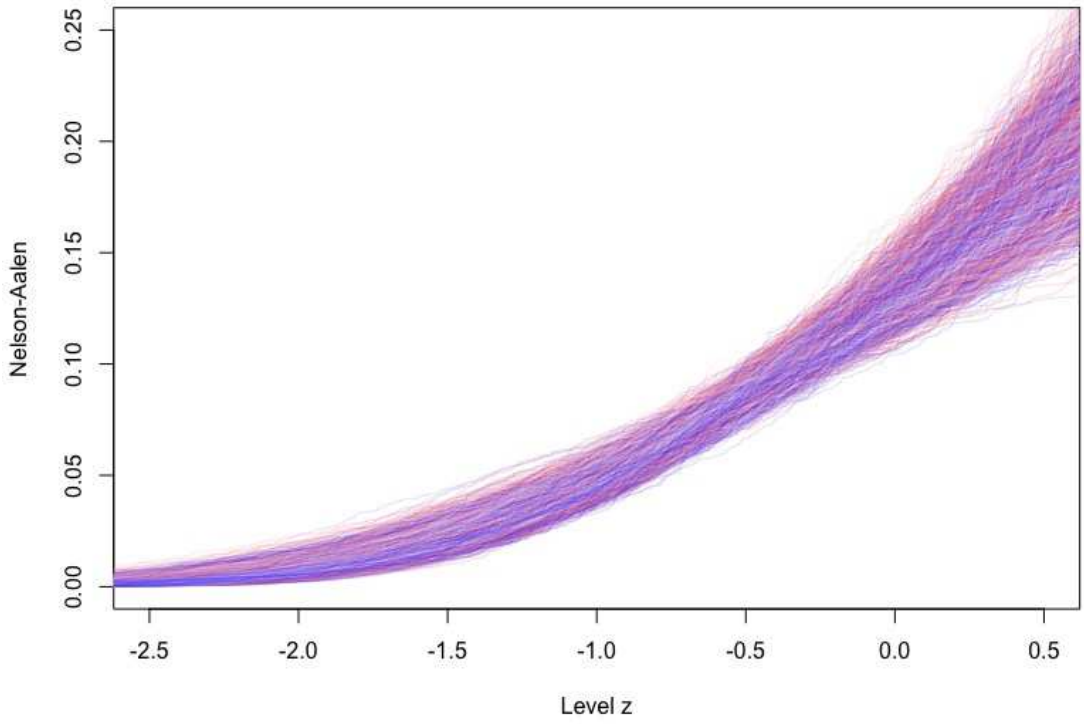

Figure 10.1: Nelson-Aalen curves for all 95 years and 30 realisations. Years are shown on a gradient colour scale, with years closer to 2006 in red and those closer to 2100 in blue.

10.1.2 Seasonal effects

As with temporal trend, it is natural to consider the presence of a seasonal effect on the Nelson-Aalen estimators. We calculate standardised residuals for monthly data, via the same process as for annual data so far. Similarly to the previous figure, in Figure 10.2 we show Nelson-Aalen curves for each of the 12 months from 2006, for all 30 realisations. Here, we use data from $t = 2006$ only for clarity. Again, we see no clear seasonal effect.

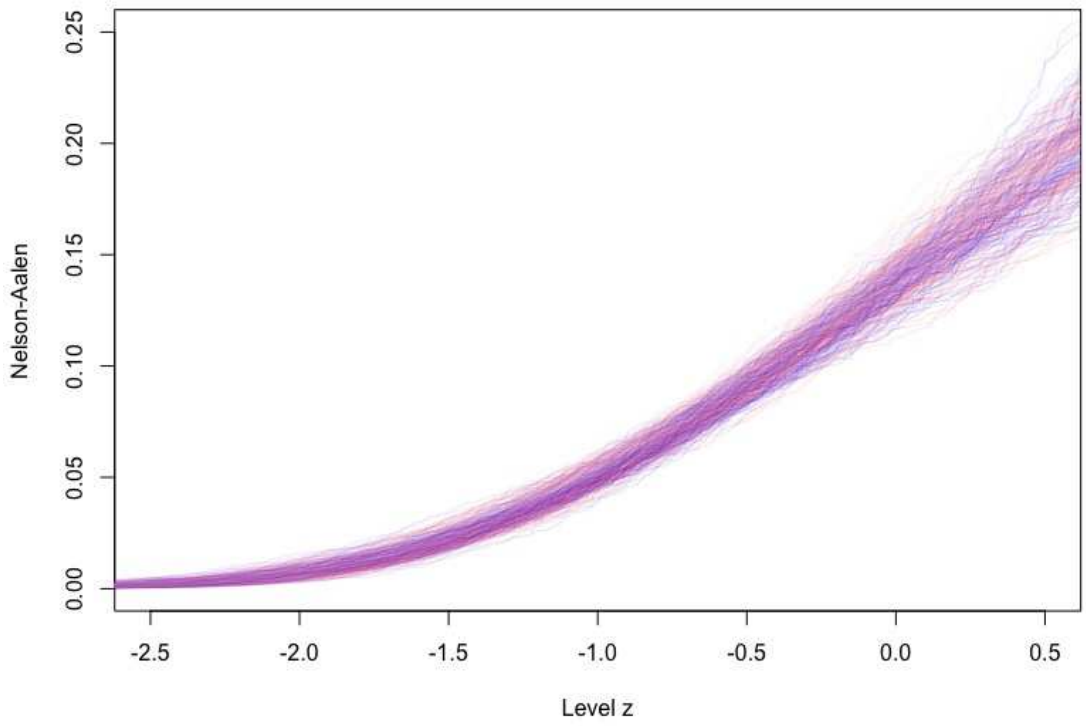

Figure 10.2: Nelson-Aalen curves for all 12 months of year $t = 2006$ and 30 realisations. Individual months are shown on a gradient colour scale, with months closer to January in red and those closer to December in blue.

10.2 Additional realisations

As we have described throughout, the LENS data set comprises 30 realisation covering many variables. Shortly after the commencement of this project, an additional three realisations were added (in fact further variables have also been added at later stages). In this section, we look at these first three additional realisations and investigate how they compare to the original set.

For realisations 31 to 33, we standardise the data using all realisations and obtain the reduced data set as in earlier work. We then calculate the Nelson-Aalen estimators for these additional realisations. Comparing expected Nelson-Aalen plots, we identify a small difference between realisation 33 and the others at very low levels where the expected cumulative hazard is noticeably higher than the other realisations, as can be seen in Figure 10.3. For realisations 31 and 32, we see no significant difference between the three new realisations and those already studied. This small difference might indicate reason for further investigation, for example examination of additional years (the figure shows $t = 2006$ only). We repeat the process for six additional years spread across the 95 and repeat the process, with results shown in Figure 10.4.

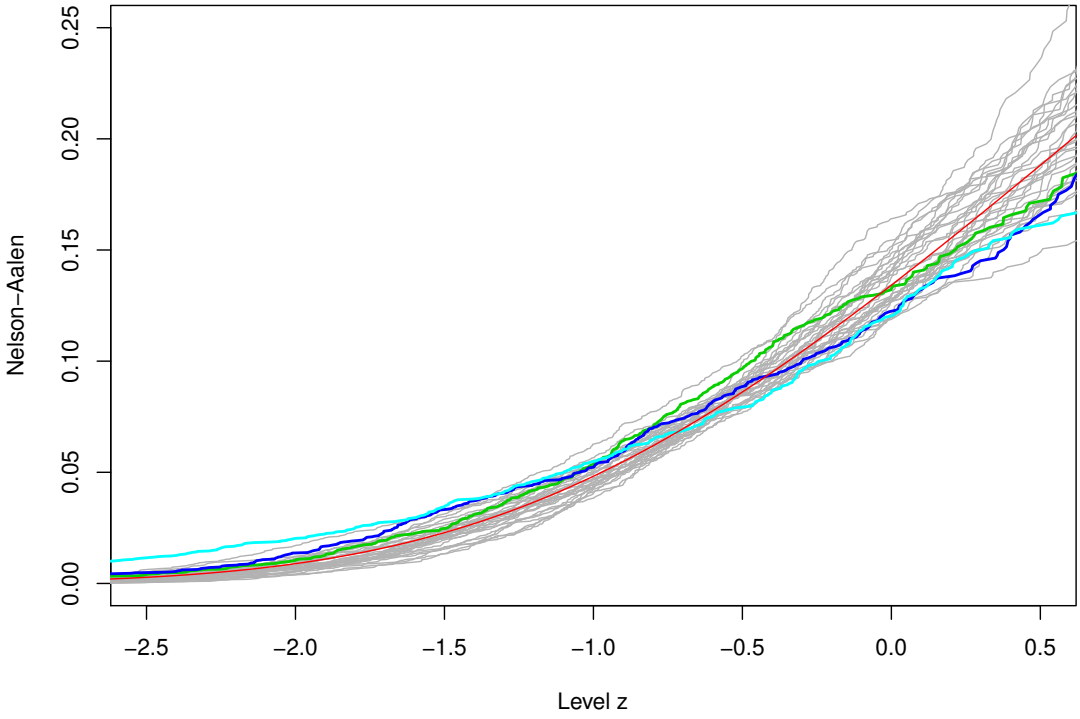

Figure 10.3: Nelson-Aalen plots for the original 30 realisations shown in grey with realisations 31 (green), 32 (dark blue) and 33 (cyan) added. As with previous work, this shows values from $t = 2006$ only.

The additional years do not show a noticeable difference between realisation 33 and the remainder, suggesting the minor difference in $t = 2006$ is an anomaly rather than indicative of a systematic difference.

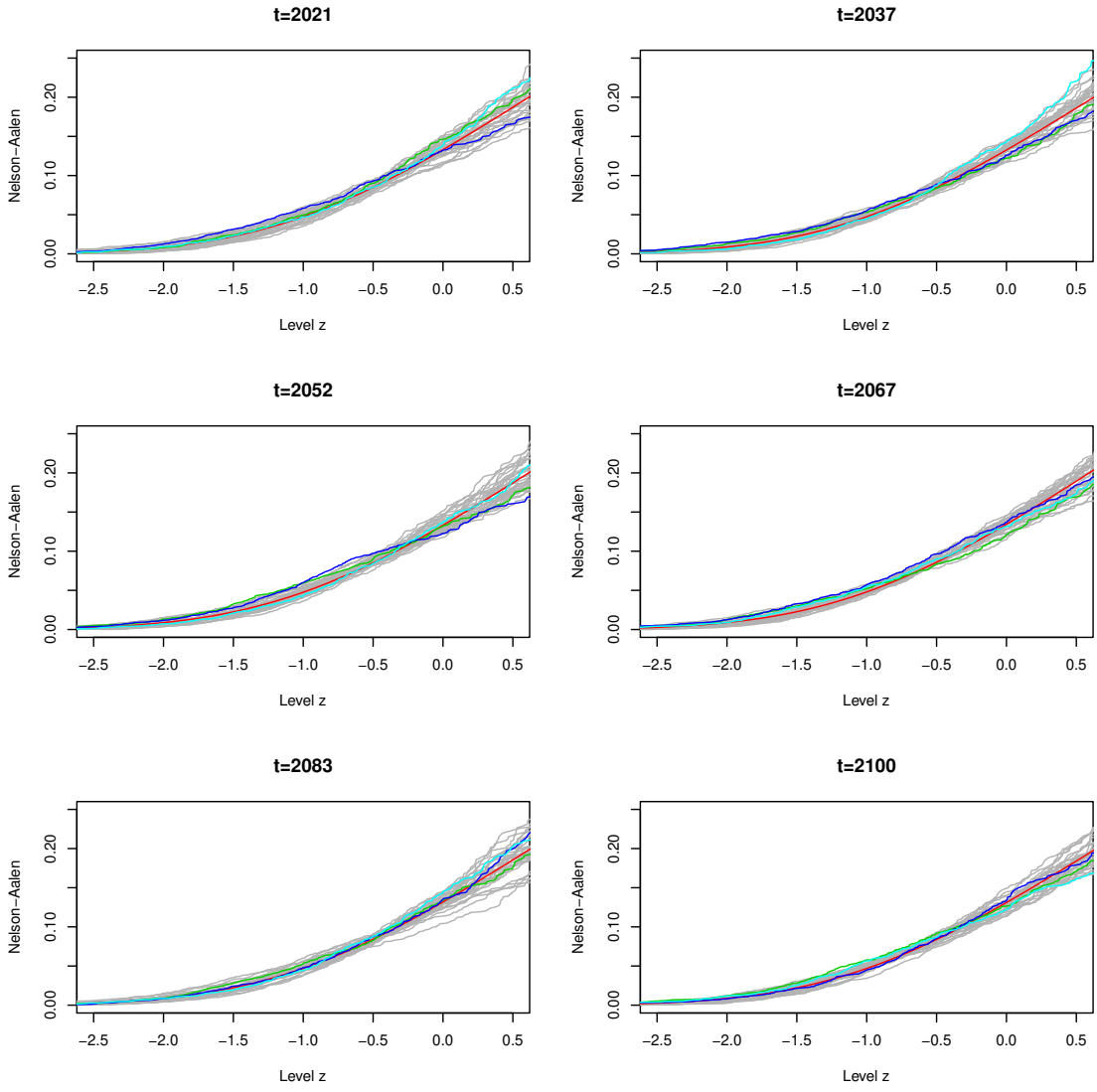

Figure 10.4: Nelson-Aalen plots for the original 30 realisations shown in grey with realisations 31 (green), 32 (dark blue) and 33 (cyan) added. Here, we see plots for $t = 2021$, 2037, 2052, 2067, 2083, 2100.

10.2.1 Additional variables

The previous research has been wholly focused on the wind variable from the LENS project (U10: wind intensities at 10m). Here, we apply our methods to a selection of additional variables, shown in Table 10.1.

Table 10.1: Variables of interest (including wind intensities)

Variable	Description	Units	Time averaging
$u10$	10m wind speed	m/s	mean
trefht	Reference height temperature	K	mean
psl	Sea level pressure	Pa	mean
ps	Surface pressure	Pa	mean
tmq	Total (vertically integrated) precipitable water	kg/m^2	mean

For each variable, we calculate the standardised residuals for $t = 2006$ using the same process we used previously for wind intensities. Figure 10.5 shows the expected cumulative hazard curves for each variable. We see that compared to the expected cumulative hazard curve for the wind data ($u10$) all other variables have lower cumulative hazard values, with total precipitable water (tmq) and reference height temperature ($trefht$) very similar, and the surface and sea level pressure (ps and psl) almost indistinguishable. We then produce figures for each showing the Nelson-Aalen estimates for $t = 2006$ of each realisation, as seen in Figure 10.6.

Each includes in grey, the original 30 realisations, with the additional three (r=31, 32, 33) shown in green, blue and cyan respectively. As with previous Nelson-Aalen figures, the expected cumulative hazard curve under Gaussianity is shown in red. Several observations can be made from these figures. Considering Nelson-Aalen figures as a way of testing for Gaussianity, realisations from both *trefht* and *tmq* centre around the expected value, indicating that the data are indeed Gaussian. For the pressure variables however, *ps* and *psl*, the realisations are almost all sitting below the expected value. This indicates possible departure from Gaussianity in these data. Further, in *ps* in particular, realisations $r = 31$ and $r = 33$ differ noticeably from the remainder. As with the *u10* variable, we repeat these figures for an additional six years, shown in Figure 10.7. Unlike wind, the difference between $r = 31$ and $r = 33$, and the remainder, is evident in additional years. This is certainly suggestive of some fundamental underlying difference between these realisations and the others.

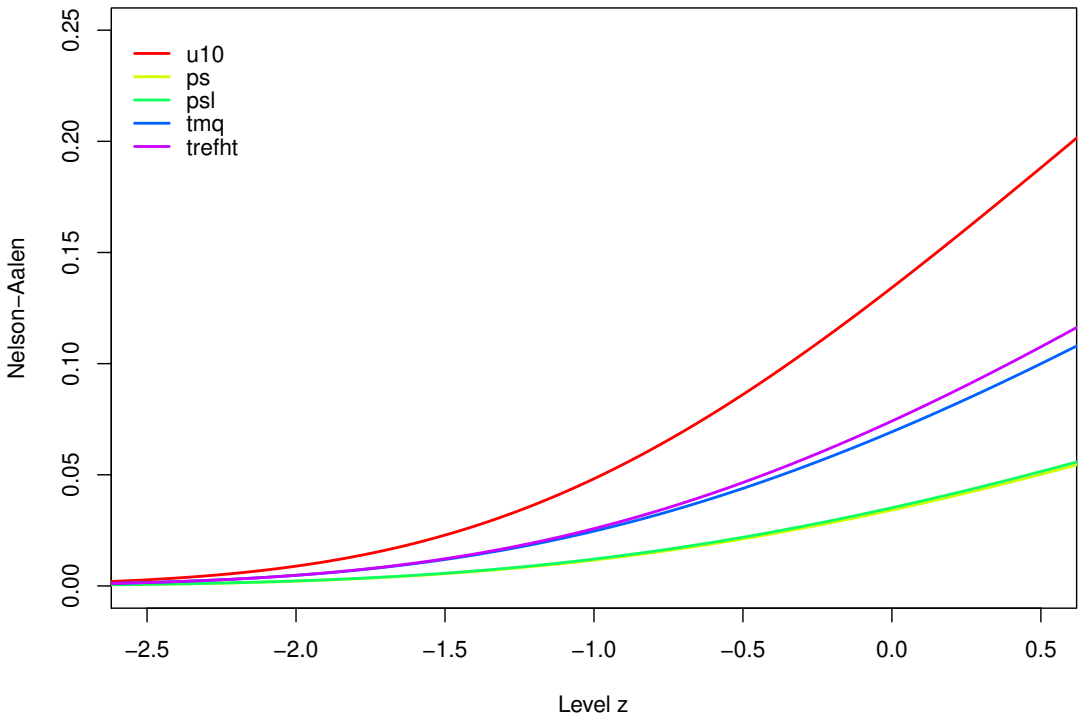

Figure 10.5: Expected cumulative hazard curves under the assumption of Gaussianity for each of our five variables. Curves for the two pressure variables, *ps* and *psl* are almost indistinguishable.

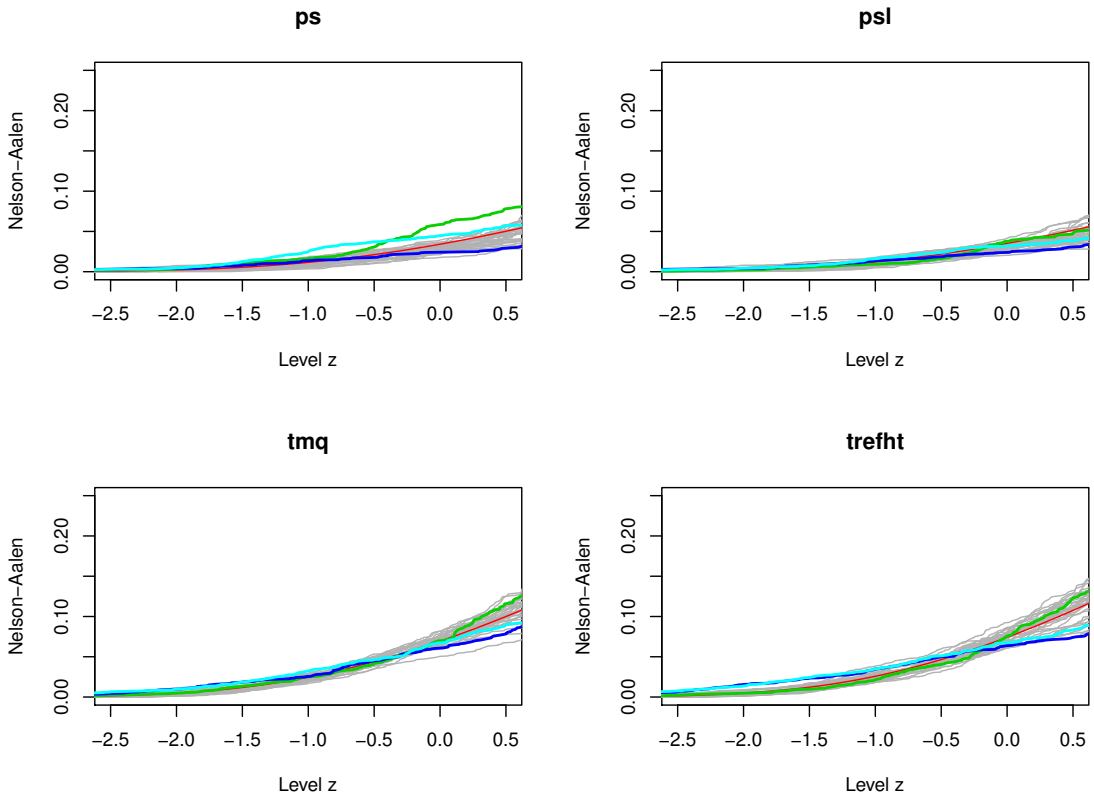

Figure 10.6: Nelson-Aalen plots for the original 30 realisations shown in grey with realisations 31 (green), 32 (dark blue) and 33 (cyan) added. As with previous work, this shows values from $t = 2006$ only.

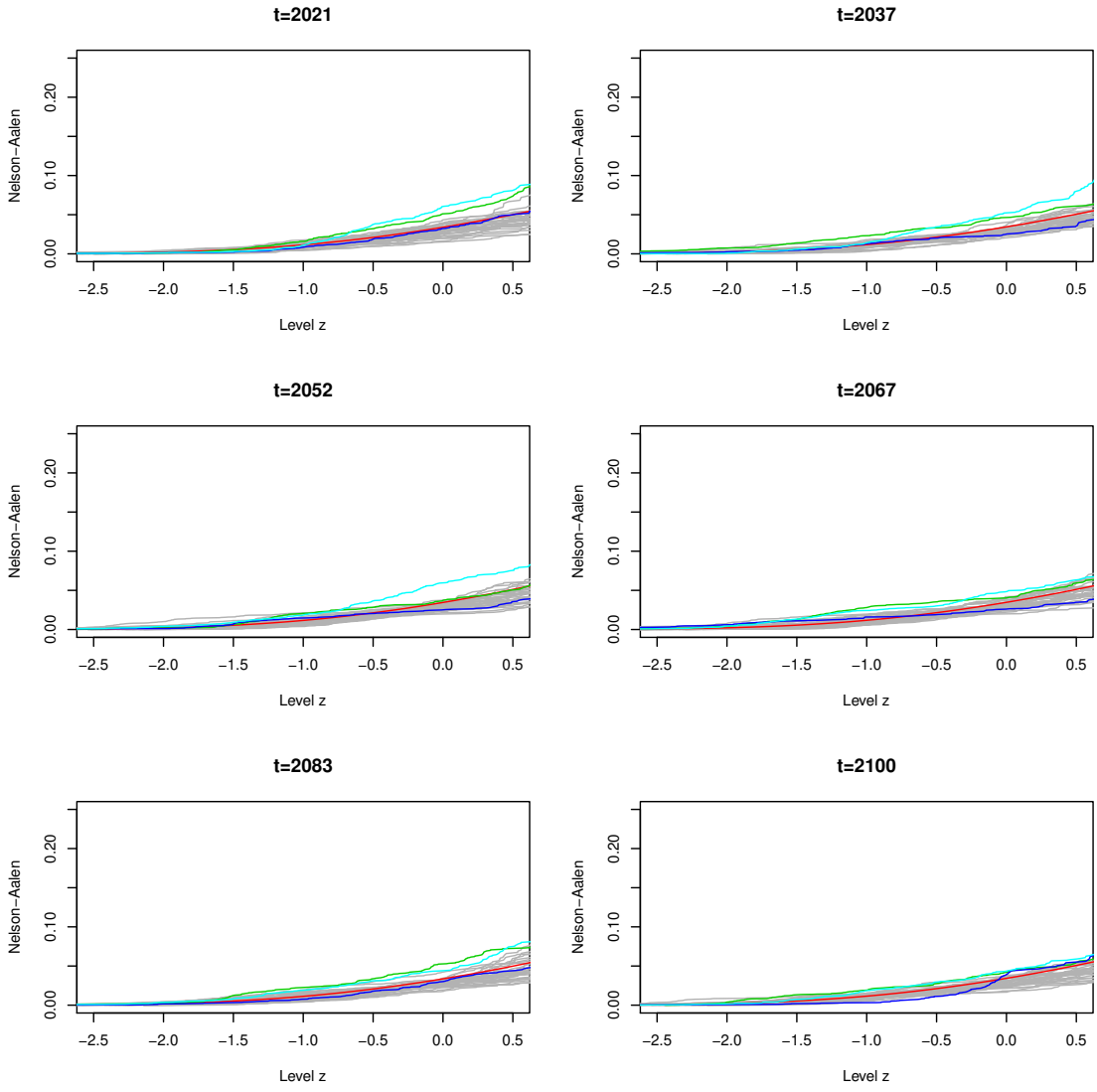

Figure 10.7: Nelson-Aalen plots for the original 30 realisations of variable *ps* shown in grey with realisations 31 (green), 32 (dark blue) and 33 (cyan) added. Here, we see plots for $t = 2021$, 2037, 2052, 2067, 2083, 2100.

Lossy data compression evaluation

Although we did not see large differences between any of the realisations for *u10* (wind), we do see noticeable differences between realisations 31 and 33 and the remainder, for *ps* (surface pressure). This is of particular interest since realisations 31 and 33 are presented in the CESM LENS project following lossy compression.

To evaluate the impact of lossy compression on model output, Kay et al. (2016) produced additional realisations that had undergone compression and presented researchers with the task of determining which had been compressed and which were the originals. Compression used the *fpzip* compression algorithm (Lindstrom and Isenburg, 2006), which discards a specified number of least significant bits, before losslessly encoding the result. Baker et al. (2014) examined in some detail a selection of different compression algorithms, results of which lead to selection of the *fpzip* algorithm for this application. For each variable (and temporal resolution), the amount of compression was decided separately, allowing maximum compression such that the replicate remained within the internal variability of the ensemble, amongst other constraints. Since a number of properties can influence the impact of compression, such as smoothness of the field or amount of internal variability present, different variables tolerate different degrees of compression.

We can see that for one of our five variables considered, topological event history analysis does indeed help identify anomalous realisations from the set. The consequences of compression on the topological features does not carry across all variables, although given the apparent similarity between *ps* and *psl* we might have expected to see similar results across both variables. Inspection of additional years for *psl* shows no noticeable difference for any of the years considered.

The above results show the potential of this method for identification of both non-Gaussian data and anomalous replicates in a set. It is important in the context of compressed realisations to make the distinction between identification and the assumption that the difference has some relevant effect. Whether the differences observed have implications for the understanding of climate variability is a question for climate scientists. We are able to say with confidence that the process of lossy compression impacts the topological features of the data for some variables.

Chapter 11

Conclusions

One of the primary aims of this thesis was to explore the potential of event history methods for topological data, particularly in the context of global climate data. In Chapter 2 we presented some of the key features of the data, showing how numerous statistical summaries of the site-wise yearly means and within-year variances demonstrate a temporal trend. For example, considering the site-wise within-year variances, the median values increase over time, yet the maximum values decrease. For yearly means, we saw how the minimum site-wise value increases over time, whilst the upper-quartile decreases. This exploratory analysis confirmed the benefit of being able to examine properties of the data via examination of all 30 replicates, allowing us to see patterns that may not be clear from a single replicate.

In Chapters 3 and 4 we demonstrated how methods from the field of statistical topology could be applied to gridded 1-d and 2-d data. We showed how examination of topological features such as local maxima and minima could distinguish between marginally-Gaussian χ^2_1 random fields and true Gaussian fields where covariance structure was matched, and derived theoretical expected results. We applied common topological data analysis techniques to wind intensities data, concluding that for some of the techniques their use presented challenges for ensemble data sets, instead having more value for investigating specific features of a single replicate.

In Chapter 5 we introduced ideas from survival analysis. We discussed non-parametric estimators for the cumulative hazard function including the Nelson-Aalen estimator which was key to the work that followed. In applying survival techniques to spatially correlated data, we addressed the challenge of estimating the variance of the Nelson-Aalen estimator, comparing several techniques including the standard 'naive' Nelson-Aalen variance estimator.

Using survival analysis for topological features in spatial data led us to propose 'topolog-

ical event history analysis' in Chapter 6. We showed how TEH could easily distinguish between different distributions, even when all are marginally Gaussian and correlation structures are matched by design. We demonstrated how this method could be valuable for global wind intensities and showed the expected cumulative hazard curve given a known correlation structure under Gaussianity. We discussed the challenge of variance estimates for Nelson-Aalen estimators, showing the insufficiency of the standard variance estimator and proposing a parametric bootstrap approach for calculation of confidence intervals.

The requirement for a known, or accurately estimated correlation structure presented no problems in a scenario such as LENS data where we have multiple replicates available. However, for data in which we may only have a single replicate, being able to accurately specify the correlation structure is important for the use of TEH methods. In Chapters 7 and 8 we examined and fit a number of stationary and nonstationary covariance models, respectively. We showed that for our application, consideration of the curvature of the earth offered no considerable improvement to the model fits so proceeded to consider the data as existing on a 2-d grid. Using a selection of assessment metrics, we compared models, demonstrating the challenge of fitting an accurate model to a single replicate. We proposed a 'regional block model', fitting stationary models separately to different regions, which although performed better than other models, did not fit the empirical covariance matrix well. Finally, we applied an unsupervised clustering algorithm to investigate differences between local correlation for sites on the grid, since our block model involved classifying sites by geographical region. This validated our approach of grouping sites by land, ocean and coast.

In Chapter 9 we showed how TEH could be a valuable tool for assessing Gaussianity of gridded data on a random field. We compared several tests which considered only marginal properties, so were unable to distinguish between the fields. We applied this method to commonly used Gaussian data simulation packages in `R`, showing the difference in data simulated with these packages. Finally in Chapter 10 we investigated other years, temporal resolutions and variables from the CESM LENS data set. We showed that for surface pressure, our method was able to identify replicates that had been subject to lossy compression, and identify the non-Gaussian nature of both surface pressure and sea-level pressure standardised residuals.

Throughout this work there have been three primary outcomes. Firstly, we have shown the value of TEH analysis in a scenario where numerous replicates exist, utilising the information available to identify anomalous replicatsd. Secondly, we have demonstrated a new approach to the use of topological methods, using concepts from survival analysis to better observe subtle differences between the rate of births of topological features. Finally, we have presented topological event history methods as a way to reflect underlying features

of Gaussian and non-Gaussian data, beyond marginal properties.

Arguably one of the greatest challenges of applying TEH methods is the requirement of an accurate covariance matrix to calculate the expected cumulative hazard curve. Clearly, this is inconsequential if applying TEH purely to identify anomalous replicates, or for classification purposes. However, to use TEH for assessment of Gaussianity this covariance matrix is essential. We showed in Chapters 7 and 8 how for a single replicate, obtaining a suitable covariance matrix is non-trivial, particularly for highly nonstationary data. For the wind intensities data (and all data from the CESM LENS project), the availability of multiple replicates allows us avoid this challenge and we can calculate expected cumulative hazard curves for all years/months/variables.

11.1 Further work

There are several interesting avenues of further work emerging from this research. Throughout this work we have used data on a grid, for the majority of which we assumed equal spacing between sites although working with a different projection does not present any significant challenges. In the extraction of the topological features and calculation of expected cumulative hazard curves, we rely on some definition of a neighbourhood of each site, in this work we use a simple cross neighbourhood. It would certainly be possible to investigate the impact of other definitions of the neighbourhood, but a more valuable extension to this work might involve looking at TEH for point cloud data, since beyond the CESM LENS project a significant amount of climate data comes in this form. In fact, gridded data of the form we have worked with is less common in climate applications. For example, a significant volume of data is collected from numerous manned and automatic surface weather stations, upper air stations, ships, buoys, weather radars and commercial aircraft and is freely available through the World Meteorological Organization. For point cloud data, rather than considering filtrations in the 'temporal' dimension of the data, filtrations can be formed through gradually increasing circles (in 2-d) around each location. Locations form connected components when these spheres touch. In a TEH context, it could be of interest to consider the rate of emergence of connected components by the size of the circles, instead of site value as in our work. Understanding TEH for point cloud data would open up these methods to many more applications.

A significant challenge of fitting covariance models, nonstationary or otherwise, is the ability to perform calculations with large matrices. Indeed, this research has required us to reduce our year-realisation data subsets by a factor of nine. In fact, our reduction is a little greater than this due to the removal of polar latitudes. For the resulting covariance matrix, this gave us a total reduction by a factor of over 125, allowing us to work with

a 4896×4896 matrix rather than a 55296×55296 matrix. Although parallel computing allows us to speed up certain computations, such as calculation of the expected NA curves or fitting models to regions independently, parallel computing methods for the inversion of large covariance matrices are non-trivial. Various methods exist for sparse matrices, and powerful graphics processing units (GPUs) can be used to handle computations beyond the scope of conventional processors. In this work, we utilised parallel computations where processes could be run synchronously without significant reworking, speeding up computations without requiring complex matrix operations. Methods for working with much larger data sets certainly exist, and a valuable extension to this work could involve using these methods to explore the models discussed with a greater volume of data. Whether this could make sufficient difference to allow use of this method on the full data set is a question to be answered. If so, it would be of interest to investigate whether the same results are seen on a full data set as on a reduced one, or whether additional information can be gained from the full data set.

Finally, of course we only touched on a small portion of the available LENS data. Many more variables and temporal resolutions are available and would absolutely be worth investigation, particularly regarding the lossy compression of realisations 31 and 33. The LENS project contains further years, however these are produced differently and so would not provide as valuable a comparison but would certainly be an interesting direction in which to take further work.

Appendix A

The Miwa algorithm in the R `mvtnorm` package

Miwa is a numerical algorithm, and its use is only advised for dimension $d < 20$ non-singular matrices. It uses quick and accurate recursive integration methods to evaluate non-centred orthoscheme probabilities, defined as non-centred orthant probabilities where the correlation matrix is tridiagonal and satisfies $\rho_{ij} = 0$ for $|i-j| > 1$. Miwa et al. (2003) showed that a non-centred m-dimensional orthant probability can be expressed as the differences between at most $(m-1)$ non-centred orthoscheme probabilities. This allows accurate evaluation of any multivariate Normal distribution function.

The GenzBretz algorithm is a quasi-randomised Monte-Carlo procedure and is suitable for arbitrary covariance structures and dimensions as high as 1000. The GenzBretz algorithm works by transforming the original integral over an arbitrary m-dimensional rectangle to an integral over the unit hypercube.(Genz and Bretz, 2009) Randomized lattice rules are applied to the transformed integral, seeking to fill the integration region evenly in a deterministic process. The algorithm constructs regular patterns such that projections of integration points onto each axis produce an equidistant subdivision of the lattice. Robust integration error bounds are then obtained by introducing additional shifts of the entire set of integration nodes in random directions. (Hothorn et al., 2001)

Based on recommendations by Mi et al. (2009) we initially ran the `mvtnorm` package using the Miwa algorithm, since we have a maximum of 4 dimensions with non-singular matrices. However, running the estimation procedure in this way, we obtained inconsistent results, giving theoretically impossible values such as large negative variances. For example, consider the two boundary cases (here each site has only three neighbours). Call the conditional covariance matrix for the right-hand boundary case A, and the equivalent for the left-hand case B. It is clear to see that matrix B is the result of exchanging position

one with position three in matrix A.

$$A = \begin{pmatrix} 9.358 \times 10^{-3} & -8.165 \times 10^{-3} & -1.689 \times 10^{-8} \\ -8.165 \times 10^{-3} & 9.358 \times 10^{-3} & 1.854 \times 10^{-8} \\ -1.689 \times 10^{-8} & 1.854 \times 10^{-8} & 9.358 \times 10^{-3} \end{pmatrix},$$

$$B = \begin{pmatrix} 9.358 \times 10^{-3} & 1.854 \times 10^{-8} & -1.689 \times 10^{-8} \\ 1.854 \times 10^{-8} & 9.358 \times 10^{-3} & -8.165 \times 10^{-3} \\ -1.689 \times 10^{-8} & -8.165 \times 10^{-3} & 9.358 \times 10^{-3} \end{pmatrix}.$$

We can see that $\det(A) = \det(B) = 1.955 \times 10^{-7}$, so although the values are small, we would expect any issues arising from the matrix being 'close' to singular to be apparent in both cases. Table A.1 shows the value of $E[I_i I_j]$ calculated using the two matrices A and B, and the two different algorithms, Miwa and GenzBretz. The results are as expected, except when using matrix A with the Miwa algorithm.

Matrix	Miwa	GenzBrentz
A	-10.766	0.041
B	0.041	0.041

Table A.1: Output from `pmvnorm` using matrices A and B and algorithms Miwa and GenzBretz

Further investigation indicates that it is the values in locations $(1, 3)$ and $(3, 1)$ of the 3×3 matrix which are causing the issues; any values in the range $(5 \times 10^-7, 5 \times 10^-9)$ and $(-5 \times 10^-9, -5 \times 10^-7)$ give unusual results. The author of the `mtvnorm` package has acknowledged the problem but at time of writing has not yet fixed it (Tetsuhisa Miwa, personal communication).

Appendix B

Commonly used covariance functions

Below are a selection of common covariance functions that have been used in this work. For all models, d is the separation distance, η is the range parameter and, where it is used, ν is the smoothness parameter.

Nugget

$C(d) = k$, where k is a constant. This corresponds to no spatial correlation.

Circular

Let $\theta = min(\frac{d}{\eta}, 1)$ and

$$g(d) = 2\frac{(\theta\sqrt{1-\theta^2} + \sin^{-1}\sqrt{\theta}}{\pi}.$$

Then

$$C(d) = \begin{cases} 1 - g(d), d < \eta \\ 0, \text{ otherwise} \end{cases}$$

Cubic

$$C(d) = \begin{cases} 1 - [7(\frac{d}{\eta})^2 - 8.75(\frac{d}{\eta})^3 + 3.5(\frac{d}{\eta})^5 - 0.75(\frac{d}{\eta})^7], d < \eta \\ 0, \text{ otherwise} \end{cases}.$$

Exponential

$$C(d) = \exp\big[-(\frac{d}{\eta})^2\big]$$

Matérn

$$C(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}\big(\frac{d\sqrt{2\nu}}{\eta}\big)^{\nu} K_{\nu}\big(\frac{d\sqrt{2\nu}}{\eta}\big)$$

Spherical

$$C(d) = \begin{cases} 1 - 1.5\frac{d}{\eta} + 0.5(\frac{d}{\eta})^3, d < \eta \\ 0, \text{ otherwise} \end{cases}$$

Powered exponential

$$C(d) = \exp\big[-(\frac{d}{\eta})^{\nu}\big], 0 < \nu \leq 2$$

Appendix C

Delta method for the variance of the Nelson-Aalen estimator

The delta method allows us to approximate the mean and variance of a function of a random variable. For large samples, one can assume that higher order terms in the approximation are small enough to be ignored. We describe the univariate case below before applying the method.

C.1 Function of one random variable

Given a random variable X we can find the mean and variance of some function $f(X)$, provided f is sufficiently differentiable and X has finite moments. Say $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

We begin by taking the Taylor expansion of $f(X)$ about the mean.

$$f(X) \approx f(\mu) + f'(\mu)(X - \mu) + \frac{1}{2}f''(\mu)(X - \mu)^2 \tag{C.1}$$

where for a large sample, $|X - \mu|$ is small and hence $|X - \mu|^p$ decreases as p increases. Hence, higher order terms are small enough to be omitted and since $\mathrm{E}[X - \mu] = 0$,

$$\begin{aligned}\mathrm{E}[f(X)] &\approx \mathrm{E}[f(\mu)] + \mathrm{E}[\frac{1}{2}f''(\mu)(X-\mu)^2] \\ &= f(\mu) + \frac{1}{2}f''(\mu)\mathrm{E}[(X-\mu)^2] \\ &= f(\mu) + \frac{1}{2}f''(\mu)\sigma^2. \end{aligned} \tag{C.2}$$

Subtracting C.2 from C.1, taking expectations and squaring, we obtain

$$\begin{aligned}\mathrm{E}[f(X) - \mathrm{E}[f(X)]]^2 &= \mathrm{Var}(f(X)) \\ &\approx \mathrm{E}[f'(\mu)(X-\mu)]^2 \\ &= [f'(\mu)]^2\mathrm{E}[(X-\mu)^2] \\ &= [f'(\mu)]^2\sigma^2.\end{aligned}$$

C.2 Variance of Nelson-Aalen

We can use a delta-method approximation to show that

$$Var(\hat{A}(\ell)) \approx \frac{\mathrm{Var}(\hat{S}(\ell))}{\hat{S}(\ell)}.$$

Let $g(t) = \hat{A}(t)$. Since $\hat{A}(t) = -\log(\hat{S}(t))$ and $\hat{S}(t) = 1 - \frac{N(t)}{n}$, we have

$$g(N) \approx -\log(1 - \frac{N}{n}),$$

where $N = N(t)$. Differentiating, we obtain

$$g'(N) \approx \frac{1}{n-N}$$

and

$$g''(N) \approx \frac{1}{(n-N)^2}.$$

Hence by the delta method above we have

$$\mathrm{E}[\hat{A}(t)] \approx -\log(1 - \frac{\mathrm{E}[N]}{n}) + \frac{\mathrm{Var}(N)}{2(n-\mathrm{E}[N])^2}$$

and

$$\begin{aligned}\mathrm{Var}(\hat{A}(t)) &\approx \left(\frac{1}{n-N}\right)^2 \mathrm{Var}(N) \\ &= \frac{\mathrm{Var}(N)}{n^2\left(1-\frac{N}{n}\right)^2}\end{aligned}$$

as required.

Appendix D

Central limit theorem

Here we assume a fixed finite discrete space $\mathcal{X}$ and that we have N independent and identically distributed replicates of the estimator $\hat{A}(t)$.

Let $W(t) = \hat{A}(t) - E\{\hat{A}(t)\}$ where $t \in (0, \tau)$. By Theorem 1 of Bloznelis and Paulauskas (1994), the following conditions confirm the existence of a central limit theorem for $\sqrt{N}\bar{A}(t)$ and verify that the limiting Gaussian process is continuous:

C1. There exists $D_1 > 0$ and $\alpha > 0.5$ such that $E[\{W(u) - W(t)\}^2] \leq D_1 \mid u - t \mid^{\alpha}$ for $0 < t \leq u < \tau$ with $u - t$ small, that is, there is an $\varepsilon > 0$ for which the result holds for all $u - t < \varepsilon$.

C2. There exists $D_2 > 0$ and $\gamma > 1$ such that

$$E[\{W(t) - W(s)\}\{W(u) - W(t)\}] \leq D_1 \mid u - t \mid^{\gamma}$$

for $0 < t \leq u < \tau$ with $u - t$ and $s - t$ both small.

For C1, let $I_x(t, u)$ be an indicator for the occurrence of a local minimum of Z at site x, where $z(x) \in (t, u)$. Let $J_x(t, u)$ be an indicator of field value $z(x)$ lying in the interval (t, u) whether it is a minimum or not. Recall that $Y(t)$ is a natural number for all t and $dN(t)/Y(t) = 0$ if $Y(t) = 0$. Then

$$\begin{aligned} \text{var}\{W(u) - W(t)\} &= \text{var}\{\hat{A}(u) - \hat{A}(t)\} \\ &\leq \text{var}\{N(u) - N(t)\} \\ &= \text{var}\{\sum_x I_x(t, u)\} \\ &\leq \mid \mathcal{X} \mid^2 \max_x [\text{var}\{I_x(t, u)\}] \\ &\leq \mid \mathcal{X} \mid^2 \max_x [\text{pr}\{I_x(t, u) = 1\}]. \end{aligned} \tag{D.1}$$

For each site x

$$\begin{aligned} \text{pr}\{I_x(t,u)=1\} &\leq \text{pr}\{J_x(t,u)=1\} \\ &= \int_t^u f_x(u)du, \end{aligned} \tag{D.2}$$

where $f_x(u)$ is the marginal density of $z(x)$. Since we have assumed $z(x)$ to be continuous, with no jump discontinuities, it follows that $f_x(u)$ is finite for $0 < u < \tau$. Hence for each x there is a value C_x such that

$$\int_t^u f_x(z)dz < C_x(u-t). \tag{D.3}$$

By (D.1)-(D.3) we have satisfied C1 with $\alpha = 1$.

For condition C2, we first observe that the moments of squared increments are dominated by the jumps in the underlying counting process $N(t)$. For example,

k	$\text{pr}\{N(u)-N(t)=k\}$	$\{W(t)-W(u)\}^2$
0	$O(1)$	$O\{(u-t)^2\}$
1	$O(u-t)$	$O(1)$
≥ 2	$o(u-t)$	$O\{(u-t)^2\}$.

Hence, we only need to consider the joint and marginal probabilities of jumps in the underlying counting process for the occurrence of events. For all sites x and y in $\mathcal{X}$ define $J_{xy}(s,t,u)$ to be an indicator that the $z(x)$ lies in the interval (t,u) and that $z(y)$ lies in the interval (t,u).

For an event to occur in the interval (t,u) it is a necessary condition that there is a site in $\mathcal{X}$ where the underlying field Z has a value in (t,u).

In turn

$$\begin{aligned} pr\{N(u)-N(t)=1, N(t)-N(s)=1\} &\leq \mid \mathcal{X} \mid^2 \max_{x\neq y}[\text{pr}\{I_{xy}(s,t,u)=1\}] \\ &\leq \mid \mathcal{X} \mid^2 C_{\mathcal{X}}(u-t)(t-s) \end{aligned}$$

where $C_{\mathcal{X}}$ depends on the well-defined and finite density of bivariate marginals of Z. Since $(u-t)(t-s) < (u-s)^2$, C2 holds with $\gamma = 2$.

Bibliography

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aalen, O. O., Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2009). History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, 5:1–28.

Adler, R. J., Bobrowski, O., Borman, M. S., Subag, E., Weinberger, S., and others (2010). Persistent homology for random fields and complexes. In *Borrowing Strength: Theory Powering Applicationsa Festschrift for Lawrence D. Brown*, pages 124–143. Institute of Mathematical Statistics.

Alegria, A., Cuevas, F., Diggle, P., Porcu, E., and others (2018). A family of covariance functions for random fields on spheres. *CSGB Research Reports, Department of Mathematics, Aarhus University*.

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, 33:1148–1159.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49:765–769.

Arnold, B. C. (2009). Flexible univariate and multivariate models based on hidden truncation. *Journal of Statistical Planning and Inference*, 139:3741–3749.

Baker, A. H., Xu, H., Dennis, J. M., Levy, M. N., Nychka, D., Mickelson, S. A., Edwards, J., Vertenstein, M., and Wegener, A. (2014). A methodology for evaluating the impact of data compression on climate simulation data. In *Proceedings of the 23rd international symposium on high-performance parallel and distributed computing - HPDC '14*, pages 203–214. ACM Press.

Barrett, J., Diggle, P., Henderson, R., and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and

exact likelihood inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:131–148.

Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). Persistent homology analysis of brain artery trees. *The Annals of Applied Statistics*, 10:198–218.

Bera, A. K. and Jarque, C. M. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 7:313–318.

Beukers, F. (2007). Gauss hypergeometric function. In Holzapfel, R.-P., Uluda, A. M., and Yoshida, M., editors, *Arithmetic and Geometry Around Hypergeometric Functions*, volume 260, pages 23–42. Birkhuser Basel.

Bhattacharya, S., Ghrist, R., and Kumar, V. (2015). Persistent homology for path planning in uncertain environments. *IEEE Transactions on Robotics*, 31(3):578–590.

Bivand, R. S., Pebesma, E., and Gmez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. Springer New York.

Blackmon, M., Boville, B., Bryan, F., Dickinson, R., Gent, P., Kiehl, J., Moritz, R., Randall, D., Shukla, J., Solomon, S., and others (2001). The community climate system model. *Bulletin of the American Meteorological Society*, 82:2357–2376.

Bloznelis, M. and Paulauskas, V. (1994). A note on the central limit theorem for stochastically continuous processes. *Stochastic Processes and their Applications*, 53:351–361.

Bohmer, P. (1912). Theorie der unabhngigen Wahrscheinlichkeiten. In *Rapports Memoires et Proces verbaux de Septieme Congres International dActuaires Amsterdam*, volume 2, pages 327–343.

Borgan, O. (1997). Three contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, and Aaalen-Johansen estimators. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420*.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16:77–102.

Bubenik, P. (2020). The persistence landscape and some of its properties. *arXiv:1810.04963*, 15:97–117.

Bubenik, P. and Dlotko, P. (2017). A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114.

Cao, G., Yang, L., and Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, 24:359–377.

Castro Morales, F. E., Gamerman, D., and Paez, M. S. (2013). State space models with spatial deformation. *Environmental and Ecological Statistics*, 20:191–214.

Castruccio, S. (2016). Assessing the spatio-temporal structure of annual and seasonal surface temperature for cmip5 and reanalysis. *Spatial Statistics*, 18:179–193.

Castruccio, S. and Genton, M. G. (2014). Beyond axial symmetry: An improved class of models for global data. *Stat*, 3:48–55.

Castruccio, S. and Genton, M. G. (2016). Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature. *Technometrics*, 58:319–328.

Castruccio, S. and Guinness, J. (2017). An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66:329–344.

Castruccio, S., McInerney, D. J., Stein, M. L., Liu Crouch, F., Jacob, R. L., and Moyer, E. J. (2014). Statistical emulation of climate model projections based on precomputed GCM runs. *Journal of Climate*, 27:1829–1844.

Castruccio, S. and Stein, M. L. (2013). Global spacetime models for climate ensembles. *The Annals of Applied Statistics*, 7:1593–1611.

Chiou, J.-M. and Muller, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association*, 104:572–585.

Chung, M. K., Bubenik, P., and Kim, P. T. (2009). Persistence diagrams of cortical surface data. In Prince, J. L., Pham, D. L., and Myers, K. J., editors, *Information Processing in Medical Imaging*, volume 5636, pages 386–397. Springer Berlin Heidelberg.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:459–513.

Crainiceanu, C. M., Staicu, A.-M., Ray, S., and Punjabi, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine*, 31:3223–3240.

Cramer, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928:13–74.

Crawford, L., Monod, A., Chen, A. X., Mukherjee, S., and Rabadn, R. (2016). Topological summaries of tumor images improve prediction of disease free survival in glioblastoma multiforme. *arXiv:1611.06818*.

Curto, C. (2016). What can topology tell us about the neural code? *Bulletin of the American Mathematical Society*, 54:63–78.

Dabaghian, Y., Brandt, V. L., and Frank, L. M. (2014). Reconceiving the hippocampal map as a topological template. *eLife*, 3:e03476.

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g1. *Biometrika*, pages 679–681.

de Silva, V. and Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358.

Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, pages 1735–1765. arXiv: 0908.1980.

DeWoskin, D., Climent, J., Cruz-White, I., Vazquez, M., Park, C., and Arsuaga, J. (2010). Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology and its Applications*, 157:157–164.

Dotko, P., Hess, K., Levi, R., Nolte, M., Reimann, M., Scolamiero, M., Turner, K., Muller, E., and Markram, H. (2017). Topological analysis of the connectome of digital reconstructions of neural microcircuits. *arXiv:1601.01580*.

Edelsbrunner, Letscher, and Zomorodian (2002). Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533.

Edelsbrunner, H. and Harer, J. (2008). Persistent homology - a survey. In Goodman, J. E., Pach, J., and Pollack, R., editors, *Contemporary Mathematics*, volume 453, pages 257–282. American Mathematical Society.

Edwards, M., Castruccio, S., and Hammerling, D. (2019). A multivariate global spatiotemporal stochastic generator for climate ensembles. *Journal of Agricultural, Biological and Environmental Statistics*, 24:464–483.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42:2301–2339.

Fisher, R. A. (1937). *The Design of Experiments*. Oliver And Boyd.

Fouedjio, F., Desassis, N., and Romary, T. (2015). Estimation of space deformation model for non-stationary random functions. *Spatial Statistics*, 13:45–61.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics New York.

Fuentes, M. (2005). A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis*, 96:30–54.

Fugacci, U., Scaramuccia, S., Iuricich, F., and Floriani, L. D. (2016). Persistent homology: a step-by-step introduction for newcomers. In *STAG*, pages 1–10. The Eurographics Association.

Furrer, R., Nychka, D., Sain, S., and Nychka, M. D. (2012). Fields: Tools for spatial data.

Gameiro, M., Hiraoka, Y., Izumi, S., Kramar, M., Mischaikow, K., and Nanda, V. (2015). A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32:1–17.

Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. Siam.

Garside, K., Gjoka, A., Henderson, R., Johnson, H., and Makarenko, I. (2020). Event History and Topological Data Analysis. *Biometrika*.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC Press.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Springer Science & Business Media.

Gidea, M. (2017). Topology data analysis of critical transitions in financial networks. *Available at SSRN 2903278*.

Gilli, M., Maringer, D., and Schumann, E. (2011). *Numerical Methods and Optimization in Finance*. Elsevier Science.

Givant, S. R. and Halmos, P. R. (2009). *Introduction to Boolean Algebras*. Springer.

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19:1327–1349.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). *A Survey of Statistical Network Models*. Now Publishers Inc.

Graham, J. W. (2012). *Missing Data*. Springer New York.

Greenwood, M. and others (1926). A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*

Guan, Y., Sherman, M., and Calvin, J. A. (2004). A nonparametric test for spatial isotropy using subsampling. *Journal of the American Statistical Association*, 99:810–821.

Guillemard, M. and Iske, A. (2017). Interactions Between Kernels, Frames, and Persistent Homology. In Pesenson, I., Le Gia, Q. T., Mayeli, A., Mhaskar, H., and Zhou, D.-X., editors, *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science: Novel Methods in Harmonic Analysis, Volume 2*, pages 861–888. Springer International Publishing.

Guinness, J. and Fuentes, M. (2015). Covariance functions for mean square dierentiable processes on spheres. *preprint*.

Guinness, J. and Fuentes, M. (2016). Isotropic covariance functions on spheres: Some properties and modeling considerations. *Journal of Multivariate Analysis*, 143:143–152.

Guinness, J. and Stein, M. L. (2013). Transformation to approximate independence for locally stationary Gaussian processes. *Journal of Time Series Analysis*, 34:574–590.

Haas, T. C. (1990). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24:1759–1769.

Haas, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association*, 90:1189–1199.

Hahn, M. G. (1977). Conditions for sample-continuity and the central limit theorem. *The Annals of Probability*, pages 351–360.

Henderson, R., Makarenko, I., Bushby, P., Fletcher, A., and Shukurov, A. (2020). Statistical topology and the random interstellar medium. *Journal of the American Statistical Association*, 115:625–635.

Herdin, M. and Bonek, E. (2004). A MIMO correlation matrix based metric for characterizing non-stationarity. In *IST Mobile & Wireless Communications Summit*.

Herdin, M., Czink, N., Ozcelik, H., and Bonek, E. (2005). Correlation Matrix Distance, a Meaningful Measure for Evaluation of Non-Stationary MIMO Channels. In *2005 IEEE 61st Vehicular Technology Conference*, volume 1, pages 136–140.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian Statistics*, 6:761–768.

Higham, N. J. (2002). Computing the nearest correlation matrix–a problem from finance. *IMA Journal of Numerical Analysis*, 22:329–343.

Hothorn, T., Bretz, F., and Genz, A. (2001). On multivariate t and Gauss probabilities in R. *sigma*, 1000:3.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., and others (2013). The community earth system model: a framework for collaborative research. *Bulletin of the American Meteorological Society*, 94:1339–1360.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, pages 163–172.

Jeong, J., Castruccio, S., Crippa, P., and Genton, M. G. (2017). Statistics-based compression of global wind fields. *arXiv:1702.01995*.

Jeong, J. and Jun, M. (2015a). A class of Matern-like covariance functions for smooth processes on a sphere. *Spatial Statistics*, 11:1–18.

Jeong, J. and Jun, M. (2015b). Covariance models on the surface of a sphere: when does it matter? *Stat*, 4:167–182.

Jones, R. H. (1963). Stochastic processes on a sphere. *The Annals of Mathematical Statistics*, 34:213–218.

Joubert, P. and Langdell, S. (2013). Modelling: Mastering the correlation matrix. *The Actuary*.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*, volume 600. Academic press London.

Jun, M. (2014). Matern-based nonstationary cross-covariance models for global processes. *Journal of Multivariate Analysis*, 128:134–146.

Jun, M. and Stein, M. L. (2008). Nonstationary covariance models for global data. *The Annals of Applied Statistics*, 2:1271–1289.

Justel, A., Pea, D., and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 35:251–259.

Kanari, L., Dotko, P., Scolamiero, M., Levi, R., Shillcock, J., Hess, K., and Markram, H. (2016). Quantifying topological invariants of neuronal morphologies. *arXiv:1603.08432*.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457.

Kay, J. E., Baker, A. H., Hammerling, D., Michelson, S. A., Xu, H., and others (2016). Evaluating lossy data compression on climate simulation data within a large ensemble. *Geoscientific Model Development*, 9:4381–4403.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M. (2015). The Community Earth System Model (CESM) Large Ensemble Project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, 96:1333–1349.

Konzen, E., Shi, J. Q., and Wang, Z. (2019). Modelling Function-Valued Processes with Nonseparable Covariance Structure. *arXiv:1903.09981 [stat]*. arXiv: 1903.09981.

Kovacev-Nikolic, V., Bubenik, P., Nikoli, D., and Heo, G. (2016). Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology*, 15:19–38.

Kuusela, M. and Stein, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474:20180400.

Leibon, G., Pauls, S., Rockmore, D., and Savell, R. (2008). Topological structures in the equities market network. *Proceedings of the National Academy of Sciences*, 105:20589–20594.

Li, Y. and Ryan, L. (2002). Modeling spatial survival data using semiparametric frailty models. *Biometrics*, 58:287–297.

Lindstrom, P. and Isenburg, M. (2006). Fast and efficient compression of floating-point data. *IEEE Transactions on Visualization and Computer Graphics*, 12:1245–1250.

Little, R. J. A. (1987). *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons.

Lo, D. and Park, B. (2018). Modeling the spread of the Zika virus using topological data analysis. *PloS one*, 13:e0192120.

Lopes, R. H., Reid, I., and Hobson, P. R. (2007). The two-dimensional Kolmogorov-Smirnov test.

Lord, L.-D., Expert, P., Fernandes, H. M., Petri, G., Van Hartevelt, T. J., Vaccarino, F., Deco, G., Turkheimer, F., and Kringelbach, M. L. (2016). Insights into brain architectures from the homological scaffolds of functional connectivity networks. *Frontiers in Systems Neuroscience*, 10:85.

Marsaglia, G., Tsang, W. W., and Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8:1–4.

Meiring, W., Guttorp, P., and Sampson, P. (1998). Computational issues in fitting spatial deformation models for heterogeneous spatial correlation. *Computing Science and Statistics*, pages 409–417.

Mi, X., Miwa, T., and Hothorn, T. (2009). mvtnorm: New numerical algorithm for multivariate normal probabilities. *The R Journal*, 1:37–39.

Miwa, T., Hayter, A., and Kuriki, S. (2003). The evaluation of general noncentred orthant probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:223–234.

Moitra, A., Malott, N. O., and Wilsey, P. A. (2018). Cluster-based data reduction for persistent homology. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 327–334.

Munch, E. (2017). A users guide to topological data analysis. *Journal of Learning Analytics*, 4:47–61.

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108:7265–7270.

Nychka, D., Hammerling, D., Krock, M., and Wiens, A. (2018). Modeling and emulation of nonstationary Gaussian fields. *Spatial Statistics*, 28:21–38.

Nychka, D., Hammerling, D., Sain, S., Lenssen, N., and Nychka, M. D. (2016). LatticeKrig: Multiresolution kriging based on Markov random fields. R package version 8.4.

Obayashi, I., Hiraoka, Y., and Kimura, M. (2018). Persistence diagrams with linear machine learning models. *Journal of Applied and Computational Topology*, 1:421–449.

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:17.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., Dubash, N. K., Edenhofer, O., Elgizouli, I., Field, C. B., Forster, P., Friedlingstein, P., Fuglestvedt, J., Gomez-Echeverri, L., Hallegatte, S., Hegerl, G., Howden, M., Jiang, K., Cisneroz, B. J., Kattsov, V., Lee, H., Mach, K. J., Marotzke, J., Mastrandrea, M. D., Meyer, L., Minx, J., Mulugetta, Y., O'Brien, K., Oppenheimer, M., Pereira, J. J., Pichs-Madruga, R., Plattner, G.-K.,

Prtner, H.-O., Power, S. B., Preston, B., Ravindranath, N. H., Reisinger, A., Riahi, K., Rusticucci, M., Scholes, R., Seyboth, K., Sokona, Y., Stavins, R., Stocker, T. F., Tschakert, P., Vuuren, D. v., and Ypserle, J.-P. v. (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17:483–506.

Pan, J. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, 90:239–244.

Patania, A., Vaccarino, F., and Petri, G. (2017). Topological analysis of data. *EPJ Data Science*, 6:1–6.

Perrin, O. and Senoussi, R. (2000). Reducing non-stationary random fields to stationarity and isotropy using a space deformation. *Statistics & Probability Letters*, 48:23–32.

Pokorny, F. T., Hawasly, M., and Ramamoorthy, S. (2016). Topological trajectory classification with filtrations of simplicial complexes and persistent homology. *The International Journal of Robotics Research*, 35:204–223.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86:677–690.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27:204–237.

R. Wadhwa, R., F.K. Williamson, D., Dhawan, A., and G. Scott, J. (2018). TDAstats: R pipeline for computing persistent homology in topological data analysis. *Journal of Open Source Software*, 3:860.

Rebonato, R. and Jaeckel, P. (2011). The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. *Available at SSRN 1969689*.

Reich, B. J., Eidsvik, J., Guindani, M., Nail, A. J., and Schmidt, A. M. (2011). A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration. *The Annals of Applied Statistics*, 5:2425–2447.

Robusto, C. C. (1957). The cosine-haversine formula. *The American Mathematical Monthly*, 64:38–40.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87:108–119.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1:27–64.

Schauf, A., Cho, J. B., Haraguchi, M., and Scott, J. J. (2016). Discrimination of economic input-output networks using persistent homology.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S., Martini, J., Ballani, F., Moreva, O., Auel, J., Menck, P. J., Gross, S., Ober, U., Ribeiro, P., Singleton, R., Pfaff, B., and R Core Team (2020). RandomFields: Simulation and analysis of fandom fields. R package version 3.3.8.

Schmidt, A. M., Guttorp, P., and O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22:487–500.

Schmidt, A. M. and Rodrguez, M. A. (2011). Modelling multivariate counts varying continuously in space. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, pages 611–638. Oxford University Press.

Schumann, E. (2020). NMOF: Numerical methods and optimization in finance. R package version 2.1.

Sethares, W. A. and Budney, R. (2014). Topology of musical data. *Journal of Mathematics and Music*, 8:73–92.

Singh, N., Couture, H. D., Marron, J. S., Perou, C., and Niethammer, M. (2014). Topological descriptors of histology images. In Wu, G., Zhang, D., and Zhou, L., editors, *Machine Learning in Medical Imaging*, volume 8679, pages 231–239. Springer International Publishing.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media. OCLC: 968504419.

Stolz, B., Harrington, H., and Porter, M. A. (2016). The topological 'shape' of brexit. *Available at SSRN 2843662*.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20:316–334.

Taylor, D., Klimm, F., Harrington, H. A., Kramr, M., Mischaikow, K., Porter, M. A., and Mucha, P. J. (2015). Topological data analysis of contagion maps for examining spreading processes on networks. *Nature Communications*, 6:1–11.

Vasudevan, R., Ames, A., and Bajcsy, R. (2013). Persistent homology for automatic determination of human-data based cost of bipedal walking. *Nonlinear Analysis: Hybrid Systems*, 7:101–115.

Ver Hoef, J. M., Cressie, N., and Barry, R. P. (2004). Flexible spatial models for kriging and cokriging using moving averages and the fast fourier transform (FFT). *Journal of Computational and Graphical Statistics*, 13:265–282.

Visser, E. (2018). Calculation of Cech homology of Hawk-Eye data. Master's thesis, Utrecht University.

Wang, J.-L., Chiou, J.-M., and Mller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Weisstein, E. W. (2020). Spherical Trigonometry. https://mathworld.wolfram.com/SphericalTrigonometry.html.

Williams, R. L. (1995). Product-limit survival functions with correlated survival times. *Lifetime Data Analysis*, 1:171–186.

Xia, K. and Wei, G.-W. (2014). Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 30:814–844.

Yoo, J., Kim, E. Y., Ahn, Y. M., and Ye, J. C. (2016). Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. *Journal of Neuroscience Methods*, 267:1–13.