THE UNIVERSITY OF CHICAGO

IRREGULAR SPACED DATA, SPATIO-TEMPORAL MODELING AND CLUSTERING OF TIME SERIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
SOUDEEP DEB

CHICAGO, ILLINOIS
AUGUST 2018

To my elder brother Sougata

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Wei Biao Wu, for his continuous support throughout my PhD career. He has introduced me to various exciting problems related to time series, and has helped me immensely in every aspect of my academic career. Without Professor Wu's extraordinary guidance, this thesis would not have been the same.

I am also grateful to Professor Ruey S. Tsay for introducing me to the problem of spatio-temporal modeling in environmental research, which culminated in Chapter 3 of this thesis. Special thanks go to Professor Michael L. Stein as well. I have learnt a lot about teaching from him, and he has been extremely helpful with his valuable feedback and suggestions about the thesis. I also want to thank Professor Mohsen Pourahmadi for collaborating with me on a project that constitutes the backbone of Chapter 2 in this thesis.

Further, I would mention some of my friends who made my stay in Chicago a memorable experience. Especially, I am grateful to my old friends Kushal Kumar Dey, Somak Dutta, Subhajit Goswami and Sayar Karmakar, who have helped a lot during these five years. I would also thank Piyush Arya, Sumitra Badrinathan, Subhadip Chowdhury, Debsouri Kundu, Krithika Mohan, Preeti Poddar and Vishwas Srivastava with whom I have spent some of my best days in the United States of America.

Finally, it goes without saying that this thesis would not be possible without the support of my parents and my elder brother. I am indebted to them for always being there by my side.

ABSTRACT

In this thesis, three different problems in time series and random field have been discussed. First, for a general class of stationary random fields, we study the asymptotic properties of different parametric and nonparametric spectral density estimators under an easily verifiable short-range dependence condition. The theory developed here allows both regular and irregular spaced data with minimal restriction on the index set and thus is applicable across a wide range of practical scenarios.

The second problem revolves around developing a spatio-temporal model with space-time interaction for air pollution data ($\mathrm{PM}_{2.5}$), which enables one to provide forecasts and insights about the air quality. The proposed model uses a parametric space-time interaction component along with the spatial and temporal components in the mean structure, and introduces a random-effects component specified in the form of zero-mean spatio-temporal processes. For application, air pollution data from Taiwan have been analyzed.

The third problem in the thesis deals with a time series clustering problem. Using $\mathbb{L}^2$ distance between nonparametric spectral density estimates, a hierarchical clustering algorithm has been developed. Simulation studies show that the power of the algorithm is very good for most scenarios. Especially, it shows much better performances for small samples. This algorithm can be extended to different practical setups and can be used on real life data obtained from various fields.

CHAPTER 1

INTRODUCTION

Time dependent or spatially dependent data arises in many different disciplines, and the analysis of such data is very important for researchers working on various topics - environmental science, medical science and finance being some popular examples. In this thesis, we are going to explore three interesting and challenging problems related to such data.

In Chapter 2, our focus will be on a very general setup where we consider random fields and explore the asymptotic properties for the discrete Fourier transform (DFT), periodogram, Whittle likelihood estimator and nonparametric spectral density estimator. Throughout this chapter, we work with the functional dependence measure developed in Wu [99]. The theory we have derived considers an easily verifiable short-range dependence condition expressed in terms of the dependence measure alluded to above. Our results are general in the sense that we allow irregularly spaced data with minimal restriction. More precisely, for sample size n, the irregular setup is indexed by the subset Γ_n of $\mathbb{Z}^d$, where d is any dimension. Then, Γ_n is only required to satisfy the condition $|\Gamma_n| \to \infty$, as $n \to \infty$ (see Assumption 2.1).

In this thesis, under the above setup, we have derived the asymptotic normality for kernel spectral density estimators and the Whittle estimator of a parameterized spectral density function. Further, we also develop asymptotic results for a covariance matrix estimate. As this restriction on the index set is very mild, our theory can be applied across variety of scenarios. As a final piece, simulation examples are provided at the end of the chapter.

Next, in Chapter 3, we work on a crucial problem related to human life. As we already know, the effect of air pollution on public health, vegetation, and more generally, on the human society and the ecosystem has been a burning issue in recent years. In fact, previous studies of spatio-temporal models showed that even a short-term exposure to high concentrations of atmospheric fine particulate matters (PM) can be hazardous to the health of ordinary people. Several epidemiological studies; for example Blangiardo et al. [13], Jerrett et al. [50], Pope III et al. [79] and Thurston et al. [92]; have established that the PM are linked to a range of serious cardiovascular, respiratory, and visibility problems. Thus, it is of utmost importance to have a clear understanding of the status of air pollution and to provide forecasts and insights about the air quality to the general public and researchers in environmental studies. Our work on this problem has more of an applied flavor and the focus has been on developing a model that can address the issue of space-time interaction in air pollution data and in turn can provide better predictive abilities for such data.

In this thesis, we work with data on PM with aerodynamic diameters less than 2.5 micron (denoted as $\mathrm{PM}_{2.5}$). The proposed model uses a parametric space-time interaction component along with the spatial and temporal components in the mean structure, and introduces a random-effects component specified in the form of zero-mean spatio-temporal processes. Further, we use a weighted least squares approach to estimate the unknown parameters and that help us in predicting the higher pollution levels with more accuracy. Asymptotic results on the estimation procedure have also been discussed in Section 3.3.

On the other hand, it is worth mention that most existing studies on related data (detailed discussion can be found in Section 3.1) concentrated on the pollution issues in different parts of Canada and the USA, while the problem is certainly not limited to this continent only. In fact, there have been only a few good studies that analyze air pollution data from countries in Asia or Africa. A noteworthy paper in this regard is Al-Awadhi and Al-Awadhi [2], who took hierarchical Bayesian approaches to develop a dynamic linear models for air pollutants in Kuwait. They dealt with the temporal and spatial effects independently, defining separate structures for the two processes. However, in general, there is a serious dearth of related papers that focus on the pollution situations in the developing countries. And that is another key contribution of this work, as we develop a model and analyze the data for Taiwan.

Finally, Chapter 4 is about a time series clustering problem, where the data are assumed to be coming from a very general class of stationary processes. Our algorithm is based on the covariance structures of the data, and it is a test-based method. We use the $\mathbb{L}^2$ distance of pairwise differences of the spectral density estimates to test if two time series objects have same spectral density. The testing procedure is then extended to multiple time series case. Then, using this testing procedure, we develop our clustering algorithm.

Throughout the study, simulation-based methods have been used to approximate the distribution of the test statistics so as to ensure much better finite-sample properties. The methods are discussed in detail in Section 4.3 and Section 4.4, while some empirical studies and an application to a real data are provided in Section 4.5.

CHAPTER 2

IRREGULAR-SPACED RANDOM FIELDS

2.1 Background

Analysis of irregularly spaced data has been attracting considerable attention from researchers in various fields, ranging from environmental science to economics. The origin of irregular data is in fact the limit theorem for random fields with continuous parameter where the sets of integration in the limit theorems approach to infinity in Van Hove sense, see for example, Ivanov and Leonenko [48]. In a broad sense, there are two different approaches to deal with the irregularly spaced spatial data. The classical and more popular Kriging or interpolation approach (Cressie [20]) is parametric in nature. A nonparametric or a frequency domain approach was considered by Fuentes [32] which revolves around the assumption that the sampled locations are fixed and not random. Vidal-Sanz [94] also considered nonparametric estimation of spectral densities for second-order stationary random fields on a d-dimensional lattice. In that paper, the author proposed modified estimator classes with improved bias convergence rate. In a much recent work, Bandyopadhyay et al. [8] formulated a spatial frequency domain empirical likelihood method for irregularly spaced data. Other related works can be found in Bandyopadhyay and Lahiri [7], Hall and Patil [40], Hall et al. [41], Im et al. [47] and the references therein.

Spectral domain methods to approximate the Gaussian likelihood for irregularly spaced datasets were proposed by Matsuda and Yajima [69] where the sampled locations are assumed random with a particular distribution having a continuous density function. Their non-parametric and parametric estimators of the spectral density function of the underlying random fields are similar to those in classical time series analysis. The parametric spectral density was estimated by minimizing the Whittle likelihood while the non-parametric spectral density estimator was a spectral window estimator, and they studied the asymptotic properties of those estimators.

In spatial data analysis one usually deals with irregularly spaced data. To set the notation, let $(\Gamma_n)_{n\geq 1}$ be a sequence of finite subsets of $\mathbb{Z}^d$ representing the sampling locations or design points. Our goal here is to work with an asymptotic regime which imposes minimal restrictions on the sampling set and its boundary:

Assumption 2.1: (Asymptotic regime) Let $\Gamma_n = \{L_{n,1}, \ldots, L_{n,n}\} \subset \mathbb{Z}^d$ be the set of sampling locations such that the choice of Γ_n satisfies the property $|\Gamma_n| \to \infty$.

For simplicity, from now on, we would write $L_k = L_{n,k}$. It is instructive to compare the above regime with the Matsuda and Yajima [69] setup where each sampling location $\mathbf{t}_i$ is obtained from a randomly generated d-dimensional vector $\mathbf{u}_i = (u_{i,1}, \ldots, u_{i,d})$ by $t_{ij} = A_j u_{ij}$ for $j = 1, 2, \ldots, d$. They further assumed that the coefficient A_j's and the sample size (n_k), if expressed as a function of k, satisfy the condition $|S_k|/n_k \to 0$ as k goes to infinity, where $|S_k|$ is the area of the rectangle $[0, A_1] \times \ldots \times [0, A_d]$.

Other examples with special constraints on the sample set can be found in Jones [51], Parzen [77], Neave [72], Neave [73], Clinger and Van Ness [19]. A related study with irregularly spaced observations for an increasing spatio-temporal domain can be found in Li et al. [63]. Similar to Matsuda and Yajima [69], they also viewed the spatial locations at which the data is observed as random in number and location; generated from a homogeneous 2-dimensional Poisson process. As an interesting feature, our theory requires a minimal condition on the set Γ_n, which is an attractive property in spatial applications in which the underlying observation domains can be quite irregular.

For a given Γ_n, the discrete Fourier transform (DFT) is defined by

$$S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{j \in \Gamma_n} X_j \exp(-\imath j' \theta) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_{L_k} \exp(-\imath L_k' \theta), \tag{2.1}$$

where $\imath = \sqrt{-1}$ and $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$. The periodogram of the data is defined by

$$I_n(\theta) = |S_n(\theta)|^2 = \frac{1}{n} \left[\left\{ \sum_{j \in \Gamma_n} X_j \cos(j' \theta) \right\}^2 + \left\{ \sum_{j \in \Gamma_n} X_j \sin(j' \theta) \right\}^2 \right]. \tag{2.2}$$

We will study aspects of both parametric and nonparametric estimators of the spectral density functions of stationary random fields. We begin by finding asymptotic results for the discrete Fourier transform (DFT) for irregularly spaced random fields, and then study the asymptotic properties of the spectral density estimates. As pointed out earlier, an important feature of our approach is that we do not impose any restriction on the index set Γ_n, other than the natural requirement $|\Gamma_n| \to \infty$.

This chapter is structured as follows: Section 2.2 presents the setup, assumptions and some preliminary results regarding short-memory stationary random fields. The discrete Fourier transform of the data and its asymptotic properties are presented in Section 2.3 while the Whittle likelihood and parametric spectral density estimator are discussed in Section 2.4. In Section 2.5, we will present the nonparametric spectral density estimator and the covariance function and its different aspects (e.g. consistency, asymptotic normality). We will also discuss the estimation of covariances matrices for an irregular set-up in Section 2.6. All proofs are provided in Appendix.

2.2 Short-Range Dependent Random Fields

We shall consider a very general class of stationary random fields which are functions of independent and identically distributed (iid) random variables. In related works, Whittle [97] considered two-dimensional linear auto-regression fields and Besag [11] discussed stationary auto-normal processes and proposed estimation methods and goodness-of-fit tests applicable to spatial Markov schemes defined over a rectangular lattice. Other noteworthy examples may be found in Guyon [37] and Kashyap [53]. Our setup is general enough to include the most common linear and nonlinear processes.

Assumption 2.2: Let $\varepsilon_j, j \in \mathbb{Z}^d$, be iid random variables. Define

$$X_i = g(\varepsilon_{i-s}; s \in \mathbb{Z}^d), \qquad i \in \mathbb{Z}^d, \tag{2.3}$$

where g is a measurable function such that X_i is well-defined.

Throughout this chapter, we work with short-range dependent stationary processes. We use the idea of coupling (Wu [99]) to define dependence measures. Let $\varepsilon'_i, \varepsilon_j, i, j \in \mathbb{Z}^d$ be iid.

Definition 2.1: (Functional dependence measure) Let $X_i \in \mathbb{L}_p$, $p \geq 1$. Define

$$\delta_{i,p} = \|X_i - X_i^*\|_p, \text{ where } X_i^* = g(\varepsilon^*_{i-s}; s \in \mathbb{Z}^d), \tag{2.4}$$

and $\varepsilon^*_j = \varepsilon_j$ if $j \neq 0$ and $\varepsilon^*_0 = \varepsilon'_0$. Also let $p' = \min\{2, p\}$ and define

$$\Theta_{m,p} = \sum_{|j|>m} \delta_{j,p} \qquad \text{and} \qquad \Psi_{m,p} = \left(\sum_{|j|>m} \delta^{p'}_{j,p} \right)^{1/p'}.$$

Definition 2.2: (Stability) The random field (X_i) defined in Equation (2.3) is said to be

p-stable if

$$\Delta_p := \sum_{i \in \mathbb{Z}^d} \delta_{i,p} < \infty.$$

Two concrete examples of such processes are given next.

Example 1: (Linear process) Let $X_i = \sum_{s \in \mathbb{Z}^d} a_s \varepsilon_{i-s}$, where $(\varepsilon_j)_{j \in \mathbb{Z}^d}$ are iid random variables with mean 0 and $\varepsilon_0 \in \mathbb{L}_p$, $p \geq 2$, and a_s are real coefficients such that $\sum_{s \in \mathbb{Z}^d} |a_s| < \infty$. Then $\delta_{i,p} = |a_i| \|\varepsilon_0 - \varepsilon_0^*\|_p = O(|a_i|)$. For the nonlinearly transformed process $Y_i = K(X_i)$, where $K(\cdot)$ is Lipschitz continuous, its functional dependence measure $\delta_{i,p}(Y)$ is also of order $O(|a_i|)$. □

Example 2: (Spatial Autoregressive Scheme) Let $\mathcal{N} \subset \mathbb{Z}^d$ be a finite set and $0 \notin \mathcal{N}$. Consider the spatial process in the form of nonlinear autoregressive scheme

$$X_i = G((X_{i-j})_{j \in \mathcal{N}}; \varepsilon_i),$$

where the function G is such that there exists nonnegative numbers $\ell_j, j \in \mathcal{N}$, with $\sum_{j \in \mathcal{N}} \ell_j < 1$ and the following holds: for all $(x_{-j})_{j \in \mathcal{N}}$ and $(x'_{-j})_{j \in \mathcal{N}}$,

$$|G((x_{-j})_{j \in \mathcal{N}}; \varepsilon_i) - G((x'_{-j})_{j \in \mathcal{N}}; \varepsilon_i)| \leq \sum_{j \in \mathcal{N}} \ell_j |x_{-j} - x'_{-j}|. \tag{2.5}$$

Also assume that there exists $(x_{-j})_{j \in \mathcal{N}}$ such that $G((x_{-j})_{j \in \mathcal{N}}; \varepsilon_0) \in \mathcal{L}^p$. Then following the argument in Shao and Wu (2004), we have $\delta_{i,p} = O(\rho^{|i|})$ for some $0 < \rho < 1$. □

2.3 Asymptotic Theory of the DFT

In this section, under some regularity conditions on the data-generating process we study the asymptotic distribution of DFT. These are the key ingredients for developing the spectral analysis of stationary processes. Peligrad and Wu [78] proved a central limit theorem for the Fourier transform of a stationary process in the regular case. Other works related to bias and variance of periodogram estimates for a regular set-up can be found in Pukkila [80] and Lin and Liu [65].

2.3.1 Asymptotic normality of the DFT

We establish the asymptotic normality of the DFT defined in Equation (2.1) using the Cramer-Wold device. To this end, instead of the DFT we consider the more general expression

$$W_n = \sum_{j \in \Gamma_n} c_j X_j \qquad \text{where } |c_j| \leq 1 \text{ for all } j \in \Gamma_n, \tag{2.6}$$

and wish to find the asymptotic joint distribution of $(Y_n(\theta), Z_n(\theta))/\sqrt{n}$ where

$$Y_n(\theta) = \sum_{j \in \Gamma_n} X_j \cos(j'\theta), \qquad Z_n(\theta) = -\sum_{j \in \Gamma_n} X_j \sin(j'\theta),$$

are the cosine and sine transforms of the data. As mentioned earlier, we are interested in linear combinations $aY_n(\theta) + bZ_n(\theta)$ (without loss of generality, we can assume that $a^2 + b^2 = 1$) which is of the form of Equation (2.6). For $k \geq 1$, let us use $N_k(0, \Sigma)$ to denote the k-variate normal distribution with 0 mean vector and variance-covariance matrix Σ.

Theorem 2.1: (Central limit theorem for DFT). Suppose $(X_i)_{i \in \mathbb{Z}^d}$ is a stationary centered random field defined by Equation (2.3) satisfying

$$\Delta_2 := \sum_{i \in \mathbb{Z}^d} \delta_{i,2} < \infty. \tag{2.7}$$

Also assume that $v_n^2 = \mathbb{E}(W_n^2) \to \infty$. Then, the following central limit theorem (CLT) holds:

$$L\left[W_n/\sqrt{n}, N\left(0, v_n^2/n\right)\right] \to 0 \; as \; n \to \infty, \tag{2.8}$$

where $L(.,.)$ is the Levy distance between distributions. Consequently, we have

$$L\left[\left(Y_n(\theta), Z_n(\theta)\right)/\sqrt{n}, N_2\left(0, \Sigma_n(\theta)/n\right)\right] \to 0, \tag{2.9}$$

where $\Sigma_n(\theta) = \mathrm{cov}((Y_n(\theta), Z_n(\theta))^T$.

The above theorem gives the joint asymptotic distribution of the discrete Fourier transform. For the regular set-up, we can further show that the two coordinates $(Y_n(\theta), Z_n(\theta))$ are asymptotically independent and the same will happen at two different frequencies.

Proposition 1: Consider the regular set-up $\Gamma_n = \prod_{l=1}^{d}\{1, 2, \ldots, n_l\}$, where $n_l \to \infty$ for all $l \leq d$. Let $\theta, \phi \in [-\pi, \pi)^d$ with $\theta, \phi \neq 0$, $\theta \neq \phi$ and $\theta + \phi \neq 0$, and suppose $f(\cdot)$ is the spectral density function. Then, $(Y_n(\theta), Z_n(\theta), Y_n(\phi), Z_n(\phi))/\sqrt{n}$ are asymptotically independent Gaussian random variables with asymptotic variances equal to $(f(\theta), f(\theta), f(\phi), f(\phi))/2$.

2.3.2 Bias of the Periodogram

It is known that for regularly observed stationary time series the periodogram is an asymptotically unbiased estimator of the spectral density function. This is not the case for irregular spatial data. In this section, we provide an expression for the bias of the periodogram.

For the process (X_i) given in Equation (2.3), assume that the mean is 0 and define the covariance function $\gamma_k = \mathbb{E}(X_0 X_k)$, $k \in \mathbb{Z}^d$. By Equation (2.7), we have $\sum_{k\in\mathbb{Z}^d} |\gamma_k| < \infty$.

Define the spectral density

$$f(\theta) = \sum_{k \in \mathbb{Z}^d} \gamma_k \cos(k'\theta). \tag{2.10}$$

In the literature the scaled form by $(2\pi)^{-d}$ is also widely used. In this study, we use the form (2.10). Recall that $\Gamma_n = \{L_1, \ldots, L_n\}$. Let $J = \{k \in \mathbb{Z}^d : \exists\, i, j \text{ with } L_i - L_j = k\}$ and $m_k = \#\{(i,j) : L_i - L_j = k\}$. We define the *location adjusted spectral density function* by

$$f_J(\theta) = \mathbb{E}[I_n(\theta)] = \frac{1}{n} \sum_{j \in J} m_j \gamma_j \cos(j'\theta). \tag{2.11}$$

The bias of $I_n(\theta)$ is

$$B_n(\theta) = f(\theta) - \mathbb{E}[I_n(\theta)] = \sum_{k \in \mathbb{Z}^d} \left(1 - \frac{m_k}{n}\right) \gamma_k \cos(k'\theta). \tag{2.12}$$

If, for any fixed k, $m_k/n \to 1$ (which holds for the regular rectangle index set $\Gamma_n = \prod_{i=1}^d \{1, 2, \ldots, J_i\}$, where $J_i \asymp n^{1/d}$ and $\prod_{i=1}^d J_i = n$), then by the Lebesgue dominated convergence theorem, $B_n(\theta) \to 0$ as $n \to \infty$. However, the same cannot be said for the irregular spatial case. In particular, if for each $k \in \mathbb{Z}^d$, the ratio m_k/n approaches r_k as $n \to \infty$ and these r_k's are constant, then the asymptotic bias is given by $B(\theta) = \sum_{k \in \mathbb{Z}^d} (1 - r_k) \gamma_k \cos(k'\theta)$.

2.4 Whittle likelihood and parametric spectral estimate

In this section we shall discuss a parametric estimator of the spectral density function $f(\theta)$. In what follows, we write the spectral density $f(\theta)$ (where $\theta \in \mathbb{R}^d$) as $f_\alpha(\theta)$ for a certain parameter vector $\alpha \in \mathbb{R}^p$. Since the spectral density governs the covariance function of a stationary process, γ_k will also be a function of α. Therefore, the location adjusted spectral density function $f_J(\theta)$ and the bias $B_n(\theta)$ discussed in the previous section are functions of both θ and α. We will denote them by $f_{J,\alpha}(\theta)$ and $B_{n,\alpha}(\theta)$, respectively.

A widely used approach to estimate the unknown parameter α is to minimize the (negative) Whittle's likelihood or an approximation to the Gaussian log likelihood which is of the form

$$p_n(\alpha) = \int_D \left\{ \log f_\alpha(\theta) + \frac{I_n(\theta)}{f_\alpha(\theta)} \right\} \, d\theta, \text{ where } D = [-\pi, \pi]^d. \tag{2.13}$$

Dahlhaus and Künsch [22] developed an asymptotic theory for Whittle estimator for regularly spaced time series data using a bias-adjusted periodogram in place of $I_n(\theta)$. We will adopt their technique to correct for the bias of the periodogram. The Whittle likelihood for the irregularly spaced data is

$$p_n(\alpha) = \int_D \left[\log f_{J,\alpha}(\theta) + \frac{I_n(\theta)}{f_{J,\alpha}(\theta)} \right] \, d\theta. \tag{2.14}$$

We denote the Whittle estimator that minimizes $p_n(\alpha)$ by $\hat{\alpha}_n$ and study the asymptotic behavior of this estimator. Suppose the parameter space for α is A. Let the true value of the parameter be α_0, and suppose that the Whittle estimate $\hat{\alpha}_n$ exists in the parameter space A for all n. Before stating the theorem for the Whittle estimator, in addition to the asymptotic regime (Assumption 2.1) and the nonlinear nature of the stationary random field (Assumption 2.2), we list a few more assumptions:

Assumption 2.3: The parameter space $A \subset \mathbb{R}^p$ is compact, and $D \subset \mathbb{R}^d$ is symmetric and compact such that the spectral density function (now denoted as $f_{J,\alpha}(\theta)$, defined on $A \times D$) is twice differentiable with respect to α and the first and second order derivatives are continuous for $\theta \in D$.

Assumption 2.4: (Identifiability condition) For $\alpha_1 \neq \alpha_2$, $f_{J,\alpha_1}(\theta) \neq f_{J,\alpha_2}(\theta)$ on a subset of D with positive Lebesgue measure.

Assumption 2.5: If ∇ denotes the first order derivative of a function, then

$$\int_D \frac{|\nabla f_{J,\alpha_0}(\theta)|}{\{f_{J,\alpha_0}(\theta)\}^2} d\theta < \infty \quad \text{and} \quad \int_D \left(\frac{|\nabla f_{J,\alpha_0}(\theta)|}{f_{J,\alpha_0}(\theta)} \right)^2 d\theta < \infty.$$

Finally, let us use $p(\alpha)$ to denote the limit of the Whittle likelihood $p_n(\alpha)$ defined by Equation (2.14). Note that $f_{J,\alpha}(\theta)$ and consequently, $p(\alpha)$ depends on the index set Γ_n. Also, for any Γ_n, we can say that $p(\alpha) > p(\alpha_0)$, for any $\alpha \neq \alpha_0$.

We now discuss the connection with the regularity assumptions. If the functional dependence measure $\delta_{i,p}$ satisfies the summability condition

$$\sum_{i \in \mathbb{Z}^d} |i|^2 \delta_{i,2} < \infty,$$

then the spectral density function is a bounded, twice partially differentiable function in view of

$$\sum_{i \in \mathbb{Z}^d} |i|^2 |\gamma_i| \leq \sum_{i \in \mathbb{Z}^d} |i|^2 \delta_{i,2} \sum_{i \in \mathbb{Z}^d} \delta_{i,2} < \infty.$$

Assumptions 2.3 to 2.5 are similar to the ones used in Matsuda and Yajima [69]. Theorem 2.2 below concerns the asymptotic properties of the Whittle estimator.

Theorem 2.2: Let $f_{J,\alpha}(\theta)$ be the location adjusted spectral density function defined in Section 2.3.2, denote the true value of the parameter α by α_0 and assume that the Whittle estimator $\hat{\alpha}_n$ exists in the parameter space A for all large n. Then, under Assumptions 2.1-2.5, the estimator $\hat{\alpha}_n$ satisfies the following:

(a) **(Consistency)** $\hat{\alpha}_n \to \alpha_0$ in probability as $n \to \infty$.

(b) **(Asymptotic normality)** Suppose, $h_\alpha(\theta) = \nabla(f_{J,\alpha}(\theta))^{-1}$. Then, the Whittle estimator $\hat{\alpha}_n$ satisfies the following ($L(.,.)$ is the Levy distance):

$$L\left(\frac{\sqrt{n}(\hat{\alpha}_n - \alpha_0)}{f_{J,\alpha_0}(0)}, N_p\left(0, 2\Gamma(\alpha_0)\Sigma_n^2\Gamma(\alpha_0)\right)\right) \to 0 \qquad \text{as } \; n \to \infty \tag{2.15}$$

where

$$\begin{aligned} n\Sigma_n^2 &= \sum_{j,k\in\Gamma_n} \left(\int_D \exp\{\imath(k-j)'\theta\} h_{\alpha_0}(\theta) d\theta\right)\left(\int_D \exp\{\imath(k-j)'\theta\} h_{\alpha_0}(\theta) d\theta\right)', \\ \Gamma(\alpha) &= \left(\int_D \nabla f_{J,\alpha}(\theta)\nabla(f_{J,\alpha}(\theta))^{-1}\, d\theta\right)^{-1}. \end{aligned}$$

Remark 1: If for any fixed k, $m_k/n \to 1$, then $f_{J,\alpha}(\theta)$ converges to $f_\alpha(\theta)$. Hence, in the regular setup, as $n \to \infty$,

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) \Rightarrow N_p\left(0, 2f_\alpha^2(0)\Gamma(\alpha_0)\Sigma_n^2\Gamma(\alpha_0)\right).$$

From a practical point of view, it is not possible to use the above theorem directly to find a confidence set for the true value of the parameter α_0, for the covariance matrix in the above theorem depends on α_0. Next, we describe a subsampling procedure to find a confidence set of α. However, for irregularly spaced stationary random field, a subsampling method will not work in most cases. Here, we describe a method only for a regular random field.

For the following discussion, let us assume that we have data $(X_i)_{i\in I}$ from a random field indexed by a rectangle I in $\mathbb{Z}^d$. For convenience, since we are going to consider regular spaced data, let us assume that $I = \{1, 2, \ldots, l\}^d$. And suppose the Whittle likelihood estimator for this data is $\hat{\alpha}_l$.

To use the subsampling procedure, we will be considering smaller blocks from I. For a point $t = (t_1, \ldots, t_d)$, $\hat{\alpha}_{l,t,b}$ will denote the Whittle likelihood estimator based on the data $(X_i)_{i\in I_{t,b}}$ where $I_{t,b} = \prod_{i=1}^d \{t_i, \ldots, t_i + b\}$. Naturally, these estimates can be obtained for all t such that $I_{t,b} \subset I$. Let us use $Q_{l,b}$ to denote all such t's.

We are going to show that an adjusted empirical distribution of the Whittle likelihood estimators for all possible blocks is essentially an approximation for the limiting distribution of $\hat{\alpha}_l$, described in the previous theorem. To do that, we will look at indicators of Borel sets. For any Borel set $A \in \mathbb{R}^p$, let us use $F(A)$ to denote the limiting value of $P[c_l(\hat{\alpha}_l - \alpha_0) \in A]$ where c_l is an appropriate scaling constant (see Remark 1). Analogously, $F_b(A)$ denotes the same for the subsamples and so, $F_b(A) \to F(A)$, as $b \to \infty$. We are going to consider the following empirical distribution of the subsampled Whittle likelihood estimators:

$$L_{l,b}(A) = \frac{1}{|Q_{l,b}|} \sum_{t \in Q_{l,b}} \mathbb{I}\{c_b(\hat{\alpha}_{l,t,b} - \hat{\alpha}_l) \in A\}. \tag{2.16}$$

Then, the following theorem proves the consistency of the above approximation and helps us determine a confidence set for α_0.

Theorem 2.3: Assume that $b \to \infty, b/n \to 0$ as $n \to \infty$. Then, $L_{l,b}(A)$ in the above definition goes to $F(A)$ in probability, for each Borel set A whose boundary has mass zero under $F(\cdot)$.

2.5 Non-parametric estimator of spectral density function

In this section, we shall study non-parametric kernel spectral density estimators of f and their large-sample properties such as consistency and asymptotic normality. For a symmetric kernel function $K(\cdot)$ and for bandwidth B_n we define the kernel spectral density estimator

$$f_n(\theta) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} X_{L_j} X_{L_k} K\left(\frac{L_j - L_k}{B_n}\right) e^{\imath (L_j - L_k)'\theta}. \tag{2.17}$$

For the consistency of the above estimator, we choose the kernel K to satisfy the following

Condition 1: The kernel function K is symmetric, has support $[-1,1]$, $K(0)=1$ and $\sup_{|x|<1}|K'(x)|<\infty$.

The above condition readily implies that $\kappa=\int_{-\infty}^{\infty}K^2(x)dx<\infty$, a quantity that will be needed later. A simple choice that satisfies the above properties is the rectangular kernel $K(x)=\mathbb{I}_{\{|x|\le 1\}}$. The following theorem asserts the consistency result of the nonparametric estimator above.

Theorem 2.4: Assume that $\mathbb{E}(X_k)=0,\ X_k\in\mathbb{L}^p,\ p\ge 2$ and $\Theta_{0,p}=\sum_{j=0}^{\infty}\delta_{j,p}<\infty$, the bandwidth $B_n\to\infty$ and $B_n=o(n)$ as $n\to\infty$. Then, under Condition 1,

$$\sup_{\theta\in\mathbb{R}^d}\|f_n(\theta)-\mathbb{E}f_n(\theta)\|_{p/2}\to 0.$$

Corollary 2.1: Assume there exists a constant $c>0$ such that $f_n(0)>c$. Let conditions in Theorem 2.4 be satisfied and let $\bar{X}_n$ be the mean of the sample $\{X_i, i\in\Gamma_n\}$. Then,

$$\frac{\sqrt{n}\bar{X}_n}{\sqrt{f_n(0)}}\Rightarrow N(0,1). \tag{2.18}$$

Proof. Let $v_n=n\mathbb{E}(\bar{X}_n^2)$. By the Lebesgue dominated convergence theorem,

$$v_n-\mathbb{E}\left(f_n(0)\right)=\frac{1}{n}\sum_{k\in\mathbb{Z}^d}m_k\left[1-K(k/B_n)\right]\gamma_k\to 0,$$

where m_k is as defined in Section 2.3.2. By Theorem 2.1, $\sqrt{n}\bar{X}_n/\sqrt{v_n}\Rightarrow N(0,1)$. Hence by Theorem 2.4 and Slutsky's theorem, Corollary 2.1 follows. □

One can apply Corollary 2.1 to construct confidence intervals for the mean μ based on irregularly spaced data $X_{L_1},\ldots,X_{L_n}$. Let $\tilde{f}_n(\theta)$ be defined as $f_n(\theta)$ in Equation (2.17) with X_j therein replaced by $X_j-\bar{X}_n$. Given $0<\alpha<1$, the $(1-\alpha)$th confidence intervals for the mean μ is $\bar{X}_n\pm z_{1-\alpha/2}\sqrt{\tilde{f}_n(0)/n}$.

Next, we discuss the asymptotic distribution of the non-parametric spectral density estimator in Equation (2.17). The proof of the theorems are given in the Appendix.

Theorem 2.5: Let Condition 1 be satisfied and define $\kappa = \int_{-\infty}^{\infty} K^2(x)dx < \infty$. Assume $\mathbb{E}(X_k) = 0$, $\mathbb{E}(X_k^4) < \infty$, $\Theta_{0,4} < \infty$, $B_n \to \infty$ and $B_n = o(n)$ as $n \to \infty$. Then, for any fixed $\theta \in \mathbb{R}^d$,

$$\sqrt{\frac{n}{B_n}} \left(\frac{f_n(\theta) - \mathbb{E}[f_n(\theta)]}{f_n(\theta)} \right) \Rightarrow N(0, \kappa) . \tag{2.19}$$

Remark 2: An interesting observation is that the variance of the estimate $f_n(\theta)$ is asymptotically equal to $\mathbb{E}(f_n(\theta))^2$, multiplied by an appropriate scaling term. Therefore, the concepts of variance stabilizing transformation and delta method tell us that we can take the logarithm of the estimate to obtain

$$\sqrt{\frac{n}{B_n}} \left(\log f_n(\theta) - \log \mathbb{E}[f_n(\theta)] \right) \Rightarrow N(0, \kappa) .$$

This result can be used to form a confidence interval for $\mathbb{E}(f_n(\theta))$, which will be of the form $\exp(\log f_n(\theta) \pm z_{1-\alpha/2}\sqrt{\kappa B_n/n})$.

2.6 Estimation of Covariance Matrices

In spatial statistics, a fundamentally important problem is to estimate the covariance matrix of the data. It is useful in many aspects of multivariate analysis including principal component analysis, linear discriminant analysis and graphical modeling. One can infer dependence structures among variables by estimating the associated covariance matrices. In this section, we will discuss the estimation of the covariance matrix of an irregular spaced data $(X_{L_1}, \ldots, X_{L_n})$. Now, to judge the quality of a matrix estimate, we will use the operator norm. For an estimate of the covariance matrix, we are going to use the l_2 “operator norm” and give an upper bound for this. Recall that l_2 norm or spectral radius of a matrix

A is defined as $\rho(A) = \max_{|x|=1} |Ax|$.

Let $\Sigma_n = (\gamma_{L_i - L_j})_{1 \le i,j \le n}$ be the covariance matrix to be estimated. Its estimator $\hat{\Sigma}_n$ is defined by

$$\hat{\Sigma}_n = (\hat{\gamma}_{L_i - L_j})_{1 \le i,j \le n} \qquad \text{where } \hat{\gamma}_k = \frac{1}{m_k} \sum_{i,j: L_i - L_j = k} (X_{L_i} - \bar{X})(X_{L_j} - \bar{X}).$$

In the above, $m_k = \#\{(i,j) : L_i - L_j = k\}$. Based on the known inconsistency results for the periodogram estimate, it can be shown that $\hat{\Sigma}_n$ is not a consistent estimate (Wu and Pourahmadi [100]). So, instead of this, we will use non-parametric kernel-based estimators, similar to what was used for the spectral density. More precisely, we define the following estimate for the covariance matrix:

$$\hat{\Sigma}_{n,B_n} = \left(\hat{\gamma}_{L_i - L_j} K \left(\frac{L_i - L_j}{B_n} \right) \right)_{1 \le i,j \le n}, \tag{2.20}$$

where K is a kernel function and B_n is a bandwidth sequence satisfying appropriate conditions, as discussed below. For this banded covariance matrix estimate, we prove the following

Theorem 2.6: Suppose $\{X_t\}_{t \in \mathbb{Z}^d}$ is a stationary process and each $X_i \in \mathbb{L}^p$ for some $p \in (2,4]$. If $m_k \asymp n$, the kernel K satisfies Condition 1, the bandwidth B_n satisfies the conditions $B_n \to \infty$ and $B_n/n^\delta \to 0$ for some $\delta > (1 - 2/p)/d$, then the estimator $\hat{\Sigma}_{n,B_n}$ in Equation (2.20) is consistent and the spectral radius $\rho(\hat{\Sigma}_{n,B_n} - \mathbb{E}(\hat{\Sigma}_{n,B_n}))$ has a convergence rate of $O_P(B_n^d n^{2/p-1})$.

Remark 3: Note that the above covariance matrix estimate is not necessarily non-negative definite. If we define a matrix $K_{n,B_n} = (K((L_i - L_j)/B_n))_{1 \le i,j \le n}$, then the covariance matrix estimate can be written as $\hat{\Sigma}_{n,B_n} = \hat{\Sigma}_n \star K_{n,B_n}$; where $\star$ is the Hadamard or Schur product, which is formed by the element-wise multiplication of matrices.

By Schur Product theorem, since $\hat{\Sigma}_n$ is already non-negative definite, the Schur product will be non-negative whenever K_{n,B_n} is non-negative definite. One particular example is the triangular kernel $K(u) = \max(0, 1 - |u|)$ which would lead to a positive definite weight matrix K_{n,B_n}. Thus, using this kernel function will give us a non-negative definite covariance matrix estimate $\hat{\Sigma}_{n,B_n}$.

2.7 A Simulation Study

In this section, we assess empirically the impact of the sampling index set Γ_n on the parameter estimation procedure. For that, we consider an isotropic spatial auto-regressive (AR) model in a two-dimensional setting. This model is similar to the isotropic spatial AR model discussed by Azomahou [4] and Lavancier [60]. More precisely, in our spatial AR model, each observation is dependent on the four neighbors in the following way:

$$X_{i,j} = c(X_{i-1,j} + X_{i+1,j} + X_{i,j+1} + X_{i,j-1}) + \varepsilon_{i,j} + \varepsilon_{i,j-1}\varepsilon_{i,j+1}, \quad \text{for} \quad i, j \in \mathbb{Z}, \tag{2.21}$$

where the parameter c gauges the strength of dependence of an observation on its four neighbors, and the innovations $\varepsilon_{i,j}$ are generated independently from a standard normal distribution. Note that this model is an example of non-linear random fields. For more detailed information on such models; see Cressie [21, Chapter 6].

We start with a full grid of size n-by-n from which we would like to generate sample data of different sizes. In order to generate the full data, we consider it in a vector form $\mathbf{X}$, such that the above model helps us write it as $\mathbf{X} = A\mathbf{X} + \mathbf{e}$, and thereby, $\mathbf{X} = (I - A)^{-1}\mathbf{e}$. Here, A an $n^2 \times n^2$ sparse matrix (non-zero elements are the parameter of the model) and $\mathbf{e}$ is a vector such that each component is of the form $\varepsilon_{i,j} + \varepsilon_{i,j-1}\varepsilon_{i,j+1}$.

In our simulation study, we generate a data-set on a grid of 75×75. Then, we take different samples of varying sizes to estimate the parameter c using the Whittle likelihood approach and calculate the mean squared error (MSE) of the estimates. For each of the sample sizes, we also choose a regular grid and compare the performances of the regular setup to that of the irregular setup. The parameters and the set-up of the simulation are described below:

- We used the parameter $c = 0.2$.

- The innovations $\varepsilon_{i,j}$ are generated independently from a $N(0, 1)$ distribution.

- The grid consists of 75×75 data points and we take samples of different sizes (between 10^2 and 40^2). First, these samples are chosen randomly from the whole grid, to ensure that we have an irregularly spaced data. And then, we choose a regular grid of similar size to compare the performances.

- For each sample size, the experiment is repeated multiple times, and the mean squared error for the parameter estimate (error calculated from the actual value $c = 0.2$) is calculate for the regular and irregular data. For the irregular set-up, the Whittle likelihood is defined using the location adjusted spectral density, as shown in Equation (2.14). In addition to that, we also perform same analysis for Whittle likelihood defined in the original way; cf. Equation (2.13). The results are shown in Table 2.1.

Table 2.1: Mean squared error in estimating parameter c using simulated data.

Sample Size	Regular	Irregular (location adjusted)	Irregular (original)
100	0.0004	0.0015	0.0239
225	0.0004	0.0012	0.0158
400	0.0001	0.0009	0.0148
625	0.0004	0.0005	0.0117
900	0.0009	0.0003	0.0109
1225	0.0004	0.0002	0.0096
1600	0.0001	0.0001	0.0081

The results of the simulation study confirm that the estimation performance in the irregular set-up is comparable to the regular set-up when the sample size is larger than 20^2. However, the regular set-up outperforms the irregular one for smaller sample sizes. On the other hand, when we perform the simulation for the original definition of the Whittle likelihood, the results are not at par with the location adjusted version. In fact, the mean squared error is very big for smaller sample sizes, but it reduces steadily as we take bigger samples. For sample size 40^2, it is about 0.0081, which clearly establishes that as more and more samples are taken, the location adjusted spectral density approaches the true value of the spectral density. This is expected, and follows what was discussed in Section 2.4.

2.8 Appendix

2.8.1 Proofs of Theorems

Proof of Theorem 2.1. It is worth mention that this proof is somewhat similar to theorem 1 in El Machkouri et al. [26]. However, for completeness of our work, we are giving a detailed proof here.

At first, assume that $\liminf_n v_n^2/n > 0$. Then, there exists a constant $c_0 > 0$ and $n_0 \in \mathbb{N}$ such that $n/v_n^2 \leq c_0$ for any $n \geq n_0$.

Let $(m_n)_{n\geq 1}$ be a sequence of positive integers going to infinity. Denote $\tilde{X}_j = \mathbb{E}(X_j \,|\, \mathcal{F}_{m_n}(j))$ where $\mathcal{F}_{m_n}(j) = (\varepsilon_{j-s}; |s| \leq m_n)$. So, there exists a measurable function h such that $\tilde{X}_j = h(\varepsilon_{j-s}; |s| \leq m_n)$. Similarly, for a coupled process (see Definition 2.1), we can write

$$\tilde{X}_j^* = h(\varepsilon_{j-s}^*; |s| \leq m_n) = \mathbb{E}\left(X_j^* \,|\, \mathcal{F}_{m_n}^*(j)\right),$$

where $\mathcal{F}_{m_n}^*(j) = (\varepsilon_{j-s}^*; |s| \leq m_n)$. Also, for any $j \in \mathbb{Z}^d$, define

$$\delta_{j,p}^{(m_n)} = \left\|(X_j - \tilde{X}_j) - (X_j - \tilde{X}_j)^*\right\|_p.$$

In order to prove the theorem, we will need the following results.

- Using Lemma 2.1, denoting $\Delta_p^{(m_n)} = \sum\limits_{j\in\mathbb{Z}^d} \delta_{j,p}^{(m_n)}$, for any $n \in \mathbb{N}$ and any $p \geq 2$,

$$\left\| \sum_{i\in\Gamma_n} a_i(X_i - \tilde{X}_i) \right\|_p \leq \left(2p \sum_{i\in\Gamma_n} a_i^2 \right)^{1/2} \Delta_p^{(m_n)}. \tag{2.22}$$

- If $\Delta_p < \infty$, then for any fixed $p \geq 0$, $\Delta_p^{(m_n)} \to 0$ as $n \to \infty$. To this end, note that

$$\begin{aligned} \delta_{j,p}^{(m_n)} &\leq \left\| X_j - X_j^* \right\|_p + \left\| \tilde{X}_j - \tilde{X}_j^* \right\|_p \\ &= \delta_{j,p} + \left\| \mathbb{E}(X_j \,|\, \mathcal{F}_{m_n}(j), \mathcal{F}_{m_n}^*(j)) - \mathbb{E}(X_j^* \,|\, \mathcal{F}_{m_n}^*(j), \mathcal{F}_{m_n}(j)) \right\| \\ &\leq 2\delta_{j,p}. \end{aligned}$$

  Since $\lim_{n\to\infty} \delta_{j,p}^{(m_n)} = 0$ and $\sum_{j\in\mathbb{Z}^d} \delta_{j,p} = \Delta_p < \infty$, using the dominated convergence theorem, we can say that $\lim_{n\to\infty} \Delta_p^{(m_n)} = 0$.

- Define $\tilde{W}_n = \sum_{j\in\Gamma_n} c_j \tilde{X}_j$. Then, using the last two results, we can write

$$\begin{aligned} \frac{\left\| W_n - \tilde{W}_n \right\|_2}{v_n} &= \frac{1}{v_n} \left\| \sum_{i\in\Gamma_n} c_i(X_i - \tilde{X}_i) \right\|_2 \\ &\leq \frac{2\Delta_2^{(m_n)}}{v_n} \left(\sum_{i\in\Gamma_n} c_i^2 \right)^{1/2} \leq \left(2\Delta_2^{(m_n)} \right) \left(\frac{n}{v_n^2} \right)^{1/2}. \end{aligned}$$

  And hence, using the assumption mentioned earlier,

$$\limsup_{n\to\infty} \frac{\left\| W_n - \tilde{W}_n \right\|_2}{v_n} = 0. \tag{2.23}$$

- We will also use the following central limit theorem, due to Heinrich [44].

  Theorem 2.7: Let $(\Gamma_n)_{n\geq 1}$ be a sequence of finite subsets of Z^d with $|\Gamma_n| \to \infty$ as $n \to \infty$ and $(m_n)_{n\geq 1}$ be a sequence of positive integers. For each $n \geq 1$, let $\{U_n(j), j \in \mathbb{Z}^d\}$ be an m_n-dependent random field with $\mathbb{E}(U_n(j)) = 0$ for all $j \in \mathbb{Z}^d$. Also, assume that $\mathbb{E}(\sum_{j\in\mathbb{Z}^d} U_n(j))^2 \to \sigma^2$ as $n \to \infty$ where σ^2 is finite. Then, $\sum_{j\in\mathbb{Z}^d} U_n(j)$ converges in distribution to $N(0, \sigma^2)$ if there exists a finite constant $c > 0$ such that for any $n \geq 1$, $\sum_{j\in\mathbb{Z}^d} \mathbb{E}(U_n^2(j)) \leq c$ and for any $\epsilon > 0$ it holds that

$$\lim_{n\to\infty} L_n(\epsilon) := m_n^{2d} \sum_{j\in\mathbb{Z}^d} \mathbb{E}\left(U_n^2(j)\mathbb{I}\{|U_n(j)| \geq \epsilon m_n^{-2d}\}\right) = 0. \tag{2.24}$$

We now apply the above results to prove our theorem. At first, define $U_n(j) := c_j\tilde{X}_j/v_n$, where c_j is same as the coefficient of X_j in W_n. Note that this definition satisfies the criteria that $\{U_n(j), j \in \mathbb{Z}^d\}$ is an m_n-dependent random field with $\mathbb{E}(U_n(j)) = 0$ for all $j \in \mathbb{Z}^d$. Further,

$$\mathbb{E}\left(\sum_{j\in\mathbb{Z}^d} U_n(j)\right)^2 = \frac{1}{v_n^2}\mathbb{E}\left(\sum_{j\in\mathbb{Z}^d} c_j\tilde{X}_j\right)^2 = \frac{\mathbb{E}\left(\tilde{W}_n^2 - v_n^2\right)}{v_n^2} + 1.$$

Now,

$$\begin{aligned}
\left|\mathbb{E}\left(\tilde{W}_n^2 - v_n^2\right)\right| &= \left|\mathbb{E}\left(\tilde{W}_n^2 - W_n^2\right)\right| \\
&\leq \left\|\tilde{W}_n + W_n\right\|_2 \left\|\tilde{W}_n - W_n\right\|_2 \\
&\leq \left\|\tilde{W}_n - W_n\right\|_2 \left(\left\|\tilde{W}_n - W_n\right\|_2 + 2\|W_n\|_2\right) \\
&\leq 2\Delta_2^{(m_n)}\left(\sum_{i\in\Gamma_n} c_i^2\right)^{1/2}\left(2\Delta_2^{(m_n)}\left(\sum_{i\in\Gamma_n} c_i^2\right)^{1/2} + 4\Delta_2\left(\sum_{i\in\Gamma_n} c_i^2\right)^{1/2}\right) \\
&= 4\Delta_2^{(m_n)}\left(\sum_{i\in\Gamma_n} c_i^2\right)\left(\Delta_2^{(m_n)} + 2\Delta_2\right)
\end{aligned}$$

Thus, using our assumptions, $|\mathbb{E}(\tilde{W}_n^2 - v_n^2)/v_n^2| \leq 4\Delta_2^{(m_n)}\left(\Delta_2^{(m_n)} + 2\Delta_2\right)(n/v_n^2) \to 0$. And hence, as $n \to \infty$, $\mathbb{E}\left(\sum_{j\in\mathbb{Z}^d} U_n(j)\right)^2 \to 1$. On the other hand, for any $n \geq n_0$,

$$\sum_{j\in\mathbb{Z}^d} \mathbb{E}(U_n^2(j)) = \frac{1}{v_n^2}\sum_{j\in\mathbb{Z}^d} c_j^2\mathbb{E}(\tilde{X}_j^2) \leq \frac{n}{v_n^2}\mathbb{E}(\tilde{X}_0^2) \leq c_0\mathbb{E}(\tilde{X}_0^2).$$

Finally, let $\epsilon > 0$ be fixed. Since $|c_j| \leq 1$, we have

$$\mathbb{I}\left\{|U_n(j)| \geq \epsilon m_n^{-2d}\right\} = \mathbb{I}\left\{|\tilde{X}_j| \geq \frac{\epsilon v_n}{|c_j|}m_n^{-2d}\right\} \leq \mathbb{I}\left\{|\tilde{X}_j| \geq \frac{\epsilon v_n}{m_n^{2d}}\right\},$$

and then, we get

$$\begin{aligned} L_n(\epsilon) &\leq \frac{m_n^{2d}}{v_n^2}\sum_{j\in\mathbb{Z}^d}\mathbb{E}\left(\tilde{X}_j^2\mathbb{I}\left\{|\tilde{X}_j| \geq \frac{\epsilon v_n}{m_n^{2d}}\right\}\right) \\ &\leq c_0 m_n^{2d}\mathbb{E}\left(\tilde{X}_0^2\mathbb{I}\left\{|\tilde{X}_0| \geq \frac{\epsilon v_n}{m_n^{2d}}\right\}\right) \\ &\leq c_0 m_n^{2d} \times \left[v_n P\left(|\tilde{X}_0| \geq \frac{\epsilon v_n}{m_n^{2d}}\right) + \mathbb{E}\left(X_0^2\mathbb{I}\left\{|X_0| \geq \sqrt{v_n}\right\}\right)\right] \\ &\leq \frac{c_0 m_n^{6d}\mathbb{E}(X_0^2)}{\varepsilon^2 v_n} + c_0 m_n^{2d}\psi(\sqrt{v_n}), \quad \text{where } \psi(x) = \mathbb{E}\left(X_0^2\mathbb{I}\{|X_0| \geq x\}\right). \end{aligned}$$

In order to ensure that $L_n(\epsilon) \to 0$, define the sequence $(m_n)_{n\geq 1}$ by

$$m_n = \begin{cases} \min\left\{\left[\psi\left(\sqrt{v_n}\right)^{-\frac{1}{4d}}\right], \left[v_n^{\frac{1}{12d}}\right]\right\} & \text{if } \psi\left(\sqrt{v_n}\right) \neq 0 \\ \left[v_n^{\frac{1}{12d}}\right] & \text{if } \psi\left(\sqrt{v_n}\right) = 0 \end{cases}$$

where [.] is the greatest integer function. Since $v_n \to \infty$ and $\psi\left(\sqrt{v_n}\right) \to 0$, it is easy to observe that

$$m_n \to \infty, \quad \frac{m_n^{6d}}{v_n} \leq \frac{1}{\sqrt{v_n}} \to 0 \qquad \text{and} \qquad m_n^{2d}\psi(\sqrt{v_n}) \leq \sqrt{\psi(\sqrt{v_n})} \to 0.$$

Hence, applying Theorem 2.7 and using Equation (2.23), we derive that

$$\frac{\tilde{W}_n}{v_n} \xrightarrow[n\to\infty]{L} N(0,1), \qquad \text{which implies that} \qquad \frac{W_n}{v_n} \xrightarrow[n\to\infty]{L} N(0,1).$$

So, we have proved the required result (2.8) assuming that $\liminf_n v_n^2/n > 0$. If this condition fails, we can get a subsequence $n' \to \infty$ such that

$$L\left[W_{n'}/\sqrt{|\Gamma_{n'}|}, N\left(0, v_{n'}^2/|\Gamma_{n'}|\right)\right] \to l, \text{ as } n' \to \infty, \tag{2.25}$$

for some $l \in [0,\infty]$. Furthermore, if $v_{n'}^2/|\Gamma_{n'}|$ does not converge to 0, we can get a further subsequence n'' such that $\liminf_{n''}(v_{n''}^2/|\Gamma_{n''}|) > 0$, implying

$$L\left[W_{n''}/\sqrt{|\Gamma_{n''}|}, N\left(0, v_{n''}^2/|\Gamma_{n''}|\right)\right] \to 0, \text{ as } n'' \to \infty. \tag{2.26}$$

The above can be shown using the method we adopted previously and this contradicts Equation (2.25). Consequently, we can say that $v_{n'}^2/|\Gamma_{n'}|$ converges to 0 and thus, $W_{n'}/\sqrt{|\Gamma_{n'}|}$ converges to 0 in probability, implying that the Levy distance between $W_{n'}/\sqrt{|\Gamma_{n'}|}$ and $N(0, v_{n'}^2/|\Gamma_{n'}|)]$ goes to 0, again contradicting Equation (2.25). So, finally, proof of the first part of the Theorem 2.1 is complete. The second part is just a corollary that follows easily from the first part. $\square$

Proof of Proposition 1. Observe that $\mathbb{E}[S_n(\theta)S_n(\phi)] = \sum_{k\in\mathbb{Z}^d} \gamma_k a_{n,k}$, where the coefficient $a_{n,k}$ is equal to $n^{-1}\sum_{j,l\in\Gamma_n:j-l=k} e^{-\imath(j'\theta+l'\phi)}$. Clearly $|a_{n,k}| \leq 1$. Under the condition $\theta \neq \phi$, $\theta, \phi \in (-\pi,\pi]^d$, $\theta+\phi \neq 0$, we have for any fixed k, $\lim_{n\to\infty} a_{n,k} = 0$. By Equation (2.7), we have $\sum_{k\in\mathbb{Z}^d} |\gamma_k| < \infty$, and by Lebesgue dominated convergence theorem, $\mathbb{E}[S_n(\theta)S_n(\phi)] \to 0$ as $n \to \infty$. Similarly, $\mathbb{E}[S_n(\theta)S_n(-\phi)] \to 0$. Note that the conjugate of $S_n(\theta)$ is $S_n(-\theta)$. Hence the covariance matrix of $(Y_n(\theta), Z_n(\theta), Y_n(\phi), Z_n(\phi))/\sqrt{n}$ is asymptotically diagonal. That the diagonal elements are equal to $f(\cdot)/2$ can be obtained by noting that $\mathbb{E}[|S_n(\theta)|^2] \to f(\theta)$ and $\mathbb{E}[Y_n(\theta)^2 - Z_n(\theta)^2] \to 0$. The proposition then follows from Theorem 2.1. □

Proof of Theorem 2.2, part (a). Define

$$p(\alpha) := \int_D \left[\log\{f_{J,\alpha}(\theta)\} + \frac{f_{J,\alpha_0}(\theta)}{f_{J,\alpha}(\theta)} \right] \, d\theta.$$

So, for any $\alpha \neq \alpha_0$, we can get

$$p(\alpha) - p(\alpha_0) = \int_D \left[\log \frac{f_{J,\alpha}(\theta)}{f_{J,\alpha_0}(\theta)} + \frac{f_{J,\alpha_0}(\theta)}{f_{J,\alpha}(\theta)} - 1 \right] \, d\theta > 0, \tag{2.27}$$

because the integrand is 0 for $\alpha = \alpha_0$ and otherwise always positive. Observe that $p_n(\alpha) \to p(\alpha)$ in probability, in view of Lemma 2.4 and the assumptions mentioned in Section 2.4. Thus, there exists a positive constant C_α such that

$$\lim_{n\to\infty} P\left[p_n(\alpha_0) - p_n(\alpha) < -C_\alpha\right] = 1. \tag{2.28}$$

Now, consider any two α_1, α_2 such that $\|\alpha_1 - \alpha_2\| < \delta$ for some small positive constant δ, fixed a priori. Then, using the continuity of the functions, we can get that

$$\begin{aligned} |p_n(\alpha_1) - p_n(\alpha_2)| \quad &\leq \quad \left| \int_D \log \frac{f_{J,\alpha_1}(\theta)}{f_{J,\alpha_2}(\theta)} \, d\theta \right| + \int_D I_n(\theta) \left| \frac{1}{f_{J,\alpha_1}(\theta)} - \frac{1}{f_{J,\alpha_2}(\theta)} \right| d\theta \\ &\leq \quad \delta \left(K_1 + K_2 \int_D I_n(\theta) d\theta \right), \end{aligned} \tag{2.29}$$

where K_1, K_2 are constants. Let us denote the above bound by $K_n(\delta)$ and observe that one can always find δ such that

$$\lim_{n\to\infty} P\left[K_n(\delta) < C_\alpha\right] = 1. \tag{2.30}$$

Now, for any particular α_1, consider all points α_2 such that $\|\alpha_1 - \alpha_2\| < \delta$. Then, using Equation (2.29) and Equation (2.30), we can say that for large n, $p_n(\alpha_1) - p_n(\alpha_2) \leq C_\alpha$ with probability going to 1. Combining this with Equation (2.28) for $\alpha = \alpha_1$, we get that for any α_1

$$\lim_{n\to\infty} P\left[\sup_{\alpha_2:\|\alpha_1-\alpha_2\|<\delta} \{p_n(\alpha_0) - p_n(\alpha_2)\} < 0\right] = 1. \tag{2.31}$$

For an α_1, let us denote the above sets of the form $\{\alpha_2 : \|\alpha_1 - \alpha_2\| < \delta\}$ by $S(\alpha_1)$. To remain consistent with out notation, for α_0, let us consider $S(\alpha_0) = \{\alpha_2 : \|\alpha_0 - \alpha_2\| < \gamma\}$. Observe that the above result is true for any possible value of γ. Now, for the whole parameter space A, if we consider the collection of subsets $\{S(\alpha) : \alpha \in A\}$, this forms an open cover of the whole parameter space and since A is compact, it will have a finite cover. Let us denote this as $\{S(\alpha_i) : i = 0, 1, 2, \ldots, m \text{ and } \alpha_i \in A \,\forall\, i\}$. Note that from Equation (2.31),

$$\lim_{n\to\infty} P\left[\sup_{\alpha\in\cup_{i=0}^m S(\alpha_i)} \{p_n(\alpha_0) - p_n(\alpha)\} < 0\right] = 1. \tag{2.32}$$

Now, $A = \cup_{i=0}^m S(\alpha_i)$ and from the above, we get that as $n \to \infty$, $p_n(\alpha_0) < \inf_A p_n(\alpha)$ with probability going to 1. Thus,

$$\lim_{n\to\infty} P\left[\inf_{\alpha\in S(\alpha_0)} p_n(\alpha) = \inf_{\alpha\in A} p_n(\alpha)\right] = 1. \tag{2.33}$$

Therefore, clearly, as $n \to \infty$, the minimizer will always be in $S(\alpha_0)$ and since we can fix γ as small as possible, we can say that $\lim_{n\to\infty} P\left[|\hat{\alpha}_n - \alpha_0| < \gamma\right] = 1$ and so, $\hat{\alpha}_n \xrightarrow{P} \alpha_0$. Hence, the Whittle likelihood estimate is consistent. □

Proof of Theorem 2.2, part (b). Since $\hat{\alpha}_n$ is the minimizer for the Whittle likelihood, we will start with the Taylor series expansion and we get

$$\hat{\alpha}_n - \alpha_0 = -\left(\frac{\partial^2 p_n(\alpha^*)}{\partial\alpha\,\partial\alpha'}\right)^{-1}\frac{\partial p_n(\alpha_0)}{\partial\alpha}. \tag{2.34}$$

Now, we would consider the two terms on the right hand side separately. For the second term, using the assumptions, we have,

$$\begin{aligned}\frac{\partial p_n(\alpha_0)}{\partial\alpha} &= \int_D \left[\frac{1}{f_{J,\alpha_0}(\theta)} - \frac{I_n(\theta)}{\{f_{J,\alpha_0}(\theta)\}^2}\right]\frac{\partial}{\partial\alpha}\{f_{J,\alpha_0}(\theta)\}\;d\theta \\ &= \int_D \{I_n(\theta) - E(I_n(\theta))\}\frac{\partial}{\partial\alpha}\{f_{J,\alpha_0}(\theta)\}^{-1}\;d\theta.\end{aligned}$$

Let us use ∇g and $\nabla^2 g$ to denote the first order derivative and the Hessian of a function g. Also, let $h_{\alpha_0}(\theta) = \nabla\{f_{J,\alpha_0}(\theta)\}^{-1}$. Note that this is non-stochastic, but depends on n and the true value of α. In order to find the asymptotic distribution of the above term, we will make use of Lemma 2.3. Observe that $I_n(\theta)$ can be written as

$$nI_n(\theta) = n\,|S_n(\theta)|^2 = \sum_{j,k\in\Gamma_n} X_j X_k \exp\{\imath(k-j)'\theta\}.$$

Now, if we consider the stochastic term in the integral above, we can write it as

$$\begin{aligned}nP_n = n\int_D I_n(\theta)h_{\alpha_0}(\theta)\;d\theta &= \sum_{j,k\in\Gamma_n} X_j X_k \int_D \exp\{\imath(k-j)'\theta\}h_{\alpha_0}(\theta)\;d\theta \\ &= \sum_{j,k\in\Gamma_n} X_j X_k \beta_{n,j-k}.\end{aligned}$$

Let us consider the asymptotic distribution of $c'P_n$ for some real-valued vector $c \in \mathbb{R}^p$. Note that the coefficients $c'\beta_{n,j-k}$ do not depend on θ. Hence, comparing the above expression with that of Lemma 2.3, we can set $b_{n,j} = a_{n,j} = c'\beta_{n,j}$ and fix $\theta = 0$ in that theorem.

Then,

$$\begin{aligned}\sigma_n^2 &= 2\sum_{j,k\in\Gamma_n}\left(\int_D \exp\{\imath(k-j)'\theta\}c'h_{\alpha_0}(\theta)\ d\theta\right)^2\\ &= 2c'\left[\sum_{j,k\in\Gamma_n}\left(\int_D \exp\{\imath(k-j)'\theta\}h_{\alpha_0}(\theta)d\theta\right)\left(\int_D \exp\{\imath(k-j)'\theta\}h_{\alpha_0}(\theta)d\theta\right)'\right]c.\end{aligned}$$

Based on the assumptions we have, we can say that the integrals are finite and on the other hand, it depends only on α_0 (through $h(.)$). Let us now denote the matrix in the middle of the above expression by $n\Sigma_n^2$. Using the assumptions in Section 2.4, one can now easily check that Equation (2.48) and other conditions, required to apply the lemma, are satisfied. So, using it, we get that

$$L\left(nc'P_n(\theta)-nE[c'P_n(\theta)], N\left(0, f_{J,\alpha_0}^2(0)\cdot 2nc'\Sigma_n^2 c\right)\right)\xrightarrow{n\to\infty} 0, \tag{2.35}$$

where $L(\cdot,\cdot)$ denotes the Levy distance. Since the above is true for all c, we can say that $L\left(nP_n(\theta)-n\mathbb{E}[P_n(\theta)], N(0, f_{J,\alpha_0}^2(0)\cdot 2n\Sigma_n^2)\right)\to 0$. Combining this with the expression obtained for $\partial p_n(\alpha_0)/\partial\alpha$, we obtain

$$L\left(\sqrt{n}\left(\frac{\partial p_n(\alpha_0)}{\partial\alpha}\right), N\left(0, 2f_{J,\alpha_0}^2(0)\Sigma_n^2\right)\right)\xrightarrow{n\to\infty} 0. \tag{2.36}$$

Now, for the first term, we need to take the second order derivative of the integrand in the expression of Whittle likelihood. Then,

$$\frac{\partial^2 p_n(\alpha^*)}{\partial\alpha\ \partial\alpha'} = \int_D\{I_n(\theta)-f_{J,\alpha^*}(\theta)\}\nabla^2\{f_{J,\alpha^*}(\theta)\}^{-1}\ d\theta-\Gamma(\alpha^*)^{-1}.$$

We already have proved that $\hat{\alpha}_n$ is consistent for α_0. Using this, along with the assumptions mentioned in Section 2.4 and Lemma 2.4, the first term above goes to 0 as $n\to\infty$ and hence,

$$-\frac{\partial^2 p_n(\alpha^*)}{\partial\alpha\;\partial\alpha'} \to \int_D \nabla f_{J,\alpha_0}(\theta)\nabla\{f_{J,\alpha_0}(\theta)\}^{-1}\;d\theta = \Gamma(\alpha_0)^{-1}. \tag{2.37}$$

Finally, the central limit theorem follows from Equation (2.34), Equation (2.36) and Equation (2.37). □

Proof of Theorem 2.3. We are going to set some notation at first.

For a Borel set A, let us use $\delta(A)$ to denote its boundary. Now, for a positive constant ϵ, set $M_{A,\epsilon} = \cup_{x\in\delta(A)} B(x,\epsilon)$, where $B(x,\epsilon)$ denotes the closed ball with center x and radius ϵ. Let $A_{+\epsilon} = A \cup M_{A,\epsilon}$, $A_{-\epsilon} = A \cap M^c_{A,\epsilon}$. Thus, if A is a ball with center z and radius r, then $A_{+\epsilon}$ denotes the closed ball with center z and radius $r+\epsilon$ while $A_{-\epsilon}$ denotes the open ball with center z and radius $r-\epsilon$.

Let us also define the following empirical distribution for $\hat\alpha_{l,t,b}$:

$$L^0_{l,b}(A) = \frac{1}{|Q_{l,b}|}\sum_{t\in Q_{l,b}} \mathbb{I}\{c_b(\hat\alpha_{l,t,b}-\alpha_0)\in A\}. \tag{2.38}$$

And suppose $E_{l,b,\epsilon}$ denotes the event $\{\|c_b(\hat\alpha_l-\alpha_0)\|\le\epsilon\}$. Because of the assumptions on b, we can easily say that $P(E_{l,b,\epsilon})\to 1$ as $n\to\infty$, for any $\epsilon>0$. Further note that

$$\begin{aligned}\mathbb{I}\{c_b(\hat\alpha_{l,t,b}-\alpha_0)\in A_{-\epsilon}\}\cdot\mathbb{I}\{E_{l,b,\epsilon}\} &\le \mathbb{I}\{c_b(\hat\alpha_{l,t,b}-\hat\alpha_l)\in A\}\cdot\mathbb{I}\{E_{l,b,\epsilon}\}\\ &\le \mathbb{I}\{c_b(\hat\alpha_{l,t,b}-\alpha_0)\in A_{+\epsilon}\},\end{aligned}$$

and hence, with probability tending to one,

$$L^0_{l,b}(A_{-\epsilon}) \le L_{l,b}(A) \le L^0_{l,b}(A_{+\epsilon}).$$

Now, if we can prove that $L^0_{l,b}(A)\to F(A)$ for any Borel set A whose boundaries have measure zero under $F(\cdot)$, then we would get that $F(A_{-\epsilon}) - \epsilon \leq L_{l,b}(A) \leq F(A_{+\epsilon}) + \epsilon$. Letting $\epsilon \to 0$ such that $A_{\pm\epsilon}$ are Borel sets whose boundaries have measure zero under $F(\cdot)$, we can then get that $L_{l,b}(A)$ is a good approximation for $F(A)$.

In order to show that $L^0_{l,b}(A) \to F(A)$, at first note that $\mathbb{E}[L^0_{l,b}(A)] = F_b(A) \to F(A)$ and thus, we only have to show that $\mathrm{Var}(L^0_{l,b}(A)) \to 0$ as $n \to \infty$.

We will be using m-dependence approximation to prove the required result. Here, we deal with the case $d = 2$ for convenience, but the proof would hold for any dimension. Let us write (i_1, i_2) or (j_1, j_2) for the 2-dimensional indexes i, j. Define the sigma field $\mathcal{F}_{i_1-m,i1} = \sigma(\varepsilon_j : j \in \mathbb{R}^d, i_1 - m \leq j_1 \leq i_1)$ and write $\mathcal{F}_{i_1}$ for $\mathcal{F}_{-\infty,i_1}$. The m-dependence approximation, for $m \geq 0$, is then defined by the following:

$$\tilde{X}_i := \mathbb{E}(X_i \,|\, \mathcal{F}_{i_1-m,i_1}). \tag{2.39}$$

Now, let us define a new quantity, as follows:

$$\tilde{L}_{l,b}(A) = \frac{1}{|Q_{l,b}|} \sum_{t \in Q_{l,b}} \mathbb{I}\{c_b(\tilde{\alpha}_{l,t,b} - \alpha_0) \in A\}. \tag{2.40}$$

Here, $\tilde{\alpha}_{l,t,b}$ is the Whittle likelihood estimator based on $\tilde{X}_t$'s. Below, for notational convenience, let us denote $\mathbb{I}\{c_b(\hat{\alpha}_{l,t,b} - \alpha_0) \in A\}$ by $Z_{t,b}$, $|Q_{l,b}|$ by q and $\mathrm{Cov}(Z_{t,b}, Z_{t+k,b})$ by τ_k. $\tilde{Z}_{t,b}$ and $\tilde{\tau}_k$ are defined analogously for $\tilde{X}_t$. Now, let $R_{s,m} = \{t \in Q_{l,b} : |\tau^{-1}(s) - \tau^{-1}(t)| \leq m\}$ and $R^c_{s,m} = Q_{l,b} - R_{s,m}$. Then,

$$\begin{aligned}
\mathrm{Var}(\tilde{L}_{l,b}(A)) &= \frac{1}{q^2} \sum_{s \in Q_{l,b}} \sum_{t \in Q_{l,b}} \tilde{\tau}_{s-t} \\
&= \frac{1}{q^2} \sum_{s \in Q_{l,b}} \sum_{t \in R_{s,m}} \tilde{\tau}_{s-t} + \frac{1}{q^2} \sum_{s \in Q_{l,b}} \sum_{t \in R^c_{s,m}} \tilde{\tau}_{s-t} = A_1 + A_2 \quad \text{(say)}
\end{aligned}$$

It is easy to note that that $A_1 = O(m^{1/d}q^{-1})$ and so, it goes to 0 as $n \to \infty$, for any fixed m. On the other hand, A_2 is exactly equal to 0, since for any s, $\tilde{\tau}_{s-t} = 0$ for all $t \in R^c_{s,m}$. The theorem is then proved in view of the fact that $\mathrm{Var}(L^0_{n,b}(A)) = \mathrm{Var}(\tilde{L}_{n,b}(A)) + o(1)$. □

Proof of Theorem 2.4. This proof will revolve around the notion of m-dependent processes, as defined by Equation (2.39). Observe that, in this theorem, we are dealing with the quantity

$$\begin{aligned} f_n(\theta) &= \frac{1}{n}\sum_{s=1}^{n}\sum_{t=1}^{n} X_{L_s}X_{L_t}K\left(\frac{L_s - L_t}{B_n}\right)e^{\imath(L_s-L_t)'\theta} \\ &= \frac{1}{n}\sum_{s=1}^{n}\sum_{t=1}^{n} a_{n,L_s-L_t}X_{L_s}X_{L_t}, \qquad \text{where } \; a_{n,r} = K(r/B_n)e^{\imath r'\theta}. \end{aligned}$$

Based on the assumptions, it is easy to observe that $\sum_{r\in\mathbb{Z}}|a_{n,r}|^2 = O(B_n)$.

Now, since we intend to approximate using the m-dependent process, we should consider the following term, which is defined in a similar way as above, but with $\tilde{X}_{L_t}$'s. So, define $Y_t = \tilde{X}_{L_t}\sum_{s=1}^{t-k} a_{n,L_s-L_t}\tilde{X}_{L_s}$. Then

$$\begin{aligned} \tilde{f}_n(\theta) &= \frac{1}{n}\sum_{s=1}^{n}\sum_{t=1}^{n} a_{n,L_s-L_t}\tilde{X}_{L_s}\tilde{X}_{L_t} \\ &= \frac{1}{n}\sum_{t=1}^{n}\tilde{X}^2_{L_t} + \frac{2}{n}\sum_{t=2}^{n}\tilde{X}_{L_t}\sum_{s=\max\{1,(t-k)\}}^{t-1} a_{n,L_s-L_t}\tilde{X}_{L_s} + \frac{2}{n}\sum_{t=k+1}^{n} Y_t. \qquad (2.41) \end{aligned}$$

The idea here is to prove the required result in three steps. At first, we consider the last term in the above expression. Now, observe that the sequence $\{Y_{t+rk}\}_{r\geq 0}$ are $\mathbb{L}^p$ martingale differences whenever $|k| > m$. For convenience, let us take $|k| = 2m$. On the other hand, using Lemma 2.2, we get that

$$\begin{aligned}\|Y_t\|_p &= \|\tilde{X}_{L_t}\|_p \left\|\sum_{s=1}^{t-k} a_{n,L_s-L_t}\tilde{X}_{L_s}\right\|_p \\ &\le \|X_0\|_p C_p \Theta_{0,p}\left(\sum_{s=1}^{t-k}|a_{n,L_s-L_t}|^2\right)^{1/2} = O(B_n^{1/2}).\end{aligned}$$

Now, we will write the last term in Equation (2.41) using the sequences of the martingale differences and then it would satisfy the following. (Here, N_j is used to denote the maximum possible index in that sequence and it will be of the order n.)

$$\left\|\frac{2}{n}\sum_{t=k+1}^{n} Y_t\right\|_p \le \frac{2}{n}\sum_{j=1}^{k}\left\|\sum_{i=1}^{N_j} Y_{j+ik}\right\|_p = \frac{2}{n}\sum_{j=1}^{k} n^{1/2}O(B_n^{1/2}) = O[(mB_n/n)^{1/2}].$$

Let $\tilde{\gamma}_k = \mathbb{E}(\tilde{X}_0\tilde{X}_k)$ and denote the last term in Equation (2.41) by W_n/n. We are now going to consider the first two terms together in view of the fact that for any l, $\|n^{-1}\sum_t \tilde{X}_t\tilde{X}_{t+l} - \tilde{\gamma}_l\|_{p/2} \to 0$ as $n \to \infty$. Observe that in the first two terms, we are essentially combining all the terms of the form $\tilde{X}_{L_t}\tilde{X}_{L_s}$ where $|s-t| \le k$. So, we can say that the following holds:

$$\left\|\tilde{f}_n(\theta) - W_n/n - \mathbb{E}[\tilde{f}_n(\theta) - W_n/n]\right\|_{p/2} \to 0.$$

Combining this with the above result that $\|W_n/n\|_p = O[(mB_n/n)^{1/2}]$, based on our assumption that $B_n/n \to 0$ as $n \to \infty$, we can conclude that

$$\|\tilde{f}_n(\theta) - \mathbb{E}[\tilde{f}_n(\theta)]\|_{p/2} \to 0 \qquad \text{as} \;\; n \to \infty. \tag{2.42}$$

Now, define the following:

$$S_n(\theta) = \sum_{j=1}^{n} X_{L_j} e^{\imath L_j' \theta} \qquad \text{and} \qquad \tilde{S}_n(\theta) = \sum_{j=1}^{n} \tilde{X}_{L_j} e^{\imath L_j' \theta}.$$

So, from the assumptions and using Lemma 2.2, we can say that $\|S_n(\theta) - \tilde{S}_n(\theta)\|_p = \Theta_{m,p} O(n^{1/2})$ and $\|S_n(\theta)\|_p + \|\tilde{S}_n(\theta)\|_p = O(n^{1/2})$.

Now, if $\hat{K}$ is the Fourier transform of K, we can write $f_n(\theta)$ with the help of S_n in the form $n f_n(\theta) = \int \hat{K}(u) |S_n(B_n^{-1}u + \theta)|^2 \, du$ and we can use a similar definition for $\tilde{f}_n(\theta)$ using $\tilde{S}_n$. And then, the following can be obtained.

$$\begin{aligned} \|f_n(\theta) - \tilde{f}_n(\theta)\|_{p/2} \quad &\leq \quad \frac{1}{n} \int_{\mathbb{R}} |\hat{K}(u)| \left\| |S_n(B_n^{-1}u + \theta)|^2 - |\tilde{S}_n(B_n^{-1}u + \theta)|^2 \right\|_{p/2} du \\ &\leq \quad \frac{1}{n} \int_{\mathbb{R}} |\hat{K}(u)| O(n) \Theta_{m,p} \, du = O(1) \Theta_{m,p}. \end{aligned} \tag{2.43}$$

Finally, observe that $\|f_n(\theta) - \mathbb{E}[f_n(\theta)]\|_{p/2} \leq \|f_n(\theta) - \tilde{f}_n(\theta)\|_{p/2} + \|\tilde{f}_n(\theta) - \mathbb{E}[\tilde{f}_n(\theta)]\|_{p/2} + |\mathbb{E}[f_n(\theta) - \tilde{f}_n(\theta)]|$.

We already have Equation (2.42) for the second term. Now, as $\Theta_{m,p}$ goes to 0 if we take $m \to \infty$, from Equation (2.43), we can say that both first and last terms in the right-hand-side of the above inequality will go to 0 and that completes the proof. □

Proof of Theorem 2.5. This theorem is a particular case of the general quadratic forms of stationary random fields and we will make use of Lemma 2.3 to prove the theorem. Note that $n f_n(\theta)$ is of the form S_n in the above-mentioned lemma with $b_{n,j} = K(j/B_n)$. Using the given conditions, one can now show that $\sum_j b_{n,j}^2 \sim B_n \kappa$ and $\sum_{i,j} b_{n,i-j}^2 \sim n^d B_n \kappa$ where $\kappa = \int_{-\infty}^{\infty} K^2(x) dx < \infty$. That means the conditions of the above lemma are satisfied and hence, we can say that

$$\sqrt{\frac{n}{B_n}} \cdot \frac{f_n(\theta) - \mathbb{E}(f_n(\theta))}{f_J(\theta)} \Rightarrow N(0, \kappa). \tag{2.44}$$

Then, in view of the fact that $\mathbb{E}(f_n(\theta)) = f_J(\theta) + o(1)$, Theorem 2.4 and a simple application of Slutsky's theorem completes our proof. □

Proof of Theorem 2.6. At first, we will define a new covariance matrix using the actual mean (unknown) of the process. Suppose, each X_i has mean μ and let us define a new banded covariance matrix estimate

$$\hat{\Sigma}^0_{n,B_n} = \left(\hat{\gamma}^0_{L_i - L_j} K\left(\frac{L_i - L_j}{B_n}\right)\right)_{1 \le i,j \le n},$$

where $\hat{\gamma}^0_k = \frac{1}{m_k} \sum_{i,j: L_i - L_j = k} (X_{L_i} - \mu)(X_{L_j} - \mu)$.

This proof will mainly use the Gershgorin circle theorem, which states that every eigenvalue of a complex $n \times n$ matrix A lies within at least one of the Gershgorin discs $D_i(a_{ii}, \sum_{j \neq i} |a_{ij}|)$, where a_{ij}'s are the elements of the matrix A. Thus, if λ_i, for $i = 1, \ldots, n$ denote eigenvalues of A, then the spectral radius satisfies the following:

$$\rho(A) = \max_i |\lambda_i| \le \max_i \left(|a_{ii}| + \sum_{j \neq i} |a_{ij}|\right) = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

Let us now consider the spectral radius of the matrix $\hat{\Sigma}^0_{n,B_n} - \mathbb{E}(\hat{\Sigma}^0_{n,B_n})$. The (i,j)-th element of this matrix is $K((L_i - L_j)/B_n)[\hat{\gamma}^0_{L_i - L_j} - \mathbb{E}(\hat{\gamma}^0_{L_i - L_j})]$. Below, we will use J_n and J_{n,B_n} to denote the following two sets:

$$\begin{aligned} J_n &= \{k: \text{ there exists } 1 \le i,j \le n \text{ satisfying } L_i - L_j = k\}, \\ J_{n,B_n} &= \{k: |k| \le B_n \text{ and there exists } 1 \le i,j \le n \text{ satisfying } L_i - L_j = k\}. \end{aligned}$$

Also, let us denote $J'_{n,B_n} = J_n - J_{n,B_n}$. Then, using the above mentioned property of spectral radius, it can be written that

$$\begin{aligned}
& \rho\left(\hat{\Sigma}^0_{n,B_n} - \mathbb{E}(\hat{\Sigma}^0_{n,B_n})\right) \\
\leq\ & \max_i \sum_{j=1}^{n} \left|\hat{\gamma}^0_{L_i-L_j} K((L_i - L_j)/B_n) - \mathbb{E}[\hat{\gamma}^0_{L_i-L_j} K((L_i - L_j)/B_n)]\right| \\
\leq\ & 2 \sum_{k \in J_n} \left|K\left(\frac{k}{B_n}\right)\right| \left\|\hat{\gamma}^0_k - \mathbb{E}(\hat{\gamma}^0_k)\right\| \\
\leq\ & 2 \sum_{k \in J_{n,B_n}} \left\|\hat{\gamma}^0_k - \mathbb{E}(\hat{\gamma}^0_k)\right\|. \qquad (2.45)
\end{aligned}$$

Note that if we relabel $X_i - \mu$ as Y_i then the above expression essentially deals with the autocovariance function and its estimate of a zero-mean stationary process. Thus, if each $X_i \in \mathbb{L}^p$ for some $p \in (2, 4]$, then based on the results obtained by Wu and Pourahmadi [100] and using the assumption $m_k \asymp n$, we can write that the above bound is of the order $O_P(B_n^d n^{2/p-1} \Theta^2_{0,p})$. If we further assume that the process is p-stable (see Definition 2.2) which tells us that the term $\Theta_{0,p}$ is finite, we can say that the bound is of the order $O_P(B_n^d n^{2/p-1})$.

Now, for the covariance matrix estimate $\hat{\Sigma}_{n,B_n}$, we can write $\rho\left(\hat{\Sigma}_{n,B_n} - \mathbb{E}(\hat{\Sigma}_{n,B_n})\right) \leq \rho\left(\hat{\Sigma}_{n,B_n} - \hat{\Sigma}^0_{n,B_n}\right) + \rho\left(\hat{\Sigma}^0_{n,B_n} - \mathbb{E}(\hat{\Sigma}^0_{n,B_n})\right) + \rho\left(\mathbb{E}(\hat{\Sigma}_{n,B_n}) - \mathbb{E}(\hat{\Sigma}^0_{n,B_n})\right)$.

We have already got the limit of the second term. For the first and third term, we can follow similar procedures as before. In case of first term, using Gershgorin Circle Theorem, we will again get a bound of a sum of terms of the form $\hat{\gamma}_k - \hat{\gamma}^0_k$, over the set J_{n,B_n}. Now, since $\bar{X} - \mu$ is $O_P(n^{-1})$, we can say that this bound is of the order $O_P(B_n^d n^{-1})$. Note that this bound is less than what we got for the second term. We can get similar results for the third term as well. And hence, the overall bound for the spectral radius of $\hat{\Sigma}_{n,B_n} - \mathbb{E}(\hat{\Sigma}_{n,B_n})$ is obtained as $O_P(B_n^d n^{2/p-1})$. □

2.8.2 Proofs of Lemmas

Lemma 2.1 (This lemma is due to El Machkouri et al. [26])**:** Following the aforementioned notation, consider Γ_n and let $(\alpha_i)_{i\in\Gamma_n}$ be a family of real numbers. Then, for any $p \geq 2$, we get

$$\left\|\sum_{i\in\Gamma_n} \alpha_i X_i\right\|_p \leq \left(2p\sum_{i\in\Gamma_n}\alpha_i^2\right)^{1/2} \Delta_p. \tag{2.46}$$

Lemma 2.2: Let $X_i \in \mathbb{L}^p$ for $p > 1$, $\mathbb{E}(X_k) = 0$, $\alpha_1, \alpha_2, \ldots \in \mathbb{C}$, $p' = \min\{2, p\}$, $A_n = (\sum_{k=1}^n |\alpha_{L_k}|^{p'})^{1/p'}$ and $C_p = 18p^{3/2}(p-1)^{-1/2}$. Then, $\left\|\sum_{k=1}^n \alpha_{L_k} X_{L_k}\right\|_p \leq C_p A_n \Theta_{0,p}$, $\left\|\sum_{k=1}^n \alpha_{L_k} \tilde{X}_{L_k}\right\|_p \leq C_p A_n \Theta_{0,p}$ and $\left\|\sum_{k=1}^n \alpha_{L_k} (X_{L_k} - \tilde{X}_{L_k})\right\|_p \leq C_p A_n \Theta_{m+1,p}$.

Proof. Let $\tau : \mathbb{Z} \to \mathbb{Z}^d$ be a bijection. For any $i \in \mathbb{Z}$, for any $j \in \mathbb{Z}^d$, define the projection operator $P_i X_j := \mathbb{E}(X_j \,|\, \mathcal{F}_i) - \mathbb{E}(X_j \,|\, \mathcal{F}_{i-1})$, where $\mathcal{F}_i = \sigma(\varepsilon_{\tau(l)}; l \leq i)$. Also, define $T^j \mathcal{F}_i = \sigma(\varepsilon_{\tau(l)-j}; l \leq i)$. Now, it is easy to note that $\|P_i X_j\| \leq \|X_{j-\tau(i)} - X^*_{j-\tau(i)}\|_p$ and hence, we get the inequality $\|P_i X_j\| \leq \delta_{j-\tau(i),p}$.

Then, noting that $X_i = \sum_{j\in\mathbb{Z}} P_j X_i$ for all $i \in \mathbb{Z}^d$, we can make use of Burkholder inequality and Cauchy-Schwarz inequality to get the first two results. The proof here is similar to Proposition 1 in El Machkouri et al. [26]. And then, the third result is a direct consequence of the first two. □

Lemma 2.3: Let $\beta_j \in \mathbb{R}$ with $\beta_j = \beta_{-j}$; $\alpha_j = \beta_j e^{\imath j'\theta}$ where $\theta \in [-\pi, \pi]^d$. Also, for any θ, define $\bar{\omega}(\theta) = 2$ if $\theta/\pi \in \mathbb{Z}^d$. Else, define it to be 1. Consider the quadratic form

$$S_n = \sum_{1\leq j,k\leq n} \alpha_{L_k - L_j} X_{L_j} X_{L_k} \qquad \text{and} \qquad \sigma_n^2 = \bar{\omega}(\theta) \sum_{1\leq j,k\leq n} \beta^2_{L_j - L_k}, \tag{2.47}$$

where we assume that $\mathbb{E}(X_0) = 0, X_0 \in \mathbb{L}^4, \Theta_{0,4} < \infty$. In addition, let $\zeta_n^2 = \sum_{1\leq t\leq n} \beta^2_{L_t}$, and let us assume that $\max_{1\le t\le n}\beta^2_{L_t}=o(\zeta^2_n), n^d\zeta^2_n=O(\sigma^2_n)$. We further consider that

$$\sum_{k=1}^{n}\sum_{t=1}^{k-1}\left|\sum_{j=1+k}^{n}\alpha_{L_k-L_j}\alpha_{n,L_t-L_j}\right|^2=o(\sigma^4_n),\ \sum_{k=1}^{n}\left|\beta_{L_k}-\beta_{L_k-1}\right|^2=o(\zeta^2_n). \tag{2.48}$$

Now, following Section 2.3.2, suppose J denotes the set $\{k\in\mathbb{Z}^d:\exists\ i,j \text{ with } L_i-L_j=k\}$ and m_k is the cardinality of the set $\{(i,j):L_i-L_j=k\}$. If $f_J(\theta)$ denotes the location adjusted spectral density, then

$$\frac{S_n-\mathbb{E}(S_n)}{\sigma_n f_J(\theta)}\xrightarrow[n\to\infty]{L}N\left(0,1\right). \tag{2.49}$$

Proof. We will make use of m-dependence approximation, defined in Equation (2.39), to prove this result. Once again, for simplicity, we prove the result for $d=2$, but the idea can be easily used for higher dimensions in a similar fashion. To begin with, note that S_n is a special case of $U_n=\sum_{s,t\in\mathbb{Z}^d_n}a_{s,t}X_sX_t$, where $\mathbb{Z}^d_n$ is a regular grid and $a_{s,t}$ are appropriate constants. Further, we have, $\left\|\sum_s X^2_s-n^d\gamma_0\right\|=O(n^{d/2})$. Then, defining $Z_s=\sum_{t\ne s}a_{s,t}X_t$ and $\tilde{Z}_s=\sum_{t\ne s}a_{s,t}\tilde{X}_t$, we write $T_n=\sum_s X_sZ_s$, $\tilde{T}_n=\sum_s\tilde{X}_s\tilde{Z}_s$ and $T^*_n=\sum_s X_s\tilde{Z}_s$.

Using the previous lemma and with similar arguments as in Proposition 1 of Liu and Wu [66], we can show that

$$\frac{\|(T_n-\mathbb{E}(T_n))-(T^*_n-\mathbb{E}(T^*_n))\|_p}{n^{d/2}\zeta_n}\le C_pd_m, \tag{2.50}$$

where $d_m\to 0$ as $m\to\infty$. We can get a similar inequality for $\left\|(\tilde{T}_n-\mathbb{E}(\tilde{T}_n))-(T^*_n-\mathbb{E}(T^*_n)\right\|^2_p$ and then, using both inequalities, we can write that $\|(T_n-\mathbb{E}(T_n))-(\tilde{T}_n-\mathbb{E}(\tilde{T}_n))\|=o(\sigma_n)$. Thus, in order to get the required result, it is enough to find the asymptotic distribution of $(\tilde{T}_n-\mathbb{E}(\tilde{T}_n))/\sigma_n$.

We write $\tilde{T}_n=\sum_{s,t:|s-t|<2m}a_{s,t}\tilde{X}_s\tilde{X}_t+\sum_{s,t:|s-t|\ge 2m}a_{s,t}\tilde{X}_s\tilde{X}_t$. It is easy to note that the first term is $O(n^{d/2} \max |a_{s,t}|) = o(\sigma_n)$. For the second term, writing it using the coordinates of the indexes and using the previously defined sigma fields, we will make use of the martingale central limit theorem (Hall and Heyde [39]) and that will give us the required result directly. The arguments here will be similar to the one dimensional case, as was done in Liu and Wu [66]. □

Lemma 2.4: Suppose $I_n(\theta)$ denotes the periodogram corresponding to the data $(X_{L_i})_{i=1(1)n}$ where each L_i is an element from $\mathbb{R}^d$. Then, for a square integrable function $f(\theta)$ defined on $D \subset \mathbb{R}^d$, variance of the quantity $\int_D I_n(\theta) f(\theta)\, d\theta$ goes to 0, as n goes to ∞.

Proof. Note that $nI_n(\theta)$ is a quadratic form in X_{L_j}'s and one can get that $\int_D nI_n(\theta) f(\theta)\, d\theta = \sum_{j,k} \alpha_{L_j - L_k} X_{L_j} X_{L_k}$ where α_k is same as the Fourier coefficients corresponding to the function $f(\theta)$. If we denote it by nT_n and define $n\tilde{T_n}$ in a similar way, but with $\tilde{X}_{L_j}$'s, then using similar arguments as in the previous lemma, one can write that

$$\left\| (S_n - \mathbb{E}(S_n)) - (\tilde{S}_n - \mathbb{E}(\tilde{S}_n)) \right\| = O_p(A_n/\sqrt{n}),$$

where $A_n = (\sum_{k=1}^{n} |\alpha_{L_k}|^2)^{1/2}$. Since α_k denotes the Fourier coefficient corresponding to $f(\theta)$ and since f is square integrable, we can say that $A_n = O_p(1)$.

Now, $n\tilde{S}_n$ is obtained using the m-dependent approximations and so, for a fixed m, the variance of $\tilde{S}_n$ goes to 0. Combining the above, it is straightforward to show that the variance of $\int_D I_n(\theta) f(\theta)\, d\theta$ (which is same as S_n) goes to 0 as $n \to \infty$. □

CHAPTER 3

SPATIO-TEMPORAL MODELING OF AIR POLLUTION DATA

3.1 Background

In this chapter, we discuss the effect of particulate matters, popularly denoted as PM. Considering the severe effect (refer to Chapter 1) of the PM into account, the Environmental Protection Agency (EPA) of the United States of America (USA) provided in 1997 new regulations that established National Ambient Air Quality Standards (NAAQS) for PM with aerodynamic diameters less than 2.5 micron. This is usually measured in units of micrograms per cubic meter (μgm^{-3}), and we will denote it as $\text{PM}_{2.5}$ henceforth. According to NAAQS, the hourly average of the $\text{PM}_{2.5}$ concentration should not be over 35 μgm^{-3}. However, the real situation often goes beyond the standard. A quick example is the data analyzed in this study. They are obtained from 66 different monitoring stations in Taiwan, for a span of 10 years (from 2006 to 2015) and the median of the $\text{PM}_{2.5}$ values is slightly above 37 μgm^{-3}. The reader is also referred to the earlier study by Mayer [70] who discussed how the air quality is deteriorating in different cities across the world. All in all, there is a growing demand to identify the main factors that contribute to the air pollution.

A brief discussion on PM is in order. Generally speaking, $\text{PM}_{2.5}$ contains different particles either emitted directly or formed in the atmosphere from gaseous emissions. The examples include sulfates formed from sulfur dioxide (SO_2) emissions, nitrates formed from NO_x emissions, and carbon formed from organic gas emissions. Now, the rates of conversion of gases to particles are often reliant on different regional and temporal factors, including the topography, the land cover and several seasonal climatic variables, and as a consequence, the $\text{PM}_{2.5}$ concentrations are also affected by these variables. This immediately increases the need for a spatio-temporal model to assess the air quality. A good model can provide better predictions and that would in turn help in determining an efficient strategy to control the air pollution level.

The spatial and spatio-temporal modeling of air pollutants started early in the past century. Elsom [27] studied the spatial correlation fields for air pollution in an urban area while different geostatistical space-time models have been applied to examine the trend in deposition of atmospheric pollutants in Eynon and Switzer [30], Bilonick [12], Rouhani and Hall [81], and Vyas and Christakos [95]. Other earlier notable works in this regard include Guttorp et al. [36], Haas [38], and Carroll et al. [17]. Most of these approaches rely on a spatio-temporal random field where the spatial or temporal dependences are incorporated in either the mean function or in the error process, and the parameters are estimated using frequentist procedures. A nice discussion on the geostatistical space-time models can be found in Kyriakidis and Journel [58].

In comparison, many 21st century studies in the related problem have made use of hierarchical Bayesian approaches for spatial prediction of air pollution. For example, Sun et al. [90] analyzed the PM_{10} (particulate matters with diameter less than 10 μ) concentrations in Vancouver and developed posterior predictive distributions using Bayesian techniques. Kibria et al. [55], on the other hand, used a multivariate setup to analyze $\mathrm{PM}_{2.5}$ concentrations in Philadelphia and developed spatial prediction methodology in a Bayesian context for this purpose. For a related problem, in order to predict PM_{10} concentrations in London, Shaddick and Wakefield [86] proposed a short-term space-time modeling technique.

In a hierarchical Bayesian setting, Sahu and Mardia [82] modeled the spatial structure using principal kriging functions and the time component by a random walk process to present a short-term forecasting analysis of $\mathrm{PM}_{2.5}$ data in New York City. There are two other notable works by the same authors. Sahu et al. [83] used indicators for urban or rural sites to employ different spatio-temporal processes in the error structure for modeling the $\mathrm{PM}_{2.5}$ series of several states in the Midwest of the United States (US). In a later paper, Sahu et al. [84]

developed a space-time model, that includes a spatially varying regression term along with an auto-regressive term in the mean structure, to analyze the ozone concentrations in the eastern states of the US. Berrocal et al. [10] extended one of their earlier works to introduce a bivariate downscaler and provided a flexible class of space-time assimilation models. On the other hand, Cameletti et al. [16] provided a nice discussion on the comparison of available space-time models using data from Piemonte, Italy.

In this study, we consider the problem of developing a new spatio-temporal model, with the main focus on identifying if there is any space-time interaction in the behavior of the $\text{PM}_{2.5}$ concentrations. An interaction means that the temporal trend of the pollution is more similar for sites closer in a spatial scale. An early study in this regard is by Wikle et al. [98]. The authors developed a hierarchical Bayesian model that allowed an interaction in space and time. However, except for that and a few related works, there have been very few efforts to quantify the space-time interaction for air pollution data, albeit there is a vast literature on the spatio-temporal modeling for the same, as we have already discussed.

Note that the problem of identifying the space-time interaction is not at all specific to the air pollution data. Rather, it has been studied in several other fields, ranging from seismology, epidemiology, criminology to transportation research. Meyer et al. [71] provided a nice discussion on the tests for space-time interaction in problems related to medical studies. The most popular techniques in this regard are Knox test, Mantel test, and space-time K-function analysis. These tests mainly revolve around test statistics of the form $T = \sum_{j \neq i} a_{ij}^s a_{ij}^t$, where a_{ij}^s and a_{ij}^t are measures of the spatial and temporal adjacency of the events i and j. Further reading on space-time interaction and related problems from other fields can be found in many articles in the literature; Kulldorff and Hjalmars [56], Legendre et al. [62], Vanem et al. [93] and the references therein being a few examples.

The rest of the chapter is organized as follows. Section 3.2 provides an exploratory analysis of the data under study. The proposed model, its properties, and related results are described in Section 3.3. Section 3.4 shows the results of detailed data analysis while Section 3.5 provides some concluding remarks and the scopes of future work.

3.2 Preliminary Analysis

3.2.1 Data

The $PM_{2.5}$ data we analyze are collected from 71 official monitoring stations across Taiwan. However, for five of those stations, we do not have the necessary information on the covariates considered and so, we decide to exclude them from this study. The remaining 66 monitoring stations are irregularly located in space, spread over the whole Island with some concentrations around big cities or industrial areas. The minimum and maximum of the pairwise distances of these stations are 0.58 kilometers and 366.7 kilometers respectively while the arithmatic mean of the same is 140.7 kilometers. One major talking point of our study is that we consider the issue of space-time interaction as a property of a region, rather than that of every individual station. For that, we divide the data into several clusters, based on the latitude and longitude of the stations. More on this is discussed in the following sections. Figure 3.1 shows the exact locations of the 66 stations. The concentration of the stations on the west coast is understandable because it is the most populated area of the Island. Different colors of the map indicate different regions (clusters) in our study.

Turn to the temporal scale of our study. The data are obtained on an hourly basis for ten years, starting from January 1, 2006 to December 31, 2015. However, the sampling frequencies vary from station to station and for some sites, there are missing values. Moreover, as we are mainly interested in identifying space-time interaction for the air pollution data, lower temporal resolution is desired and simpler. Thus, in this analysis, we decide to aggregate the hourly data into weekly averages based on all available measurements within a week. It

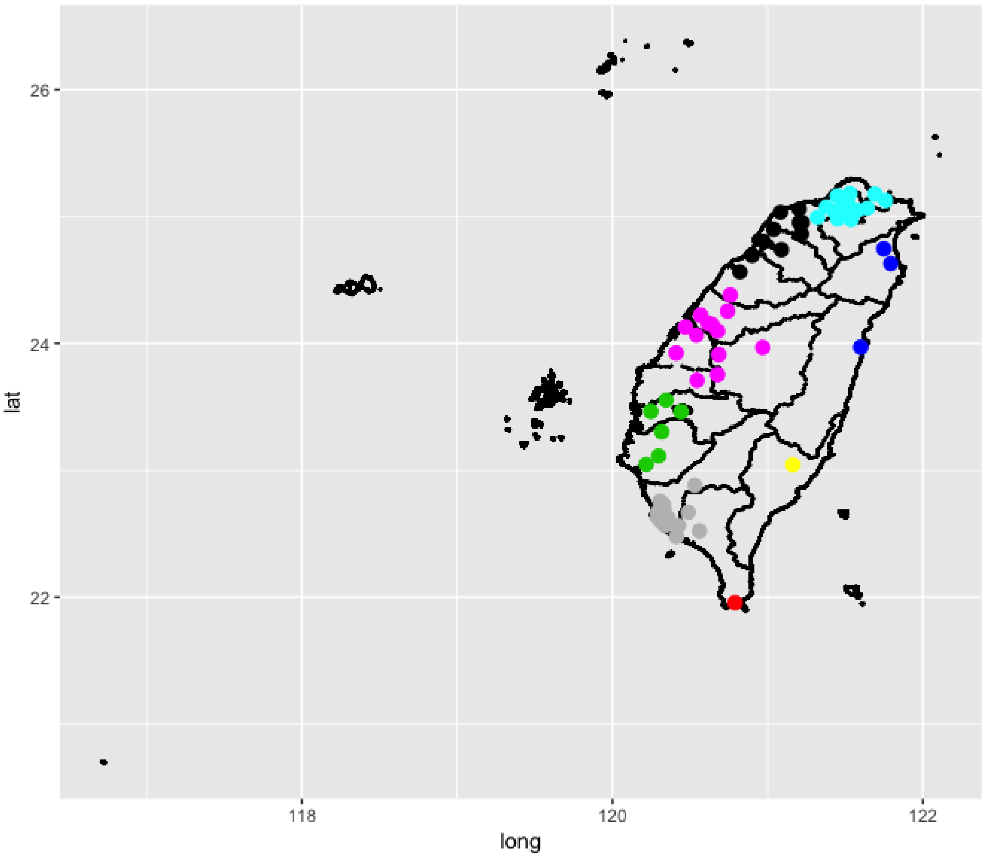

Figure 3.1: Map of the locations and clustering for the 66 monitoring stations on Taiwan.

is worth mentioning that this is a common practice while dealing with monitoring data, as discussed in Smith et al. [88]. In order to maintain continuity in the time scale, we consider the whole set of 3652 days (from 2006 to 2015) together and divide it into 522 weeks. Hence, the total number of data points considered in this study is $66 \times 522 = 34452$. Throughout the study, whenever needed, we would use one year's (2015) data, i.e. $52 \times 66 = 3432$ points (approximately the last 10% for each site), for validation purposes. A detailed discussion on this is presented in the following sections.

To begin, we present some exploratory analysis on the data. As has been done in several related studies, we convert the $PM_{2.5}$ values to the square root scale and the rest of the study is performed with the transformed data. This type of transformation is a common practice for air pollution data, cf. Smith et al. [88]. The graph in Figure 3.2 shows the overall means and the variances of the transformed $PM_{2.5}$ observations for different stations and different months. The top left panel describes the variances against the means of the stations. The top right panel shows the same, but for different months in the study. The bottom two plots show the behavior of the means and variances corresponding to different months.

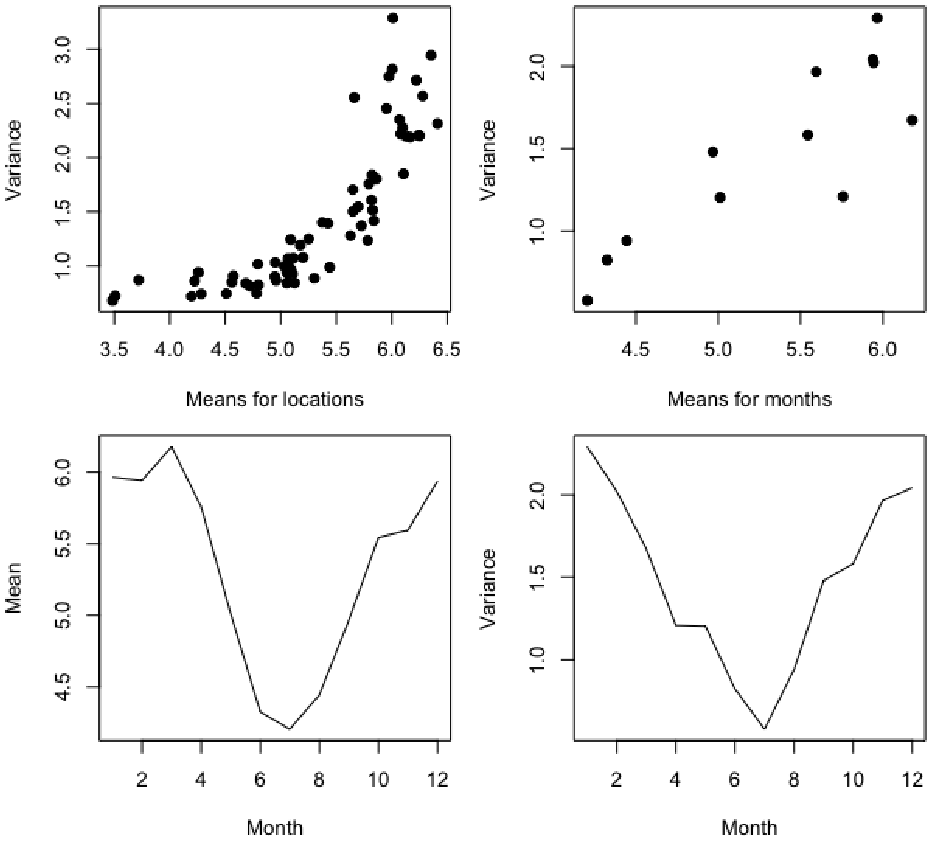

Figure 3.2: (Top left) Sample variances versus means of the weekly $\sqrt{PM_{2.5}}$ observations for 66 stations, (Top right) Means versus variances of the $\sqrt{PM_{2.5}}$ observations for 12 months, (Bottom left) Means of the $\sqrt{PM_{2.5}}$ observations corresponding to different months, (Bottom right) Variances of the $\sqrt{PM_{2.5}}$ observations corresponding to different months.

From the top two plots, it is clear that the variance increases in a nonlinear manner as the mean increases. On the other hand, the bottom left plot establishes that there is seasonality in the data, which is expected for most related time series problems. Moreover, interestingly, the bottom right plot of the variance corresponding to the months suggests that the variance cannot be assumed to be homoskedastic. In fact, it varies seasonally, and that motivates us to consider a heteroskedastic nature for the error variance in the Gaussian process of the proposed model.

On the other hand, we also show a heat map (Figure 3.3) of the weekly average of $PM_{2.5}$ observations for all the locations. In the map, we show the averages for different seasons. It is evident that the spatial pattern of the weekly averages changes according to seasons, and that shows the need to consider the space-time interaction coefficient in the model, as described in Section 3.3.

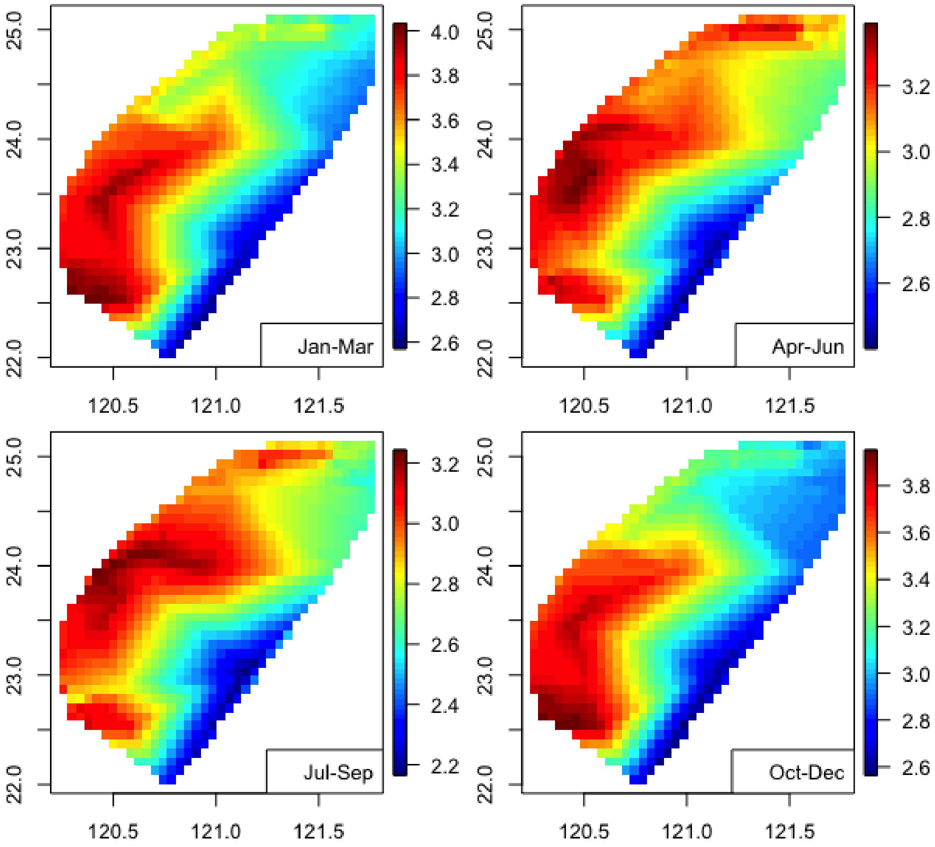

Figure 3.3: Seasonal heat map of the $PM_{2.5}$ levels on Taiwan.

3.2.2 Covariates used in the study

For each station, along with the observations of $\mathrm{PM}_{2.5}$, data were collected on temperature, relative humidity, and wind speed and direction. Similar to the air pollution observations, these data were also collected on an hourly basis from January 1, 2006 to December 31, 2015. We aggregate them per week and use the representative values as covariates in our study.

At first, we examine the relationship of the pollution data with the relative humidity and the temperature. The scatter plots of the square root transformations of $\mathrm{PM}_{2.5}$ against these covariates are shown in Figure 3.4. From a quick glance, it seems that the air pollution is not so significantly affected by the temperature, but is dependent on the humidity. It shows a slight decrease in pollution with the increase in the relative humidity. The Spearman correlation coefficient between transformed $\mathrm{PM}_{2.5}$ and humidity was found out to be -0.326, while the same for the temperature was -0.254.

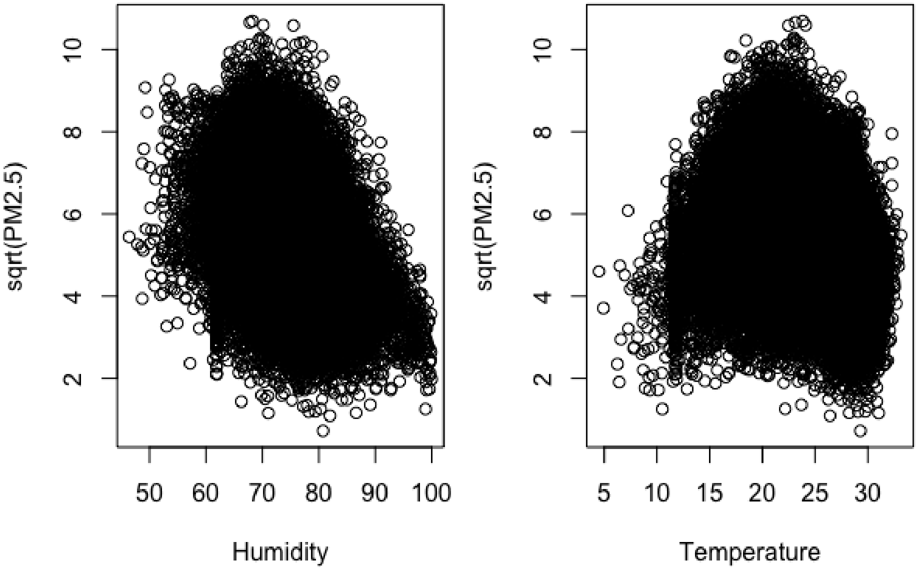

Figure 3.4: Scatter plots of the square roots of the $\mathrm{PM}_{2.5}$ observations, with respect to relative humidity (left) and temperature (right)

The wind speed is another variable that may have some effects on the air pollution. As before, the data on wind speed and wind direction were available on an hourly basis. It was noted that the behavior of the wind speed and direction varied widely for different locations. To prove this point, boxplots of the daily wind speed, corresponding to all the locations, are displayed in Figure 3.5.

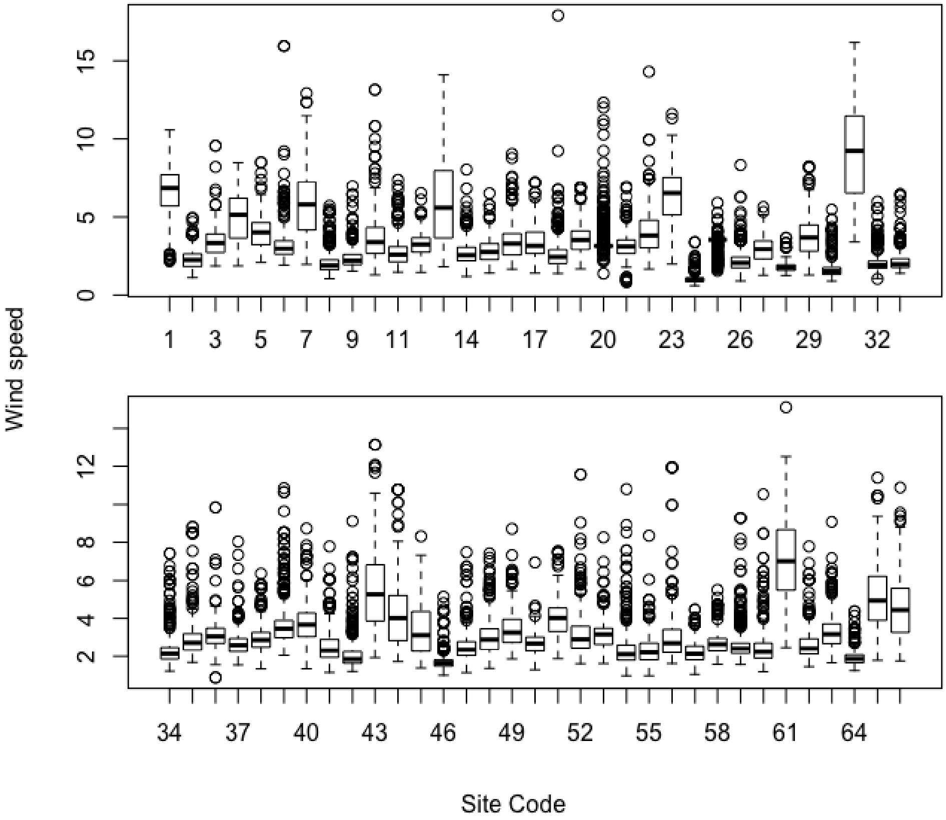

Figure 3.5: Boxplots of the wind speed for 66 stations

The boxplots shows clearly that the ranges of the wind speed vary markedly for different locations, while the means are not so much. Naturally, in order to account for the effect of the wind, unlike the previous covariates, it will be wise not to take the weekly average. Instead, we decided to use the maximum daily wind speed for every week. It is also worth mentioning that for most weeks, the maximum wind speed occurred on same day in all the locations, which again justifies why the maximum in lieu of the average is a good measure in this regard. The sample correlation coefficient between transformed $PM_{2.5}$ and the maximum wind speeds is -0.146.

3.3 Methods

3.3.1 The proposed model

The proposed model is in some sense inspired by the one studied by Sahu et al. [83]. The key feature of our model is that it accounts for possible space-time interactions in the air-pollution data. We describe the model in a general setup, and discuss the particular case of Taiwan data in Section 3.4.2.

Suppose the data are collected for n locations and over T consecutive time points. Let us denote the square root of the $\mathrm{PM}_{2.5}$ at time t_j and location s_i by $Z(s_i, t_j)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, T$. Then, we assume that the overall mean values for different locations are different, and we subtract the location-wise overall means from the actual values of the transformed $\mathrm{PM}_{2.5}$ values to obtain the mean-adjusted numbers. These mean-adjusted values are denoted by $Y(s_i, t_j)$. For convenience, we drop the subscripts whenever not needed. In the proposed model, we consider the following hierarchical structure:

$$Y(s,t) := U(s,t) + \epsilon(s,t), \tag{3.1}$$

where $U(s,t)$ describes a spatio-temporal process and $\epsilon(s,t)$ denotes a white noise process, to account for the measurement errors. We assume the white noise process to follow heteroskedastic $N(0, \sigma_i^2)$ distributions independently, where σ_i^2 is chosen according to different seasons. $U(s,t)$, on the other hand, assumes the following structure:

$$U(s,t) := \mu(s,t) + v(s,t), \tag{3.2}$$

where $\mu(s,t)$ stands for the mean of the $U(s,t)$ process and $v(s,t)$ denotes a zero-mean spatio-temporal process.

The aforementioned mean function is considered to be an additive combination of the effects of the covariates, the seasonal effects, and a spatio-temporal interaction effect. In order to capture the effects of available covariate information, we consider a term of the form $B\alpha$ where B is the design matrix for p covariates. The population densities, temperature, humidity, number of factories, number of cars, etc can be taken as the covariates for the analysis. The seasonal variation is captured by introducing indicators for different seasons. Throughout this study, we have considered monthly indicators, but the method can be used for other cases too. In general, let us use J to denote the number of seasons in a year.

The spatio-temporal effect we use in the mean structure allows us to test for space-time interaction, if any. In principle, the effect is described as a linear function of the form $\gamma_0 + \gamma_s t$, thereby allowing us to test if the γ_s values are same across different locations. However, this would increase the complexity and computational burden dramatically, if we have a lot of sites. To address this issue, we make a logical assumption that the coefficients γ_s for locations which are close to each other are the same, and that it should be a characteristic of a region rather a station. So, based on the coordinates of the stations, we divide the stations into clusters first and employ γ_k for all stations in the kth cluster. The number of clusters we choose is based on the number of locations we have. If n, the number of locations, is 5 or less, we do not use clustering. But for $n > 5$, we will use $[\sqrt{n}]$ clusters ($[\cdot]$ denotes the greatest integer function) to divide the locations into different regions. In practice, we use the k-means clustering method, based on the latitude and longitude, for this purpose.

Further, we propose to use an intercept-free model, and so, along with the response variable, we also subtract the overall means of the covariates from the respective observations. Thus, the mean structure of the process can be described as:

$$\mu(s,t) = c(s,t)'\alpha + \sum_{j=2}^{J} \beta_j \; m(t,j) + \sum_{k=1}^{K} \gamma_k \; r(s,k)t. \tag{3.3}$$

In the above, $c(s,t)$ is a column vector with the values of the covariates and $m(t,j)$ and $r(s,k)$ are the indicators for the season and location region, respectively, where J and K are the seasonality and the number of regions. That is, $m(t,j) = 1$ if time t is in the jth season and 0 otherwise. Similarly, $r(s,k) = 1$ if location s is in kth region and 0 otherwise. For identifiability purposes, we would take $\beta_1 = 0$. Note that because of using mean-adjusted values for the response and the covariates, we do not have any intercept term in the model. On the other hand, we would scale down the time points (t) to equi-spaced points in the interval $[0,1]$.

The term $v(s,t)$ in Equation (3.2) can be treated as a spatially varying temporal trend. Averaging over different sites in a region, we can get the adjustment to the regional trends and averaging over time can help us obtain the adjustment at the temporal scale. For convenience, we consider a separable structure for the covariance of this process. Moreover, we assume that the locations in different cluster are independent of each other. In particular, the covariance between $v(s_1,t_1)$ and $v(s_2,t_2)$, when s_1, s_2 are in the same cluster, is taken as a product of the spatial dependence and temporal dependence and we write it as

$$\mathrm{Cov}(v(s_1,t_1), v(s_2,t_2)) = \sigma_v^2 \, \rho(\|s_1 - s_2\|, \phi_s) \cdot \rho(\|t_1 - t_2\|, \phi_t) \cdot \mathbb{I}\{s_1, s_2 \in \mathcal{C}_k\}, \qquad (3.4)$$

where $\rho(x,d)$ denotes the exponential covariance function e^{-dx}, and $\mathcal{C}_k$ denotes a particular cluster. The distance functions $\|s_1 - s_2\|$ or $\|t_1 - t_2\|$ are taken as the Euclidean distance of the two points.

Throughout this chapter, Y would denote the vector of $N = nT$ data points, arranged according to clusters at first, time points next, and sites then. Thus, if $\{s_1, s_2\}$ form a cluster, then the first few observations will be $Y(s_1,t_1), Y(s_2,t_1), Y(s_1,t_2), Y(s_2,t_2)$, etc. The vector $v = (v(s_i,t_j))$ is formed in a similar way. We denote the full covariance matrix of v by Σ_v. Note that Σ_v is a block diagonal matrix, where each block corresponds to the covariance matrix of a cluster of locations, and is of the form $\sigma_v^2(\Sigma_t \otimes \Sigma_s)$ such that $\Sigma_t(i,j) = \rho(\left\|t_i - t_j\right\|, \phi_t)$ and $\Sigma_s(i,j) = \rho(\left\|s_i - s_j\right\|, \phi_s)$. Further, note that when the sampling design considers equally-spaced time points, the common spacing being d_t, one can write $\Sigma_t(i,j) = \psi^{|i-j|}$, where $\psi = e^{-\phi_t d_t}$.

On the other hand, as mentioned before, we entertain a heteroskedastic error function for $\epsilon(s,t)$. The exploratory analysis suggests that the variances are different for different seasons, and so, we assume $\epsilon(s,t) \sim N(0, \sigma^2_{m(t)})$, where $m(t)$ denotes the season the time t is in. If we use ϵ to denote the vector of $\epsilon(s,t)$, arranged similarly as Y and w, then the above discussion implies that $\epsilon \sim N(0, \sigma^2 D)$, where D is a diagonal matrix such that the diagonal element corresponding to $\epsilon(s,t)$ is $\sigma^2_{m(t)}/\sigma^2$. We denote these parameters by $\tau_1^2, \ldots, \tau_J^2$, where τ_i^2 is the variance parameters associated with the ith season. Also, we set $\tau_1^2 = 1$ to avoid any potential identifiability problem.

Now, to write the full model in a vector-matrix notation, recall that Y is the observed vector of dependent varaible, and v and ϵ denote the corresponding vectors of the zero mean spatio-temporal process and the Gaussian error process, respectively. We can write the mean function as the sum of $c(s,t)'\alpha$, $m_t'(\beta_j)$ and $r_{st}'(\gamma_k)$, where m_t and r_{st} are the column vectors corresponding to the parameter vectors $(\beta_j)_{2\leq j\leq J}$ and $(\gamma_k)_{1\leq k\leq K}$, respectively. Then, denoting the vector of all the parameters by θ and letting X be a design matrix such that each row is of the form $X(s,t)' = (c(s,t)', m_t', r_{st}')$, the model can be written as:

$$\begin{aligned} Y &= X\theta + v + \epsilon, \qquad (3.5)\\ \text{where } \ v &\sim N(0, \sigma_v^2 \Sigma_v), \\ \text{and } \ \epsilon &\sim N(0, \sigma^2 D). \end{aligned}$$

It is evident that there are $(p + J + K)$ components in the parameter vector θ if we consider $p + 1$ covariates and K regions for the locations. On the other hand, we write the two variance components σ^2, σ_v^2 to be equal. Here, the estimate of σ^2 would give us an idea about the variance explained by the spatio-temporal process while the estimates of the diagonal elements of D tell us how much is explained by the pure error process.

Finally, the best estimates for ϕ_s and ϕ_t are obtained using a cross validation scheme. The validation scheme considers prediction for all the sites at some time points and obtains the mean squared error for those predictions. In this study, for every site, we take the first 80% of the time points under consideration (e.g. in case of weekly data for 10 years, we would consider the first 8 years or 418 weeks) for model fitting and then make predictions for the last 20% (104 weeks for the aforementioned data) to see which combination of (ϕ_s, ϕ_t) would work the best. The possible choices for ϕ_s used in the study were (0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5) while the choices for ϕ_t were (0.5, 0.75, 1, 1.5). To find out the optimal choice, we find the combination with the smallest mean squared error of the predictions, which is calculated as follows. For each site s_i, let us denote the validation time points by $t_1, \ldots, t_b$ and the predictions by $\hat{Y}(s_i, t_j)$ for $i = 1, \ldots, n$; $j = 1, \ldots, b$. Then, for each of the $5 \times 4 = 20$ combinations, the prediction mean squared error is computed as:

$$\text{MSE} = \frac{1}{nb} \sum_{i=1}^{n} \sum_{j=1}^{b} \{Y(s_i, t_j) - \hat{Y}(s_i, t_j)\}^2. \tag{3.6}$$

The prediction procedure is discussed in Section 3.3.3.

3.3.2 Testing for interaction and parameter estimation

The model described in Equation (3.5) allows us to test for space-time interaction. The model includes no interaction term if $\gamma_1 = \ldots = \gamma_K = 0$. We employ the Lagrange multiplier (LM) test for this purpose. Recall that the main advantage of the LM test or the score test is that it, unlike Wald test or likelihood ratio test, does not require an estimate of the information under the alternative hypothesis or unconstrained maximum likelihood. The LM test uses only the assumptions in the null hypothesis to get the maximum likelihood estimates and then calculates the value of the test statistic (which follows a chi-squared distribution with appropriate degrees of freedom) to make a decision.

Once we perform the LM test, we can make a decision about the model to use. We use the full model (Equation (3.5)) if the decision is to reject the null hypothesis that there is no space-time interaction.

To estimate the parameters, we employ the generalized least squares techniques, with little modifications. Observe that the proposed model can be thought of in the form $Y = X\theta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2\Omega)$ such that $\Omega = \Sigma_v + D$. This is in the setup of a generalized least squares problem. Further, note that the number of unknown parameters in the error covariance matrix Ω is $J + 2$, namely $\sigma^2, \tau_2^2, \ldots, \tau_J^2; \phi_s, \phi_t$. As we decided to get the optimal choices for the last two parameters by a cross-validation procedure, that leaves us the task of estimating J additional variance parameters from the model.

From a practical point of view, it is more important to identify the cases where the air pollution is hazardous for health and environment, rather than the ones when the situation is less harmful. According to the standards set by the EPA, the average for the fine particulate matters should not exceed 35 μgm^{-3}. So, while minimizing the squared sum of residuals, we put more weight on the cases where the actual PM$_{2.5}$ values are more than 35. Thus, the loss function that we want to minimize is of the form

$$L = \sum_{i=1}^{n}\sum_{j=1}^{T} w(s_i, t_j)\hat{e}(s_i, t_j)^2. \tag{3.7}$$

Here, $w(s_i, t_j)$ should be taken higher for $Z(s_i, t_j) \geq \sqrt{35}$. We take it to be $(1+2/\log N)$ for $Z(s_i, t_j) \geq \sqrt{35}$ and $(1-2/\log N)$ otherwise, where $N = nT$. An attractive feature of this is that the weights will approach 1 as $N \to \infty$, thereby establishing that the importance will be approximately equal on all observations when the sample size is huge. $\hat{e}(s_i, t_j)$, on the other hand, is the standardized residual for location s_i and time t_j. Note that it can be written as $\hat{\Omega}^{-1/2}\hat{\varepsilon}/\hat{\sigma} = \hat{\Omega}^{-1/2}(Y - X\hat{\theta})/\hat{\sigma}$. Thus, assuming that $\hat{\Omega}$ and $\hat{\sigma}$ are known to us, we can say that minimizing L is equivalent to minimizing $(Y - X\theta)'\hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}(Y - X\theta)$ with respect to θ. Here W is a diagonal matrix with the diagonal elements being the same as $w(s_i, t_j)$. It is evident that the minimizer is $\hat{\theta} = (X^TVX)^{-1}(X^TVY)$, where $V = \hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}$. In light of the above discussion, we propose to use the following procedure to estimate the parameters in the model.

Parameter estimation procedure:

1. Set $\hat{\tau}_j^2 = 1$ for $j = 1, \ldots, J$.

2. Evaluate $\hat{\Omega}$ and $V = \hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}$.

3. Compute $\hat{\theta} = (X^TVX)^{-1}(X^TVY)$ and set $\hat{\varepsilon} = Y - X\hat{\theta}$.

4. Compute $\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\Omega}^{-1}\hat{\varepsilon}/N$, where $N = nT$ is the total number of observations.

5. Let $\hat{\varepsilon}_j$ be the error corresponding to the jth season, for $j = 1, \ldots, J$.

6. For $j = 2, \ldots, J$, note that $\varepsilon_j \sim N(0, \sigma^2(\Sigma_v^{(j)} + \tau_j^2 I))$, where $\Sigma_v^{(j)}$ is the submatrix of Σ_v corresponding to the jth month.

7. Use optimization methods to compute the MLE $\hat{\tau}_j^2$ using the above.

8. Repeat steps 2 to 7 until convergence.

In the above procedure, in order to reduce the computational burden, we exploit the block-diagonal structure of the matrix Ω. Note that, if each block of Ω is denoted by B_i, then $\Omega^{-1/2}$ can be written as a block diagonal matrix, where the blocks are $B_i^{-1/2}$. Using this and writing W and X accordingly in the above steps, we substantially reduce the computational burden in the estimation.

The following theorem describes the asymptotic properties of the above estimators. Proof of the theorem is provided in Section 3.6.

Theorem 3.1: The estimate $\hat{\theta}$ obtained from the above procedure is consistent, in the sense that as the number of locations (n) and the number of time points (T) approach infinity, $\hat{\theta} \to \theta$ in probability. Furthermore, $\sqrt{nT}(\hat{\theta} - \theta) \to N(0, \sigma^2 Q^{-1})$, where Q is the limit of $(X'\Omega^{-1}X)/nT$ as $n, T \to \infty$.

3.3.3 Future prediction

To make a new prediction for site s' at time t', we use the parameter estimates obtained from the method mentioned in the previous section. Let us use $X(s', t')$ to denote the new set of covariate vectors, similar to what we did in Section 3.3.1. Then, it can be said that $\hat{Y}(s', t') = X(s', t')'\theta + \varepsilon(s', t')$. However, in view of the fact that Ω is not a constant multiple of the identity matrix, we cannot simply assume that the prediction error is going to be independent of the sample disturbances. The prediction procedure has to take the dependence into account.

If ε is the error vector corresponding to the original data, let us denote $\text{Cov}[\varepsilon(s', t'), \varepsilon]$ by w, which is going to be a nT-dimensional column vector. It is evident that w and Ω depend on the values of ϕ_s, ϕ_t, σ^2 and τ_j^2 for $j = 1, \ldots, J$. We use the estimates of these parameters to get $\hat{w}$ and $\hat{\Omega}$. Then, following Goldberger [35], we can say that the best linear unbiased predictor (BLUP) is

$$\hat{Y}(s', t') = X(s', t')'\hat{\theta} + \frac{1}{\hat{\sigma}^2} \cdot \hat{w}'\hat{\Omega}^{-1}(Y - X\hat{\theta}). \tag{3.8}$$

Because of using $\hat{\theta}$ instead of θ, one can also call the above EBLUP, a term coined by Zimmerman and Cressie [105]. Here, the E stands for 'estimated' or 'empirical', cf. Harville and Jeske [43]).

Finally, we add the overall average (estimated from historical data) of that particular site and transform the predictions to get the estimate of the actual air pollution measurement. Here, it should be mentioned that this method poses a problem for predicting the future pollution level at out-of-sample sites. The main caveat is that in order to predict $Y(s', t')$ for an out-of-sample site s', both $X(s', t')$ and the average pollution level are needed. However, without historical data, it will not be possible to estimate with efficiently.

3.4 Detailed Analysis

3.4.1 Simulation studies

To begin, we present some simulation studies to show that our methods are capable of capturing the space-time interaction that might be present in the empirical application we are interested in.

We perform the simulation experiment in two stages - first with a linear space-time interaction term and then with a non-linear one. Throughout this study, we consider weekly average of the air pollution data. And then, for different time intervals, we evaluated the type-I error and the power for different values of n. In the simulation study, we considered four different values of T: 52 (1 year), 104 (2 years), and 261 (5 years). We worked with different numbers of locations ($n = 10, 20, 50$) to understand how the proposed method performs as we have more data. For n locations, the xy-coordinates were generated randomly from a $(0, n^2) \times (0, n^2)$ grid. The values of the variance parameters σ^2, τ_j^2 (see Section 3.3.1) were generated from an inverse-gamma distribution with parameters $(4, 3)$. All the coefficients in the model were simulated from independent normal distributions with mean 0 and standard deviation 1. Finally, we used $\phi_s = 0.1, \phi_t = 0.75$ for the covariance matrix of the spatio-temporal process $w(s,t)$.

For the case of linear space-time interaction, the observations (square root of the $\text{PM}_{2.5}$ data) were obtained from the model (Equation (3.5)). However, for the nonlinear case, we replaced the linear term with the quadratic term $(t/T)^2$, and assumed that the space-time interaction coefficients are negative. This type of interaction is common, and usually the coefficients are not big in magnitude. We wanted to observe whether a linear approximation works well to capture this type of interaction as well. The results for different cases under the linear time dependence are displayed in Table 3.1, while those for the nonlinear time dependence are shown in Table 3.2. The results reported are based on 500 iterations of simulation.

Table 3.1: Results for Different Cases, When the Dependence on Time is Linear. 5% critical values and 500 iterations are used.

	Type-I error			Power		
Number of locations	10	20	50	10	20	50
Weekly (1 year)	0.052	0.050	0.048	0.224	0.512	0.786
Weekly (2 years)	0.056	0.058	0.042	0.312	0.578	0.820
Weekly (5 years)	0.054	0.044	0.046	0.428	0.702	0.868

While the type I error remains under control across all scenarios, the power improves significantly for 20 or more locations and this is true both for the linear dependence and for the nonlinear dependence. The power also increases with sample size for all cases. The results confirm that the proposed testing procedure does not have large size distortions and, as expected, fares well for larger sample sizes.

We further extended our simulation studies to see how the proposed method behaves for larger data. To understand this, we concentrated on weekly data from 3 years (157 observations for each site) and took different values for n, ranging from 10 to 50. The results of 500

Table 3.2: Results for Different Cases, When the Time-Dependence is Nonlinear. 5% critical values and 500 iterations are used

	Type I error			Power		
Number of locations	10	20	50	10	20	50
Weekly (1 year)	0.058	0.046	0.046	0.232	0.394	0.678
Weekly (2 years)	0.048	0.042	0.062	0.268	0.478	0.714
Weekly (5 years)	0.058	0.050	0.066	0.358	0.500	0.794

iterations for different cases are shown in Figure 3.6. It is evident that, when the dependence on time is linear, the proposed testing procedure fares nicely for large number of sites. The testing procedure also works well for the nonlinear case as its power also increases, reaching about 0.8 when the number of locations is 50.

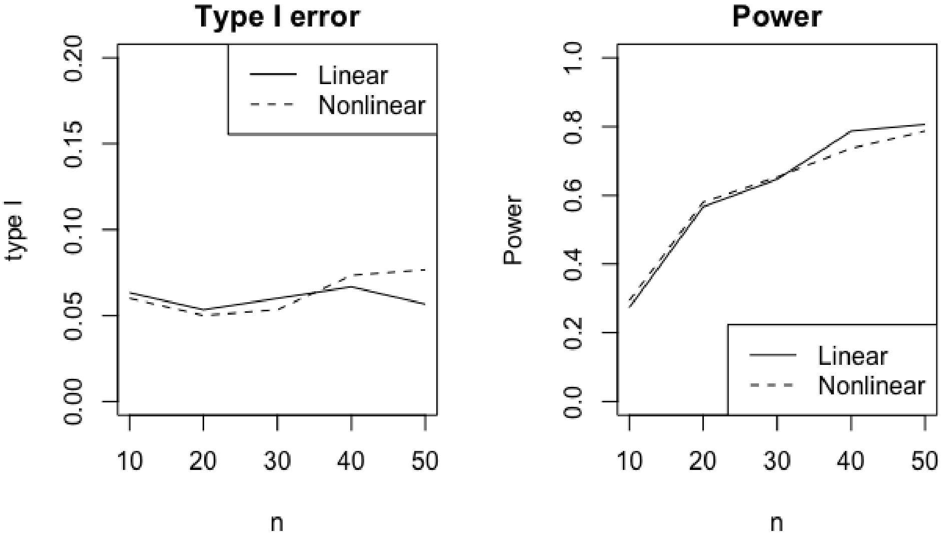

Figure 3.6: Type-I error (on left) and power (on right) for linear (solid) and nonlinear (dotted) cases, corresponding to different number of locations. $T = 157$ (3 years) and 500 iterations are used for all cases.

Finally, we present a particular case to show that the parameter estimation process indeed works well to identify the effect of the factors. In this example, we used 20 different locations, two covariates (humidity and temperature) and simulated data from our model to get weekly averages for 5 years (divided in 12 seasons). Thus, the number of observations in the data were $20 \times 261 = 5220$. We then estimated the parameters using the procedure described in Section 3.3.2. In Figure 3.7, the true values and the estimates of all the parameters in the model are plotted. We can see that most points lie along the line $y = x$ (displayed in the figure), thereby showing that the estimates are not too different from the true values. This study confirms that the proposed iterative estimation procedure is reliable.

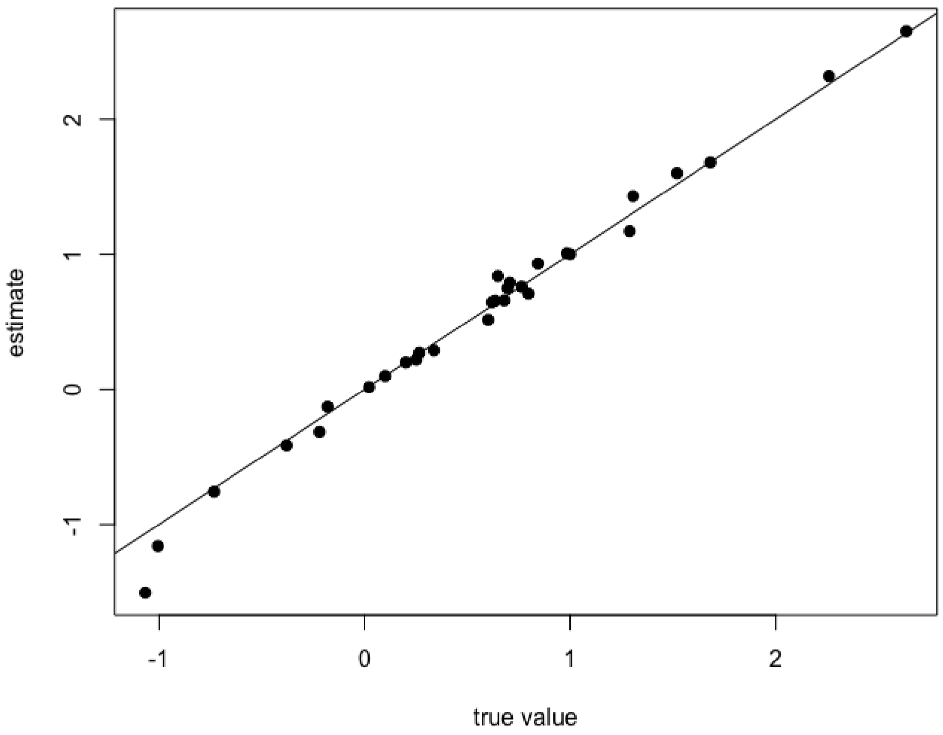

Figure 3.7: True value and estimated value for the parameters in the model, where data is generated for 20 locations and 5 years

3.4.2 Model selection

The first task to implement the proposed method is to choose proper decay parameters and to identify if there exists any space-time interaction. For the decay parameters, we searched in a two-dimensional array to find out which combination gives the least mean-squared error (refer to Equation (3.6)). The choices for ϕ_s were $(0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5)$ while the same for ϕ_t were $(0.5, 0.75, 1, 1.5)$. We used 80% of the data as our training data and the rest 20% were used for validation purposes in this regard.

We found that the combination of $\phi_s = 0.01$ and $\phi_t = 1$ was the best one for the data at hand. To put it into the perspective of the actual spatial and temporal scale, we can say that these choices correspond to a significant correlation in an approximate range of 300 kilometers and in a time span of 3 weeks, respectively. On the other hand, the LM test returned a p-value less than the level of significance (0.05), establishing that a space-time interaction effect is indeed present in the data. Thus, the model we decide to fit for our empirical analysis is the same as Equation (3.5), which is shown below for ease in reference.

$$Y(s,t) = c(s,t)^T\alpha + \sum_{j=2}^{12} m(t,j)\beta_j + \sum_{k=1}^{K} r(s,k)\ \gamma_k t + v(s,t) + \epsilon(s,t),$$

for $s = 1, \ldots, n;\ t = 1, \ldots, T$. Recall that $c(s,t)$ describes the covariates in the study, and we use relative humidity, temperature, and wind speed in the analysis. Furthermore, $T = 522$, $n = 66$ and thus, the number of clusters (K) is taken to be 8.

3.4.3 Parameter estimates

Next, we fitted the above model to Taiwan data to obtain parameter estimates. In this particular application, we have 22 parameters in the mean structure and 12 more in the variance structure. In what follows, we discuss these estimates step-by-step. Table 3.3 shows the coefficient estimates corresponding to the covariates and the estimates of the seasonal effects.

From the table, it is seen that almost all of the estimates show a significant effect. The $PM_{2.5}$ is higher whenever the humidity is less and the temperature is higher. The effect of humidity is possibly because of the fact that higher humidity is associated with higher precipitation, which is responsible for washing out particles from the air, thereby reducing the

Table 3.3: Estimates of the Parameters in the Mean Structure. β_j represents the estimated deviation from January in the seasonality pattern.

Parameter	Estimate	Standard error	Confidence interval
α_1 (Humidity)	-0.0318	0.0008	$(-0.0334, -0.0302)$
α_2 (Temperature)	0.0203	0.0024	$(0.0156, 0.0250)$
α_3 (Wind speed)	-0.0779	0.0038	$(-0.0853, -0.0705)$
β_2 (February)	0.0100	0.0257	$(-0.0403, 0.0603)$
β_3 (March)	0.0682	0.0267	$(0.0159, 0.1206)$
β_4 (April)	-0.3526	0.0296	$(-0.4105, -0.2946)$
β_5 (May)	-1.1230	0.0342	$(-1.1900, -1.0561)$
β_6 (June)	-1.7756	0.0378	$(-1.8497, -1.7016)$
β_7 (July)	-1.9105	0.0395	$(-1.9879, -1.8331)$
β_8 (August)	-1.5696	0.0390	$(-1.6460, -1.4932)$
β_9 (September)	-0.9832	0.0375	$(-1.0568, -0.9096)$
β_{10} (October)	-0.4402	0.0330	$(-0.5049, -0.3756)$
β_{11} (November)	-0.3213	0.0294	$(-0.3790, -0.2636)$
β_{12} (December)	0.0494	0.0257	$(-0.0010, 0.0998)$

pollution level. Further, as expected, a higher wind-speed reduces the amount of pollution significantly.

On the other hand, there exists a strong seasonal pattern. Recall that all estimates of the seasonality components are the deviations from the same in January. The winter months (from December to February) do not show a significant change in the pollution. March shows a significant increase from January level whereas the $PM_{2.5}$ decreases over the summer months, especially between June and August. This is understandable given the geographical location and climate pattern of Taiwan. It is more likely to rain over the summer in Taiwan and the wind tends to be from the south-east and the speed can be high with possible strong typhoons. In comparison, winter months tend to be dry with wind from north-west in Taiwan. This points to the further need to understand and analyze the effect of wind flux, a direction we are interested to pursue in future.

Next, we plotted the space-time interaction coefficients to see how they spread across the whole Island. Figure 3.8 shows these estimates on a spatial scale. It was found that all of the regions showed significantly negative estimates for the space-time interaction coefficients, the extent of interaction being different from region to region. Note that a negative value indicates that the overall pollution level decreases with time, and it is evident that this extent of decrease is very different for different regions. In absolute sense, it is the least for stations Hengchun (denoted by red) and Guanshan (denoted by yellow) while it is the most for stations around Taipei (light blue), Taichung (magenta), Taoyung (black), and Kaohsiung (grey) of Taiwan. This is understandable as Hengchun is at the southern tip of Taiwan and Guanshan is less populated and is without any heavy industry. In a similar line, we can also observe that the most significant stations are around the four largest cities in Taiwan.

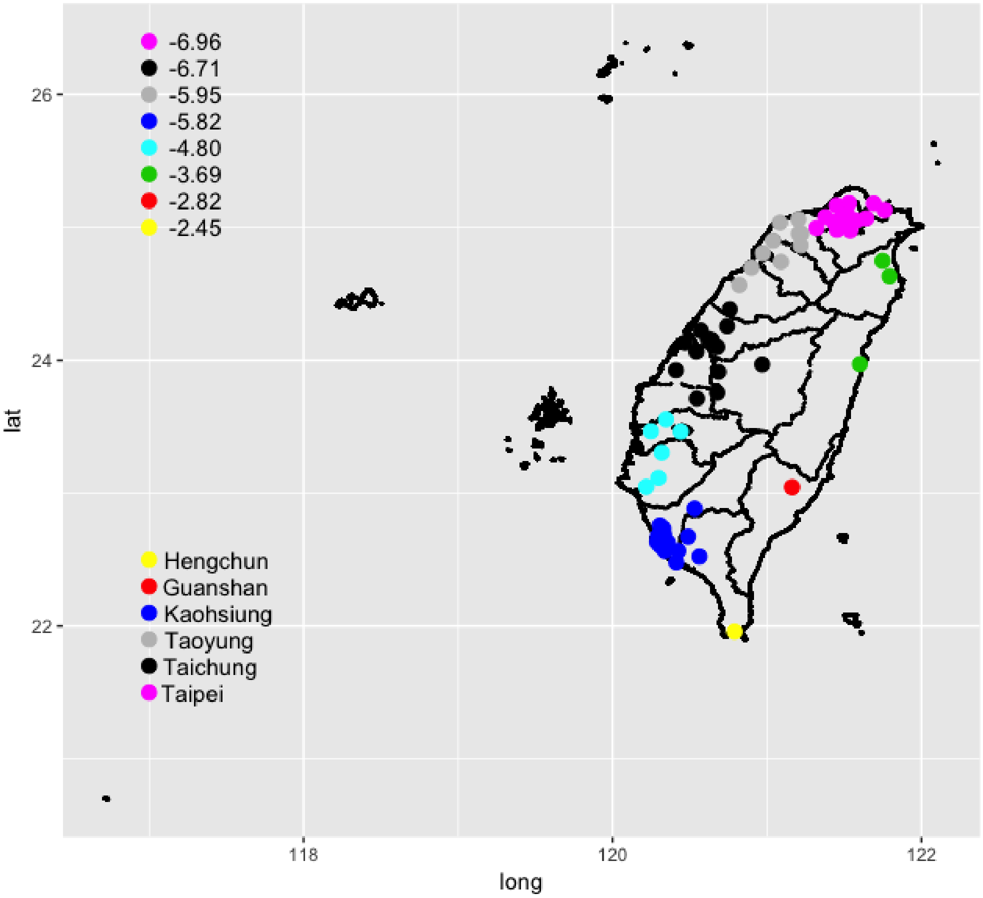

Figure 3.8: Estimates of the space-time interaction terms for different regions

The estimates of the variance parameters for different months are shown in Figure 3.9. They show an oscillatory behavior across different months, and the magnitudes of the variances range between 0.4 and 1. These estimates can explain the extent of the heteroskedasticity in the white noise process in our model. So far as the spatio-temporal process is concerned, its variance σ^2 was estimated at 0.8462, which is more than most variance terms of the white noise. This indicates that the white noise process can explain less variability in the data, and that the spatio-temporal process is more significant in this aspect.

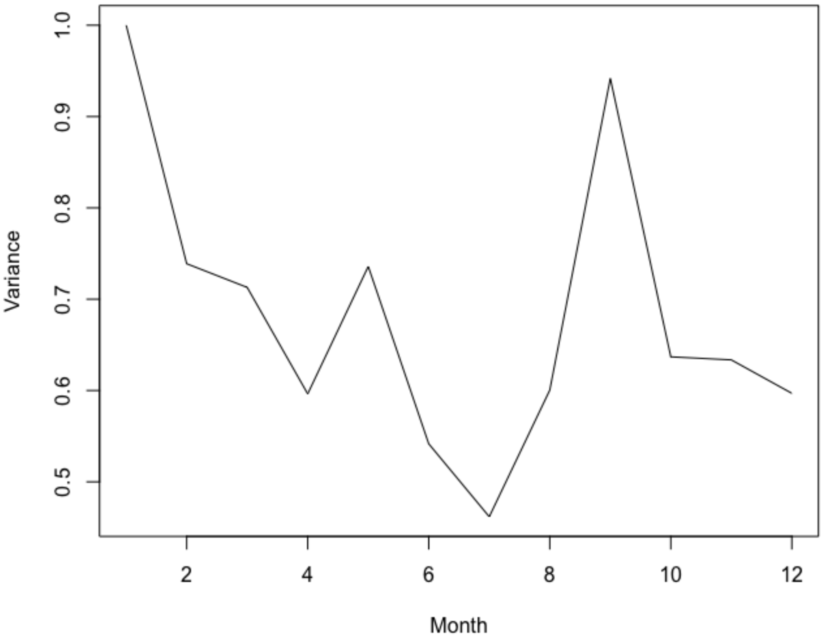

Figure 3.9: Estimates of the variance term for different months

3.4.4 Model diagnostics

In this section, we present some residual plots to show that the proposed model fares well to capture the effects of Taiwan $PM_{2.5}$ data. We evaluate the residuals after fitting the model and plot them using standard statistical procedures. First, the top panel in Figure 3.10 shows that there is no particular pattern in the residuals, thereby establishing that the residuals can be assumed to be uncorrelated. Next, the plot of the residuals corresponding to different months are presented in the bottom panel of the same figure. It shows that the variances are more or less evenly distributed across the months, thereby showing that the issue of heteroskedasticity has been taken care of.

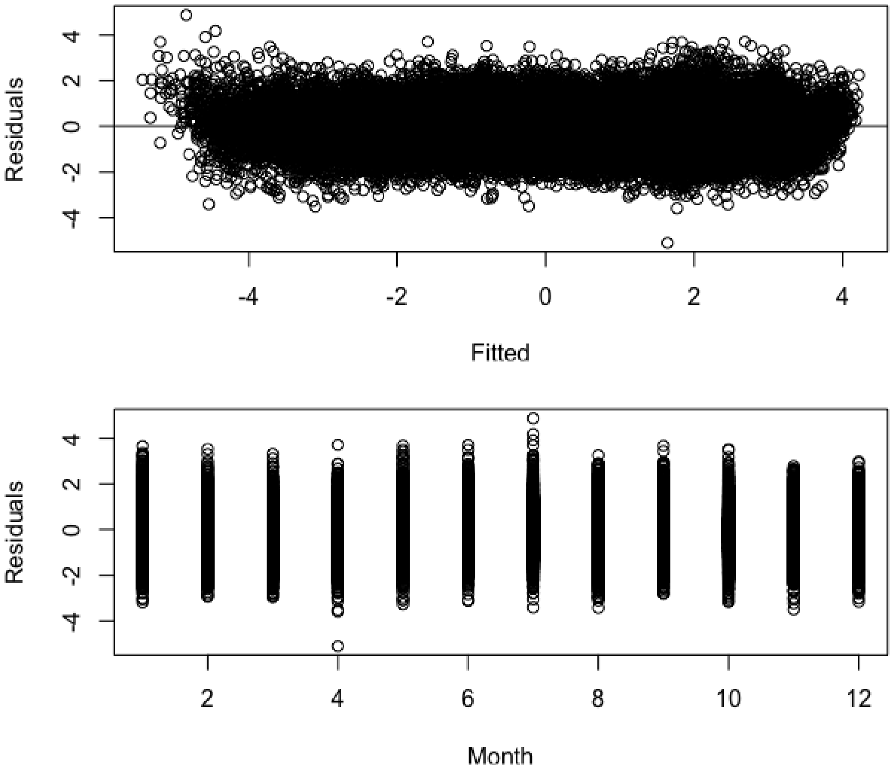

Figure 3.10: (Top) Standardized residuals are plotted against fitted values; (Bottom) Standardized residuals are plotted corresponding to different months

Furthermore, the left panel of Figure 3.11 shows that the histogram is approximately normally shaped, while the QQ plot presented in the right panel of the same figure corroborates that as well. Consequently, the introduction of heteroskedasticity and the spatio-temporal process works well for the data under study.

Next, we checked the autocorrelation function (ACF) for the residuals. For example, these plots for two locations (Guanyin and Tucheng) are displayed in Figure 3.12. It establishes that after fitting the model, the residuals do not show significant temporal dependence.

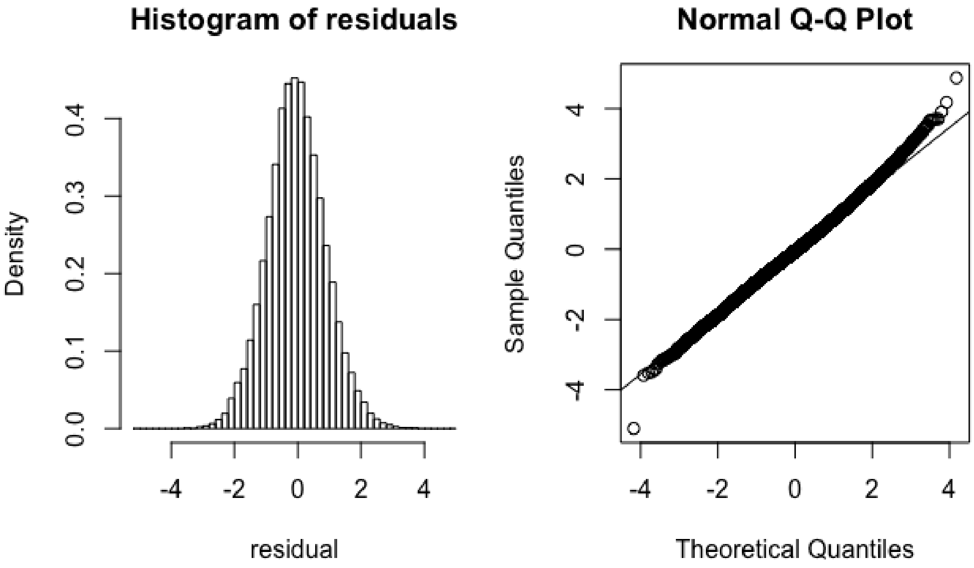

Figure 3.11: (Left) Histogram and (Right) QQ plot of the standardized residuals

There is a slight hint of autocorrelation at lags 12 and/or 24 (more clear for the right plot), and it suggests that the seasonal cycle might vary a bit with different spatial locations. This perhaps is another direction worth looking into in future.

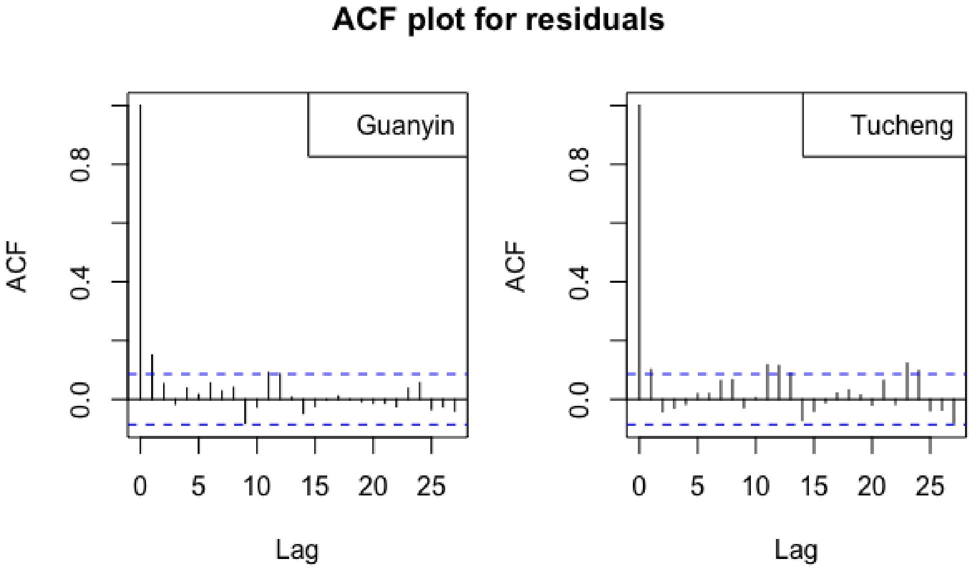

Figure 3.12: Autocorrelation function of the standardized residuals for two different locations

Finally, we check the prediction abilities of the proposed model, and we use cross validation techniques. For this, 90% of the available data (from 2006 to 2014, for all the stations) are used to train the model and then we predict the $PM_{2.5}$ levels for the year 2015 for all the stations. In order to evaluate how good the predictions are, we have calculated the root mean squared error for the transformed $PM_{2.5}$ concentrations, and it is approximately 1.712. In the original scale, the same was found out to be around 13.089 units. In order to understand the effect of the space-time interaction term in our model, we also measured the prediction abilities of the same model, but without the interaction component. In that case, the root mean squared error was found out to be approximately 2.448 in the transformed scale, and approximately 31.341 units in the original scale. To put it into perspective, in the transformed data, this is about 43% more (about 140% more in the original scale) than the root mean squared error for the proposed model.

3.5 Conclusion

In this chapter, we have developed a new spatio-temporal modeling technique, with an aim to identify the space-time interaction. The simulation studies and the data analysis confirm that the method performs well. In particular, we have showed that the estimates obtained by the proposed method are consistent, and have used standard diagnostic techniques to establish that the model assumptions are reasonable. This modeling technique can successfully detect and estimate the space-time interaction for air pollution data. Further, because of the weighting scheme we use in the method, it has the potential to predict higher level of pollution with more precision. This is going to be useful from a practical point of view.

We finish with some notes on future studies. An important potential future direction of this work is to consider a more generalized framework in the spatio-temporal process. In particular, we consider a separable structure for the spatial and temporal dependence, and that condition can be relaxed to address a more general setup. Moreover, in the aforementioned data analysis example, it was found that maximum wind speed at every location plays an important role in the air pollution, while there is significant space-time interaction effect as well. Combined, they raise an important question - how much effect does the wind flux, calculated from the wind speed, wind direction and the coordinates of different locations, have on the pollution? In order to address that, it will be necessary to know about the physical behavior of the wind, and that process can be incorporated in the model to develop more efficient techniques.

3.6 Appendix

Proof of Theorem 3.1. Note that the set-up of our problem is similar to a generalized least squares (GLS) problem, where $Y = X\theta + \varepsilon$, such that $\varepsilon \sim N(0, \sigma^2\Omega)$. Following our previous notations, $\Omega = (\Sigma_v + D)$, where D is a diagonal matrix with diagonal elements equal to some τ_j^2.

Now, for proving the required result, we define three different estimators of θ. Below, $\hat{\theta}$ is the estimator we are considering in this study, $\hat{\theta}_G$ denotes the usual GLS estimator, and $\hat{\theta}_F$ is a feasible GLS estimator.

$$\begin{aligned}
\hat{\theta} &= (X'\hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}X)^{-1}(X'\hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}Y) \\
\hat{\theta}_G &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y) \\
\hat{\theta}_F &= (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}Y)
\end{aligned}$$

In the above, W is the weight matrix as defined in Section 3.3.2 and $\hat{\Omega}$ is our estimate of the covariance matrix. For convenience, as use $N = nT$ hereafter. Following Baltagi [6, Chapter 9], we know that $\sqrt{N}(\hat{\theta}_G - \theta)$ and $\sqrt{N}(\hat{\theta}_F - \theta)$ have the same asymptotic distribution

$N(0, \sigma^2 Q^{-1})$, where $Q = \lim(X'\Omega^{-1}X/N)$ as $N \to \infty$, if $X'(\hat{\Omega}^{-1} - \Omega^{-1})X/N \xrightarrow{P} 0$ and $X'(\hat{\Omega}^{-1} - \Omega^{-1})\varepsilon/N \xrightarrow{P} 0$. Further, a sufficient condition for this to hold is that $\hat{\Omega}$ is a consistent estimator for Ω and that X has a satisfactory limiting behavior.

Let us now assume that the estimate $\hat{\tau}_j^2$ is consistent for τ_j^2, for all j. That would automatically ensure the consistency of $\hat{\Omega}$ and thereby we can conclude that $\hat{\theta}_F$ and $\hat{\theta}_G$ have same asymptotic distribution. Further, note that $X'\hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}X - X'\hat{\Omega}^{-1}X = X'\hat{\Omega}^{-1/2}(W - I)\hat{\Omega}^{-1/2}X$. Taking any appropriate norm (2-norm, for example) on both sides, we can argue that $\left\| X'\hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}X - X'\hat{\Omega}^{-1}X \right\| \to 0$ as $N \to \infty$, in view of the fact that $\|W - I\| = 2/\log N$, and that $\hat{\Omega}$ is a consistent estimator for Ω, the population covariance matrix. In a similar fashion, we can show that $\left\| X'\hat{\Omega}^{-1/2}W\hat{\Omega}^{-1/2}\varepsilon - X'\hat{\Omega}^{-1}\varepsilon \right\| \to 0$ as $N \to \infty$, and thus we can conclude that $\sqrt{N}(\hat{\theta} - \theta)$ and $\sqrt{N}(\hat{\theta}_F - \theta)$ have the same asymptotic distribution.

Clearly, all we need to prove is that $\hat{\tau}_j^2$ is a consistent estimator for τ_j^2 for all j. To this end, recall that $\hat{\tau}_j^2$ is the maximum likelihood estimator (MLE) of τ_j^2 for the problem $\hat{\varepsilon}_j \sim N(0, (\Sigma_v^{(j)} + \tau_j^2 I))$, where $\hat{\varepsilon}_j$ is the vector of scaled residuals corresponding to the jth season and $\Sigma_v^{(j)}$ is the submatrix of Σ_v corresponding to the same. It is well-known that MLE is a consistent estimator for such problems. Let n_j be the length of ε_j. Since $T \to \infty$, it is clear that the number of observations per season will also approach infinity, and thus $n_j \to \infty$. Hence, $\hat{\tau}_j^2$ is consistent for τ_j^2 and that ends our proof for the asymptotic normality of $\hat{\theta}$. The consistency result follows automatically from the above. □

CHAPTER 4

CLUSTERING OF TIME SERIES DATA

4.1 Background

The clustering of time series data is very important in many fields of scientific research, popular examples being medial science, climate studies and geographical scenarios. While there have been thousands of literature on the classification and clustering of independent point objects, the same, unfortunately, cannot be said when each observation is a sequence of dependent data.

A common way to deal with this problem is to extend the popular clustering techniques to the time series setup. A few early examples in this regard are the hierarchical clustering (Ward [96]), the k-means clustering (MacQueen [67]) and the EM algorithm on mixture models (Dempster et al. [24]).

In more recent studies, the paper by Smyth et al. [89] is one of the first works. Here, the authors considered hidden Markov models for classifying sequences and took a Monte-Carlo cross-validation approach to select the appropriate number of clusters. Kumar et al. [57] estimated the seasonality component of sales data and developed a clustering algorithm based on those estimates. Abraham et al. [1] proposed to first fit the functional data by B-splines and then apply the k-means algorithm on the fitted coefficients to find clusters.

On the other hand, for classification of time series data, one of the common practice in recent times is to use dimension reduction techniques, for example, the principal component analysis. The conventional methods are usually coupled with such techniques when the objects of interest are functions, which are intrinsically infinite-dimensional. In one such work, Hall et al. [42] proposed to focus on some summary statistics, quantities that are nonlinear functions of the data curves, to avoid the difficulty of interpreting higher-order principle components. Antoniadis et al. [3] used wavelet thresholding and Neyman truncation for dimension reduction and the EM algorithm on Gaussian mixture models for clustering. Some other contributions can be found in Chiou and Li [18], Díaz and Vilar [25], Garcí a Escudero and Gordaliza [34], Maharaj and D'Urso [68], Tarpey and Kinateder [91] and the references therein.

Further, Hu et al. [46] used subsequences of the time series objects and then used a nearest neighbor technique to classify the objects. Degras et al. [23] used the concept of parallelism to test if the trends of two time series are same. Then, using a simulation-based approximation to the distribution of the test statistic, they designed a clustering algorithm. Zhang [104] followed it up with another method, which was also based on the concept of parallelism. He used semi-parametric estimator for the trend of each of the time series and then employed a greedy algorithm, in conjunction with a Bayesian information criterion, to cluster the objects under consideration.

For the multivariate case, using appropriate discrimination criteria for the covariance structures of different time series, Kakizawa et al. [52] performed a cluster analysis. Yang and Shahabi [102] used PCA-based similarity measure in this regard. Singhal and Seborg [87] used two different similarity factors, one based on PCA and another based on Mahalanobis distance, to develop clustering algorithms for multivariate time series data. Few other notable works in this regard are Gao et al. [33], Izakian et al. [49], Liao [64] and Oates et al. [75]. The reader is further instructed to read Santos and Kern [85] to know more about different time series classification techniques.

In this work, we are going to use a distributional approach to develop a new and efficient algorithm. Our work revolves around the spectral density of the time series objects. The algorithm we develop is based on the distribution of the estimate of the $\mathbb{L}^2$ distance between two spectral densities.

This chapter is structured as follows. In Section 4.2, we provide our notations, preliminary assumptions and definitions. Then, in Section 4.3, the testing procedure is discussed and the clustering algorithm and related results are described in Section 4.4. A detailed simulation study and a real-life application are provided in Section 4.5, while some concluding remarks and future scope are provided in the final section.

4.2 Preliminaries

Throughout this chapter, $\mathbf{X}_i = \{X_{i1}, \ldots, X_{in}\}$ denotes a stationary time series. We use $\mathbf{X}_1, \ldots, \mathbf{X}_m$ to denote the set of m objects (each one is a time series) in the study. Here, so far as each time series is concerned, we shall consider a very general class of stationary time series which are functions of independent and identically distributed (iid) random variables, cf. Wu [99]:

Assumption 4.1: Let $\varepsilon_j, j \in \mathbb{Z}$, be iid random variables. Define $\mathbf{X} = \{X_1, \ldots, X_n\}$, satisfying

$$X_i = g(\varepsilon_{i-s}; s \in \mathbb{Z}), \qquad i \in \mathbb{Z}, \tag{4.1}$$

where g is a measurable function such that X_i is well-defined. Further, for a time series $\mathbf{X}$, γ_j is the autocovariance at lag j and for $0 \leq \theta < 2\pi$, $f(\theta) = \sum_{j \in \mathbb{Z}} \gamma_j e^{\imath j \theta}$ is the spectral density.

Now, before developing a clustering algorithm, it is a necessary and crucial step to decide on a measure that can reflect how similar two observations - in this case, two time series - are. As mentioned before, in this chapter, we propose to cluster the objects based on the covariance structures, rather than the means; and that motivates us to use the spectral densities of the time series. In the current study, we shall use the $\mathbb{L}_2$ distance between the logarithms of the spectral densities to classify the time series objects. The key idea of our clustering algorithm is to test if the spectral densities of the time series are same, and then, based on the p-value of the test, we put them in same or different clusters.

Definition 4.1: For two time series $\mathbf{X}_i$ and $\mathbf{X}_j$, let f_i and f_j denote the spectral densities. Then, the pairwise similarity measure is defined as

$$d(\mathbf{X}_i, \mathbf{X}_j) = \left[\int_0^{2\pi} \left\| \log f_i(\theta) - \log f_j(\theta) \right\|^2 d\theta \right]^{1/2}, \tag{4.2}$$

Remark 4: Our algorithm is based on the premises that two time series $\mathbf{X}_i, \mathbf{X}_j$ belong to the same cluster if and only if they have same spectral densities, and thus $d(\mathbf{X}_i, \mathbf{X}_j)$ has to be 0. Further, it is easy to note that the above-mentioned measure satisfies the three required properties for a distance metric: (i) $d(\mathbf{X}_i, \mathbf{X}_i) = 0$, (ii) $d(\mathbf{X}_i, \mathbf{X}_j) = d(\mathbf{X}_j, \mathbf{X}_i)$ and (iii) $d(\mathbf{X}_i, \mathbf{X}_k) \leq d(\mathbf{X}_i, \mathbf{X}_j) + d(\mathbf{X}_j, \mathbf{X}_k)$.

Having discussed the pairwise similarity measure that we will be using in this study, we proceed to define an estimate of the same. It is obvious that the estimate for $d(\mathbf{X}_i, \mathbf{X}_j)$ will involve estimates of the spectral density function for the two time series objects. One common way of estimating the spectral density is to use the periodogram. Recall that, for $\mathbf{X}_i$, the periodogram $I_n^{\mathbf{X}_i}(\theta)$ is defined as

$$I_n^{\mathbf{X}_i}(\theta) = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} X_{ij} X_{ik} \exp\{\imath (j-k)\theta\}. \tag{4.3}$$

It is known that an appropriately scaled periodogram is asymptotically unbiased, but not a consistent estimate of the spectral density. Moreover, we know that for different Fourier frequencies, the scaled periodograms are asymptotically independent and follow exponential distribution with parameter equal to the values of the spectral densities at those points. Our testing procedure and subsequently the clustering algorithm will rely on this result, as discussed in the following sections.

4.3 Testing procedure

4.3.1 Developing the test statistic

As the first step towards developing a clustering algorithm, we need to define an estimate for the distance measure, and we can do that using smoothed periodograms for the time series objects. The periodogram is not a consistent estimate of the spectral density, and is a wildly fluctuating estimate of the spectrum with high variance. Thus, in order to obtain a stable estimate, the periodogram must be smoothed. In short, the smoothed periodogram $\hat{f}(\theta)$ can be calculated by $\int I_n(\lambda)K((\lambda-\theta)/B_n)d\lambda$, where $I_n(\lambda)$ is the value of the actual periodogram at λ, $K(\cdot)$ is a kernel function and B_n is a bandwidth sequence satisfying $B_n \to \infty$, $B_n/n \to 0$ as $n \to \infty$. The kernel function is considered to be even, defined on $[-1,1]$, and should satisfy the conditions $\int K(\lambda)d\lambda = 1$, $\int K^2(\lambda)d\lambda < \infty$. Refer to Bloomfield [14] for a detailed discussion on this.

In this work, we would use the rectangular kernel as the smoothing filter for generating an estimated spectrum from the periodogram. This corresponds to the kernel function $K(u) = \mathbb{I}\{-1 \leq u \leq 1\}$.

Using the above idea, we can now define an estimate for the pairwise similarity measure.

Definition 4.2: Suppose $\mathbf{X}_i$ is a time series of length n. Let $\hat{g}_{i,n}(\theta)$ be the logarithm of the smoothed periodogram for $\mathbf{X}_i$. First, define an estimate of the similarity measure between $\mathbf{X}_i$ and a given spectral density $f_0(\theta)$ as

$$\hat{d}(\mathbf{X}_i, f_0)^2 = \int_0^{2\pi} \left\| \hat{g}_{i,n}(\theta) - \log f_0(\theta) \right\|^2 d\theta. \tag{4.4}$$

And then, for two series $\mathbf{X}_i, \mathbf{X}_j$, an estimate for the measure $d(\mathbf{X}_i\mathbf{X}_j)$ [cf. Equation (4.2)] is defined by

$$\hat{d}(\mathbf{X}_i, \mathbf{X}_j)^2 = \int_0^{2\pi} \left\| \hat{g}_{i,n}(\theta) - \hat{g}_{j,n}(\theta) \right\|^2 d\theta. \tag{4.5}$$

We can use the above to test if the two spectral densities are the same. Now, we propose another statistic to test if a group of time series objects are in the same cluster $\mathcal{C}$, that is, if they have the same spectral density $f_{\mathcal{C}}(\theta)$. First, for $\mathcal{C} = \{\mathbf{X}_1, \ldots, \mathbf{X}_r\}$, a natural estimate for the logarithm of the common spectral density is developed using the average of the logartthms of the estimated spectral densities. Let $\hat{g}_{\mathcal{C},n}(\theta) = (1/r)\sum_{i=1}^{r} \hat{g}_{i,n}(\theta)$, where $\hat{g}_{i,n}(\theta)$ is as defined above. And then, the following gives the required statistic to test if $\mathbf{X}_1, \ldots, \mathbf{X}_r$ have the same spectral density.

Definition 4.3: Suppose $\mathbf{X}_i$, $i = 1, \ldots, r$, are time series objects under consideration, and let $\hat{g}_{i,n}(\theta)$, $\hat{g}_{\mathcal{C},n}(\theta)$ be as above. Then, define the following measure to test if the objects belong to same cluster:

$$\hat{d}(\mathcal{C})^2 = r \sum_{i=1}^{r} \int_0^{2\pi} \left\| \hat{g}_{i,n}(\theta) - \hat{g}_{\mathcal{C},n}(\theta) \right\|^2 d\theta. \tag{4.6}$$

It is easy to note that for $r = 2$, if we consider $\mathcal{C} = \{\mathbf{X}_1, \mathbf{X}_2\}$, then the estimate $\hat{d}(\mathcal{C})^2$ is exactly equal to $\hat{d}(\mathbf{X}_1, \mathbf{X}_2)^2$, as defined earlier. Thus, the notations and definitions are consistent. Further, one can think of the above measure as an estimate for

$$d(\mathcal{C})^2 = r \sum_{i=1}^{r} \int_0^{2\pi} \left\| \log f_i(\theta) - \frac{1}{r} \sum_{j=1}^{r} \log f_j(\theta) \right\|^2 d\theta.$$

Under the null hypothesis that all time series in $\mathcal{C} = \{\mathbf{X}_1, \ldots, \mathbf{X}_r\}$ have the same spectral density, $d(\mathcal{C})^2 = 0$, and thus, we should expect $\hat{d}(\mathcal{C})^2$ to be small. In fact, we can show the following.

Theorem 4.1: Let $\mathcal{C}$ be a set of time series $\{\mathbf{X}_1, \ldots, \mathbf{X}_r\}$, all coming from a process with true spectral density $f(\theta)$, such that $\inf_\theta |f(\theta)| \geq \delta$, for some $\delta > 0$. Then, $\hat{d}(\mathcal{C})^2 \to 0$, as the number of observations in each time series (denoted as n) goes to infinity.

Proof. It is known that the smoothed periodogram estimate $\hat{f}_i(\theta)$ is a consistent estimate for $f(\theta)$. In fact, following similar idea as in Liu and Wu [66], we can argue that for all $1 \leq i \leq r$, $\sup_\theta \|\hat{f}_i(\theta) - f(\theta)\| \to 0$, as $n \to \infty$.

Now, using Taylor series expansion for $\hat{g}_{i,n}(\theta) = \log \hat{f}_i(\theta)$ and following the assumption above, we can say that, as $n \to \infty$, $\sup_\theta \|\hat{g}_{i,n}(\theta) - g(\theta)\| \to 0$. Here $g(\theta) = \log f(\theta)$. This in turn implies that $\sup_\theta \|\hat{g}_{i,n}(\theta) - \hat{g}_{\mathcal{C},n}(\theta)\| \to 0$. Thus, for all $1 \leq i \leq r$, as $n \to \infty$,

$$\int_0^{2\pi} \left\|\hat{g}_{i,n}(\theta) - \hat{g}_{\mathcal{C},n}(\theta)\right\|^2 d\theta \to 0.$$

This completes the proof. □

Hence, one should reject the aforementioned null hypothesis when a large value of $\hat{d}(\mathcal{C})^2$ is observed. Now, in order to perform this test, as mentioned before, we employ the idea of simulation based approximation to the distribution of the test statistic. Degras et al. [23] discussed that this type of procedure has better finite sample properties and hence can be used efficiently to formulate the clustering algorithm.

For the simulation-based method, we start with the idea that for the set $\mathcal{C} = \{\mathbf{X}_1, \ldots, \mathbf{X}_r\}$, $\log(I_n^{\mathbf{X}_i}(\theta)/2\pi)$ [see Equation (4.3)] can be approximated as $\log f_i(\theta) + \log Z_i$ where Z_i, for $i = 1, \ldots, r$, are independent and identically distributed Exponential(1) random variables. Under the null hypothesis $H_0 : d(\mathcal{C})^2 = 0$, which corresponds to $f_i(\theta) = f_j(\theta)$ for all $i, j \in \{1, \ldots, r\}$, the distribution of $\hat{d}(\mathcal{C})^2$ can be assessed by simulating $d^*(\mathcal{C})^2$, which is computed using the values of $(\log Z_i - \log Z_j)$. Specifically, one can generate many (throughout this study, we would use $M = 10000$) realizations of Z_i's and can compute $d^*(\mathcal{C})^2$ from each such set of realizations. Let us denote these simulated values by $d_1^*(\mathcal{C})^2, \ldots, d_M^*(\mathcal{C})^2$. Then, the empirical p-value for test can be evaluated as

$$\hat{p} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\{d_1^*(\mathcal{C})^2 \geq \hat{d}(\mathcal{C})^2\}. \tag{4.7}$$

We reject the null hypothesis when $\hat{p}$ is less than the level of significance α. Alternately, one can compute the $(1-\alpha)$th quantile q_α from the simulated values $d^*(\mathcal{C})^2$, and can reject H_0 if $\hat{d}(\mathcal{C}) > q_\alpha$. Unless otherwise specified, we have used $\alpha = 0.05$ everywhere in this chapter.

4.3.2 Bandwidth selection

As it has been established in many literature, choice of kernel $K(\cdot)$ is generally only of second importance to the choice of the bandwidth function B_n. There have been many studies who focus on finding the most appropriate bandwidth function for smoothing the periodogram. Here, we follow the idea laid down by Ombao et al. [76] and use the cross-validation techniques. Earlier works of BeltraTo and Bloomfield [9] and Lee [61] are also noteworthy in this regard.

To evaluate the dependence on the bandwidth, for $i = 1, \ldots, r$, let us denote the smoothed spectral density estimate for $\mathbf{X}_i$ by $\hat{f}_i(\theta; B_n)$. We consider the values of $\hat{f}_i(\theta; B_n)$ and that of the periodogram $I_n^{\mathbf{X}_i}(\theta)$ at different Fourier frequencies $\theta_j = 2\pi j/n$, for $j = 0, \ldots, n-1$. Let $n_0 = [n/2]+1$, where $[\cdot]$ denotes the floor function. Now, the generalized cross validation deviance function is defined by

$$\mathrm{GCV}(B_n) = \frac{1}{r} \sum_{i=1}^{r} \frac{n_0^{-1} \sum_{j=0}^{n_0-1} D(I_n^{\mathbf{X}_i}(\theta_j), \hat{f}_i(\theta_j; B_n))}{(1 - K_{B_n}(0))^2}. \tag{4.8}$$

In the above equation, $K_{B_n}(0)$ is the weight at the center of the smoothing window. $D(\cdot,\cdot)$

is the deviance function and an appropriate choice is

$$D(I_n^{\mathbf{X}_i}(\theta_j), \hat{f}_i(\theta_j; B_n)) = q_j[-\log\{I_n^{\mathbf{X}_i}(\theta_j)/\hat{f}_i(\theta_j; B_n)\} + I_n^{\mathbf{X}_i}(\theta_j)/\hat{f}_i(\theta_j; B_n) - 1],$$

where $q_j = 1 - 0.5\mathbb{I}\{j = 0 \text{ or } j = n_0 - 1\}$. This scheme of assigning different weights accounts for the fact that the distribution of the periodogram is different for these two cases.

Finally, for different choices of B_n, one can evaluate $\text{GCV}(B_n)$ and the optimal bandwidth is the choice that minimizes this quantity.

4.4 Clustering algorithms

In this section, we propose two clustering procedures for a set of time series objects $S = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$. First one works under the assumption that the number of clusters (k) is known, and it is similar to the idea of classical k-means clustering. The second algorithm is more practical, and works when k is unknown. Both of the methods would make use of the test statistic $\hat{d}(\mathcal{C})$ for different subsets $\mathcal{C} \subset S$.

4.4.1 When the number of clusters is known

In this scenario, one can use the idea of k-means clustering. Specifically, we can find the farthest centers at first and then will assign each of the remaining time series objects to the center closest to it. All of these distance measures are calculated using the test statistics we mentioned in the previous section. Below, we present a short description of the algorithm and then it is formally written as Algorithm 1.

The algorithm starts with assigning the sample $\mathbf{X}_{j_1}$ as the first cluster centre. Then, running through the rest of the points, another sample $\mathbf{X}_{j_2}$ is found which is farthest away from $\mathbf{X}_{j_1}$, that is the value of $\hat{d}(\mathbf{X}_{j_1}, \mathbf{X}_i)$ is the most for $i = j_2$. This point is then set as the second cluster center. Next, we find the third center $\mathbf{X}_{j_3}$ by maximizing the minimum distance of $\mathbf{X}_i$ from the existing cluster centers. This procedure is repeated until we have k centers. And then, for the remaining points, we find their distance from each of the clusters and assign them to the one showing minimum overall distance.

Algorithm 1 Known number of clusters

```
input: $m$ time series $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$, number of clusters $k$.
output: partition of the set $\{1, \ldots, m\}$ into $k$ clusters.
1. initialize $S = \{1, \ldots, m\}$, $S_{\text{rest}} = \phi$, $j_1 = 1$;
2. set $C = \{C_1, \ldots, C_k\}$, $C_1 = \{j_1\}$, $C_i = \phi$ for $i \neq 1$;
3. for $i = 2, \ldots, k$
       - find $j_i = \underset{r \notin \cup_{l=1}^{i-1} C_l}{\text{argmax}} \left\{ \min_{l \in C_s; 1 \leq s \leq i-1} \hat{d}(\mathbf{X}_l, \mathbf{X}_r)^2 \right\}$;
       - set $C_i = \{j_i\}$;
4. set $S_{\text{rest}} = S \backslash (\cup_{i=1}^{k} C_i)$;
5. for $r \in S_{\text{rest}}$
       - set $C_{j,r} = C_j \cup \{r\}$ for $j = 1, \ldots, k$;
       - find $i = \text{argmin } \hat{d}\left(\{\mathbf{X}_l : l \in C_{j,r}\}\right)^2$;
       - set $C_i = C_i \cup \{\mathbf{X}_r\}$;
6. return $C$ as the final set of clusters.
```

We now define a notion of consistency for the clustering algorithm, and would prove that the above algorithm is consistent. This definition and the subsequent discussion are inspired from the work of Khaleghi et al. [54].

Definition 4.4: Let h be a clustering algorithm, that takes the set of observations S and for a given k (number of clusters) produces a partition $\mathcal{C}$ of length k. We denote it as $h(S, k) = \mathcal{C}$. Let $\mathcal{B}$ be the ground truth, that is the actual clustering of the set S. Then, we say that h is consistent if $P(h(S, k) = \mathcal{B})$ goes to 1, as the number of points in each time series object goes to infinity.

Theorem 4.2: Suppose that S, a set of m time series objects, each of length n, are coming from k (known) different processes with spectral densities $f_1, \ldots, f_k$. Let $d(f_i, f_j) =$

$\int \left\|\log f_i(\theta) - \log f_j(\theta)\right\|^2 d\theta$ denote the distance between the spectral densities, and assume that $\min_{i \neq j} d(f_i, f_j)$ is greater than some $\delta > 0$. Then, Algorithm 1 is consistent, in the sense of Definition 4.4.

Proof. The proof works with the idea that the time series objects coming from same process are closer to each other than to the rest of the series in the data.

In order to show that, let us use $\mathcal{B}_1, \ldots, \mathcal{B}_k$ to denote the true clusters, corresponding to the k spectral densities $f_1, \ldots, f_k$. Following the discussion in the previous section [cf. Theorem 4.1], we can say that, for any r, for any subset $\mathcal{C} \subset \mathcal{B}_r$, $\hat{d}(\mathcal{C})^2$ goes to 0, as the number of points in each series goes to infinity. It further tells us that for some $\epsilon > 0$, $\sup \hat{d}(\mathcal{C})^2 \leq \epsilon$, where the supremum is taken over all $\mathcal{C} \in \mathcal{B}_r$, $r \in \{1, \ldots, k\}$. In a similar way, it can be shown that $\inf \hat{d}(\mathcal{C})^2 > \delta - \epsilon$, where the infimum is taken over all $\mathcal{C}$, such that $\mathcal{C} \subset S$, but $\mathcal{C} \not\subset \mathcal{B}_r$ for any $r \in \{1, \ldots, k\}$.

Since the above is true for any fixed ϵ, we can choose $\epsilon < \delta$. This ensures that the cluster centers obtained from the algorithm correspond to different spectral densities. Further, using the same idea outlined above, we can argue that the rest of the time series objects in the data will be clustered with the center that corresponds to the true spectral density of each object. Hence, the algorithm above will be able to identify the true clustering for sufficiently large samples. And that proves the theorem. □

4.4.2 When the number of clusters is unknown

One way to deal with this case is to implement the above algorithm for different values of k and then choose the one for which the value of $\sum \hat{d}(\mathcal{C}_i)^2$ is minimum. However, that would involve huge computational burden. That is why we propose to use a different algorithm when the number of clusters is unknown.

This iterative algorithm would start with the whole set and would use $\hat{d}(S)^2$ to test if all of the objects have same spectral density. The aim will be to find the maximal subset of S for which the test will not reject the null hypothesis. Thus, if the test on S is rejected, we would remove the object showing maximum deviation from others, and would perform the test again on the resulting subset. Note that, for a set S, the deviation for $\mathbf{X}_i$ is calculated by

$$D_i^2 = |S| \int_0^{2\pi} \left\| \hat{g}_{i,n}(\theta) - \hat{g}_{S,n}(\theta) \right\|^2 d\theta. \tag{4.9}$$

At this point, it is worth mention a lot of iterations might be required if we have to remove one object at a time and rerun the test on the resulting set, time and again. Thus, we propose to remove multiple objects at each step. Experimentally, we found that the results are not affected if we choose to remove approximately $n/20$ objects at each step, and so, we would follow that in the algorithm. Now, this procedure is done sequentially until the test is accepted; which would give us our first cluster C_1. These same steps are then repeated for $S_{\text{rest}} = S \backslash C_1$, and we continue the iterations until all objects are put in some cluster.

The procedure is formally presented as Algorithm 2. Here, $|\cdot|$ denotes the cardinality of a set, and $[\cdot]$ denotes the floor function.

Note that the algorithm has the attractive property that it does not fix the number of clusters beforehand. In general, this method would compute the similarity measure for the subsets and then put similar time series objects in one cluster. However, for some examples, it is always possible that the m time series are coming from processes with different spectral densities, and thus will belong to m different clusters. In such cases, this approach has a decided advantage over the methods which fix the number of clusters in advance.

Algorithm 2 Unknown number of clusters

```
input: $m$ time series $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$, significance level $\alpha$.
output: a partition of the set $\{1, \ldots, m\}$.
1. initialize $C$ as an empty list, $k = 0$;
2. initialize $S = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$, $S_{\text{rest}} = \phi$;
3. compute $\hat{d}(S)^2$ from the data and perform the simulation based test to obtain the empirical $p$-value $\hat{p}$.
3. Check,
    (a) If $\hat{p} < \alpha$:
        – find $D_i^2$ for all $i \in S$ [see (Equation (4.9))], sort them in a decreasing sequence as $D_{k_1}^2, D_{k_2}^2, \ldots$;
        – set $L = \max\{1, [|S| /10]\}$;
        – set $S = S \backslash \{\mathbf{X}_{k_1}, \ldots, \mathbf{X}_{k_L}\}$;
        – set $S_{\text{rest}} = S_{\text{rest}} \cup \{\mathbf{X}_{k_1}, \ldots, \mathbf{X}_{k_L}\}$;
        – return to step 3;
    (b) else if $\hat{p} \geq \alpha$:
        – set $k = k + 1$, $C_k = S$;
        – If $|S_{\text{rest}}| > 0$, set $S = S_{\text{rest}}$, return to step 3;
        – else, return $C$ as the list of clusters..
```

4.5 Results

4.5.1 Accuracy of the testing procedure

First, we present some simulation results on the testing procedure. For different values of m, time series observations $\mathbf{X}_1, \ldots, \mathbf{X}_m$ are generated and then $\hat{d}(\mathcal{C})^2$ [cf. Equation (4.6)], where $\mathcal{C} = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$, is used to test if the time series have same spectral densities.

In order to estimate the type I error, we simulate time series from same auto-regressive moving-average (ARMA) processes, with appropriate auto-regressive and moving-average coefficients, such that the stationarity assumption is satisfied. And then, we test if the null hypothesis is rejected or not. In this study, the orders for the ARMA process are taken to be $p = 2$, $q = 3$, with the AR coefficients $(0.89, -0.48)$, the MA coefficients $(-0.23, 0.25, 0.18)$, and the innovations are generated from a standard normal distribution. This experiment was repeated 1000 times and then we evaluated type I error using the proportion of cases the test rejected.

Next, we evaluate the power of the test procedure. For that, we consider two different ARMA models. One is same as the above, while the other is an ARMA(1,2) model with AR coefficient (-0.34), MA coefficients $(0.61, -0.45)$, and the innovations were generated from a t distribution with degrees of freedom 5. Then, half of the time series in $\mathcal{C}$ are generated from either of these two models. Following that, the testing procedure is performed on the generated data and the power is computed using the proportion of times (among 1000 experiments) the test rejected the null hypothesis. These results are presented in Table 4.1.

Table 4.1: Simulation results for testing groups of ARMA models for different values of m (number of time series) and n (number of observations per series). $\alpha = 0.05$ was used in all cases.

	Type I error				
$n \backslash m$	2	6	10	20	50
50	0	0	0	0	0
100	0	0	0	0	0
300	0	0	0	0	0
500	0	0	0	0	0
	Power				
$n \backslash m$	2	6	10	20	50
50	0.493	0.607	0.793	0.988	1
100	0.589	0.714	0.877	0.999	1
300	0.735	0.761	0.978	1	1
500	0.807	0.816	0.994	1	1

Overall, the results showed that the test procedure works very well. It was interesting that there was no false positive in any of the simulation studies. Further, for large number of observations, the power showed incredible perfection, when 50% of the observations are coming from each of the two models. Motivated by this, we further explore how the test performs when the frequencies of the two models in the data are different. For $\phi = 0.05$ to 0.95, ϕ proportion of a set of $m = 50$ time series objects are taken from the first model while the rest are taken from the other model. Then, for different values of n (number of observations per series), we empirically evaluate the power of the test based on 1000 experiments. These results are displayed in Figure 4.1.

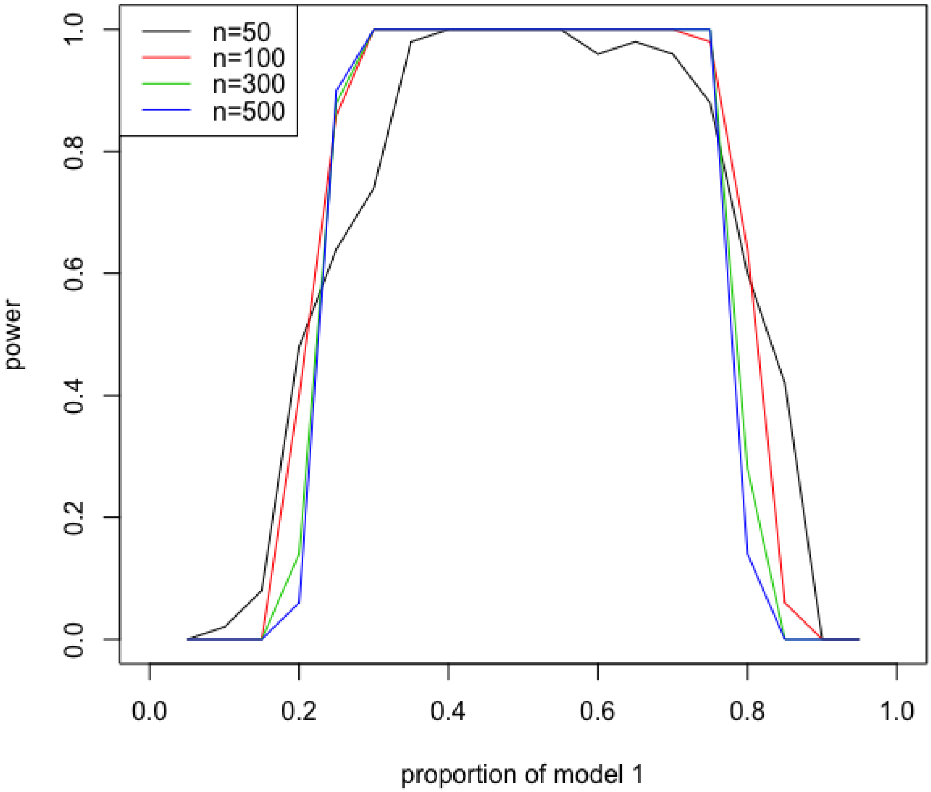

Figure 4.1: Empirical power of the testing procedure, corresponding to proportion of time series coming from model 1. n denotes the number of observations per series and a total of 50 series is used.

It can be seen that the performance improves at first, and as expected, reaches its peak when 50% series are generated from either model. However, even with ϕ close to 0.25, the power is nearly 1 for $n \geq 100$, showing that as the number of observations grows, the test procedure is able to detect any non-similarity in the groups of time series observations.

4.5.2 Accuracy of the clustering algorithms

As a last piece of this simulation study, we want to see how the clustering algorithms perform. In order to evaluate that, we compute the *purity index*, as defined below.

Definition 4.5: 'Purity index' is an external evaluation criterion of cluster quality. It calculates the percent of the total number of objects (time series observations in our case) that were classified correctly. Suppose, there are N objects coming from k clusters $\mathcal{B} = \{\mathcal{B}_1, \ldots, \mathcal{B}_k\}$ and the algorithm returns classes $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_l\}$. Then, define

$$\text{Purity} = \frac{1}{l} \sum_{i=1}^{k} \frac{1}{|\mathcal{C}_i|} \max_j \left|\mathcal{C}_i \cap \mathcal{S}_j\right| \tag{4.10}$$

It is easy to observe that the index lies in the interval $[0, 1]$. It will be 1 if every object is classified correctly, that is, if there is no pair $(\mathbf{X}_r, \mathbf{X}_s)$ that is in same cluster in $\mathcal{C}$, but are in different classes in $\mathcal{S}$. On the other hand, in the worst possible classification, the value of the index will be 0.

Performance evaluation for ARMA models:

In the simulation study for evaluating the performance of the algorithms, we again use different ARMA(p, q) models. A set of m time series, each of length 300, are generated from k (which is the actual number of clusters) different ARMA models, each equally likely. In Table 4.2, the set of AR and MA coefficients and the innovation distributions for these models (denoted as $M_1, \ldots, M_5$) are displayed.

Then, we apply both of the aforementioned clustering algorithms on the set of simulated data and evaluate the accuracy of the method by finding out the purity index. This is done for different m and k and the results are presented in Table 4.3 below.

We find that the performance is near perfect so far as the binary classification (2 classes) case

Table 4.2: List of ARMA(p, q) models $(M_1, \ldots, M_5)$ used for the simulation study. Here, ϵ denotes the innovation process and Z stands for the standard normal distribution.

	AR coef	MA coef	ϵ
M_1	$(0.89, -0.48)$	$(-0.23, 0.25, 0.18)$	t_5
M_2	(-0.34)	-	Z
M_3	-	$(0.57, -0.9, -0.8)$	t_{10}
M_4	$(-0.1, 0.2, 0.45)$	$(-0.21, -0.37)$	t_{50}
M_5	0.7	$(0.61, -0.45)$	Z

Table 4.3: Mean of the Purity index values for the clustering algorithms, for different values of m (number of time series) and k (number of true clusters). Each time series is generated from one of the models in Table 4.2 and has 300 observations.

	Algorithm 1 (k known)			
$k \backslash m$	20	50	100	200
2	0.991	0.992	0.995	0.995
3	0.889	0.916	0.930	0.962
4	0.805	0.819	0.851	0.872
5	0.727	0.759	0.778	0.795
	Algorithm 2 (k unknown)			
$k \backslash m$	20	50	100	200
2	0.835	0.851	0.876	0.881
3	0.791	0.834	0.853	0.864
4	0.693	0.718	0.747	0.759
5	0.654	0.670	0.669	0.692

is concerned. In general, the purity index increases as the average number of observations in each group increases. Especially, if we know the true number of clusters, the performance is very good, and usually remains above 0.85. For Algorithm 2, the mean purity index is close to 0.9 when the ratio m/k is 70 or more. Thus, empirically, we can argue that both the algorithms perform brilliantly when the average number of observations per group is large. Another interesting thing is that these results are obtained for time series of moderate length, but the performance improves further if there are more observations in each series.

Performance evaluation for GARCH models:

Next, in order to understand how the algorithms perform across different situations, we considered generalized autoregressive conditional heteroskedastic (GARCH) models under different settings. Autoregressive conditional heteroskedastic (ARCH) models were first introduced by Engle [29], and ARCH/GARCH models are extremely useful in the context of applied econometrics (cf. Engle [28]).

In this part, we consider five different ARCH/GARCH models, as discussed in Table 4.4 below. There, α and β coefficients denote the parameters for the GARCH model. The autoregressive coefficients are denoted using AR while the other parameters (δ, μ, shape, skew) are as used by Wuertz et al. [101] in the fGarch package in R. On the other hand, the last column represents the conditional distribution in generating the time series. Here, GED, STD, Z and SSTD denote generalized error distribution, standard student's t distribution, standard normal distribution and skewed student's t distribution, respectively. More details about the properties and use of the models used in this simulation study can be found in Fernández and Steel [31], Hentschel [45], Laurent [59] and Nelson [74].

Table 4.4: List of GARCH models ($M_1, \ldots, M_5$) used for the simulation study.

	Description	Parameters	Distribution
M_1	ARCH(2)	$\alpha = (0.1, 0.3)$	GED
M_2	AR(5)-GARCH(2,1)	$\alpha = (0.1, 0.1), \beta = 0.75$ $\text{AR} = (0.5, 0, 0, 0, 0.1)$	STD
M_3	Taylor Schwert GARCH(1,1)	$\alpha = 0.1, \beta = 0.1$ $\delta = 1$	Z
M_4	GARCH(3,2)	$\alpha = (0.1, 0.1, 0.2), \beta = (0.1, 0.4)$	Z
M_5	AR(1)-t-APARCH(2, 1)	$\alpha = (0.10, 0.05), \beta = 0.8$ $\delta = 1.8, \mu = 0.0001$ $\text{AR} = 0.5, \text{shape} = 4, \text{skew} = 0.85$	SSTD

Similar to the previous simulation study, for different values of k (number of true clusters), m time series data were generated from the above five models, each equally likely. Then, both algorithms are implemented on the data. The results are presented in Table 4.5.

Table 4.5: Mean of the Purity index values for the clustering algorithms, for different values of m (number of time series) and k (number of true clusters). Each time series is generated from one of the models from Table 4.4 and has 300 observations.

	Algorithm 1 (k known)		
$k\backslash m$	50	100	200
2	0.930	0.951	0.964
3	0.932	0.957	0.966
4	0.819	0.838	0.857
5	0.766	0.786	0.795
	Algorithm 2 (k unknown)		
$k\backslash m$	50	100	200
2	0.938	0.949	0.958
3	0.959	0.961	0.973
4	0.853	0.861	0.870
5	0.757	0.762	0.782

We see that the performance of the algorithms are very good, similar to the previous case. The mean purity index values are more than 0.9 when number of classes is less and they increase steadily as the number of series increases. An interesting observation is that the algorithm with unknown number of clusters is actually performing better than the first algorithm when the number of classes is less. Also, the accuracy is a bit better than the ARMA models. Both of these phenomena are possibly due to the fact that the models under this study are more different from each other in nature, than the previous scenario.

4.5.3 Analyzing real data

As a real-life example, we analyze a data obtained from www.timeseriesclassification.com, an open source website.

The data we consider has 258 time series observations, each corresponding to an individual from the *Caenorhabditis elegans* species, a roundworm commonly studied as a model organism in the study of behavioral genetics. Detailed discussion on these worms and related time series classification methods can be found in Bagnall et al. [5], Brown et al. [15] and Yemini et al. [103]. For our purpose, a quick description on the data is warranted at this point.

It has been shown that the space of shapes this organism adopts can be represented by combinations of four base shapes, known as eigenworms. After extracting the outline of the worm, each frame of its motion is captured by four scalars representing the amplitudes along the dimensions when the shape is projected onto the four eigenworms. Then, the main purpose of the study is to classify individual worms based on the time series of the first eigenworm, down-sampled to second-long intervals, as either wild-type or one of four mutant types: goa-1; unc-1; unc-38 and unc-63. For each individual under consideration, there is a time series of length 900.

First, we test for stationarity using Box-Ljung test for all time series observations. It was found that the stationarity assumption holds good for all series in the data. Then, using the information that there are 5 possible clusters we employ Algorithm 1 with $k = 5$. In Table 4.6, the sizes of the estimated clusters are shown.

Table 4.6: Results for the Worms data with known number of clusters (Algorithm 1).

Cluster number	1	2	3	4	5
No. of observations	76	72	57	43	10

Next, we implemented Algorithm 2 on the data. In Table 4.7, we present the size of the clusters, along with the optimal bandwidth function and the significance level of the clusters.

It can be observed that nearly 96.1% of the time series objects are assigned to the first five clusters, which is in line with the theory that there are five true clusters in the data. Moreover, the significance levels tell us about how close the time series are, and that is an added advantage of our algorithm.

On the other hand, we compared the performance of the first algorithm to the second, and wanted to see how close the results are. The purity index [cf. Equation (4.10)] with the results of the two algorithms taken together, was approximately 0.742, showing that the

Table 4.7: Clustering results for the Worms data with unknown number of clusters (Algorithm 2). Bandwidth value d corresponds to the function $B_n = n^d$.

Cluster	Size	Bandwidth	Significance
1	125	0.30	0.050
2	61	0.35	0.063
3	36	0.30	0.084
4	16	0.30	0.060
5	10	0.35	0.073
6	2	0.35	0.434
7	2	0.15	1
8	2	0.20	1
9 to 12	1	-	-

performances are not too different for the two methods.

4.6 Concluding remarks

In this study, we have presented testing procedures to identify the similarity of pair or groups of time series objects, based on their spectral densities. Simulation studies show that the methods work very well. Moreover, it has nice finite sample properties, as we use simulation based approximations. This testing method is of independent interest as well, and can be used in other statistical problems.

On the other hand, as the main contribution of this work, we have developed two clustering algorithms for different setups. It has been established, both theoretically and empirically, that the efficiency of the algorithm when the number of clusters is known is very good and thus, can be used in real life problems, where one knows the number of possible classes. A particular example is the one with the worms data we have discussed here. The other method, in contrast, can be used when there is no a priori information on the possible number of clusters. In Section 4.5, using simulated data and with the real data, we showed that the performance of this algorithm is good. This method has two attractive features in particular. First, since it does not fix the number of clusters, it allows atypical time series to be set apart and hence, can be used across various setups, where the objects need not be coming from similar processes. Second, it readily provides the significance level of the clusters, and thus, it gives a quantitative sense of how similar the clusters are. Finally, the implementation of both algorithms are simple and that is an added advantage as well.

A primary future direction is to consider dependence among the time series observations in the study. Throughout, we assumed that the objects are independent and our methods are developed based on that. However, as a future endeavor, we want to relax the assumption and deal with a broader class of problems. Another interesting direction would be to extend the algorithms to non-stationary and multivariate time series data as well.

REFERENCES

[1] C. Abraham, P. A. Cornillon, E. Matzner-Lø ber, and N. Molinari. Unsupervised curve clustering using B-splines. *Scand. J. Statist.*, 30(3):581–595, 2003. ISSN 0303-6898. doi: 10.1111/1467-9469.00350.

[2] Fahimah A Al-Awadhi and Shafiqah A Al-Awadhi. Spatial-temporal model for ambient air pollutants in the state of kuwait. *Environmetrics*, 17(7):739–752, 2006.

[3] Anestis Antoniadis, Jérémie Bigot, and Rainer von Sachs. A multiscale approach for statistical characterization of functional images. *J. Comput. Graph. Statist.*, 18(1): 216–237, 2009. ISSN 1061-8600. doi: 10.1198/jcgs.2009.0013.

[4] Théophile T Azomahou. Memory properties and aggregation of spatial autoregressive models. *Journal of Statistical Planning and Inference*, 139(8):2581–2597, 2009.

[5] Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. Time-series classification with cote: the collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2522–2535, 2015.

[6] Badi H. Baltagi. *Econometrics*. Springer Texts in Business and Economics. Springer, Heidelberg, fifth edition, 2011. doi: 10.1007/978-3-642-20059-5.

[7] S. Bandyopadhyay and S. N. Lahiri. Asymptotic properties of discrete Fourier transforms for spatial data. *Sankhyā*, 71(2, Ser. A):221–259, 2009. ISSN 0972-7671.

[8] Soutir Bandyopadhyay, Soumendra N. Lahiri, and Daniel J. Nordman. A frequency domain empirical likelihood method for irregularly spaced spatial data. *Ann. Statist.*, 43(2):519–545, 2015. ISSN 0090-5364. doi: 10.1214/14-AOS1291.

[9] Kaizô I BeltraTo and Peter Bloomfield. Determining the bandwidth of a kernel spectrum estimate. *Journal of time series analysis*, 8(1):21–38, 1987.

[10] Veronica J Berrocal, Alan E Gelfand, and David M Holland. A bivariate space-time downscaler under space and time misalignment. *The annals of applied statistics*, 4(4): 1942, 2010.

[11] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, 36:192–236, 1974. ISSN 0035-9246. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

[12] Richard A Bilonick. The space-time distribution of sulfate deposition in the northeastern united states. *Atmospheric Environment (1967)*, 19(11):1829–1845, 1985.

[13] Marta Blangiardo, Francesco Finazzi, and Michela Cameletti. Two-stage bayesian

model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and Spatio-temporal Epidemiology*, 18:1–12, 2016.

[14] Peter Bloomfield. *Fourier analysis of time series: an introduction.* John Wiley & Sons, 2004.

[15] André EX Brown, Eviatar I Yemini, Laura J Grundy, Tadas Jucikas, and William R Schafer. A dictionary of behavioral motifs reveals clusters of genes affecting caenorhabditis elegans locomotion. *Proceedings of the National Academy of Sciences*, 110(2): 791–796, 2013.

[16] Michela Cameletti, Rosaria Ignaccolo, and Stefano Bande. Comparing spatio-temporal models for particulate matter in piemonte. *Environmetrics*, 22(8):985–996, 2011.

[17] RJ Carroll, R Chen, EI George, TH Li, HJ Newton, H Schmiediche, and N Wang. Ozone exposure and population density in harris county, texas. *Journal of the American Statistical Association*, 92(438):392–404, 1997.

[18] Jeng-Min Chiou and Pai-Ling Li. Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(4):679–699, 2007. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2007.00605.x.

[19] William Clinger and John W. Van Ness. On unequally spaced time points in time series. *Ann. Statist.*, 4(4):736–745, 1976. ISSN 0090-5364.

[20] Noel Cressie. Spatial prediction and ordinary kriging. *Math. Geol.*, 20(4):405–421, 1988. ISSN 0882-8121. doi: 10.1007/BF00892986. New advances in geostatistics (Redwood City, CA, 1987).

[21] Noel Cressie. *Statistics for spatial data.* John Wiley & Sons, 2015.

[22] R. Dahlhaus and H. Künsch. Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, 74(4):877–882, 1987. ISSN 0006-3444. doi: 10.1093/biomet/74.4.877.

[23] David Degras, Zhiwei Xu, Ting Zhang, and Wei Biao Wu. Testing for parallelism among trends in multiple time series. *IEEE Trans. Signal Process.*, 60(3):1087–1097, 2012. ISSN 1053-587X. doi: 10.1109/TSP.2011.2177831.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. ISSN 0035-9246. With discussion.

[25] Sonia Pértega Díaz and José A Vilar. Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. *Journal of classification*, 27 (3):333–362, 2010.

[26] Mohamed El Machkouri, Dalibor Volný, and Wei Biao Wu. A central limit theorem for stationary random fields. *Stochastic Process. Appl.*, 123(1):1–14, 2013. ISSN 0304-4149. doi: 10.1016/j.spa.2012.08.014.

[27] DM Elsom. Spatial correlation analysis of air pollution data in an urban area. *Atmospheric Environment (1967)*, 12(5):1103–1107, 1978.

[28] Robert Engle. Garch 101: The use of arch/garch models in applied econometrics. *Journal of economic perspectives*, 15(4):157–168, 2001.

[29] Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[30] Barrett P Eynon and Paul Switzer. The variability of rainfall acidity. *Canadian Journal of Statistics*, 11(1):11–23, 1983.

[31] Carmen Fernández and Mark FJ Steel. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[32] Montserrat Fuentes. Approximate likelihood for large irregularly spaced spatial data. *J. Amer. Statist. Assoc.*, 102(477):321–331, 2007. ISSN 0162-1459. doi: 10.1198/016214506000000852.

[33] Zhong-Ke Gao, Yu-Xuan Yang, Peng-Cheng Fang, Yong Zou, Cheng-Yi Xia, and Meng Du. Multiscale complex network for analyzing experimental multivariate time series. *EPL (Europhysics Letters)*, 109(3):30005, 2015.

[34] Luis Angel Garcí a Escudero and Alfonso Gordaliza. A proposal for robust curve clustering. *J. Classification*, 22(2):185–201, 2005. ISSN 0176-4268. doi: 10.1007/s00357-005-0013-8.

[35] Arthur S Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298):369–375, 1962.

[36] Peter Guttorp, Wendy Meiring, and Paul D Sampson. A space-time analysis of ground-level ozone data. *Environmetrics*, 5(3):241–254, 1994.

[37] Xavier Guyon. Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1):95–105, 1982. ISSN 0006-3444. doi: 10.1093/biomet/69.1.95.

[38] Timothy C Haas. Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association*, 90(432): 1189–1199, 1995.

[39] P. Hall and C. C. Heyde. *Martingale limit theory and its application.* Academic Press,

Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980. ISBN 0-12-319350-8. Probability and Mathematical Statistics.

[40] Peter Hall and Prakash Patil. Properties of nonparametric estimators of autocovariance for stationary random fields. *Probab. Theory Related Fields*, 99(3):399–424, 1994. ISSN 0178-8051. doi: 10.1007/BF01199899.

[41] Peter Hall, Nicholas I Fisher, and Branka Hoffmann. On the nonparametric estimation of covariance functions. *The Annals of Statistics*, pages 2115–2134, 1994.

[42] Peter Hall, Young K. Lee, and Byeong U. Park. A method for projecting functional data onto a low-dimensional space. *J. Comput. Graph. Statist.*, 16(4):799–812, 2007. ISSN 1061-8600. doi: 10.1198/106186007X257296.

[43] David A Harville and Daniel R Jeske. Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, 87(419): 724–731, 1992.

[44] Lothar Heinrich. Asymptotic behaviour of an empirical nearest-neighbour distance function for stationary Poisson cluster processes. *Math. Nachr.*, 136:131–148, 1988. ISSN 0025-584X. doi: 10.1002/mana.19881360109.

[45] Ludger Hentschel. All in the family nesting symmetric and asymmetric garch models. *Journal of Financial Economics*, 39(1):71–104, 1995.

[46] Bing Hu, Yanping Chen, and Eamonn Keogh. Time series classification under more realistic assumptions. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 578–586. SIAM, 2013.

[47] Hae Kyung Im, Michael L Stein, and Zhengyuan Zhu. Semiparametric estimation of spectral density with irregular observations. *Journal of the American Statistical Association*, 102(478):726–735, 2007.

[48] AA Ivanov and Nicolai Leonenko. *Statistical analysis of random fields*, volume 28. Springer Science & Business Media, 2012.

[49] Hesam Izakian, Witold Pedrycz, and Iqbal Jamal. Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, 39:235–244, 2015.

[50] Michael Jerrett, Richard T Burnett, Bernardo S Beckerman, Michelle C Turner, Daniel Krewski, George Thurston, Randall V Martin, Aaron van Donkelaar, Edward Hughes, Yuanli Shi, et al. Spatial analysis of air pollution and mortality in california. *American journal of respiratory and critical care medicine*, 188(5):593–599, 2013.

[51] Richard H. Jones. Spectral analysis with regularly missed observations. *Ann. Math. Statist.*, 33:455–461, 1962. ISSN 0003-4851.

[52] Yoshihide Kakizawa, Robert H Shumway, and Masanobu Taniguchi. Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441):328–340, 1998.

[53] Rangasami L. Kashyap. Characterization and estimation of two-dimensional ARMA models. *IEEE Trans. Inform. Theory*, 30(5):736–745, 1984. ISSN 0018-9448. doi: 10.1109/TIT.1984.1056955.

[54] Azadeh Khaleghi, Daniil Ryabko, Jérémie Mary, and Philippe Preux. Consistent algorithms for clustering time series. *J. Mach. Learn. Res.*, 17:Paper No. 3, 32, 2016. ISSN 1532-4435.

[55] BM Golam Kibria, Li Sun, James V Zidek, and Nhu D Le. Bayesian spatial prediction of random space-time fields with application to mapping pm2. 5 exposure. *Journal of the American Statistical Association*, 97(457):112–124, 2002.

[56] Martin Kulldorff and Ulf Hjalmars. The knox method and other tests for space-time interaction. *Biometrics*, 55(2):544–552, 1999.

[57] Mahesh Kumar, Nitin R Patel, and Jonathan Woo. Clustering seasonality patterns in the presence of errors. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 557–563. ACM, 2002.

[58] Phaedon C Kyriakidis and André G Journel. Geostatistical space–time models: a review. *Mathematical geology*, 31(6):651–684, 1999.

[59] Sébastien Laurent. Analytical derivates of the aparch model. *Computational Economics*, 24(1):51–57, 2004.

[60] Frédéric Lavancier. Aggregation of isotropic autoregressive fields. *Journal of Statistical Planning and Inference*, 141(12):3862–3866, 2011.

[61] Thomas CM Lee. A simple span selector for periodogram smoothing. *Biometrika*, 84 (4):965–969, 1997.

[62] Pierre Legendre, Miquel De Cáceres, and Daniel Borcard. Community surveys through space and time: testing the space–time interaction in the absence of replication. *Ecology*, 91(1):262–272, 2010.

[63] Bo Li, Marc G. Genton, and Michael Sherman. On the asymptotic joint distribution of sample space-time covariance estimators. *Bernoulli*, 14(1):228–248, 2008. ISSN 1350-7265. doi: 10.3150/07-BEJ6196.

[64] T Warren Liao. Clustering of time series data?a survey. *Pattern recognition*, 38(11): 1857–1874, 2005.

[65] Zhengyan Lin and Weidong Liu. On maxima of periodograms of stationary processes. *Ann. Statist.*, 37(5B):2676–2695, 2009. ISSN 0090-5364. doi: 10.1214/08-AOS590.

[66] Weidong Liu and Wei Biao Wu. Asymptotics of spectral density estimates. *Econometric Theory*, 26(4):1218–1245, 2010. ISSN 0266-4666. doi: 10.1017/S026646660999051X.

[67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif., 1967.

[68] Elizabeth Ann Maharaj and Pierpaolo D'Urso. Fuzzy clustering of time series in the frequency domain. *Information Sciences*, 181(7):1187–1211, 2011.

[69] Yasumasa Matsuda and Yoshihiro Yajima. Fourier analysis of irregularly spaced data on $\mathbb{R}^d$. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(1):191–217, 2009. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2008.00685.x.

[70] Helmut Mayer. Air pollution in cities. *Atmospheric environment*, 33(24):4029–4037, 1999.

[71] Sebastian Meyer, Ingeborg Warnke, Wulf Rössler, and Leonhard Held. Model-based testing for space–time interaction using point processes: An application to psychiatric hospital admissions in an urban area. *Spatial and spatio-temporal epidemiology*, 17: 15–25, 2016.

[72] Henry R. Neave. Spectral analysis of a stationary time series using initially scarce data. *Biometrika*, 57:111–122, 1970. ISSN 0006-3444.

[73] Henry R. Neave. An improved formula for the asymptotic variance of spectrum estimates. *Ann. Math. Statist.*, 41:70–77, 1970. ISSN 0003-4851.

[74] Daniel B Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370, 1991.

[75] Tim Oates, Laura Firoiu, and Paul R Cohen. Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, pages 17–21. Sweden Stockholm, 1999.

[76] Hernando C Ombao, Jonathan A Raz, Robert L Strawderman, and Rainer Von Sachs. A simple generalised crossvalidation method of span selection for periodogram smoothing. *Biometrika*, 88(4):1186–1192, 2001.

[77] Emanuel Parzen. On spectral analysis with missing observations and amplitude modulation. *Sankhyā Ser. A*, 25:383–392, 1963. ISSN 0581-572X.

[78] Magda Peligrad and Wei Biao Wu. Central limit theorem for Fourier transforms of stationary processes. *Ann. Probab.*, 38(5):2009–2022, 2010. ISSN 0091-1798. doi: 10.1214/10-AOP530.

[79] C Arden Pope III, Michael J Thun, Mohan M Namboodiri, Douglas W Dockery, John S Evans, Frank E Speizer, and Clark W Heath Jr. Particulate air pollution as a predictor of mortality in a prospective study of us adults. *American journal of respiratory and critical care medicine*, 151(3_pt_1):669–674, 1995.

[80] T. Pukkila. The bias in periodogram ordinates and the estimation of ARMA models in the frequency domain. *Austral. J. Statist.*, 21(2):121–128, 1979. ISSN 0004-9581.

[81] Shahrokh Rouhani and Timothy J Hall. Space-time kriging of groundwater data. In *Geostatistics*, pages 639–650. Springer, 1989.

[82] Sujit K Sahu and Kanti V Mardia. A bayesian kriged kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):223–244, 2005.

[83] Sujit K Sahu, Alan E Gelfand, and David M Holland. Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1):61–86, 2006.

[84] Sujit K Sahu, Stan Yip, and David M Holland. Improved space–time forecasting of next day ozone concentrations in the eastern us. *Atmospheric Environment*, 43(3): 494–501, 2009.

[85] Tiago Santos and Roman Kern. A literature survey of early time series classification and deep learning. In *SAMI@ iKNOW*, 2016.

[86] Gavin Shaddick and Jon Wakefield. Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):351–372, 2002.

[87] Ashish Singhal and Dale E Seborg. Clustering multivariate time-series data. *Journal of chemometrics*, 19(8):427–438, 2005.

[88] Richard L Smith, Stanislav Kolenikov, and Lawrence H Cox. Spatiotemporal modeling of pm2. 5 data with missing values. *Journal of Geophysical Research: Atmospheres*, 108(D24), 2003.

[89] Padhraic Smyth et al. Clustering sequences with hidden markov models. *Advances in neural information processing systems*, pages 648–654, 1997.

[90] Li Sun, James V Zidek, Nhu D Le, and Halûk Özkaynak. Interpolating vancouver's daily ambient pm10 field. *Environmetrics*, 11(6):651–663, 2000.

[91] Thaddeus Tarpey and Kimberly K. J. Kinateder. Clustering functional data. *J. Classification*, 20(1):93–114, 2003. ISSN 0176-4268. doi: 10.1007/s00357-003-0007-3.

[92] George D Thurston, Richard T Burnett, Michelle C Turner, Yuanli Shi, Daniel Krewski, Ramona Lall, Kazuhiko Ito, Michael Jerrett, Susan M Gapstur, W Ryan Diver, et al. Ischemic heart disease mortality and long-term exposure to source-related components of us fine particle air pollution. *Environmental Health Perspectives (Online)*, 124(6):785, 2016.

[93] Erik Vanem, Arne Bang Huseby, and Bent Natvig. Bayesian hierarchical spatio-temporal modelling of trends and future projections in the ocean wave climate with a co_2 component. *Environmental and ecological statistics*, 21(2):189–220, 2014.

[94] Jose M. Vidal-Sanz. Automatic spectral density estimation for random fields on a lattice via bootstrap. *TEST*, 18(1):96–114, 2009. ISSN 1133-0686. doi: 10.1007/s11749-007-0059-5.

[95] Vikram M Vyas and George Christakos. Spatiotemporal analysis and mapping of sulfate deposition data over eastern usa. *Atmospheric Environment*, 31(21):3623–3633, 1997.

[96] Joe H. Ward, Jr. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.*, 58:236–244, 1963. ISSN 0162-1459.

[97] P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954. ISSN 0006-3444.

[98] Christopher K Wikle, L Mark Berliner, and Noel Cressie. Hierarchical bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154, 1998.

[99] Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154, 2005.

[100] Wei Biao Wu and Mohsen Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statist. Sinica*, 19(4):1755–1768, 2009. ISSN 1017-0405.

[101] Diethelm Wuertz, Yohan Chalabi, and M Miklovic. fgarch: Rmetrics-autoregressive conditional heteroskedastic modelling. *R package version*, 2110, 2009.

[102] Kiyoung Yang and Cyrus Shahabi. A pca-based similarity measure for multivariate time series. In *Proceedings of the 2nd ACM international workshop on Multimedia databases*, pages 65–74. ACM, 2004.

[103] Eviatar Yemini, Tadas Jucikas, Laura J Grundy, André EX Brown, and William R Schafer. A database of caenorhabditis elegans behavioral phenotypes. *Nature methods*, 10(9):877–879, 2013.

[104] Ting Zhang. Clustering high-dimensional time series based on parallelism. *J. Amer. Statist. Assoc.*, 108(502):577–588, 2013. ISSN 0162-1459. doi: 10.1080/01621459.2012.760458.

[105] Dale L Zimmerman and Noel Cressie. Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the institute of statistical mathematics*, 44(1):27–43, 1992.