Application of Time Series Testing and Clustering Framework to Detecting Abnormal Personal Weather Stations

Bo Yang

B.S., Zhejiang University, P.R.China, 2014

A Dissertation Presented to the Graduate Faculty
of University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
July, 2020

Abstract

Personal weather stations (PWSs) empower people to monitor the real-time temperature, humidity, wind speed, wind direction and rainfall of any personalized locations. Currently, there are more than 250,000 personal weather stations across the globe, providing rich and hyperlocal weather data. With the availability of higher spatial and temporal resolution of rainfall and temperature measurements, the question comes that should we count on them without any doubt? Even though PWSs are user-friendly and affordable, they are owned by amateur citizens that might be lack of professional knowledge or guidance on installation or maintenance. As a consequence, there are 'untrustworthy' personal weather stations, and we want to find out which weather station we can trust and which is unreliable. We want to detect abnormal behaviors with the data, especially rainfall and temperature data.

In this thesis, we present comprehensive methods to identify 'untrustworthy' PWSs effectively. Our methods include two perspectives. One approach is to conduct hypothesis testing under time series context, which means we test each PWS data against ground truth. Weather measurements reported by PWS can be of different statistical properties. For example, daily maximum temperature has a seasonal pattern and can be converted to stationary time series, so we propose the theory for testing periodograms by smoothing, which turns out to outperform all the existing methods. Precipitation measurement, however, has its unique property of zero-inflated distribution and non-stationarity. Tweedie distribution is introduced for the modelling purpose, and we also convert the rainfall data to the truncated stationary process to implement our proposed testing procedure. Based on the practical problem that PWS has multiple weather measurements, we also propose to apply tapered test statistic to the setting of multivariate time series. Besides, we introduce a clustering-

based time series abnormal detection method, which is based on the assumption that we do not have access to the outside data source as the ground truth. We apply the proposed tapered test statistic to hierarchical clustering as modified dissimilarity measure.

With proposed methods, as PWSs produce real-time climatological data daily, we can generate Bayes Factors with the input data and consistently be used to quantify the 'trustworthiness' of PWS, as well as use clustering results for the situation where surrounding NCDC stations are not available.

Acknowledgements

First and foremost, I would like to express my sincere thanks to my advisor Prof. Spitzner, for his continuous support, supervision and guidance during the past years. He has been an integral part of my development and the development of this dissertation. Without his guidance and advice this dissertation would not be possible. Prof. Spitzner himself acts as role model for me in various aspects: his great commitment for teaching, enthusiasm about research, his rigorous scientific attitudes and patience in mentoring students. His passion has inspired me to work hard and remain focused throughout my PhD candidacy.

I would like to thank my dissertation committee members, Prof. Zhou and Prof. Keenan have taught me many statistics courses which I benefited a lot. They have always been inspiring and motivating, encouraged me to overcome obstacles and move forward to pursue my goal. I'm very grateful for their generous support and help during my five years at UVA. My special thanks goes to Prof. Tian, for her kind help in finding the application part of the work and her insightful feedback and suggestions on my research, I'm very fortunate to have her support in my Ph.D. studies. I would like to thank Prof. Li and Prof. Tang for their helpful feedback and suggestions on my Ph.D. research and dissertation work.

I'm grateful to have my fellow PhD students and all the friends in Charlottesville. Thank you all for the friendship and support.

Finally and above all, I owe my deepest gratitude to my parents for their unconditional love and support throughout my life.

Contents

Chapter 1

Introduction

1.1 Objective and Outline

At present, the National Climatic Data Center (NCDC) has achieved the real-time monitoring of the essential meteorological factors such as temperature, humidity, air pressure, wind speed, and wind direction. Despite the authority and reliability, however, there are a limited amount of NCDC stations in each city, for densely populated areas, this coverage is far from enough. For modern weather monitor and forecasts, it is necessary to have enough accurate weather information for improving decision-making quality related to operation and forecast.

Personal weather stations (PWSs) provide a higher spatial and temporal resolution of temperature and rainfall measurements, empowering people to monitor the real-time temperature, humidity, wind speed, wind direction and rainfall of any personalized locations. This hyper-local weather management for the environment made it possible for more accurate flood prediction. Thus, it can be deployed in a large area to improve monitoring density.

Currently, more than 250,000 personal weather stations across the globe already

send data to Weather Underground. Personal weather station (PWS) is user-friendly and affordable, as well as easy to install and manage, which enables real-time data observation and transmission.

In our research, we pay special attention to PWS in Norfolk in Virginia state. In general, there are 2778 stations in Virginia of the area around 42,775 square miles; the density is one station per 15 square miles. For PWS in Norfolk area, however, the density is much higher of 1 station per 2.90 square miles, since there are 33 stations in 96 square miles area. Among them, there are 21 (63.6%) stations started from 2015-2018 (7 stations/year, compared to 0.5 stations/year from 2003-2015). The spread-out plot of PWS is shown in Figure 1.3.

The motivating example comes from the underground weather station daily rainfall data and temperature data in Norfolk, Virginia. Norfolk is located on the coast, and the city has a long history as a strategic military and transportation point. The largest Navy base in the world, Naval Station Norfolk, is located in Norfolk along with one of NATO's two Strategic Command headquarters. However, it suffers from coastal flooding due to land subsidence and sea-level rise, making it an ideal motivating example for testing the proposed methods. We can access and extract real-time and historical data through websites or API of the personal weather station. However, with the availability of user-friendly and affordable weather stations and rapid growth of weather enthusiasts, challenges arise concerning the trustworthiness issues. Low-cost meteorological devices are vulnerable to external variables, which might lead to discrepancies between the measured data and the standard data because of its unstable performance. Unlike the professional stations owned and managed by experts, there are 'untrustworthy' personal weather stations. We intend to detect abnormal behaviors with the data, especially the rainfall and temperature data. Currently, there lack well-defined tools and methods to measure the trustworthiness of

Figure 1.1: Personal Weather Station

the crowdsourced data in real-time, and barriers exist in capturing the integrity of published data by PWSs.

In the thesis, we propose to come up with methods of standardized evaluation of PWSs at large scales. To begin with, the crucial step is to define the anomaly of PWSs. Two approaches to capturing the abnormal behavior will be provided in our research. The first approach is based on the comparison of PWS weather measurement with official data sources. National Climatic Data Center (NCDC) is the most comprehensive, accessible, and trusted source of state-of-the-art historical weather data for climate monitoring, which provides quality controlled daily, monthly, seasonal, and yearly historical measurements such as temperature, precipitation as a baseline. Despite its authority and reliability, however, there are a limited amount of NCDC stations in each city. Thus the weather measurements can not achieve the granularity as PWSs. With the belief that similar locations should share similar weather patterns, one way of detecting suspicious PWS is to conduct time-series hypothesis testing with NCDC data, and for those labelled as 'alternative hypothesis' time-series data, it will be labelled as 'untrustworthy' ones.

The second approach, however, can be applied to the scenario that we do not have

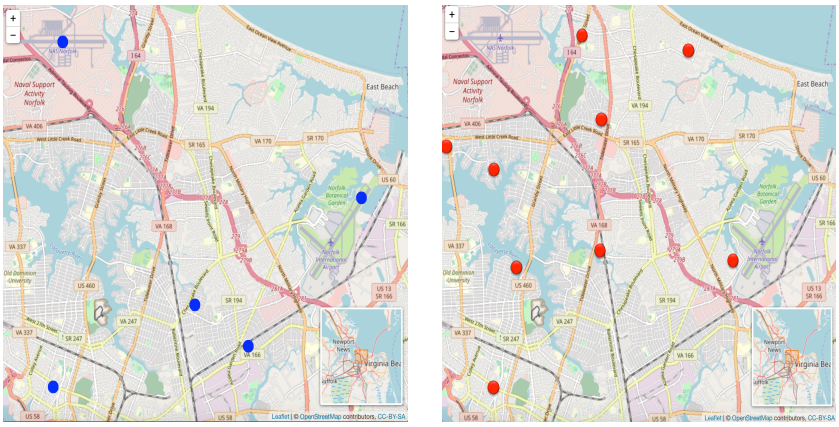

Figure 1.2: Geographical distribution of PWS and NCDC in Norfolk

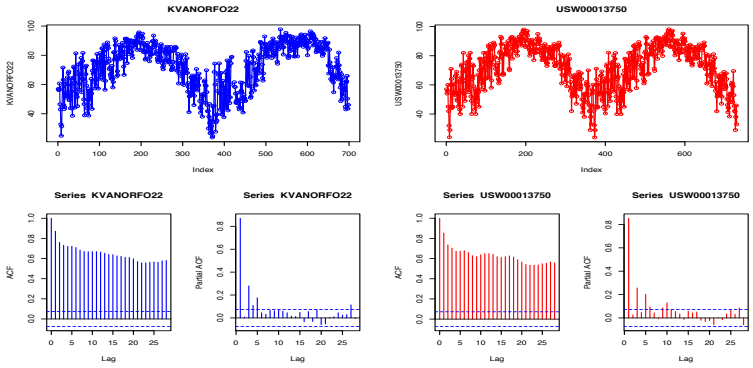

Figure 1.3: PWS and NCDC in Norfolk

access to the outer data source as the ground truth. Instead of comparison of time series, we implement time series clustering of the weather time series historical data in the same region, with the belief that stations within the cluster, if appropriately worked, should report similar weather measurements. Then we investigate the pattern within each cluster and detect abnormal ones. Ideally, if an abnormality exists, they should be clustered apart from the majority of the data. Existing methods also have limitations for this problem, for the reason that clustering methods on time series are different from other types of data since time-series data has strong dependencies and unique properties.

The distribution of PWS and NCDC weather stations is shown in Figure 1.2 ,

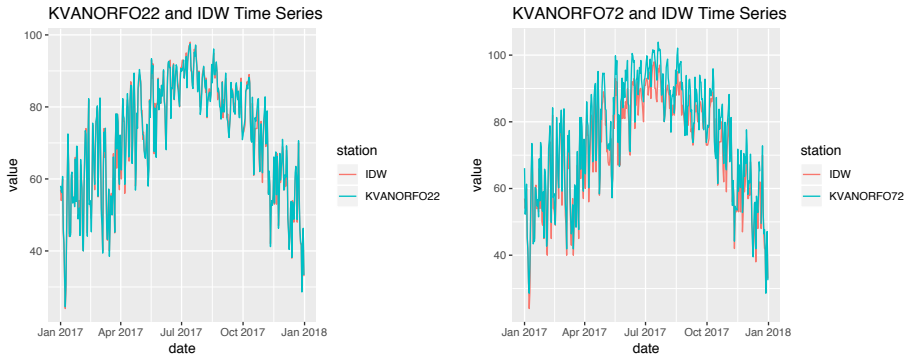

Figure 1.4: Time Series Plot of PWS vs IDW for Max Temperature (Green for PWS, Red for IDW)

which illustrates they share very similar geographical locations. In Figure 1.3, we take PWS KVANORFO22 and NCDC USW00013750 as example, and draw the time series plot and ACF, PACF plots to visualize the comparison of two weather stations. In our research, geographical difference is also handled by means that observations of NCDC stations being interpolated at gridded locations using technique such as Inverse Distance Weighting (IDW), so that we can derive the ground truth based on NCDC for testing the reliance of PWSs. Figure 1.4 shows the time series plot of PWS compared with interpolated IDW time series.

In this work, to address the first problem, we propose the theory for testing periodograms by smoothing and introduce Bayesian testing framework to time series setting, and also propose to apply our tapered test statistic to the setting of multivariate time series. We want to come up with an end-to-end framework for hypothesis testing problems in the time series context. This approach also brings in the challenge that existing testing procedures for assessing whether two stationary time series have the same dynamics lose power for the periodogram-based test, for the reason that the periodogram is not a consistent estimator of the spectral density. It also brings in the issue of high-dimensionality, which motivates incorporating tapered weights, em-

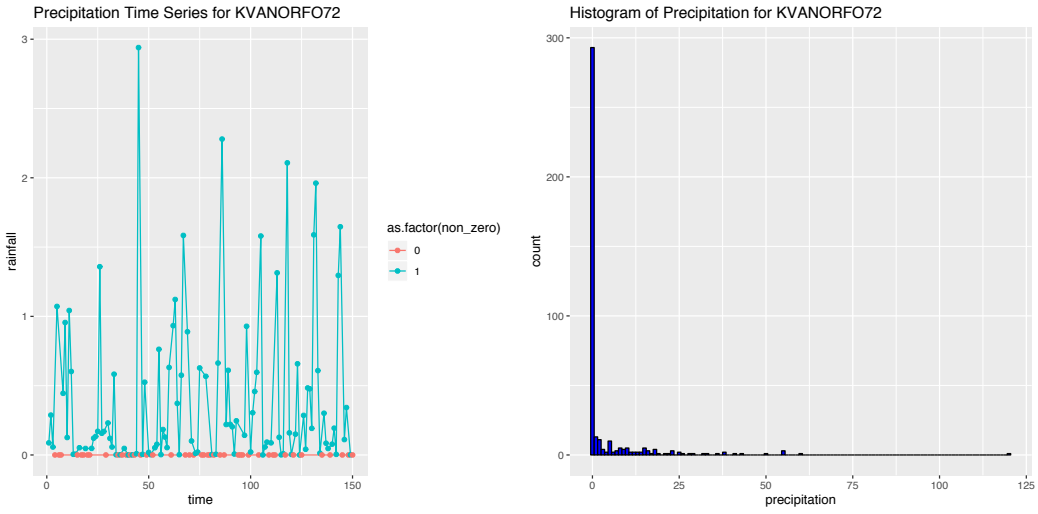

Figure 1.5: Time Series Plot of Precipitation of KVANORFO72

phasizing low-frequency periodic shapes over high-frequency periodic shapes, which is critical for avoiding accumulating stochastic errors for the high-dimensional data. Weather measurements such as daily maximum temperature and daily minimum temperature have strong seasonality and can be easily converted to stationary time series, so we propose the theory for testing periodograms by smoothing for temperature-related measurements. Precipitation measurement, however, has its unique property of zero-inflated distribution and non-stationarity, as shown in Figure 1.5, so in this case, the two-step hypothesis testing procedure will be introduced. In the first step, we reduce data to indicators of the presence or absence of precipitation and carry out analysis for binary time series. On condition that the binary time series can pass the test, we will apply our analysis for the non-zero subspace of precipitation data.

To solve the second problem, we propose methods by applying the proposed tapered test statistic to hierarchical clustering as modified dissimilarity measure to identify the cluster with abnormal behaviors.

With proposed methods, in real settings, the goal is to assign each PWS a label to distinguish whether it is trustworthy or not based on the data stream it provides

to the network. Once data has been collected, preprocessing methods will be applied, such as spatial interpolation and outlier detection. There are three testing problems and a clustering problem:

1. Hypothesis testing on temperature data
2. Hypothesis testing on precipitation data
3. Hypothesis testing on multivariate data
4. Time-series clustering-based abnormal detection

The first three approaches are established on the grounds that the underlying truth on the PWS is given, and Bayesian testing procedure is carried out to generate Bayes Factors individually, which can 'quantify' the evidence of whether or not supporting the null hypothesis for each PWS. Hypothesis testing on temperature data lays out the foundation of this thesis, for the reason that temperature data, after preprocessing steps of interpolation and outlier removal as well as transformation, can be converted to stationary time series, thus we can use periodogram-based methods with tapering technique to conduct the hypothesis testing framework. Hypothesis testing on precipitation data is more relevant to the problem of our concern, since it directly apply to rainfall data and can give us informational result on flood monitoring and prediction. Hypothesis testing on multivariate data is motivated by the practical situation that weather stations report multiple weather measurements such as minimum temperature, maximum temperature and average temperature, and this approach can compare multiple weather measurements jointly and test whether their underlying spectral matrices are the same. Bayesian approach to the correction of multiplicity is proposed to handle the multiple testing problem. For each of these three tests, Bayes Factors (BF) are calculated to 'quantify' our belief for each scenerio. For those classified as 'alternative hypothesis' time-series data, it will be labelled as 'untrustworthy'. The fourth approach, time series clustering based abnormal de-

tection, models PWSs in the same district jointly and output multiple clusters so that we can further investigate into the pattern within each cluster. If abnormality exists, they should be clustered apart from the majority of the data.

Our methods can be used for consistently producing real-time Bayes Factors to achieve a standardized evaluation of PWSs at large scales. The entire toolbox works in this fashion: first of all, historical time series data of PWSs of our interest and surrounding NCDC stations within the same region is collected, along with their geographical information for the data interpolation purpose. Once the data has been collected, preprocessing techniques such as missing value imputation, outlier detection and changepoint detection is applied. In the next step, three hypothesis testing processes are implemented, and Bayes Factors are calculated correspondingly. Furthermore, clustering-based abnormal detection as complementary methods. This process can help us achieve real-time monitoring of the integrity of published data by PWSs. The workflow is shown in Figure 1.6.

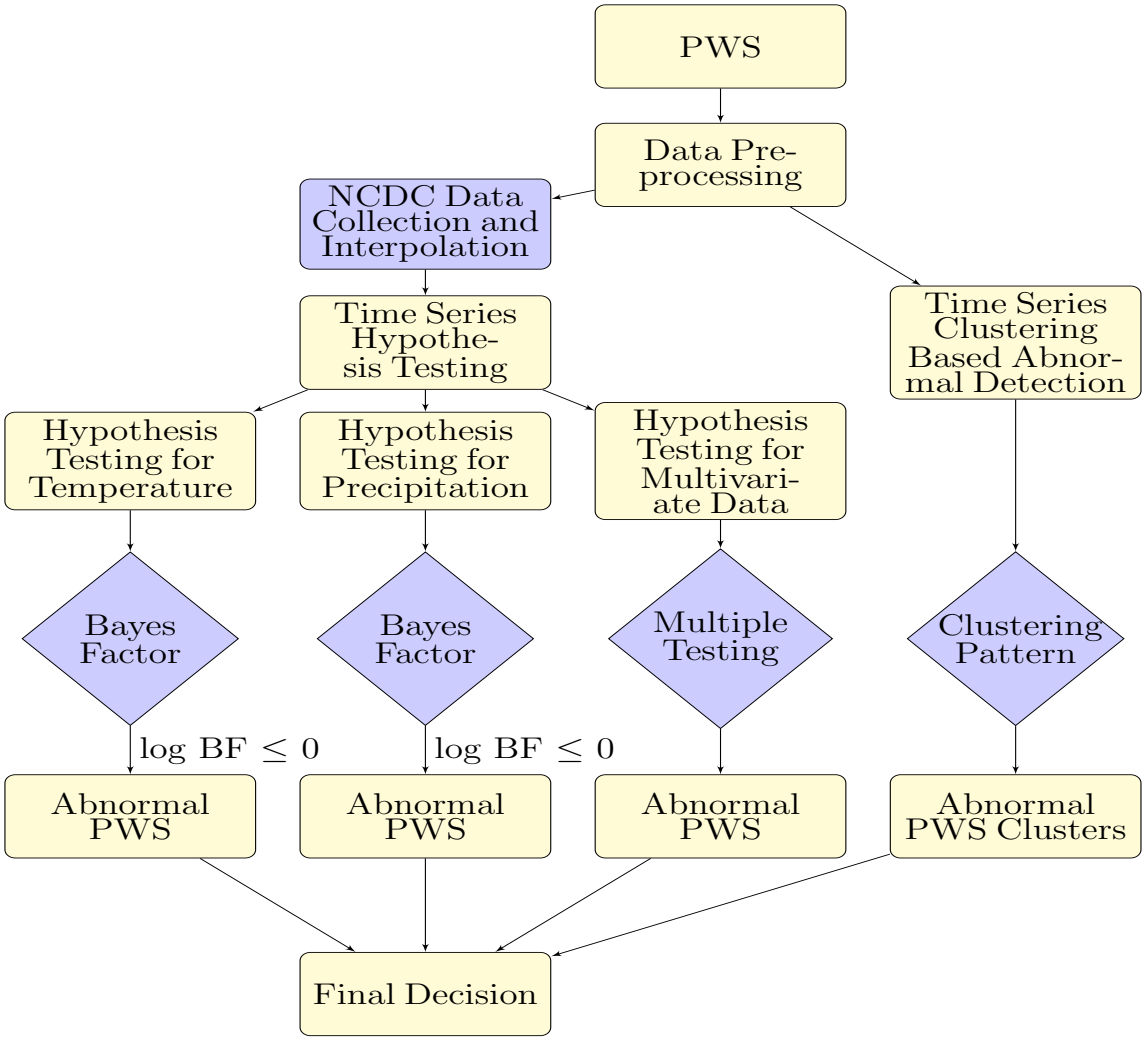

Figure 1.6: Personal Weather Station (Overview of the workflow: Temperature and precipitation data from both PWS and NCDC in Norfolk area are collected from the network jointly; the data is then preprocessed to get interpolated with the NCDC data; the system computes the Bayes factor based on temperature and rainfall data for each PWS individually; the system conducts the time series clustering for all the neighbouring PWSs jointly.)

Chapter 2

Personal Weather Station Temporal Data

From the previous chapter, we introduce the testing framework of comparing Personal Weather Stations (PWS) with the National Climatic Data Center (NCDC) as the accurate underlying distribution. However, before we jump to the modelling part, there are several preprocessing steps we have to follow along, including the spatial interpolation and outlier detection, as well as change-point detection.

2.1 Hypothesis Testing With Spatial Interpolation

In the proposal, among more than 40 NCDC stations in Norfolk area, we choose one of the NCDC as the ground truth and implement our proposed model, with the belief that weather measurements within the same region should share a similar pattern. However, such approach did not take into consideration of the fact that geographical difference between given PWS and NCDC might introduce some random error which will cause bias to the hypothesis testing results, for the reason that there is the spatial

difference among the stations, there should be minor difference in their underlying distribution.

Multiple NCDC stations are distributed in various locations in Norfolk, providing real-time climatological and meteorological data. PWS covers the personal residences of finer granularity, in order to derive the time series for any location of PWS as similar as the accurate underlying distribution, the observations of NCDC stations need to be interpolated at gridded locations using a technique such as Inverse Distance Weighting (IDW).

2.1.1 Inverse Distance Weighting (IDW)

Introduced by Burrough and McDonnell (1998), inverse distance weighting (IDW) method, with the advantage of easy to implement and fast computation, as well as straightforward to interpret, has been widely used in the context of spatial interpolation. Based on the assumption that the attribute value of an unsampled point is the weighted average of known values within the neighbourhood, the weights are inversely related to the distances between the prediction location and the sampled locations. Mathematically, the unknown value $\hat{y}(S_0)$ is represented by a linear combination of the weights and observed y values at sampled locations S_i:

$$\hat{y}(S_0) = \sum_{i=1}^{n} \lambda_i y(S_i)$$

where λ_i is defined as:

$$\lambda_i = d_{0i}^{-\alpha} / \sum_{i}^{n} d_{0i}^{-\alpha}$$

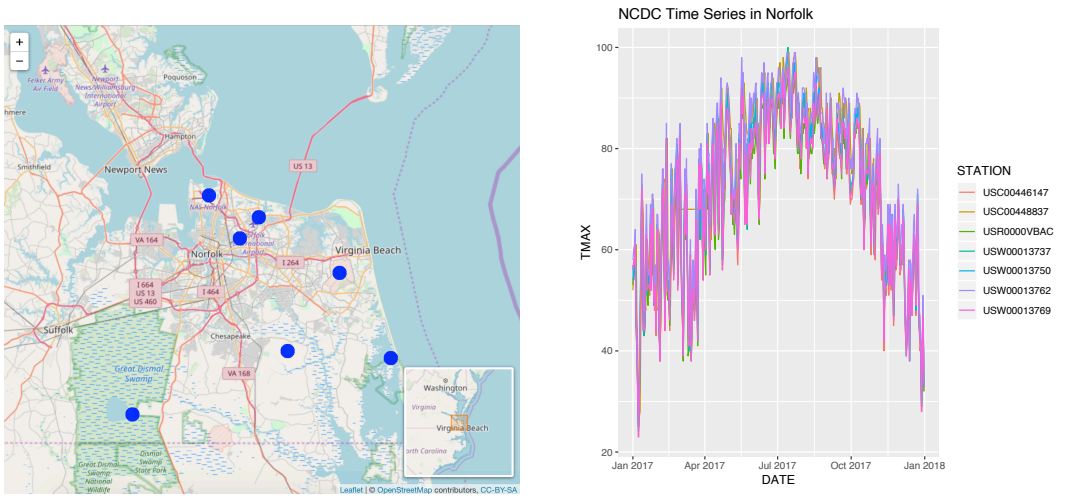

Figure 2.1: Geographical distribution and Time Series Plot of NCDC in Norfolk

with $\sum_i^n \lambda_i = 1$. In which S_0 denotes the location of the PWS where we desire to test the 'trustworthiness', which can be regarded as an interpolated (arbitrary) point, S_i is the location of NCDC, and is an interpolating (known) point, d is the distance measure from the known point S_i to the unknown point S_0, n is the total number of NCDC stations used in interpolation and α is a positive real number, called the power parameter. With the distance d increases from the interpolated point, the weight parameter λ_i decreases. The greater value of α assigns a higher weight to stations closer to the interpolated point.

In our research, since multiple stations can be observed with the latitude and longitude information, which makes it feasible to apply IDW based on neighbouring stations to make an estimate of the measurement at the station of interest, with the belief that weather measurement values at nearby locations will be similar.

In the case study of Norfolk area, take the station KVANORFO72 as a motivating example, the neighbouring NCDC stations is distributed as on the left side of Figure 2.1, while the corresponding time series plot is on the right side. In the year 2017, the NCDC stations share a quite similar pattern with a small variation. To

get the interpolated value for PWS KVANORFO72 with longitude 36.938 and latitude -76.262, we apply IDW to interpolate those NCDC stations to the coordinates of PWS in order to get the derived time-series data at the given spatial grid. The time series plot of interpolated value by IDW and the several typical PWS is shown in Figure 2.2. Based on visual inspection, it seems that PWS KVANORFO22 is consistent with IDW time series, while there is some discrepancies with KVANORFO72. KVANORFO69, however, seems hard to draw a decision based on eyeballing. Thus our modelling procedure will take into effect for such scenarios.

2.2 Formulation on 'Abnormal' PWS

In this section, we will take a closer look at the data, to understand the pattern and look into outliers. Figure 2.3 shows the daily maximum temperature time series plot of PWS in the Norfolk area compared with NCDC data. It is noteworthy that some stations such as KVANORFO20, KVANORFO72 and KVANORFO305 have several very distinct outliers, which can significantly influence the outcome of abnormal detection, even though the rest part seems consistent with NCDC time series data.

Figure 2.3 is the plot on a daily scale, still taking PWS KVANORFO72 as an example, since there are several suspicious spikes among the data for the daily maximum temperature, to figure out mechanics for those outliers, we also extract the data from Weather Underground API, which provides real-time measurement for every 15 minutes. In Figure 2.4, the top plot is the time series of maximum temperature for every 15 minutes over the period from 2017 January to 2019 March. There are two discrepant parts which look suspicious, and the left bottom and right bottom plots are the enlarged plots of those two time periods with abnormalities. We discover that during February 2018, there are two measurements significantly low compared

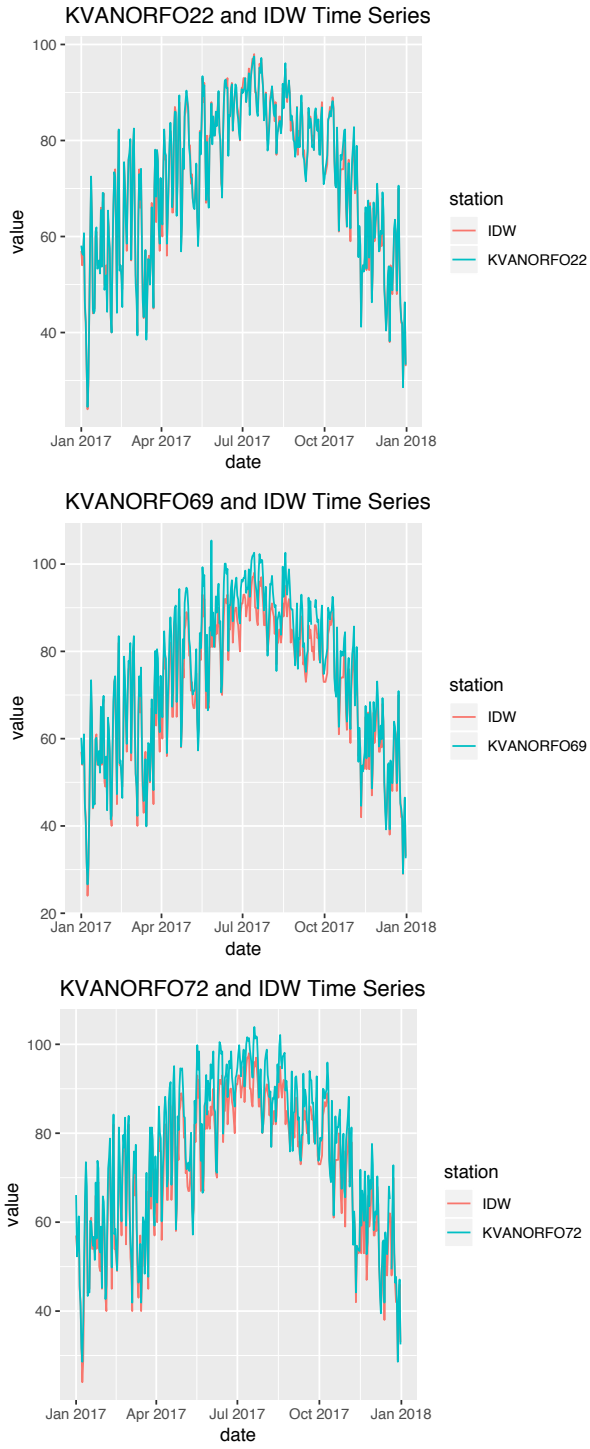

Figure 2.2: Time Series Plot of PWS vs IDW for Max Temperature (Green for PWS, Red for IDW)

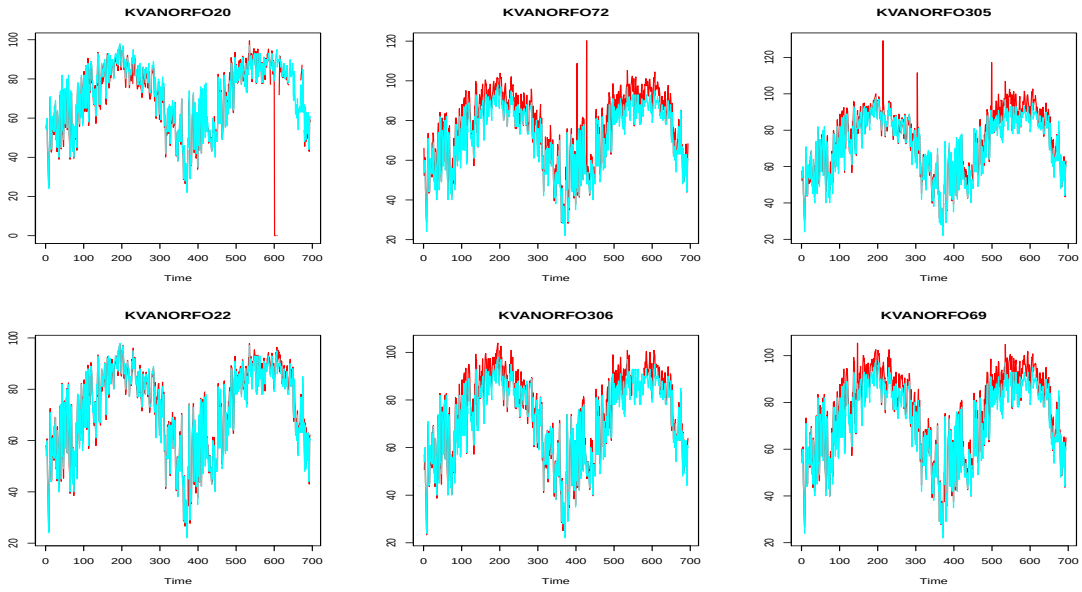

Figure 2.3: Time Series Plot of PWS vs NCDC for Max Temperature (Blue for NCDC, Red for PWS)

with neighbouring points, and during 2018 December and 2019 January, there are several suspiciously high and low spikes, as well as a portion of missing values. In a word, the existence of values that deviate a lot from neighbouring points and missing values makes it essential to add data preprocessing part before the modelling part, to successfully identify the outliers and diminish the bias of the model introduced by the suspicious outliers.

2.2.1 Outlier Detection

In the domain of time series outlier detection, the challenge arises concerning treating data with temporal dependencies and seasonal variations, as well as how to find surprising instances efficiently. Many temporal anomaly detection techniques have been proposed in a variety of research disciplines including data mining, environmental science and spatio-temporal mining.

Grubbs (1950) introduced the Grubbs test for the univariate sample set to identify

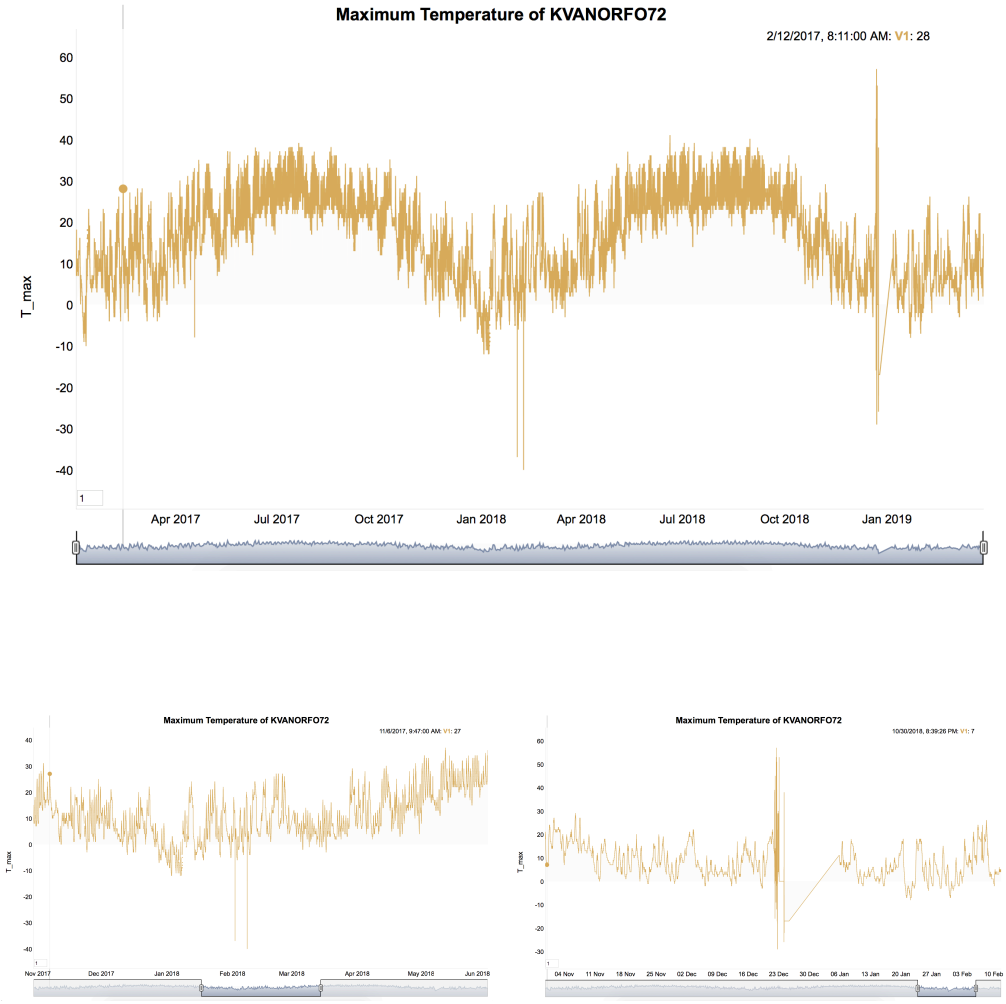

Figure 2.4: Time Series Plot of KVANORFO72 (Top) and Outlier Parts (Bottom)

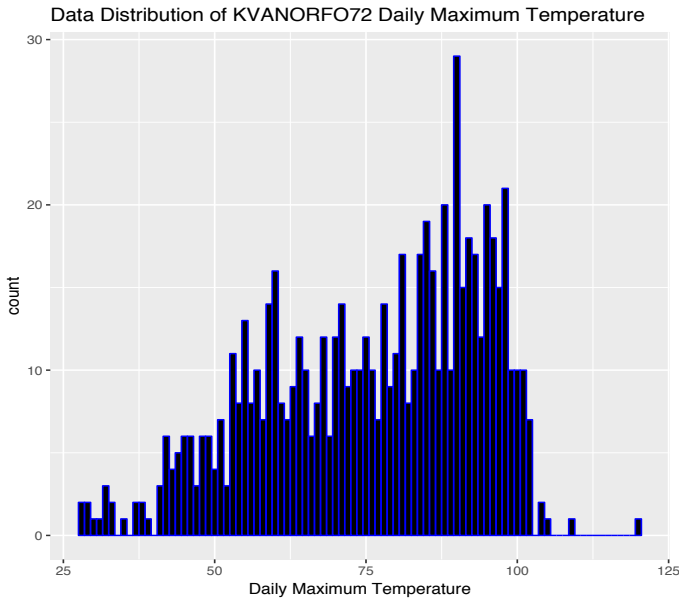

Figure 2.5: Data Distribution of KVANORFO72 Daily Maximum Temperature

the largest anomaly. The basic assumption is that the underlying data distribution is normal, and the hypothesis is :

$$H_0 : \text{There are no outliers in the data set}$$

vs

$$H_A : \text{There is at least one outlier in the data}$$

The Grubbs' test statistic is defined as follows:

$$C = \frac{\max_t |x_t - \bar{x}|}{s}$$

in which $\bar{x}$ is the mean and s is the variance of time series X. The hypothesis is

rejected at α if:

$$C > \frac{N-1}{\sqrt{N}} \sqrt{\frac{(t_{\alpha/(2N),N-2})^2}{N-2+(t_{\alpha/(2N),N-2})^2}}$$

In other words, Grubbs' test measures whether the minimum or maximum deviates from the mean value with a suspicious extent. However, in the practical context, instead of a single outlier, quite often there are multiple outliers in the given time series. Rosner (1975) introduced the Extreme Studentized Deviate test (ESD), which computes the following test statistic for the k most extreme values in the data set:

$$C_k = \frac{\max_k |x_k - \bar{x}|}{s}$$

The test statistic is then compared with a critical value:

$$\lambda_k = \frac{(n-k)t_{p,n-k-1}}{\sqrt{(n-k-1+t^2_{p,n-k-1})(n-k+1)}}$$

ESD repeats this process k times, with the number of anomalies equal to the largest k such that $C_k > k$.

In practice, when we want to apply the abnormal detection techniques to weather data, two significant challenges arise: The first challenge is that weather data, especially temperature, displays substantial seasonality. Grubbs and ESD testing procedure can potentially be used for detecting multiple anomalies. However, it is ill-suited for capturing seasonal anomalies, for the reason that ESD test is used to detect one or more outliers in a univariate data set that follows an approximately normal distribution, so it can not deal with seasonal data successfully. Another challenge is that temperature data exhibits multimodal data distribution, as shown in Figure 2.5,

which violates the normal distribution assumed for the ESD test.

To address those challenges, Hochenbaum (2017) adopted the time series decomposition, to decompose the given time series X into three components to compute the residuals:

$$R_X = X - T_X - S_X$$

in which R_X is the residual, S_X and T_X denotes seasonal component and trend component respectively. ESD is then applied to the residual component to detect anomalies, which is referred to as Seasonal-ESD (S-ESD). Seasonal decomposition is employed to filter the trend and seasonal components of the time series, followed by the use of robust statistical metrics – median and median absolute deviation (MAD) to accurately detect anomalies, including some unusual noise and breakdown.

Figure 2.6 is the outlier detection result by applying Hochenbaum (2017)'s method to KVANORFO72 on both daily scale and 15-minute scale of the maximum temperature. The top plot shows that the two abnormal periods at the end of the year 2017 and 2018 can be captured, the bottom plot shows more outliers on the finer granularity.

2.3 Changepoint Detection

In our proposed methods, we come up with the theory for testing periodograms by smoothing and introduce Bayes Factor to measure our 'belief' on the likelihood of PWS abnormality. However, this process is to use all of the time-series data generated by each PWS as input, although in the previous section, the influence of existence of outliers can be removed by adding step for outlier detection. Such modelling

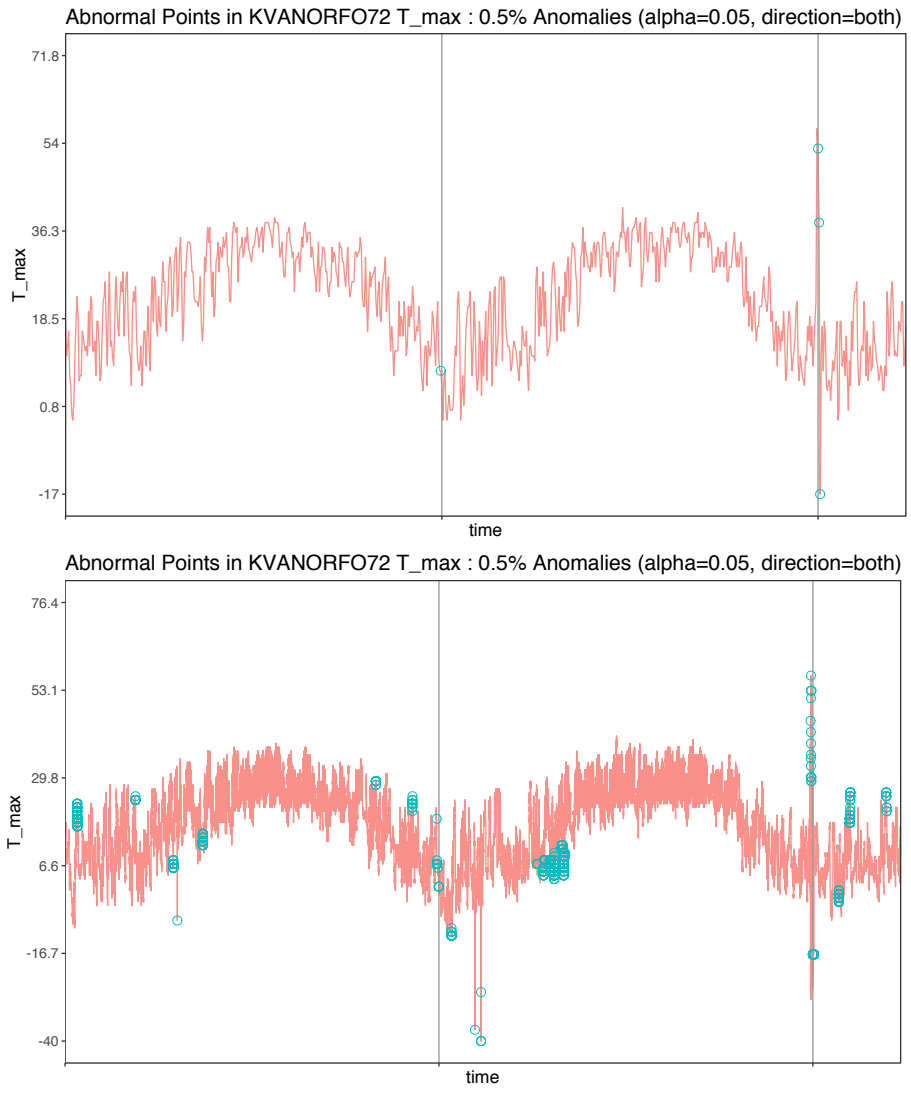

Figure 2.6: Outlier Detection on Daily Maximum Temperature (top) and 15-min Maximum Temperature (bottom)

procedure may lose some information on the time series pieces. For example, when we look into the case for KVANORFO305 from Figure 2.3, it's noteworthy that some pieces of time series data seem to be consistent with interpolated NCDC time series, and some part of the data in the latter part of the year 2018 seem to be higher than NCDC data. It motivates our curiosity of developing the changepoint detection procedure to better capture the dynamics of the time series.

Bayesian partitioning models partition the entire sequentially observed data into disjoint clusters (temporal clusters). The timestamp when a cluster ends and another one starts is called a 'change point'. Within each cluster, the sequential observations can be treated as independent and identically distributed random variables, following a single distribution such as the normal or Poisson distribution. The main inferential problem is the estimation of the number and locations.

2.3.1 Changes In Mean

We can analyze time-series data by partitioning the observations into disjoint contiguous segments, referred to as temporal clusters. Within each cluster, it is assumed that the data are independent and identically distributed random variables with a normal distribution.

Barry and Hartigan (1992) proposed the product partition model (PPM) to identify multiple change points in normal means with common variance. The setting is to consider the time series $\{Y_1, Y_2, ..., Y_n\}$, given a partition $\pi = \{C_1, ..., C_c\}$, $c \in I$, there are common parameters μ_{Cj} for $j = 1, ..., c$ which indexes the conditional density of Y_{Cj}. G_{Cj} denotes the prior cohesion function associated with the block C_j, then the

prior distribution of $\pi = \{C_1, ..., C_c\}$ following the distribution:

$$P(\pi = \{C_1, ..., C_c\}) = \frac{K \prod_{j=1}^{c} G_{Cj}}{\sum_C K \prod_{j=1}^{l} G_{Cj}} \tag{2.1}$$

Conditional on π, the sequence Y_1,...,Y_n has the joint density given by

$$P(y|\pi = \{C_1, ..., C_c\}) = \prod_{i=1}^{c} f(y_{Ci}) \tag{2.2}$$

Under the PPM, the posteriors distribution can be written as:

$$P(\pi = \{C_1, ..., C_c\}|y) = \prod_{i=1}^{c} G_{Cj} f(y_{Ci}) \tag{2.3}$$

In the PPM model, the key assumption is that with $\mu_1, ..., \mu_n$ given, random variables $Y_1, ..., Y_n$ are independently distributed. However, this assumption might be too restricted for the problem. Given the fact that there exists a minor difference in the geographical locations, instead of fixing a constant value for each μ_i within a cluster, we want to model it as vary smoothly within one cluster.

Monteiro, etal. (2011) extended the product partition model (PPM) by introducing a one-dimensional version of the intrinsic conditional autoregressive (ICAR) model as prior distribution:

$$f(\mu|\pi, \tau_\mu) \propto \tau_\mu^{(n-c)/2} exp\{\frac{\tau_\mu}{2} \sum_{i=1}^{n-1} \delta_i (\mu_i - \mu_{i+1})^2\} \prod_{i=1}^{n} 1_S(\mu_i) \tag{2.4}$$

The spatial difference is modelled conditionally in ICAR model that random effect in a given area depends on a small set of neighbouring values. Monteiro, et al. (2011) introduced the framework to both Poisson and normal cases. In the context of abnormal weather station detection, since the difference between PWS and NCDC

may have non-positive values so that we would adopt the scenario of Normal distribution for the sample distributions. The set up is to suppose given $\mu = \{\mu_1, ..., \mu_n\}$ and τ_y, the random variables Y_1,...,Y_n follows that $Y_i|\mu_i, \tau_y \sim N(\mu_i, \tau_y^{-1})$. With the prior specification that $\tau_y \sim Gamma(t, \mu)$, the posterior distribution for parameters $\theta = (\mu, \tau_y, \pi, p, \tau_\mu)$ is calculated as:

$$\begin{aligned} P \propto & f(Y|\mu, \tau_y) f(\mu|\tau_\mu, \pi) P(\pi|p) f(\tau_y) f(p) f(\tau_\mu) \\ & \propto \exp\{-\frac{\tau_y}{2}(\sum_{i=1}^{n}(y_i - \mu_i)^2 + 2\mu)\} \exp\{-\tau_\mu(\sum_{i=1}^{n-1} \delta_i(\mu_i - \mu_{i+1})^2 + s)\} \\ & * \tau_y^{(n+2t-2)/2} \tau_\mu^{(n-c+2r-2)/2} p^{n+a-c-1}(1-p)^{c+b-2} \end{aligned} \tag{2.5}$$

Since the above posterior distributions do not have a closed form, Markov chain Monte Carlo (MCMC) methods have been applied. A Gibbs sampler approximates the posteriors for the other parameters.

Figure 2.7 is the result by applying Monteiro, etc.(2011)'s approach to the difference between KVANORFO72 and IDW data to detect the change points. Several distinct points can be detected as changepoint, such as the point on timestamp 146.

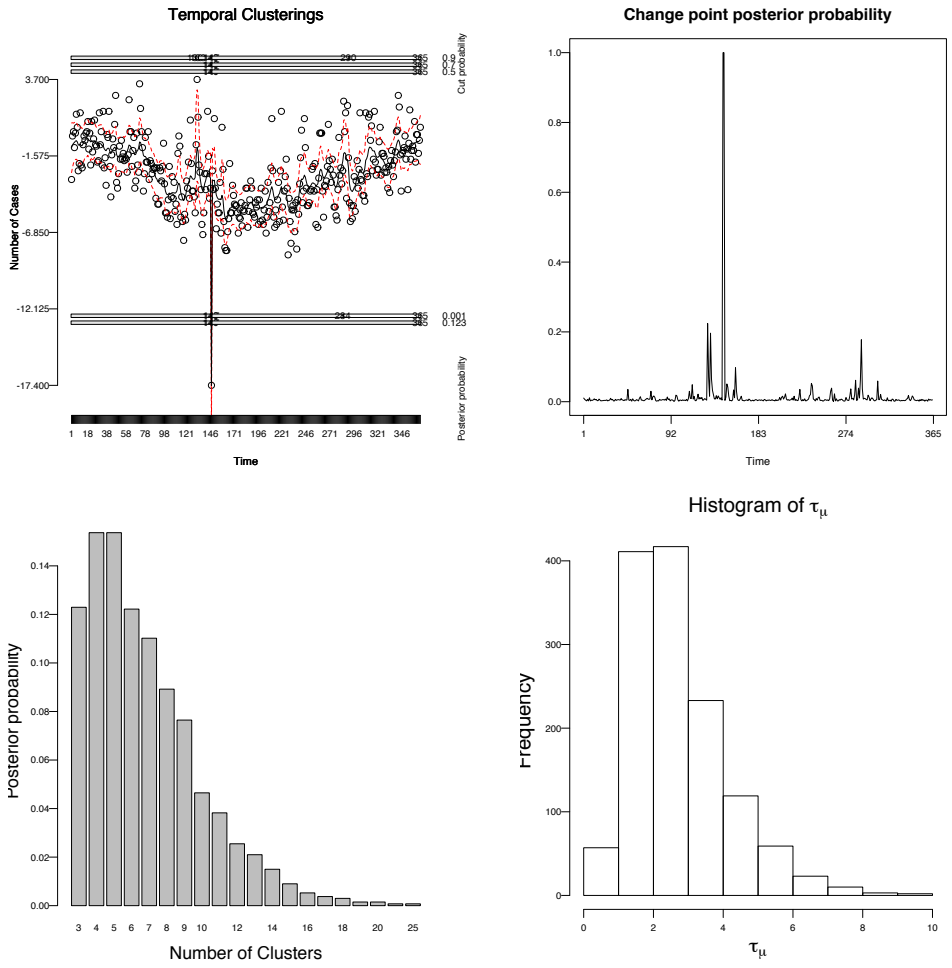

Figure 2.7: Product Partition Model for Change Point Detection for KVANORFO72

Chapter 3

Models and Asymptotics for Time Series Testing Problem

3.1 Introduction

The comparison and clustering of different time series is an important topic in statistical data analysis. In this section, the methodology of PWS-NCDC comparison has been developed. Fundamentally, time series data of PWS and NCDC stations will be compared through hypothesis testing procedure. Our research proposes to identify similarities or dissimilarities between PWS-NCDC time series data by comparing the entire auto-covariance structure of two time series, which can be effectively done in the frequency domain by comparing their spectral characteristics.

In the field of economy, finance and healthcare, e.g. inflation, interest rates, EEG data, exhibit slow decay in correlation, so parametric model may lose power, since in general, parametric approach is likely to give a more powerful test when the parametric assumptions are approximately valid and it is hard to be achieved with complex data patterns. Thus, the tendency is to use a flexible, non-parametric

approach. Therefore, we devise an "optimal tapering test" based on kernel smoothing model using the "optimal bandwidth".

This chapter overviews testing procedures for assessing whether two stationary time series have the same dynamics (or equivalent autocovariances at all lags). The context is as follows:

Consider two stationary and independent series $\{X_t\}$ and $\{Y_t\}$, with autocovariance functions $\gamma_X(h)$, $\gamma_Y(h)$ respectively at lag h. The spectral density function $f_X(\omega)$ defined by $f_X(\omega) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\gamma_X(h)e^{-i\omega h}$, in which ω is the frequency, $\gamma_X(h)$ is the auto-covariance function at lag h, and $e^{-i\omega h}$ denotes complex exponentiation. Spectral density $f_1(\omega)$ is known to characterize the process's auto-covariance function and the entirety of its distributional properties. We can regard spectral density as Fourier transformation of autocovariance. To test for equality or non-equality of the spectral densities across all frequencies, we have:

$$H_0 : f_1(\omega) = f_2(\omega), \text{for all } \omega \in (0, \pi)$$

$$H_1 : f_1(\omega) \neq f_2(\omega), \text{for some } \omega \in (0, \pi)$$

where $f_1(\omega)$ and $f_2(\omega)$ are spectral density functions of these two time series process $\{X_t\}$ and $\{Y_t\}$, respectively.

Since the spectral density is unknown, we need to estimate it. The periodogram is a commonly used way of estimating the spectral density. The periodogram of a time-series is the sample analogue of the spectral density. For time-series $\{X_i\}$, the periodogram at frequency $\omega_{ij} = 2\pi j/n_i$, for $j = 0, \ldots, [n_i/2]$, where $\omega_{i,j}$ is a uniform grid of points in $(-\pi, \pi)$, is $I_i(\omega_{ij}) = \sum_{|k|<n_i}\hat{\gamma}_i(k)e^{-ik\omega_{ij}}$, in which $\hat{\gamma}_i(k) =$

$n_i^{-1}\sum_{t=1}^{n_i-k}(X_{it+k}-\overline{X}_i)(X_{it}-\overline{X}_i)$ for $k \geqslant 0$, $\hat{\gamma}_i(k) = \hat{\gamma}_i(-k)$ for $k < 0$, and $\overline{X}_i = n_i^{-1}\sum_{t=1}^{n_i} X_t$.

Though a periodogram intuitively estimates its underlying spectral density, it is well known that variability in $I_i(\omega_j)$ does not vanish as n_i grows large, which means the periodogram is an unbiased but non-consistent estimate of spectral density, hence it cannot estimate $f_i(\omega_j)$ with asymptotic consistency. A common technique to adjust this shortcoming in the development of nonparametric time-series procedures is to substitute a smoothed periodogram for the raw periodogram. We explore this technique within an adaptation of nonparametric testing methodology developed in Spitzner (2008), in which tuning parameters are determined using asymptotic "rates-of-testing" theory. We will refer to Spitzner's (2008) procedure as a "tapering" test, motivated by its form, which will be described soon in detail.

The remainder of the chapter is organized as follows. In Section 3.2 we discuss briefly previous test statistics in time series and present Spitzner's tapering test and rate-of-testing theory, which lay out the theorectical ground for our proposed method. In Section 3.3 we discuss the models and asymptotics for the proposed block average model and kernel smoothing model, and derive the optimal tapering test based on the optimal bandwidth. In Section 3.4 we present results from comprehensive power study.

3.2 Literature Review

3.2.1 Test Statistics

The problem of identifying similarities or dissimilarities in time series data has been widely studied in the statistical research. Early work on this topic is described in

Coates and Diggle (1986), they considered a number of periodogram-based tests of the hypthesis that two independent time series $\{X_t : t = 1, ..., n\}$ and $\{Y_t : t = 1, ..., n\}$ are realizations of the same stationary process. Their motivation arises from the context to test whether or not there are significant differences between the wall thickness of a gas pipe at two different locations. Their test is based on the range of periodogram ratios with the test statistics:

$$R = \max\{\log \frac{I_1(\omega_j)}{I_2(\omega_j)}\} - \min\{\log \frac{I_1(\omega_j)}{I_2(\omega_j)}\}$$

This test statistic is extremely weak and can not provide an accurate solution to the time series testing problem. A semiparametric procedure is also proposed by Coates and Diggle (1986), which is based on the model that the underlying spectral densities $f_X(\omega)$ and $f_Y(\omega)$ are related via the equation:

$$f_Y(\omega) = f_X(\omega) \exp(\alpha + \beta\omega + \gamma\omega^2)$$

However, semi-parametric approach is likely to give a more powerful test when the parametric assumptions are approximately valid.

Motivated by the problem in the analysis of hormonal time series data, Diggle and Fisher (1991) developed graphical techniques to test the hypothesis that the two underlying spectral densities are the same. They adapted techniques such as the Kolmogorov-Smirnov and Cramer-Von Mises tests to the analysis of periodograms to access departure, which is based on empirical spectral distribution as follows:

$$D_m = \sup |F_1(\omega) - F_2(\omega)|$$

where

$$F_1(\omega_j) = \sum_{i=1}^{j} I_1(\omega_i) / \sum_{i=1}^{p_n} I_1(\omega_i)$$

$$F_2(\omega_j) = \sum_{i=1}^{j} I_2(\omega_i) / \sum_{i=1}^{p_n} I_2(\omega_i)$$

The use of normalized cumulative periodograms implies that they wish to identify only shape difference between the two underlying spectral. Although this approach refrains from the need for the subjective smoothing, it only uses the normalized cumulative periodogram, so they can only detect shape rather than scale difference between the two underlying spectral. Diggle and Fisher (1991) deliberately used test statistics which are insensitive against alternatives of the form $_X(\omega) = cf_Y(\omega)$ for all ω, where c. When pure scale difference between two underlying process is to be detected, however, the test statistic D_m then seems less natural.

Driven by the need to identify a good climatological reference series for a given station, Lund, et al. (2009) considered the test statistics $D_l = \ln(I_X(\omega_l)) - \ln(I_Y(\omega_l))$. Under the null hypothesis H_0, the D_l follow log-logistic distribution. The sample average of absolute deviation is:

$$\overline{D} = \frac{1}{n/2-1} \sum_{l=1}^{n/2-1} |D_l|$$

Large values of $\overline{D}$ support rejection of equivalent autocovariances. However, Lund, et al. (2009) ponits out that the test statistic $\overline{D}$ has relatively low power, so they considered test statistics in the time domain to access whether two stationary and independent time series have the same autocovariance at all lags. Lund. et al. (2009)

demonstrated that the above time domain based test statistic C performs better than frequency domain test statistics $\overline{D}$ in terms of power. Nevertheless, C depends on the selection of smoothing parameter L, which is subjective.

Lu and Li (2013) developed new adaptive tests which do not depend on the choices of unknown parameters. The work is inspired by Fan (1996)'s Adaptive Neyman Test (ANT). The original context of ANT is goodness-of-fit testing, nonparametric regression, and functional data analysis. Lu and Li (2013) applied ANT to the log-ratio periodogram test for assessing whether two stationary and independent time series have the same spectral densities. The frequency domain test is defined through the test statistic:

$$\overline{T_{A,N}}^{*} = \max_{1 \leqslant k \leqslant N_M} \frac{1}{2k\hat{\sigma^4}} \sum_{i=1}^{k} ((\overline{D}_i)^2 - \hat{\sigma}^2)$$

where

$$\hat{\sigma}^2 = \frac{1}{N_m - I_m} \sum_{i=I_{N_m}+1}^{N_m} (\overline{D}_i)^2 - (\frac{1}{N_m - I_m} \sum_{i=I_{N_m}+1}^{N_m} \overline{D}_i{}^2)$$

Based on this statistic, they normalize the test statistic to obtain:

$$\overline{T_{A,N}} = \sqrt{2 \log\log(N_m)} \overline{T_{A,N}}^{*} - \{2 \log\log(N_m) + .5 \log\log\log(N_m) - .5 \log(4\pi)\}$$

Therefore, for the hypotheses H_0 and H_1, large values of $\overline{T_{A,N}}$ tend to reject H_0. The main contribution of Lu and Li (2013)'s approach is that the test statistics are completely data-driven, since the approach does not crucially depend on any parameter selection. However, for the reason that periodogram is not consistent estimator, the lack of asymptotic consistency of the periodogram is mishandled by Lu and Li (2013), thus preventing the ANT from carrying over at full strength in their

adaptation. A revised procedure derived from a smoothed periodogram is proposed as a more suitable analogue, and we propose to adapt Spitzner's tapering technique to the time series setting with the use of rates-of-testing theory as a guide. In the following part I'll review Spitzner's tapering test and rates-of-testing theory.

3.2.2 Spitzner's Tapering Test

Spitzner (2008)'s tapering test is developed under the similar context as of Fan (1996)'s method, i.e., goodness-of-fit testing, nonparametric regression, and functional data analysis. Under suitable stationarity assumptions, asymptotic theory establishes an approximate model for this scenario given by

$$Y_j \quad = \quad \sqrt{n}\theta_j + \sigma\epsilon_j,$$

in which the θ_j are the parameters of interest, σ is a fixed scale parameter, and the ϵ_j are zero-mean, unit-variance, independent and identically distributed random variables. The objective of inference is to test the hypotheses:

$$H_0 : \theta_j = 0 \text{ for every } j = 1, \ldots, p$$

$$H_1 : \theta_j \neq 0 \text{ for some } j = 1, \ldots, p$$

The test statistic based on the quadratic form are constructed as:

$$Q \quad = \quad \sum_{j=1}^{p} w_j Y_j^2, \tag{3.1}$$

whose weights, w_j, are presumed to "taper" to zero, in the sense that $w_j \to 0$ as $j \to \infty$. The value Y_j is the j'th of a set of Fourier coefficient summaries. The impact of incorporating tapered weights is to emphasize low-frequency periodic shapes over high-frequency periodic shapes, which is critical for avoiding accumulating stochastic errors for the high-dimensional data (*i.e.*, when p is large). Spitzner (2008) identifies a preferred choice of the weights, given by $w_j = j^{-1/2}$ under the rates-of-testing theory, whose details are described below.

Our interest in comparing spectral density thus originated from how well Spitzner's tapering test would adapt to the time-series context.

3.2.3 Rate of Testing Theory

The rates-of-testing theory provides a formal setup for deducing the asymptotic performance of models. A geometric constraint is imposed on the parameter space that may be interpreted as a "smoothness constraint" on the functional mean parameter. Suppose the sequence $\theta = (\theta_1, \theta_2, ...)$ is bounded with respect to a Sobolev-type norm:

$$\mathbb{B}_{s,M} = (\theta_1, \theta_2, ...) : \sqrt{\sum_{j=1}^{\infty} j^{2s}\theta_j^2} \leqslant M \tag{3.2}$$

Within this framework, asymptotic performance is evaluated through sequences δ_n such that $\delta_n \to 0$ as $n \to \infty$. A satisfactory procedure must retain statistical power across all parameter values at minimum distance δ_n from the null hypotheses. Within this framework, asymptotic performance is evaluated through sequences δ_n such that $\delta_n \to 0$ as $n \to \infty$. A satisfactory procedure must retain statistical power across all parameter values at minimum distance δ_n from the null hypotheses. In frequentist

terms, this requirement is precisely stated according to

$$\inf_{\boldsymbol{\theta}\in H_1(\delta_n/\delta_n^\star;s,M)} P_{\boldsymbol{\theta}}[\text{“reject } H_0\text{”}] \to 1 \text{ for every } \delta_n^* \to 0 \tag{3.3}$$

where

$$H_1(\delta; s, M) = \{\boldsymbol{\theta} : \sum_j \theta_j^2 \geqslant \delta^2 \text{ and } \sqrt{\sum_j j^{2s}\theta_j^2} \leqslant M\} \tag{3.4}$$

Stronger performing procedures are those satisfying the requirement for sequences, δ_n, that tend to zero at faster *rates*. Performance limits of various types are deduced in Ingster (1993), Spokoiny (1996), and Spitzner (2008).

Earlier papers by Ingster (1993) and by Lepski and Spokoiny (1995b) evaluated the minimax rate of testing for this problem:

$$\widehat{\delta_n^M}(s) = n^{-2s/(4s+1)} \tag{3.5}$$

$\widehat{\delta_n^M}(s)$ is commonly expressed as the minimax rate for geometry $B_{s,M}$ at specific s. However, it was shown that both the optimal rate of testing and the structure of optimal (in rate) tests depend on smoothness parameters which are usually unknown in practical applications.

To address this issue, Spokoiny (1996) considered the problem of adaptive (assumption free) testing, and he came up with the optimal adaptive rate defined based on the adaptive factor. Spokoiny (1996) showed that for the problem under consideration the minimal adaptive factor is $(\log\log\epsilon^{-2})^{1/4}$, and he defined the optimal adaptive rate of testing as below:

$$\widehat{\delta_n^{AM}}(s) = [n^2(\log\log n)^{-1}]^{4s/(4s+1)} \text{ across } s_* < s < s^* \tag{3.6}$$

This optimal adaptive rate of testing is restricting the unknown smoothness parameter s to the interval (s_*, s^*), in which s_* and s^* are two known constants. It is shown that adaptive testing without loss of efficiency is impossible. An extra log log-factor is inessential but unavoidable payment for the adaptation. A simple adaptive test based on wavelet technique is constructed which is nearly minimax for a wide range of Besov classes. The rate $\widehat{\delta_n^{AM}}(s)$ is commonly expressed as the adaptive minimax rate for geometry $B_{s,M}$ across $s_* < s < s^*$.

Spitzner (2008) proved that no tapering test with rate δ_n goest to 0 faster than

$$\widehat{\delta_n^Q}(s) = [n^2(\log n)^{-1}]^{-s/(4s+1)} \text{ across } s_* < s < s^* \tag{3.7}$$

Although compared with Ingster's minimax rate in (3.5) and Spokoiny's adaptive minimax rate in (3.6), the rate (3.7) is slower than both of them. But $\widehat{\delta_n^Q}(s)$ already represents an adaptively optimal configuration among tests based on tapering. Spitzner (2008) shows that the test Q_n^{OPT} has superior power against a class of alternatives in a non-asymptotic context.●

3.3 Models and Asymptotics

3.3.1 Models and Asymptotics: Kernel Smoothing Periodogram

The goal of our proposed method is to develop a new powerful test based on 'rate of testing' theory. Although Spitzner's tapering test was originally proposed in functional data analysis framework, we can still borrow the idea of tapering and try to identify suitable definitions of those quantities for time series problems. Thus, our testing procedures will be derived from Spitzner's tapering test based on certain transformation of periodograms.

For the moment, let us suppose the lengths of series are equal, so $n = n_1 = n_2$, and the raw periodogram values of both time series are calculated at the same set of frequencies, $\omega_j = 2\pi j/n$ for $j = 0, \ldots, [n/2]$. It is then possible to define summary coefficients according to

$$Y_j \quad = \quad \log I_1(\omega_j) - \log I_2(\omega_j), \tag{3.8}$$

Statistical properties of the Y_j in (3.8) may be understood from asymptotic analysis of the periodogram: as $n \to \infty$ and $\omega_j \to \omega \in (0, 2\pi)$, it is known that

$$\log I_2(\omega_j) \quad \to \quad \pi f_i(\omega_j)\chi_2^2, \tag{3.9}$$

where convergence is "in distribution" and χ_2^2 is a chi-square random variable with two degrees of freedom. The coefficients are moreover asymptotically independent when the corresponding ω_j have distinct limits. An approximate model for (3.8) is therefore given by

$$Y_j \quad = \quad \theta_j + \sigma\epsilon_j, \tag{3.10}$$

where $\theta_j = \log f_1(\omega_j) - \log f_2(\omega_j)$, $\sigma^2 = \pi^2/3$, and the ϵ_j are zero-mean, unit-variance, independent and identically distributed random variables.

Pan (2016) explored the tapering test based on the model of raw log-periodograms as estimates for log-spectrums, in which the tapering test statistic is constructed in the form of a tapered sum of squared summary coecients Y_js as in (11). Pan (2016) proved it will never satisfy the rate of testing criteria, since the tapering test based on non-consistent estimates of log-spectrum will always be sub-optimal to the test based on model based on consistent estimates.

To avoid the deficiency when the Y_j are formulated from raw periodograms, the proposed approach instead reformulates the Y_j in a manner directly paralleling (3.8), but on smoothed periodograms. The advantage over using raw periodograms is twofold: primarily, a smoothed periodogram, when suitably constructed, is an asymptotically consistent estimate of its underlying spectral density (provided that density is also smooth), which admits consideration of asymptotic performance concepts used in rates of testing theory; secondly, a smoothed periodogram may be calculated at *any* frequency, ω, which provides a means of extension to the case of unequal time-series lengths, $n_1 \neq n_2$.

To take a closer look at the "rate of testing" criteria, let's firstly write it into two equivalent conditions involving the asymptotic mean and variance of Q_n.

Theorem 3.3.1. *Suppose the hypotheses H_0 vs H_1 are to be tested under the Neyman-Pearson framework on the basis of a statistic Q_n and rejection rule such that H_0 is rejected when Q_n is large. For fixed smoothness parameter $s > 1/2$ and $M > 0$, and some positive sequence $\delta_n \to 0$, the rate of testing criteria holds for the test statistic Q_n if and only if*

$$\inf_{\boldsymbol{\theta} \in H_1(\delta_n/\delta_n^\star; s, M)} E_{H_{1-0}}(Q_n)/\sqrt{Var_{H_0}}(Q_n) \to \infty$$

and

$$\inf_{\boldsymbol{\theta} \in H_1(\delta_n/\delta_n^\star; s, M)} E_{H_{1-0}}(Q_n)/\sqrt{Var_{H_{1-0}}}(Q_n) \to \infty$$

where

$$E_{H_{1-0}}(Q_n) = E_{H_1}(Q_n) - E_{H_0}(Q_n)$$

$$Var_{H_{1-0}}(Q_n) = Var_{H_1}(Q_n) - Var_{H_0}(Q_n)$$

Theorem 3.3.1 provides the equivalent conditions of "rate of testing" criteria represented in the form of Q_n, which will be used for getting the optimal bandwidth as will be discussed below.

In this section, we are trying to develop the tapering test under time series context. Since the model based on the raw periodogram as the estimate for spectrum suffers from the inefficiency problem because of the in-consistency of the periodogram. We propose to overcome this obstacle by block average and kernel smoothing models where the estimate is consistent. We further prove that if the number of blocks of the block average model and the bandwidth of the kernel satisfy certain conditions, the "rate of testing criteria" can be satisfied. Corresponding "optimal bandwidth" and "optimal adaptive rate of testing" are derived.

3.3.2 Discussion on the Blocked Average Setting

Our main goal is to develop a new powerful test based on "rate of testing" theory which can provide guidance for practical appliance. We intend to construct model based on kernel smoothing model. As a simplier version of kernel smoothing model, we'll explore the case for the blocked average model to gain insight of how selection of block numbers will influence the power of test.

For now, let's suppose the length of these two time series $\{X_{1,t_1}, t_1 = 1, ..., n_1\}$ and

$\{X_{2,t_2}, t_2 = 1, ..., n_2\}$ are equal, i.e. $n_1 = n_2 = n$. Define

$$Y_j = \log I_1(\omega_j) - \log I_2(\omega_j), \tag{3.11}$$

and

$$\theta = \log f_1(\omega_j) - \log f_2(\omega_j), \tag{3.12}$$

for $j = 1, ...p$. p is the number of frequencies and $p = n/2$ if n is even, $p = n/2 - 1$ if n is odd. The hypothesis is equivalent to

$$H_0 : \theta_j = 0 \text{ for all } j = 1, ..., p_n$$

$$H_0 : \theta_j \neq 0 \text{ for some } j = 1, ..., p_n$$

For time series $\{X_{1,t_1}\}$ and $\{X_{2,t_2}\}$, the periodogram is calculated at frequency $\omega_j = 2\pi j/n$, for $j = 0, 1, ..., p$, where ω_j is a uniform grid of points in $(-\pi, \pi)$. The setting of blocked average model is that we evenly divide the n points into m_n blocks with $[n/m_n]$ points within each block, and denote the blocks as $S_i, i = 1, ..., m_n$. We take the average of the Y_j s within each block as the modified $\bar{Y_i}$, this procedure can be represented as:

$$\bar{Y_i} = \frac{1}{|S_i|} \sum \{Y_j : j \in S_i\} \tag{3.13}$$

Then it's easy to prove that:

$$\bar{Y_i} = \bar{\theta_i} + \frac{\sigma}{\sqrt{m_n}} \epsilon \tag{3.14}$$

and $m_n \to \infty$ as $n \to \infty$.

We construct the block average tapering statistics:

$$\bar{Q}_n = \sum_{j=1}^{m_n} \frac{1}{j^{1/2}} \bar{Y}_j^2 \tag{3.15}$$

Our goal is to set a setting of m_n under the rates-of-testing criterion. The block average settings would be formulated as functions of dimension, $\hat{m}_n = \hat{m}_n(p)$. We also try to identify the high-performance, as characterized via $\hat{\delta}_n$.

An accompanying theorem would have the following form.

Theorem 3.3.2. *For arbitrary m_n, suppose the statistic $\bar{Q}_n = Q$ is defined from the block averaged $\bar{Y}_j$ in (3.15), and the decision rule is to "reject H_0" precisely when $\bar{Q}_n$ exceeds a specified cutoff. Define known constant $s > 1/2$. Under the model (3.10), the following properties hold:*

- *Define the sequence $\bar{\delta}_n = (p^{1/4-3s} \log p^{1/4})^{\frac{2s}{4s+1}}$. For arbitrary m_n, and $\delta_n = o(\bar{\delta}_n)$, the rate-of-testing criterion is not satisfied. i.e. $\bar{\delta}_n$ is the optimal adaptive rate of testing for the block average tapering mechanism. Additionally, when $m_n = \hat{m_n}$, then the test $\bar{Q}_n$ will satisfy the rate of testing criteria for $\bar{\delta}_n$.*
- *If $\hat{m_n} \asymp p^{s+3/4} \log p^{1/4}$ for $i = 1, 2$, and δ_n is such that $\bar{\delta}_n = O(\delta_n)$, then the rate-of-testing criterion is satisfied.*

The setting $\bar{m}_n$ would thus offer a specific setting of the bandwidth parameters that is expected to yield good performance in a test based on the tapering statistic.

3.3.3 Discussion on the Kernel Smoothing Model

Blocked average statistic explores the most basic setting which can give us the criteria of selecting the number of intervals m_n for time series of different lengths. The

intuitive explanation is that if m_n is very large, it means that very few points with each block is averaged, which is quite close to the setting of raw periodogram and may suffer from low power in detecting the difference in spectral density. When m_n is relatively small, however, then the raw periodogram is averaged within a very broad block, the testing framework may also lose power due to the loss of information when averaging over too many points. The idea of how to select the number of blocks in the block average tapered statistic motivates us to find the optimal bandwidth under the kernel smoothing model.

Let us denote by $\tilde{I}_i(\omega)$ a smoothed periodogram calculated at frequency ω by kernel weighing of raw periodogram values on a logarithmic scale.

Let's firstly define the kernel smoothing function as:

$$C_{i,j,k_i} = K(\frac{w_{ik_i} - w_j}{b_i}) / \sum_{k_i=1}^{p_i} K(\frac{w_{ik_i} - w_j}{b_i}), \tag{3.16}$$

Then define the kernel smoothed log-periodogram:

$$\log \tilde{I}_i(\omega_j) = \sum_{k_i=1}^{p_i} C_{i,j,k_i} \log I_i(\omega_j), \tag{3.17}$$

In the proposed methodology, the Y_j are defined according to

$$\tilde{Y}_j \quad = \quad \log \tilde{I}_1(\omega_j) - \log \tilde{I}_2(\omega_j), \tag{3.18}$$

By Brockwell and Davis (1991), it is easy to prove that $\tilde{Y}_j \longrightarrow N(\theta_j, V_{n,j})$, in which

$$V_{n,j} \quad = \quad \frac{\sigma^2}{2} \left\{ \sum_{k_1=1}^{[n_1/2]-1} C_{1,j,k_1}^2 + \sum_{k_1=1}^{[n_2/2]-1} C_{2,j,k_2}^2 \right\}, \tag{3.19}$$

A central objective of this project is to a settings, $\hat{b}_{n,i}$, of the bandwidth parameters

that performs optimally with respect to some rate, $\hat{\delta}_n$, denoting high-performance (or optimality) under the rates-of-testing criterion. The bandwidth settings would be formulated as functions of dimension, $\hat{b}_{n,i} = \hat{b}_{n,i}(p_n)$, and might also depend on the characteristics of the kernel. High-performance, as characterized *via* $\hat{\delta}_n$, is to be identified as the investigation proceeds.

In the first stage, we would consider the case when $n_1 = n_2 = n$ for the two time series to be compared, and in this case, the kernel smoothing funciton can be simplified as:

$$C_{j,k} = K(\frac{w_k - w_j}{b_n}) / \sum_{k=1}^{p_n} K(\frac{w_k - w_j}{b_n}), \tag{3.20}$$

Thus, the kernel smoothed log-periodogram is defined as:

$$\log \tilde{I}_i(\omega_j) = \sum_{k=1}^{p_n} C_{j,k} \log I_i(\omega_j), \tag{3.21}$$

which leads to:

$$V_{n,j} = \frac{\sigma^2}{2}[\sum_{k=1}^{[n_1/2]-1} C_{j,k}^2 + \sum_{k=1}^{[n_2/2]-1} C_{j,k}^2] = \sigma^2[\sum_{k=1}^{[n/2]-1} C_{j,k}^2]$$

In this section, we define the kernel smoothing of log-periodogram as estimated as:

$$\tilde{Y_i} = \tilde{\theta}_i + \sigma_n \epsilon \tag{3.22}$$

in which $\sigma_n^2 = \sigma^2 \sum_{k=1}^{[n/2]-1} C_{j,k}^2$

We construct the tapering statistics

$$\tilde{Q}_n = \sum_{j=1}^{p} \frac{1}{j^{1/2}} \tilde{Y}_j^{\,2} \tag{3.23}$$

Our goal is to set a setting of b_n under the rates-of-testing criterion. The block average settings would be formulated as functions of dimension, $\hat{b}_n = \hat{b}_n(p)$. We also try to identify the high-performance, as characterized via $\tilde{\delta}_n$.

Uniform Kernel Smoothing Model

To begin with, we adopt the most straightforward kernel smoother, the uniform kernel smoother, for which the kernel function is defined as below:

$$K(\mu) = \frac{1}{2} \tag{3.24}$$

for $|\mu| \leqslant 1$.

Pan (2016) derive the "optimal" bandwidth parameter under the uniform kernel smoothing setting with the bandwidth parameter $b_{n,j}$ such that

$$\min_{i,j} C_{i,j,k_i} = \{p_n^{4s+1} \log p_n\}^{-1/4}$$

We get the optimal bandwidth under the rate of testing framework and an accompanying theorem would have the following form.

Theorem 3.3.3. *For arbitrary b_n, suppose the statistic $\hat{Q}_n = Q$ is defined from the block averaged $\hat{Y}_j$ in (3.15), and the decision rule is to "reject H_0" precisely when $\hat{Q}_n$*

exceeds a specified cutoff. Define known constant $s > 1/2$. Under the model (3.10), the following properties hold:

- *Define the sequence $\hat{\delta}_n = (p^{(4s+1)}logp)^{\frac{2s}{4s+1}}$. For arbitrary b_n, and $\delta_n = o(\hat{\delta}_n)$, the rate-of-testing criterion is not satisfied. i.e. $\hat{\delta}_n$ is the optimal adaptive rate of testing for the block average tapering mechanism. Additionally, when $b_n = \hat{b_n}$, then the test $\hat{Q}_n$ will satisfy the rate of testing criteria for $\hat{\delta}_n$.*
- *If $\hat{b_n} \asymp p^{\frac{4s+1}{2}}$ for $i = 1, 2$, and δ_n is such that $\hat{\delta}_n = O(\delta_n)$, then the rate-of-testing criterion is satisfied.*

The setting $\hat{b}_n$ would thus offer a specific setting of the bandwidth parameters for unifrom kernel that is expected to yield good performance in a test based on the tapering statistic.

Gaussian Kernel Smoothing Model

Gaussian kernel is also considered for the tapered statistic, which takes the following form:

$$K(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{3.25}$$

An accompanying theorem would have the following form.

Theorem 3.3.4. *For arbitrary b_n, suppose the statistic $\tilde{Q}_n = Q$ is defined from the block averaged $\tilde{Y}_j$ in (3.15), and the decision rule is to "reject H_0" precisely when $\tilde{Q}_n$ exceeds a specified cutoff. Define known constant $s > 1/2$. Under the model (3.10), the following properties hold:*

- *Define the sequence* $\tilde{\delta}_n = p^{4s} log p^{4s}$. *For arbitrary* b_n, *and* $\delta_n = o(\tilde{\delta}_n)$, *the rate-of-testing criterion is not satisfied. i.e.* $\tilde{\delta}_n$ *is the optimal adaptive rate of testing for the block average tapering mechanism. Additionally, when* $b_n = \hat{b_n}$, *then the test* $\tilde{Q}_n$ *will satisfy the rate of testing criteria for* $\tilde{\delta}_n$.
- *If* $\tilde{b_n} \asymp \frac{n^2-1}{\log p^{4s}}$ *for* i = *1, 2, and* δ_n *is such that* $\tilde{\delta}_n = O(\delta_n)$, *then the rate-of-testing criterion is satisfied.*

The setting $\tilde{b}_n$ *would thus offer a specific setting of the bandwidth parameters for Gaussian kernel that is expected to yield good performance in a test based on the tapering statistic.*

Theorem 3.3.2, Theorem 3.3.3 and Theorem 3.3.4 provide the criteria for choices of m_n and b_n under blocked average and kernel smoothing settings. The optimal m_n is relatively smaller than the optimal b_n, which is as expected, for the reason that when we try to implement the blocked average model, a broader interval selection will result in the loss of individual information, since we just did the average over each block. However, for the kernel smoothing case, the choice of b_n can be relatively large, because we did the kernel smoothing for each individual point, so a broader uniform or Gaussian kernel will take into consideration of the neighboring points as well as achieve the goal of smoothing.

In a word, existing testing procedures for assessing whether two stationary time series have the same dynamics lose power for the raw periodogram-based test, for the reason that periodogram is not a consistent estimator of the spectral density. In this thesis, we apply multiple smoothing methods to tackle this issue, it will integreate the information from neighboring points to achieve the goal of smoothing, however, it will introduce dependencies among different frequencies. Rate-of-testing framework has the assumption of non-dependencies, which is violated in this setting. However, in

the simulation studies, this tapered test statistic shows superior performance among the existing methods.

3.4 Power Study

3.4.1 Simulation for Choosing Number of Blocks for the Blocked Average

In this section, we want to present a comprehensive simulation study of the blocked average model. We present different choices for the number of blocks and assess the empirical power under each scenerio to find the optimal bandwidth.

Test Statistics To Be Compared

- Proposed blocked average test statistic $\bar{Q}_n$ of different lengths m_n:

$$\bar{Y}_i = \frac{1}{|S_i|} \sum \{Y_j : j \in S_i\}$$

and $m_n \to \infty$ as $n \to \infty$.

The tapering statistics is constructed as:

$$\bar{Q}_n = \sum_{j=1}^{m_n} \frac{1}{j^{1/2}} \bar{Y}_j^2$$

The choice of m_n, however, might affect the power of the testing problem, so in the first simulation studies, we investigate in how the choice of m_n affect the model performance.

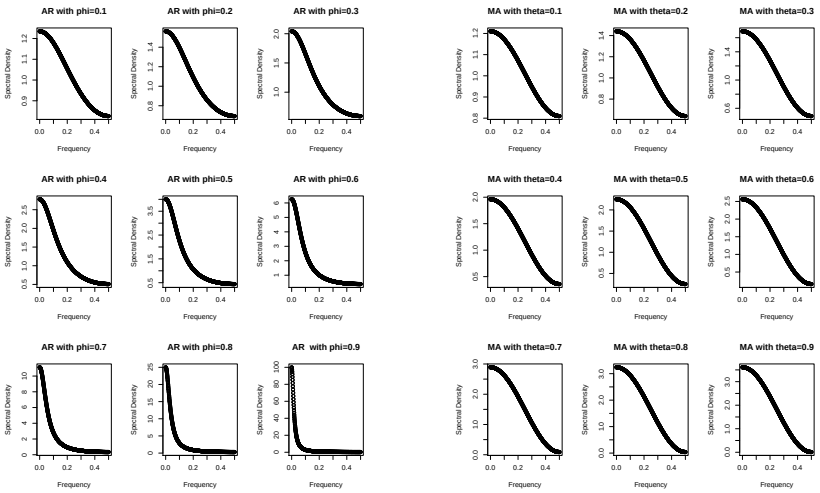

Figure 3.1: Spectral Density of AR(1) and MA(1) with Different Parameter Values

Simulation Design

The simulation design is such that the models examined have dimensionality p_n. Two scenarios are considered for the length of the time series, short time series with $n = 256$ and long time series with $n = 1024$. We compare the power of the tests by applying them to simulations of AR(1) and MA(1) time series of various coefficients.

- Setting 1: AR(1) with $\phi_1 = 0.1$ versus AR(1) with

  $\phi_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

- Setting 2: MA(1) with $\theta_1 = 0.1$ versus MA(1) with

  $\theta_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$

Figure 3.1 illustrates the spectral density distribution of AR(1) and MA(1) models of various parameter values as listed above, which indicate that they have different underlying spectral densities.

Empirical Comparisons

This section explores the empirical critical values of the suggested testing procedures for finite sample performance. First, two independent Gaussian AR(1) and MA(1) series of length $n = 256$ and $n = 1024$ with the same autoregressive parameter ϕ and θ are generated, and we make sure that $|\phi| < 1$ and $|\theta| < 1$. Based on the fact that both time series share the same parameter, hence they have equivalent spectral densities. In the first step, we get the empirical critical values for the proposed blocked average test statistics with the significance level 0.05, for the length of the time series n =256 and 1024, and the parameter $\phi_1 = 1$ for AR(1) and $\theta = 1$ for MA(1).

Power Comparisons

In this section, we want to obtain estimates of the power when the two underlying processes are different. 1000 replicate simulations of each pair are performed.

We try to look at different scenarios as m_n varies. As shown in Figure 3.1 there is difference among the underlying spectral density for time series with various parameters. We compare the powers of the test statistics listed above, and the power of each test is evaluated at the alternative given in the simulation design. Basically, we calculated the empirical power (the proportion of values of tests exceeding their critical values) for each choice of m_n under different scenerios. The empirical powers are summarized in the following plots.

Figure 3.2 shows the empirical power for AR(1) short and long time series of different bandwidth selection. In this setting, I mainly look into the case of comparing with ϕ equal to 0.2, 0.3 and 0.4, for the reason that when ϕ takes a large value such as $\phi = 0.6, 0.7, 0.8$ and 0.9, the empirical power of the tests is quite close to 1.0 for all of the selection of m_n, especially for the case of long series, so it's hard to get

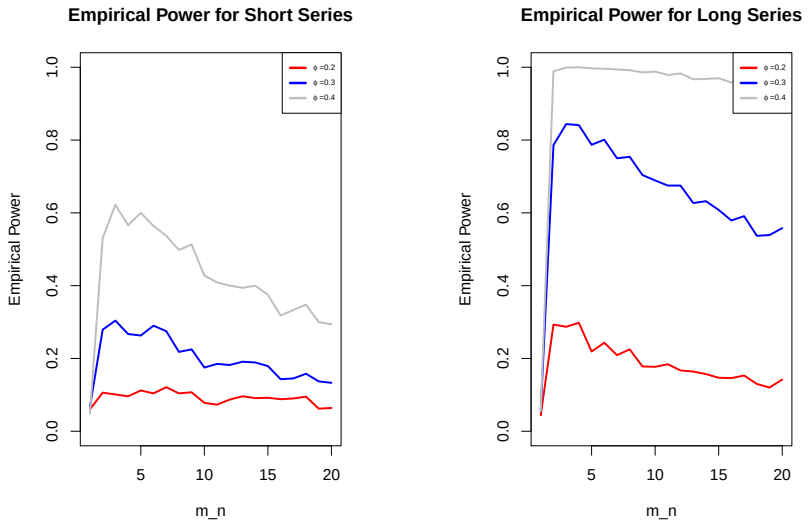

Figure 3.2: Empirical Power for Short and Long AR(1) Series of Different Bandwidth

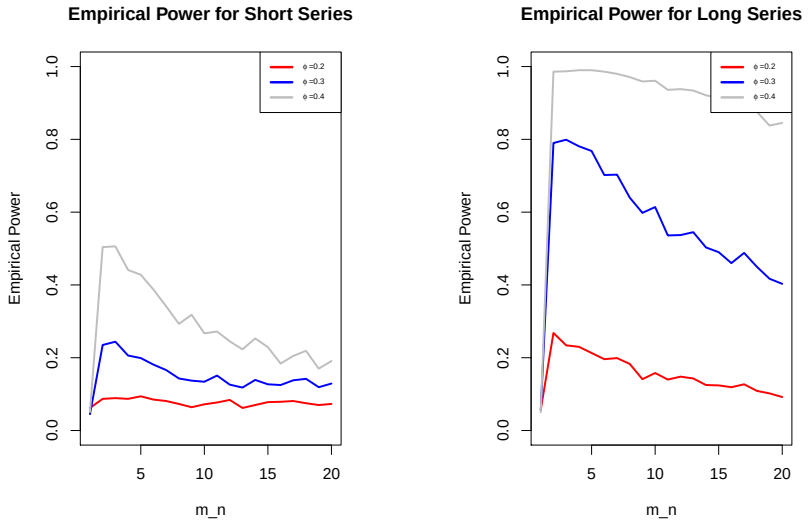

Figure 3.3: Empirical Power for Short and Long MA(1) Series of Different Bandwidth

any conclusion on the optimal choice of m_n. However, for the case of ϕ equal to 0.2, 0.3 and 0.4, it's straightforward that for the short time series $n = 256$, $m_n = 3$ is the optimal selection, and for the long time series $n = 1024$, $m_n = 4$ is the optimal selection.

Figure 3.3 is the similar power study for MA(1) short and long time series. Same findings hold that for the short time series $n = 256$, $m_n = 3$ is the optimal selection for both AR(1) and MA(1) process, and for the long time series $n = 1024$, $m_n = 4$ is the optimal selection for both AR(1) and MA(1) process. The results consolidate our conclusion that regardless of the underlying distribution of the data (in this case, whether the data has AR or MA representation), the optimal selection of bandwidth m_n only depends on the length of time series.

3.4.2 Simulation for Choosing Bandwidth for Uniform Kernel

In this section, we also conduct the simulation study for the tapered test based on the uniform kernel smoothing. We present different choices for the number of bandwidths and assess the empirical power under each scenerio to find the optimal bandwidth.

Test Statistics To Be Compared

The tapering statistics is constructed as as proposed in the uniform kernel smoothing model setting, with test statistic:

$$\hat{Q}_n = \sum_{j=1}^{m_n} \frac{1}{j^{1/2}} \hat{Y}_j^2$$

The choice of bandwidth b_n, however, might affect the power of the testing problem,

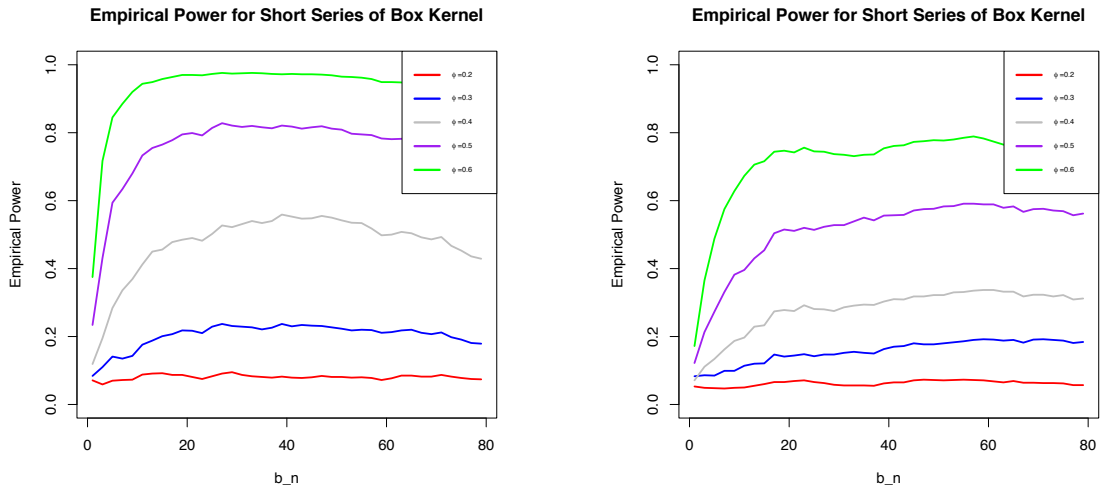

Figure 3.4: Empirical Power for Short and Long AR(1) Series of Different Bandwidth

so in the simulation studies, we investigate in how the choice of b_n affect the model performance.

3.4.3 Simulation for Choosing Bandwidth for Gaussian Kernel

In this section, we also conduct the simulaiton study for the tapered test based on Gaussian kernel smoothing. We present different choices for the number of blocks and assess the empirical power under each scenerio to find the optimal bandwidth.

Test Statistics To Be Compared

The tapering statistics is constructed as proposed in the Gaussian kernel smoothing model setting, with test statistic:

$$\tilde{Q}_n = \sum_{j=1}^{m_n} \frac{1}{j^{1/2}} \tilde{Y}_j^2$$

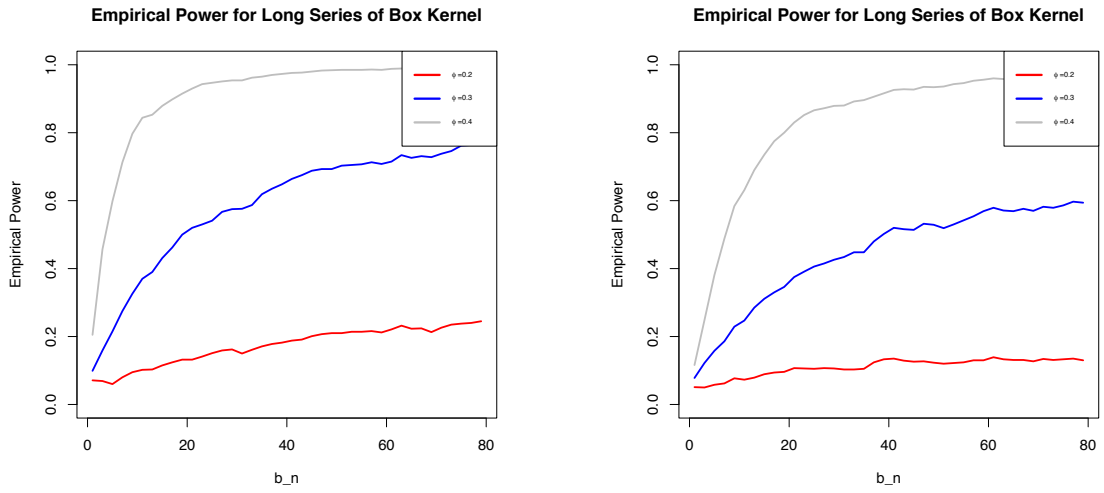

Figure 3.5: Empirical Power for Short and Long MA(1) Series of Different Bandwidth

The choice of bandwidth b_n, however, might affect the power of the testing problem, so in the simulation studies, we investigate in how the choice of b_n affect the model performance.

3.4.4 Simulation Design for Power Comparison

Test Statistics To Be Compared

In this section, empirical power of the tests based on different test statistics are compared for existing tests. For our proposed blocked average statistic, we'll use the one with the optimal bandwith m_n.

We want to carry out empirical comparison of the following test statistics:

- Coates and Diggle (1986)'s test based on the range of periodogram ratios. with the test statistic:

$$R = \max\{\log \frac{I_1(\omega_j)}{I_2(\omega_j)}\} - \min\{\log \frac{I_1(\omega_j)}{I_2(\omega_j)}\}$$

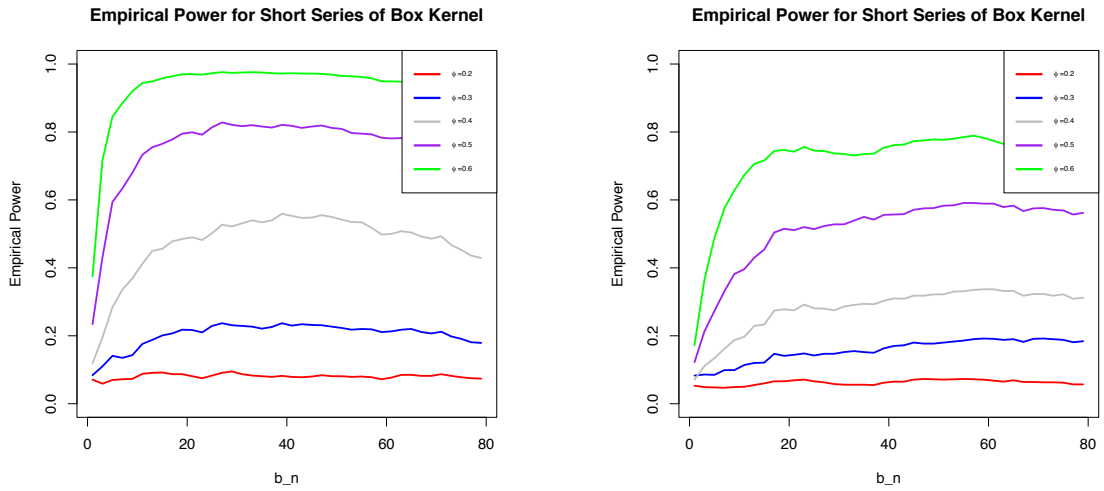

Figure 3.6: Empirical Power for Short and Long AR(1) Series of Different Bandwidth

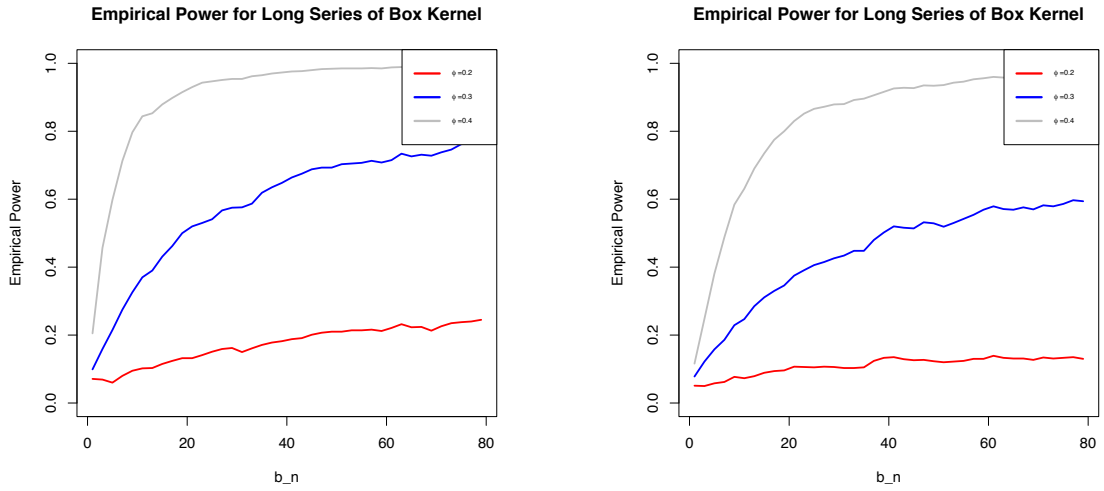

Figure 3.7: Empirical Power for Short and Long MA(1) Series of Different Bandwidth

- Lund et al. (2009)'s frequency domain test based on the average of log ratio of periodograms, with the test statistic:

$$\overline{D} = \frac{1}{n/2-1} \sum_{j=1}^{n/2-1} |\log \frac{I_1(\omega_j)}{I_2\omega_j)}|$$

- Lu and Li (2013)'s frequency domain test applying Fan's adaptive Neyman tests idea to the Lund's test, with the test statistic:

$$\overline{T_{A,N}}^* = \max_{1 \leqslant k \leqslant N_M} \frac{1}{2k\hat{\sigma^4}} \sum_{i=1}^{k} ((\overline{D}_i)^2 - \hat{\sigma^2})$$

where

$$\hat{\sigma^2} = \frac{1}{N_m - I_m} \sum_{i=I_{N\,m}+1}^{N_m} (\overline{D}_i)^2 - (\frac{1}{N_m - I_m} \sum_{i=I_{N\,m}+1}^{N_m} \overline{D}_i{}^2)$$

- Pan (2017)'s tapering test based on uniform kernel smoothing with test statistic:

$$\hat{Q}_n = \sum_{j=1}^{p} \frac{1}{j^{1/2}} \hat{Y_j}^2$$

with the optimal bandwidth: Pan (2016) derive the "optimal" bandwidth parameter under the uniform kernel smoothing setting with the bandwidth parameter $b_{n,j}$ such that

$$\min_{i,j} C_{i,j,k_i} = \{p_n^{4s+1} \log p_n\}^{-1/4}$$

- Proposed blocked average test statistic $\bar{Q}_n$ over averaging interval of optimal m_n proportion to $p^{s+3/4} \log p^{1/4}$, and as shown in the above section, for the short

time series $n = 256$, $m_n = 3$ is the optimal selection and for the short time series $n = 1024$, $m_n = 4$ is the optimal selection.

$$\bar{Y}_i = \frac{1}{|S_i|} \sum \{Y_j : j \in S_i\}$$

The tapering statistics is constructed as:

$$\bar{Q}_n = \sum_{j=1}^{m_n} \frac{1}{j^{1/2}} \bar{Y}_j^2$$

- Proposed tapering test based on Gaussian kernel smoothing with test

$$\tilde{Q}_n = \sum_{j=1}^{p} \frac{1}{j^{1/2}} \tilde{Y}_j^{\,2}$$

with the optimal bandwidth:

$$\hat{b_n} = B[\frac{1}{2} p_n^{3/4} (\log p_n)^{1/4} - \frac{1}{2}]$$

Simulation

The simulation design is such that the models examined each have dimensionality p_n. The length of the time series are specified into several scenarios, short time series with $n = 256$ and long time series with $n = 1024$.

The MA(q) refers to the moving average model of order q :

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

where μ is the mean of the series, the θ_1, ..., θ_q are the parameters of the model and the ϵ_t, ..., ϵ_{t-q} are white noise error terms.

The AR(p) indicates an autoregressive model of order p. The AR(p) model is defined as

$$X_t = c + \varphi_1 X_{t-1} + ... + \varphi_p X_{t-p} + \varepsilon_t$$

where ϕ_1, ... ,ϕ_p are parameters of the model, and ϵ_t is white noise.

In a seasonal model, seasonal AR and MA terms predict using data values and errors at times with lags that are multiples of S (the span of the seasonality). An $AR(1)_{12}$ process $\{X_t\}$, which is a generalized exponential smoothing models that can incorporate long-term trends and seasonality, is defined by

$$X_t = \phi_{12} X_{t-12} + \epsilon_t$$

where ϕ_{12} is the parameter of the model, ϵ_t is the white noise error term.

We compared the powers of different tests by applying them to simulations of AR(1) and MA(1) as well as seasonal AR(1) and MA(1) time series of various coefficients.

- Setting 1: AR(1) with $\phi_1 = 0.1$ versus AR(1) with

  $\phi_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

- Setting 2: MA(1) with $\theta_1 = 0.1$ versus MA(1) with

  $\theta_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

- Setting 3: Long-order Gaussian seasonal $MA(1)_{12}$ with $\theta_1 = 0.1$ versus Gaussian seasonal $MA(1)_{12}$ of $\theta_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

- Setting 4: Long-order Gaussian seasonal $AR(1)_{12}$ with $\phi_1 = 0.1$ versus Gaussian seasonal $AR(1)_{12}$ of $\phi_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

Empirical Comparisons

We explore the empirical critical values of the suggested testing procedures for finite sample performance. First, two independent Gaussian AR(1) and MA(1) with the same autoregressive parameter ϕ series of length $n = 256$ and $n = 1024$ are generated, and we make sure that $|\phi| < 1$ and $|\theta| < 1$. Based on the fact that both time series share the same parameter, hence they have equivalent spectral densities. In the first step, we get the empirical critical values for the proposed blocked average test statistics and the other comparing test statistics with the significance level 0.05, for the length of the time series n =256 and 1024, and the parameter ϕ_1 for AR(1) and θ_1 for MA(1).

Power Comparisons

In this section, we want to obtain estimates of the power when the two underlying processes are different. 1000 replicate simulations of each pair are performed.

We try to look at different scenarios as introduced in Section 2.4.2. We compare the powers of the test statistics listed above, and the power of each test is evaluated at the alternative given in the simulation design. Basically, we calculated the empirical power (the proportion of values of tests exceeding their critical values) for each scenerio. The empirical powers are summarized in the following figures.

It's noteworthy that the test statistic proposed by Coates and Diggle (1986) has the poorest performance among others, and the statistic proposed by Lu and Li (2013) has relative high performance in both short and long time series. The block average statistic exceed the existing methods in different scenerios, Pan (2016)'s method based on the uniorm kernel smoothing is relatively better, and the proposed method based on the Gaussian kernel smoothing turns out to have the best performance.

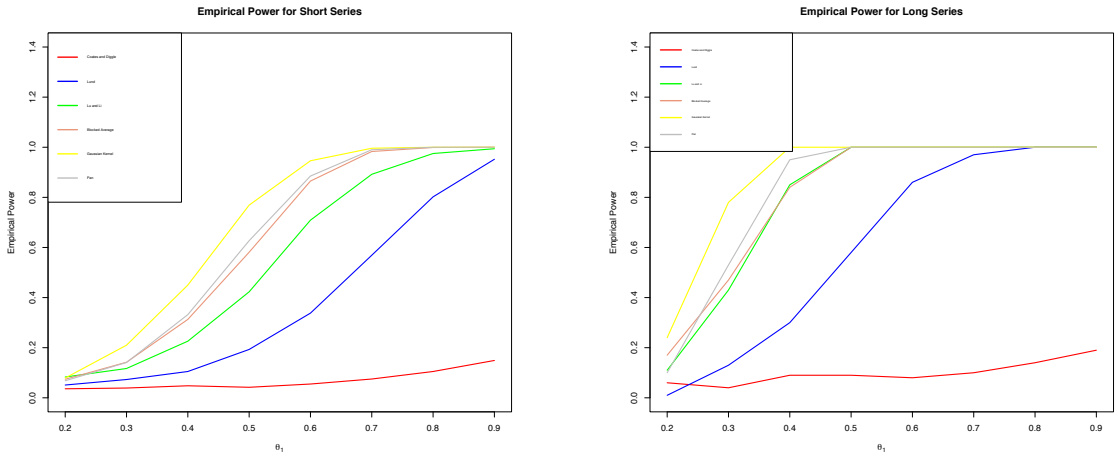

Figure 3.8: Empirical Power for Short and Long AR(1) Series

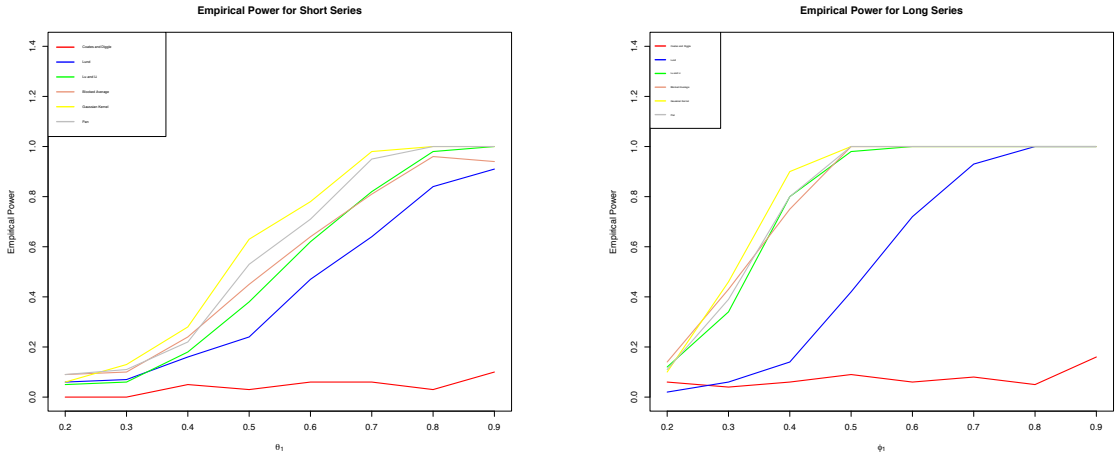

Figure 3.9: Empirical Power for Short and Long MA(1) Series

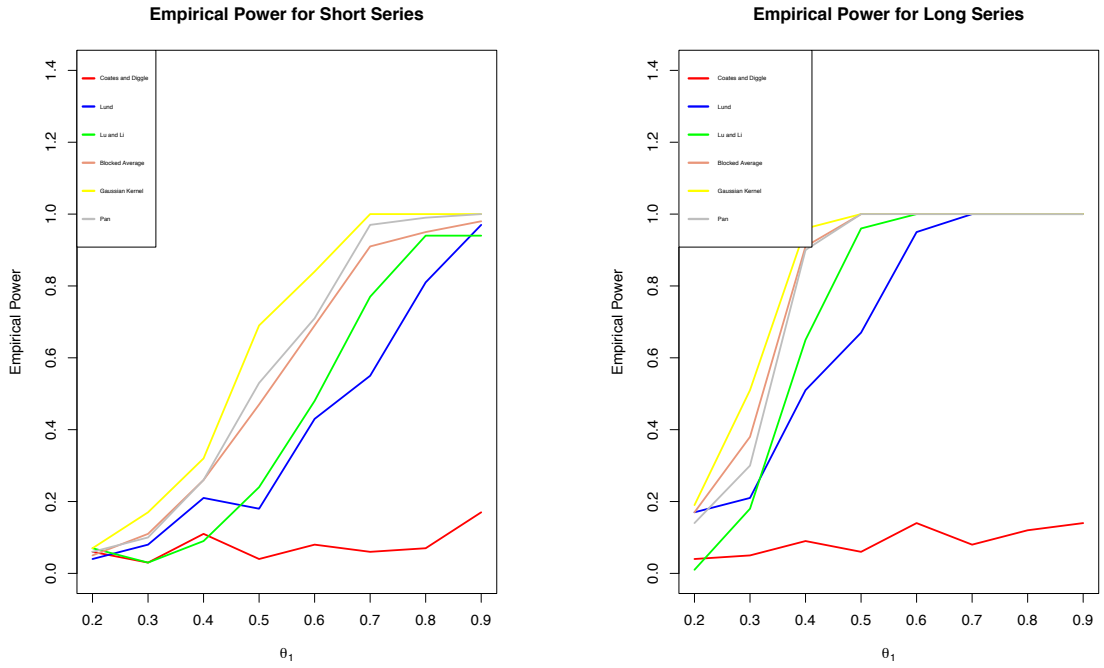

Figure 3.10: Empirical Power for Short and Long Seasonal MA(1) Series

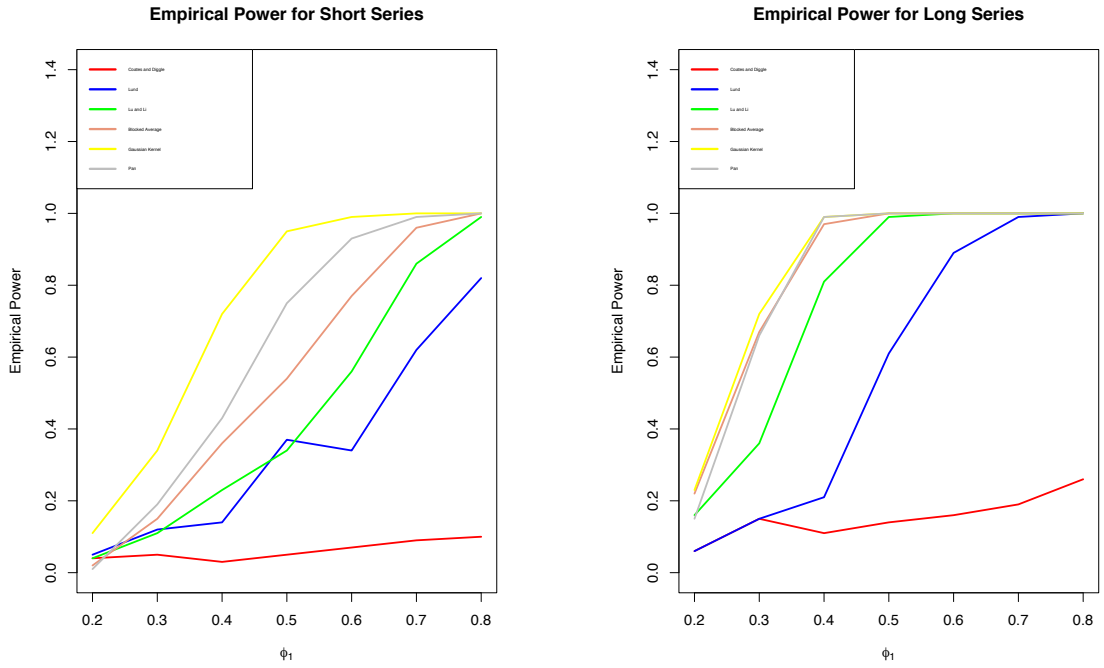

Figure 3.11: Empirical Power for Short and Long Seasonal MA(1) Series

Motivated by our proposed method, which prove to be very powerful and effective, we are interested in applying this proposed statistic to the practical case of untrustworthy personal weather station detection in Norfolk area, under the context of Bayesian testing.

3.5 Modeling Procedure

3.5.1 Critical Value Generation for Test Statistic

In our research, we have demonstrated that blocked average tapering statistic outperforms existing test statistics in the time series context, which can be a good approach for detecting abnormal PWS problem. However, under practical settings, it's hard to get the empirical critical value for two time series with unknown spectral densities, and we'll tackle this problem from two approaches.

In the first approach, since we are conducting the hypothesis between Personal Weather Stations (PWS) and National Climatic Data Center (NCDC). Based on the fact that there are around 40 NCDC stations located in Norfolk area. Assuming that they come from the same underlying distribution, then we can use these time series data to generate the critical value, which is 0.1613813 at the significance level 0.05. Then we can conduct the hypothesis testing procedure.

Table 3.1 is the test statistic for each PWS compared with NCDC, which shows that PWS KVANORFO72, KVANORFO305 and KVANORFO69 is coming from different distributions from NCDC.

Table 3.1: Tapered Statistic for PWS

PWS	Tapered Statistic
KVANORFO2	0.1205419
KVANORFO20	0.03537467
KVANORFO72	0.2666065
KVANORFO305	0.2063066
KVANORFO306	0.1161478
KVANORFO22	0.03651071
KVANORFO28	0.02810953
KVANORFO42	0.05843504
KVANORFO66	0.05786412
KVANORFO69	0.2773826

3.5.2 Bayes Factor and Bayesian Setup

In the first approach, based on the fact that there are multiple NCDC weather stations distributed in Norfolk area, we can get the empirical critical value based on the pairwise comparison distribution of the tapered statistics. In this section, we plan to solve the problem based on Bayesian perspective.

From frequentist perspective, p-value is the proportion of unexpected outcomes under repeated tests. From Bayesian view, however, probability quantifies our subjective belief about the likelihood of an outcome, and is rationally updated given new data. The Bayes factor is the ratio of the data likelihoods, compares the marginal likelihood of the data under competing models:

$$BF_0(Y) = \frac{m(Y|H_0)}{m(Y|H_1)} \tag{3.26}$$

where $m(Y|H_0)$ is the marginal density of Y under model H_0 and $m(Y|H_1)$ for the marginal density under model H_1. Thus, the Bayes Factor is the measurement, by considering the collected data at hand, transforms the prior opinion, hence repre-

Table 3.2: Decision Rules for Bayes Factor

log BF	Evidence for H_0
0 to 1	Not worth more than a bare mention
1 to 3	Positive
3 to 5	Strong
greater than 5	Very strong

sents the evidence provided by the data. Consequently, the Bayes Factor is a more informative measure than the posterior odds ratio of the evidence for a given model.

In particular, Bayes factor determines the decision rules for Bayesian hypothesis testing problems as discussed in Kass and Raftery (1995) and Jeffreys (1961). The higher the Bayes factor value supports, the more evidence in favor of H_0. In Kass and Raftery (1995), a rough descriptive statement of some decision rules regarding Bayes factors was empirically addressed in table 3.2. More discussions and empirical examples regarding Bayes factor are provided in Kass and Raftery (1995). Basically, If this ratio is large, we can conclude that there is strong evidence for the null hypothesis. In contrast, if the inverse of this ratio is large, we have evidence supporting the alternative hypothesis.

Based on the definition of the kernel smoothing of log-periodogram as:

$$\tilde{Y}_i = \tilde{\theta}_i + \tilde{\sigma_n}\epsilon \tag{3.27}$$

in which $\sigma_n^2 = \sigma^2 \sum_{k=1}^{[n/2]-1} C_{j,k}^2$

for $j = 1, ..., p_n$. We have the asymptotically independent property that:

$$\tilde{Y}_i \rightarrow N(\tilde{\theta}_i, \tilde{\sigma_n}^2) \tag{3.28}$$

The objective of inference is to test the hypotheses:

$$H_0 : \theta_j = 0 \text{ for every } j = 1, \ldots, p_n$$

$$H_1 : \theta_j \neq 0 \text{ for some } j = 1, \ldots, p_n$$

Essentially, we can treat this problem as a Gaussian means problem: Suppose a sample of n independent p-dimensional measurements, $Y = (Y_1, ..., Y_p)$ is observed. The data are generated from Gaussian distributions, $Y_i|\Sigma \sim G(\theta, \Sigma)$. Under H_0, the mean vector parameter is restricted to $\theta = 0$ while unrestricted under H_1. Suppose further that the prior distribution under H_1 is such that $\theta|\Sigma \sim G(0, \tau^2\Sigma^{1/2}\Delta\Sigma^{1/2})$, where Δ is a symmetric, positive-definite matrix. It follows that the conditional Bayes factor for the model-comparison H_0 vs. H_1 is:

$$\begin{aligned}
BF_{01} &= \frac{m(\tilde{Y}|H_0)}{m(\tilde{Y}|H_1)} \\
&= N(0,\Sigma)/N(0,\Sigma+\tau^2\Sigma^{1/2}\Delta\Sigma^{1/2}) \\
&= (\tau^2 n)^{v/2}|\Delta|^{1/2}|W|^{-1/2} exp\{-\frac{1}{2}Z^T W Z\} \\
&= \prod_{j=1}^{p_n}\Big\{(1-v_{n,j})^{-\frac{1}{2}}\exp\Big[-\frac{1}{2\sigma_n'^2}(\tilde{Y_j}^2-(1-v_{n,j})\tilde{Y_j}^2)\Big]\Big\} \\
&= \prod_{j=1}^{p_n}\exp\Big\{-\frac{1}{2\sigma_n'^2}v_{n,j}\tilde{Y_j}^2-\frac{1}{2}log(1-v_{n,j})\Big\} \\
&= \exp\Big\{-\frac{1}{2\sigma_n'^2}\sum_{i=1}^{p_n}v_{n,j}\tilde{Y_j}^2-\frac{1}{2}log(1-v_{n,j})\Big\} \qquad (3.29)
\end{aligned}$$

where $Z = n^{1/2}\Sigma^{-1/2}\bar{Y}$, $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, and $W = \{I + \Delta^{-1}/(\tau^2 n)\}^{-1}$.

The choice of τ, however, can be subjective. Under this setting, we choose $\tau = 1$. Note that in equation (3.29), the quadratic term $\sum_{i=1}^{p_n} v_{n,j}\tilde{Y_j}^2$ in the Bayes Factor can

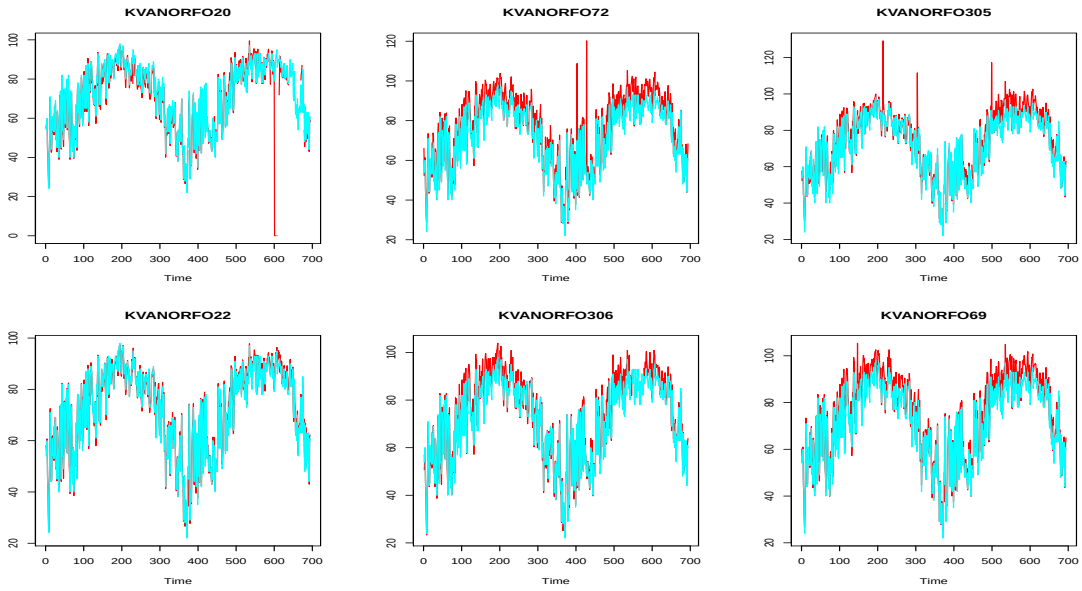

Figure 3.12: Time Series Plot of PWS vs NCDC for Max Temperature (Blue for NCDC, Red for PWS)

Table 3.3: log BF for PWS

PWS	log BF
KVANORFO2	43.74921
KVANORFO20	-8.630177
KVANORFO72	47.90501
KVANORFO305	49.67174
KVANORFO306	-203.297
KVANORFO22	43.74921
KVANORFO28	49.04882
KVANORFO42	2.686782
KVANORFO66	-17.72713
KVANORFO69	46.34115

relate to the optimal test statistic $\tilde{Q_n} = \sum_{i=1}^{p_n} j^{-\frac{1}{2}} \tilde{Y_j}^2$, which leads us to set the value of $v_{n,j}$ such that

$$v_{n,j} = j^{-\frac{1}{2}}, j = 1, 2, ..., p_n$$

and borrow the idea of proposed Gaussian smoothed term with the test statistic

$$\tilde{Q_n} = \sum_{j=1}^{p} \frac{1}{j^{1/2}} \tilde{Y_j}^2$$

of the optimal bandwidth:

$$\hat{b_n} = B[\frac{1}{2} p_n^{3/4} (\log p_n)^{1/4} - \frac{1}{2}]$$

As discussed in the previous part in this chapter, this rate is optimal among tests based on quadratic forms.Thus this weight setting achieves the prior that, from an adaptive "rate of testing" viewpoint, reduce the indistinguishable region to the greatest extent.

With this method applied, the result of the logrithm of Bayes Factoris shown in Table 3.3, which shows that PWS KVANORFO20, KVANORFO306 and KVANORFO66 turn out to be abnormal.

Chapter 4

Problem Formulation on Precipitation Data

Time series hypothesis testing on precipitation data, unlike the temperature data, can be more challenging due to the difficulty in applying standard tests given the zero-inflated and non-stationary process.

Hamill (1999) performed the lag-one Spearman rank correlation analysis, showing that there is no statistically significant correlation. Hence, it can be assumed that rainfall occurred each day may effectively be treated as independent from previous and subsequent days. Thus, Hamill (1999) evaluated several possible hypothesis test methods, including the paired t-test, the nonparametric Wilcoxon signed-rank test, and two resampling tests.

Besides, Hamill (1999) also introduced the equitable threat score, and the problem can be summarized to the contingency table. Threat score can be regarded as evidence of a difference in forecast skill between competing forecasts, which are generated from a contingency table of hits, misses, false alarms, and correct no forecasts:

In our research, we are very interested in whether PWS can successfully capture

Table 4.1: Contingency table of possible events

	Observed	
Forecast	Yes	No
Yes	a	b
No	c	d

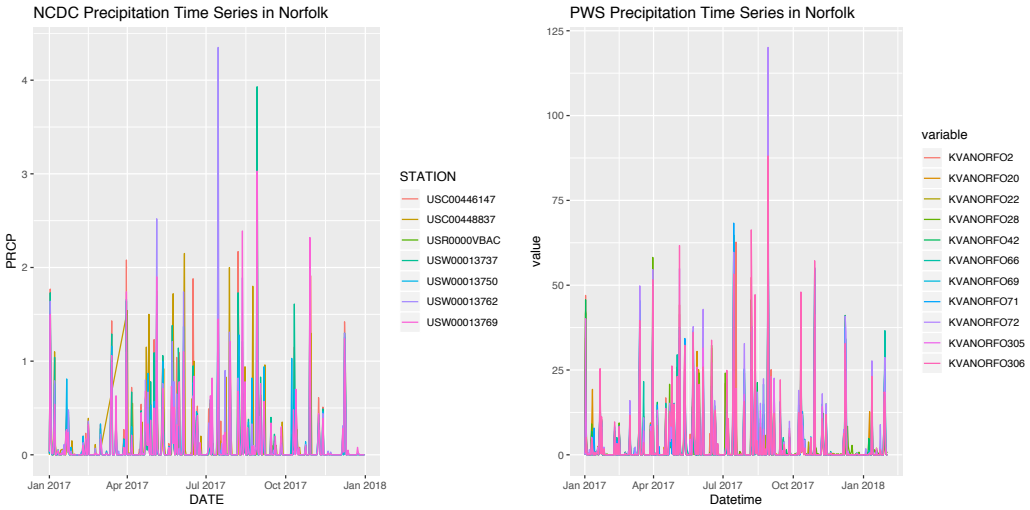

Figure 4.1: Time Series Plot of NCDC and PWS Rainfall in Norfolk

the flood event. Thus this idea can be borrowed to generate a contingency table of hits, misses, false alarms, and correct no flood. Figure 4.1 illustrates the rainfall time series plot of both PWS and NCDC stations in the Norfolk area, and we are interested in proposing the method to detect the abnormal PWS concerning precipitation.

The left side of Figure 4.2 shows that the precipitation time series can be either zero, which means there is no rainfall on a given day, or there might be positive continuous values on a rainy day. The right side of Figure 4.2 is the histogram of precipitation data.

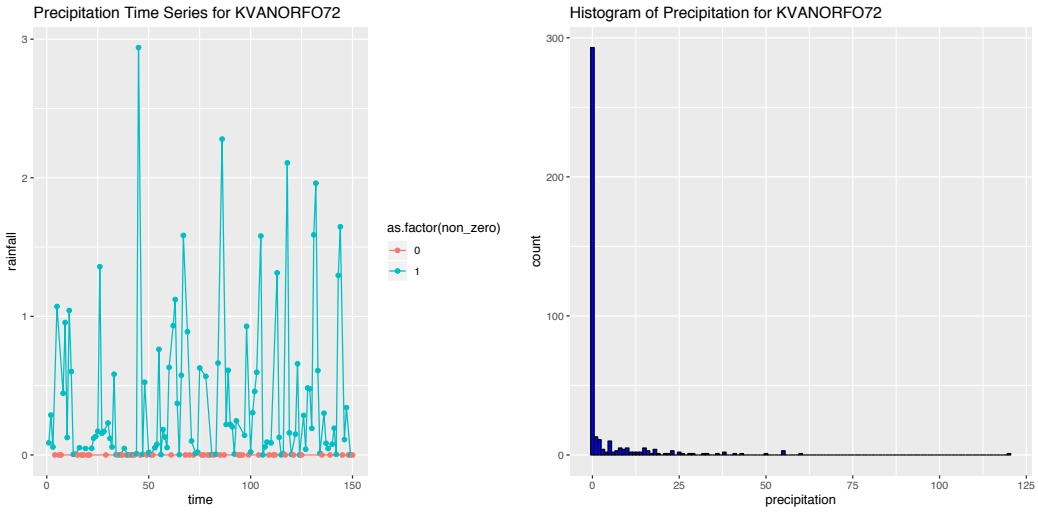

Figure 4.2: Time Series Plot of Precipitation of KVANORFO72

4.1 Modeling Precipitation Data as Tweedie Family

Rainfall data is generally zero-inflated in that the amount of rainfall received on a day can be zero with a positive probability but continuously distributed otherwise, which makes it challenging to transform the data to normality by power transforms or to model it directly using continuous distribution. The Poisson-Gamma distribution is an appropriate way of modelling the rainfall distribution. However, it has a complicated probability density function, making parameters challenging to estimate. Thus, it is easier to estimate the parameters in terms of a Tweedie distribution.

Jørgensen (1987) established the exponential dispersion model (EDM), which is a two-parameter family of distributions consisting of a linear exponential family with an additional dispersion parameter. EDMs are important in statistics because they are the response distributions for generalized linear models. An EDM is characterized by its variance function $V()$, which describes the mean-variance relationship of the distribution when the dispersion is held constant. If Y follows an EDM distribution

with mean μ and variance function $V()$, then

$$var(Y) = \phi V(\mu)$$

where ϕ is the dispersion parameter.

Of special interest are EDMs with power mean-variance relationships,

$$var(\mu) = \mu^p$$

This class of EDM is named after Tweedie. Some very familiar distributions fall into the Tweedie family. Setting $p = 0$ gives a normal distribution. $p = 1$ is Poisson. $p = 2$ gives a gamma distribution and $p = 3$ yields an inverse Gaussian. The Tweedie model distributions for $1 < p < 2$ can be represented as Poisson mixtures of gamma distributions. They are mixed distributions with mass at zero but are otherwise continuous on the positive real values.

McCullagh and Nelder (1989) showed rainfall distributions belong to the exponential family of distributions based on generalized linear models. Let R_i denote the amount of precipitation on the i-th event, and suppose it follows a gamma distribution with parameters (α, γ), which tells us that the mean value for the precipitation is $-\alpha\gamma$ and variance is $-\alpha\gamma^2$. We want to model precipitation for N days and assume that the number of precipitation events in any given day has a Poisson distribution, which implies that there are days with no precipitation events. The total daily precipitation Y can be found as the Poisson summation of the gamma random variables, which means

$$Y = R_1 + R_2 + ... + R_N$$

where $N \sim Poisson(\lambda)$.

The resulting distribution has been called a compound Poisson distribution. However, numerical methods are needed as no closed forms exist for evaluating the density function or cumulative distribution functions, except in special cases. The resulting probability function is complicated, and can be written as:

$$\log f_p(y; \mu, \phi) = \begin{cases} -\lambda & y = 0 \\ -y/\gamma - \lambda - \log y + \log W(y; \phi, p) & y > 0 \end{cases}$$

where $\gamma = \phi(p-1)\mu^{p-1}$, $\lambda = \mu^{2-p}/[\phi(2-p)]$, and W is an example of Wright's generalized Bessel function (Wright, 1933), which is expressed as:

$$W(y, \phi, p) = \frac{y^{j\alpha}(p-1)^{\alpha j}}{\phi^{j(1-\alpha)}(2-p)^j j!\Gamma(-j\alpha)}$$

Importantly, the probability of recording no precipitation is given by:

$$P(Y = 0) = exp(-\lambda) = exp(-\frac{\mu^{2-p}}{\phi(2-p)})$$

In theory, the Tweedie distribution can be characterized by the parameters (μ, ϕ, p), however, in the real world under the climatological setting, we would prefer the parameterization in terms of $(\lambda, \gamma, \alpha)$. In the context of precipitation, λ refers to the mean number of precipitation events per year, γ refers to the shape of the precipitation events, and $-\alpha\gamma$ refers to the mean quantity of precipitation per event.

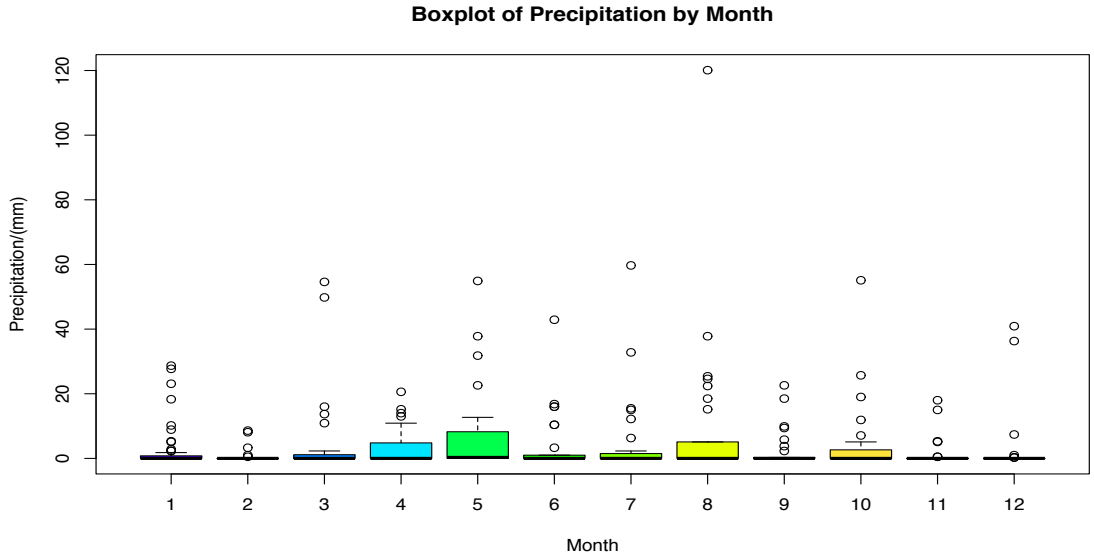

Figure 4.3: Boxplot of Precipitation by Month

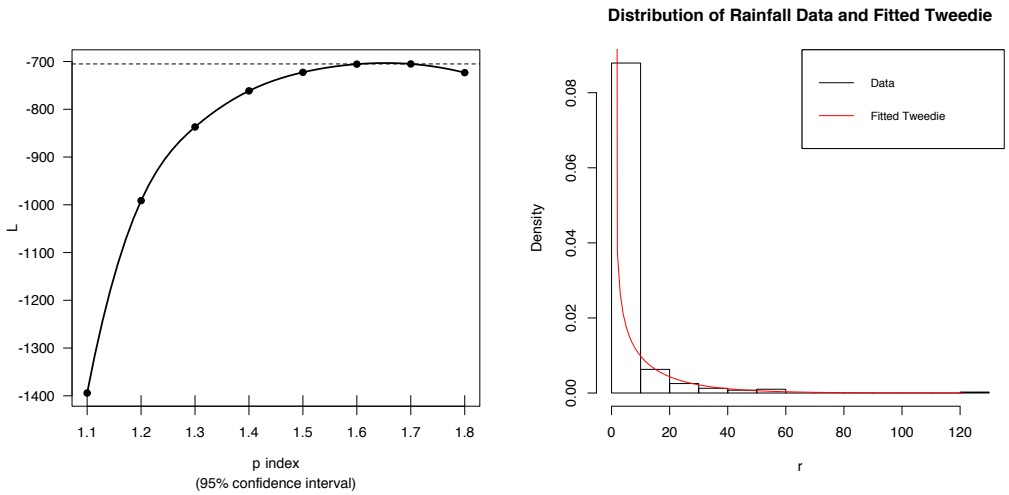

Figure 4.4: Time Series Plot of Precipitation of KVANORFO72

Based on the previous discussion, the Tweedie family is a three-parameter family: μ (the mean), ϕ (the dispersion parameter) and p. The estimation procedure is to plot a profile log-likelihood based on the maximum likelihood value, then get the estimate of p. Once the value of p is found, estimates of and can be computed correspondingly, so the distribution within the class of Tweedie distribution is identified.

The methods for estimating the parameters of Tweedie distributions are discussed. The estimation of μ is straightforward since we restrict ourselves to using no covariates, μ can be estimated by the sample mean. ϕ can be estimated using the maximum likelihood estimate. Estimating the maximum likelihood value of p is performed using a profile log-likelihood plot, which requires the computation of the density. For a given fixed value of p, estimates of μ and ϕ can be computed. Nominal confidence intervals for p can also be found, using that $2[\log L(\hat{p}) - \log L(p_0)]$ has χ_1^2 distribution asymptotically, where p_0 is the true parameter value.

The left side of Figure 4.4 shows the maximum likelihood estimation of the Tweedie index parameter power p, basically, by estimating the power parameter p by maximizing the profile log-likelihood across the grid values of p in the range of 1.1 to 1.8, the value of p corresponding to the MLE is approximately estimated at 1.65 in this case, and the estimate of ϕ is 10.96688.

Tweedie distribution, with its property of taking care of zero-inflated data distribution, is suitable for modelling the distribution of precipitation data. However, it has a complicated probability density function, making it hard to be applied to the rate-of-testing theory, so it has not been incorporated into the current work yet.

4.1.1 Hypothesis Testing For Precipitation Data

Precipitation, unlike temperature data, has its unique property of inflated with zero values, thus making it difficult to be converted into stationary time series, which makes our previous approach of hypothesis testing by smoothing not applicable here. However, we come up with two approaches to deal with this issue. The first approach is to smooth the precipitation data with the surrounding values, making it more stationary. The second approach is a two-step process. In the first step, we reduce the data to indicators of the presence or absence of precipitation and carry out analysis for binary time series. On condition that the binary time series can pass the test, we carry forward to the next step of applying time-series hypothesis testing analysis for the non-zero subspace of precipitation data.

Approach one: Time Series Hypothesis Testing For Smoothed Precipitation Data

A straightforward way to deal with the non-stationary time series is to convert it to stationary process, so in the first approach, we smooth the precipitation data with the surrounding values, to eliminate the presence of zero values. Figure 4.5 is an illustration of PWS KVANORFO20 when smoothing technique is applied, we are making the rainfall data stationary.

For the PWSs in Norfolk area, the smoothing method is applied to both PWS and NCDC precipitation data, then we carry out the testing procedure based on the previous chapter, and the consequent results table is shown in Table 4.2. Figure 4.6 shows the smoothed time series of PWS KVANORFO20 and KVANORFO72 compared with smoothed series of interpolated NCDC data. From the visualization we can see that the smoothed data of PWS KVANORFO20 matches quite well with

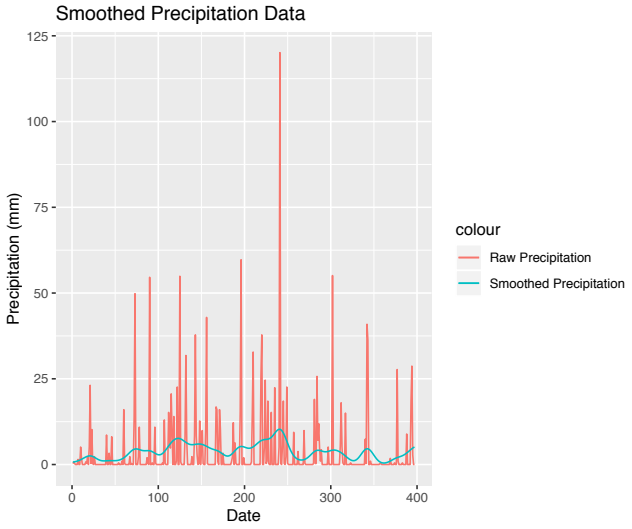

Figure 4.5: Raw Data and Smoothed Value of Precipitation

smoothed interpolated NCDC data, which agrees with the result that the log Bayes Factor for this station is 5.49975, meaning there is the strong evidence that the null hypothesis is true. PWS KVANORFO72, however, as shown on the right side of Figure 4.6, exhibit distinct discrepancies from the NCDC data, with -1908.528 as the log Bayes Factor, so we can get the conclusion that the null hypothesis is rejected in this case, indicating the station PWS KVANORFO72 is not trustworthy.

Approach two: Two-Step Testing Procedure For Zero-Inflated Data

Although the smoothing method is an effective way of converting the time series into the stationary process, however, it might introduce extra noise to the raw data. So the second approach only deals with the raw data itself, with the treatment of only extracting the non-zero values to avoid the effect of zero-inflated distribution.

- Pass One: Reduce the data to indicators of the presence or absence of precipitation and carry out analysis for binary time series.

- Pass Two: On condition that the binary time series can pass the test, we will

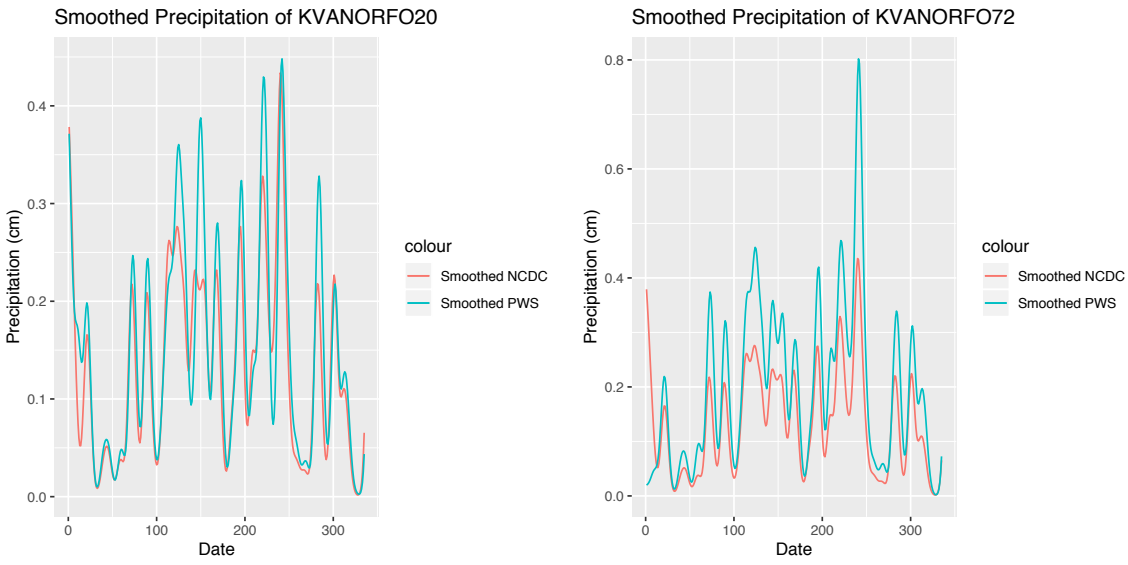

Figure 4.6: Raw Data and Smoothed Value of Precipitation

Table 4.2: log BF for PWS

PWS	log BF
KVANORFO2	-333.4697
KVANORFO20	5.49975
KVANORFO22	-87.52388
KVANORFO28	-22.98318
KVANORFO42	-685.3949
KVANORFO66	18.40082
KVANORFO69	-172.1699
KVANORFO72	-1908.528
KVANORFO305	-203.2546
KVANORFO306	-666.7461

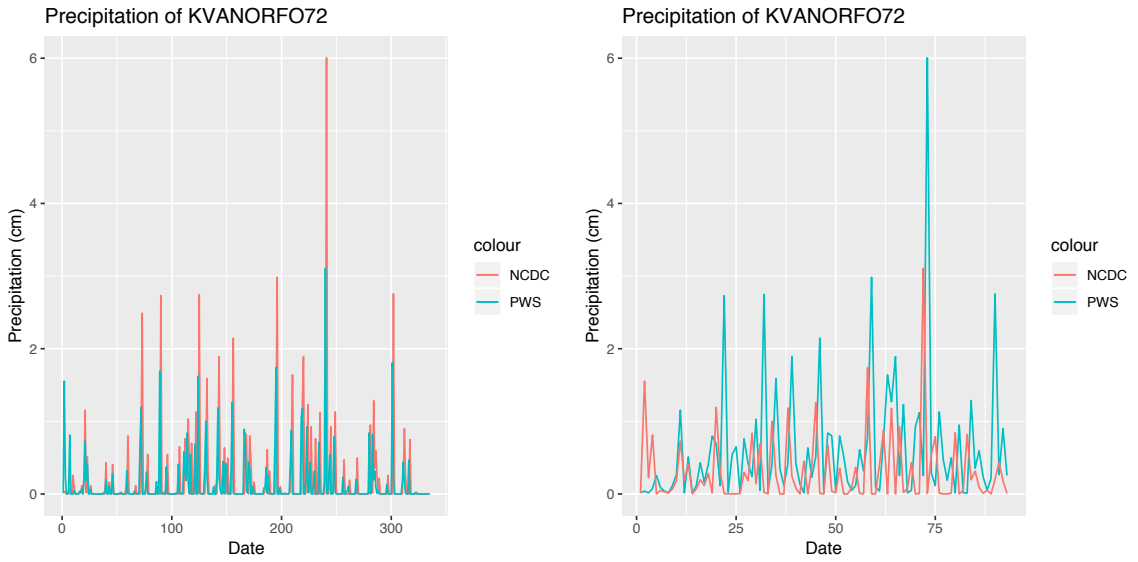

Figure 4.7: Emperical Power for Tweedie Family

apply our analysis for the non-zero subspace of precipitation data.

Once we have the all-year precipitation data, we can truncate the non-zero rainfall data and find the overlap of both PWS and NCDC. Figure 4.7 and Figure 4.8 illustrate rainfall data on the left side and the truncated non-zero data on the right side for both PWS KVANORFO20 and KVANORFO72. After carrying out the testing procedure based on the previous chapter, the following results table is shown in Table 4.3. PWS KVANORFO20 matches quite well with interpolated NCDC data, which agrees with the result that the log Bayes factor for this station is 26.53325, meaning there's the strong evidence that the null hypothesis is true. For PWS KVANORFO72, however, as shown on the right side of Figure 4.8 , exhibits several obvious discrepancies from the NCDC data, with -1326.598 as the log Bayes Factor, so we can get the conclusion that the null hypothesis is rejected in this case, indicating the station PWS KVANORFO72 is not trustworthy.

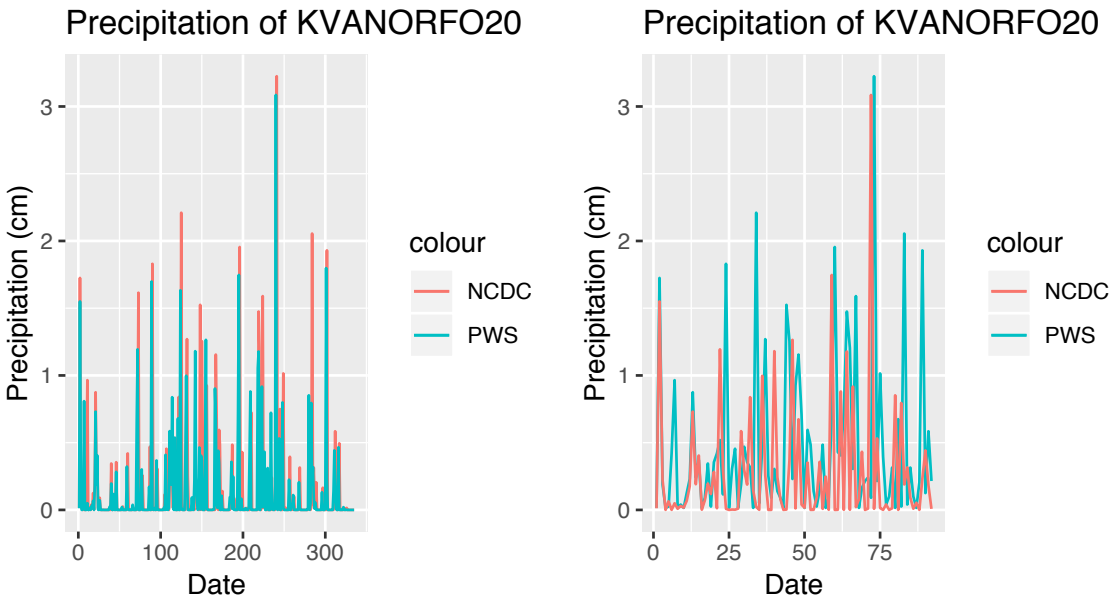

Figure 4.8: Emperical Power for Tweedie Family

Table 4.3: log BF for PWS

PWS	log BF
KVANORFO2	-49.06801
KVANORFO20	26.53325
KVANORFO22	-44.09912
KVANORFO28	-22.98318
KVANORFO42	-338.7884
KVANORFO66	-12.02769
KVANORFO69	19.94753
KVANORFO72	-1326.598
KVANORFO305	21.78035
KVANORFO306	20.64477

Chapter 5

Spectrum-Based Comparison of Multivariate Time Series

5.1 Multivariate Time Series Setting

In the previous work, we come up with a new non-parametric testing approach to test the hypothesis that the two underlying spectra are the same, which is only applied to stationary univariate time series. In this section, however, motivated by our problem that weather stations report multiple weather measurements, we would like to explore the tapering test based on the smoothed periodogram for multivariate time series context. Figure 5.1 shows the time series structure plot of minimum temperature, maximum temperature and average temperature of both PWS and NCDC, and we are motivated by the question of comparing multiple weather measurements jointly and test whether their underlying spectral matrices are the same.

Consider two stationary and independent p-dimensional processes $\{X_t\}$ and $\{Y_t\}$. Let $x_1, ..., x_T$ and $y_1, ..., y_T$ denote T observations from $\{X_t\}$ and $\{Y_t\}$, μ_x and μ_y denote their respective means, let $\lambda_X(\omega_l)$ and $\lambda_Y(\omega_l)$ denote their respective autocovari-

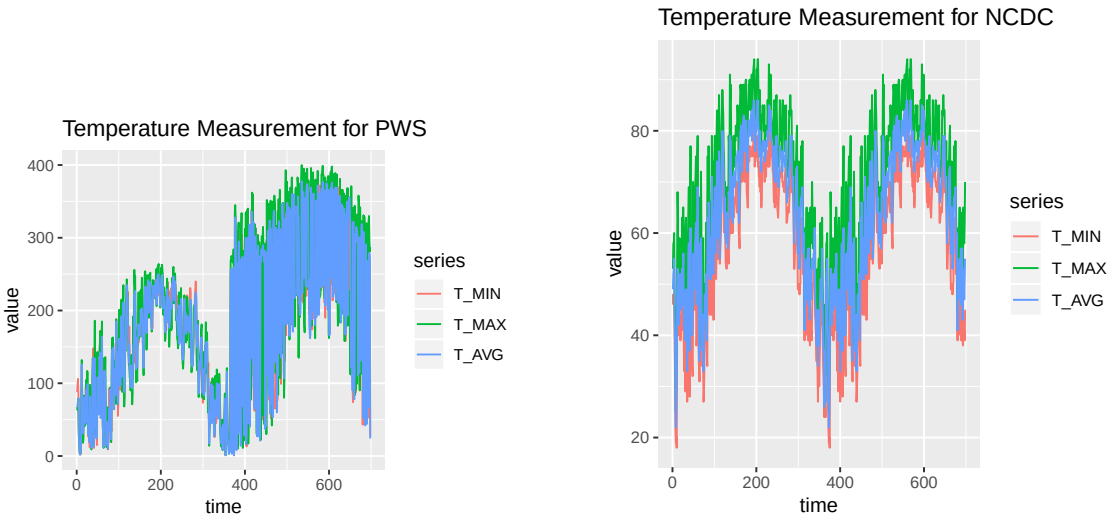

Figure 5.1: Multiple Time Series Plot of PWS and NCDC

ance matrices at lag r, and let$\boldsymbol{\lambda}_X(\omega_l) = \{\lambda_{X,jk}(\omega_l)\}$ and $\boldsymbol{\lambda}_Y(\omega_l) = \{\lambda_{Y,jk}(\omega_l)\}$ denote their respective spectral density matrices at Fourier frequency ω_l where $\omega_l = 2\pi l/T$, $l = 1, 2, ..., [T/2]$. The components of the spectral matrices are defined as

$$\lambda_{X,jk}(\omega) = 1/2\pi \sum_{r=-\infty}^{\infty} e^{-ir\omega}\gamma_{X,jk}(r), j, k = 1, ..., p \tag{5.1}$$

and

$$\lambda_{X,jk}(\omega) = 1/2\pi \sum_{r=-\infty}^{\infty} e^{-ir\omega}\gamma_{X,jk}(r), j, k = 1, ..., p \tag{5.2}$$

where $i = \sqrt{-1}$.

To assess whether the two multivariate series are similar, the approach is based on a test of the equality of their spectral density matrices, i.e., consider the null hypothesis:

$$H_0 : \lambda_X(\omega) = \lambda_Y(\omega), \forall\omega \in (-\pi, \pi)$$

where ω_X, ω_Y are the spectral density matrices of multivariate time series $\{X_t\}$,$\{Y_t\}$. This formulation assumes the variances are the same, so we will standardize the data.

Since the spectral density matrice is unknown, we need to estimate it. The periodogram is a commonly used way of estimating the spectral density. The periodogram of a time-series is the sample analogue of the spectral density. The $p * p$ raw periodogram matrix of $\{X_t\}$ at the Fourier frequency ω_l is $\boldsymbol{I}_X(\omega_l) = \{I_{X,jk}(\omega_l)\}$ where

$$I_{X,jk}(\omega) = \frac{1}{2\pi T}(\sum_{i=1}^{T} x_{tj}e^{-it\omega_l})(\sum_{i=1}^{T} x_{tk}e^{-it\omega_l}) \tag{5.3}$$

and

$$I_{Y,jk}(\omega) = \frac{1}{2\pi T}(\sum_{i=1}^{T} y_{tj}e^{-it\omega_l})(\sum_{i=1}^{T} y_{tk}e^{-it\omega_l}) \tag{5.4}$$

Similar to the case of univariate time series, under multivariate time series settings, raw periodogram matrix is asymptotically unbiased but not consistent of the spectral density matrix. Brockwell and Davis (1991) came up with the smoothed periodogram estimate using the Daniell window at ω_l, then the average of the raw periodogram ordinates in a neighborhood of ω_l is defined as:

$$I_X^{\star}(\omega_l) = (2m+1)^{-1} \sum_{|k| \leqslant m} I_X(\omega_{l+k}) \tag{5.5}$$

for $\omega \in (-\pi, \pi)$, where $2m+1$ is a positive integer related to the degrees of freedom of a complex Wishart distribution. It is known that

$$I_X^{\star}(\omega) \approx \frac{1}{2m+1} W_c(2m+1, p, \lambda_X(\omega)) \tag{5.6}$$

5.1.1 Likelihood Ratio Test Based on Cross-Spectra

Ravishanker et al. (2010) defined a similarity measure using the smoothed periodogram estimates of the spectral matrices for two stationary multivariate time series. The likelihood functions $L(\lambda_X(\omega))$, $L(\lambda_Y(\omega))$ and $L(\lambda_0(\omega))$ can be derived as:

$$L(\lambda_X(\omega), \lambda_Y(\omega)) = \frac{1}{\Gamma_p(2m+1)^2} |I_X^\star(\omega) I_Y^\star(\omega)|^{2m+1-p} |\lambda_X^\star(\omega)\lambda_X^\star(\omega)|^{-(2m+1)} \\ *exp[-tr\{\lambda_X^{-1}(\omega) I_X^\star(\omega) + \lambda_Y^{-1}(\omega) I_Y^\star(\omega)\}]$$

Under the null hypothesis, define:

$$Q(\omega) = \frac{L_0(\lambda(\omega))}{L_X(\lambda(\omega)) L_Y(\lambda(\omega))} = 2^{2p(2m+1)} \frac{|I_X^\star(\omega)|^{2m+1} |I_X^\star(\omega)|^{2m+1}}{|I_X^\star(\omega) + I_Y^\star(\omega)|^{2m+1}} \tag{5.7}$$

An effective overall test statistic for H_0 is the unweighted average of the M smallest $Q(\omega_j)$. The quasi-distance $Q_{XY}^\star$ is defined as the average of the M smallest values of the likelihood ratio:

$$Q_{XY}^\star = \frac{1}{M} \sum_{j=1}^{M} Q_{XY}(\omega_{(j)}) \tag{5.8}$$

Ravishanker et al. (2010) give a subjective guideline for selecting the value of M in the definition of $Q_{XY}^\star$, where M is chosen large enough to provide some smoothing of Q without being too large to miss details of the spectra. The idea is to choose M sufficiently large such that local variations in the ratios $Q(\omega_j)$ over the Fourier frequencies should be smoothed out, but also to keep it small enough so that we can uncover sufficient detail in the overall spectra, and be sensitive to relatively narrow peaks in the spectra. Ravishanker et al. (2010) point out that a practical guideline is to take $\sqrt{T} \leqslant M \leqslant T/10$. However, the author did not provide a specific criterion

of selecting the value of M, in our research, we use the optimal M based on the rate-of-testing theory as the guideline of the smoothing methods.

5.1.2 Tapered Test for the Multivariate Settings

As discussed earlier, Ravishanker et al. (2010) propose a quasi-distance between the series based on the pairwise comparison between two multivariate stationary time series via a likelihood ratio test based on the estimated cross-spectra of the series yields. Ravishanker et al. (2010) give a subjective guideline for selecting the value of M in the definition of $Q^{\star}_{XY}$. In this section, for the goal of comparing two spectral matrices, we borrow the idea of tapered testing developed for the univariate time series and apply to the multivariate scenario. With our proposed tapering test based on Gaussian kernel smoothing of test

$$\tilde{Q}_n = \sum_{j=1}^{p} \frac{1}{j^{1/2}} \tilde{Y_j}^2$$

with the optimal bandwidth:

$$\hat{b_n} = B[\frac{1}{2} p_n^{3/4} (\log p_n)^{1/4} - \frac{1}{2}]$$

In the simulation study, we apply this kernel smoothing method to each time series of bivariate time series simultaneously and compare the performance with the approach by Ravishanker et al. (2010).

5.1.3 Simulation Design for Multivariate Time Series Setting

In this section, we present a comprehensive simulation study to illustrate the performance of our proposed tapered test under the multivariate time series setting.

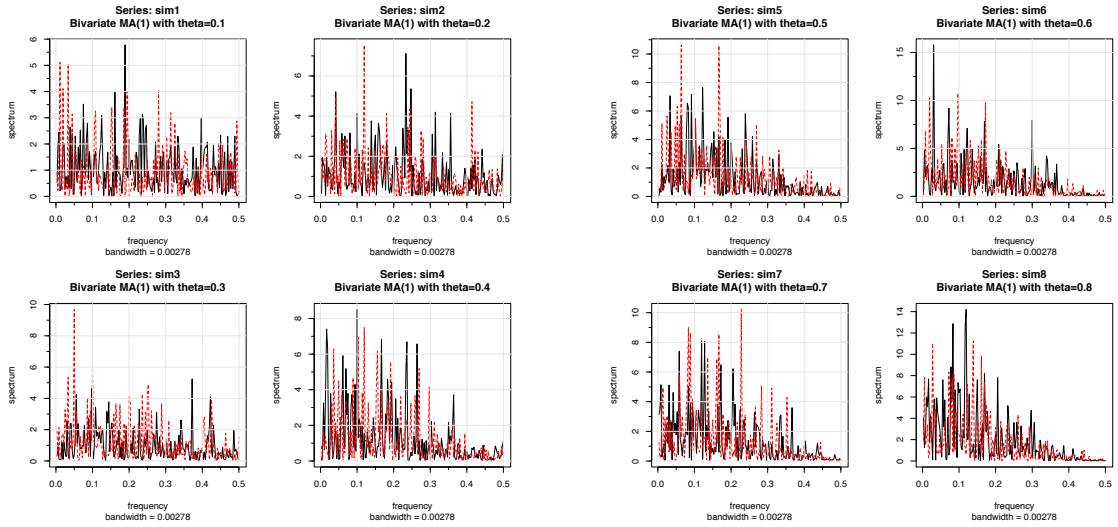

Figure 5.2: Raw periodogram of Bivariate MA(1) models of various parameter values

The simulation design is based on a bivariate VAR time series under three different scenarios:

- Setting 1: Bivariate first-order MA(1) with series generated from each of two different distribution, the null distribution is defined by MA(1) $\theta_1 = 0.1$ versus MA(1) with $\theta_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

- Setting 2: Bivariate first-order AR(1) with series generated from each of two different distribution, the null distribution is defined by AR(1) $\phi = 0.1$ versus AR(1) with $\phi_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

- Setting 3: Bivariate first-order seasonal $MA(1)_{12}$ with series generated from each of two different distribution, the null distribution is defined by$MA(1)_{12}$ with $\theta_1 = 0.1$ versus $MA(1)_{12}$ with $\theta_1^* = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

Figure 5.2 illustrates the raw periodogram of bivariate MA(1) models of various parameter values as listed above, which implies that they have different underlying spectral densities. Figure 5.3, Figure 5.4 and Figure 5.5 show that by applying the

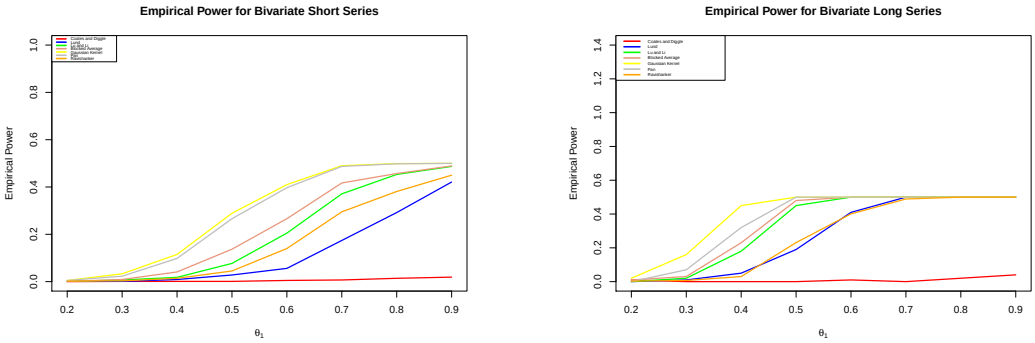

Figure 5.3: Empirical Power for Short and Long Bivariate MA(1) Series

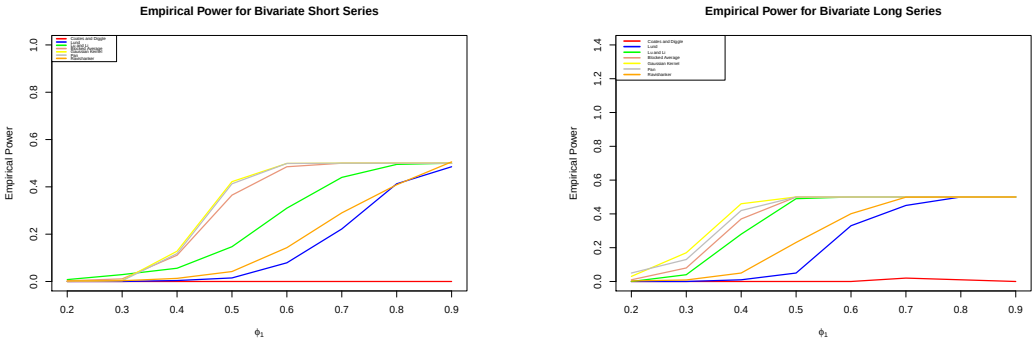

Figure 5.4: Empirical Power for Short and Long Bivariate AR(1) Series

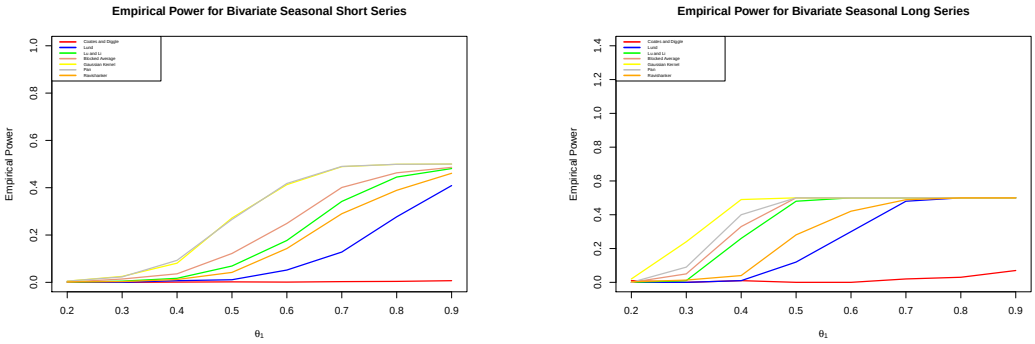

Figure 5.5: Empirical Power for Short and Long Bivariate Seasonal MA(1) Series

optimal bandwidth we derive from the previous chapter under the univariate time series setting, it is also superior to Ravishanker et al. (2010)'s methods under the multivariate settings.

5.2 Bayesian Approach to the Correction of Multiplicity

In the case of variable selections, each of m variables can be chosen independently from huge possible model space with $2m$ model forms. If we assign 2^{-m} to each variable, which is equivalent to assigning each variable the prior of $\frac{1}{2}$ to be included in the final model, as a consequence, no multiplicity control is carried out. Scott and Berger (2010) define a hierarchical model structure for the variable selection in the regression analysis. The context is given that the response vector y is of length n, and the input matrix X of size $n * m$, the regression model is represented as:

$$y_i = \beta_0 + X_{ij}\beta_1 + ... + X_{im}\beta_m + \epsilon_i$$

in which $j = 1, ..., m$, and $i = 1, .., n$. ϵ_i denotes noise of distribution $N(0, \sigma^2)$. The null model is denoted with only the intercept term, represented as M_0, and the full model is defined with all covariates included as M_m. Within the model space, each model is indexed by a length m binary vector γ indicating whether the corresponding coefficient is included or excluded in the regression:

$$\gamma_i = \begin{cases} 0, & if \beta_j = 0 \\ 1, & if \beta_j \neq 0 \end{cases}$$

Scott and Berger (2010) treat the inclusion of every β_j as Bernoulli trial with success probability p, so the overall probability of the included coefficients can be represented as:

$$p(M_\gamma|q) = \prod_{j=1}^{m} q^{\gamma_j}(1-q)^{1-\gamma_j} = q^{k_\gamma}(1-q)^{m-k_\gamma}$$

We have $p \sim Uniform(0,1)$ on p, and the likelihood is

$$f(Y|y) = \prod_{k=1}^{M} f(Y_k|\gamma_k) = \prod_{\gamma_k=1} N(0, \Sigma_n^{H_0}) \prod_{\gamma_k=0} N(0, \Sigma_n^{H_1})$$

The MCMC algorithm can be found as below:

Algorithm 1 Metropolis-Hastings MCMC Algorithm

1: **procedure** RUN_METROPOLIS_MCMC({γ},{L},{BF}) ▷ MCMC Input: Initial State, Iterations, Bayes Factor
2: Define the initial state $\boldsymbol{\gamma^{(0)}}=\left(\gamma_1^{(0)}, ..., \gamma_K^{(0)}\right)$
3: **for** i in 1:L **do**
4: **for** j in 1:K **do** ▷ Update all K elements of $\boldsymbol{\gamma}$, repeat for L times
5: Generate a candidate $\boldsymbol{t_k^{(l)}}$ from the proposal distribution Bernoulli(0.5).
6: The current state $\boldsymbol{S_k^{(l)}}=\left(\boldsymbol{\gamma_1^{(l+1)}, ..., \gamma_{k-1}^{(l+1)}, \gamma_k^{(l)}, \gamma_{k+1}^{(l)}, ..., \gamma_K^{(l)}}\right)$ the candidate state $\boldsymbol{T_k^{(l)}}=\left(\boldsymbol{\gamma_1^{(l+1)}, ..., \gamma_{k-1}^{(l+1)}, t_k^{(l)}, \gamma_{k+1}^{(l)}, ..., \gamma_K^{(l)}}\right)$ with the acceptance ratio $\boldsymbol{\alpha = \min\left\{1, \frac{p(T_k^{(l)}|Y)}{p(S_k^{(l)}|Y)}\right\}}$

The posterior ratio in the algorithm can be derived from the settings of priors:

$$\frac{p(T_k^{(l)}|Y)}{p(S_k^{(l)}|Y)} = BF_{TS_k} * \frac{p(T_k^{(l)})}{p(S_k^{(l)})} = BF_{TS_k} * \frac{n_{T^{(l)}}!(M - n_{T^{(l)}})!}{n_{S^{(l)}}!(M - n_{S^{(l)}})!}$$

In the context of PWS abnormal detection, the multiple measurements are recorded for each PWS, and we take the variables of minimum temperature, maximum temper-

ature to model jointly using the Bayesian approach to correct multiplicity, it will be a promising future direction to access the Integrity of the PWSs based on the overall performance of various variables.

Chapter 6

Time Series Clustering Based Abnormal Detection

6.1 Time Series Clustering

In recent years, there has been increasing research works on unsupervised solutions like clustering algorithms to extract knowledge from a huge amount of data. Clustering is a solution for classifying enormous data when there is not any early knowledge about classes. Clustering is an unsupervised data mining technique where similar data are placed into related or homogeneous groups without previous knowledge of the groups information.

Specifically, time-series clustering has been widely used in diverse areas such as financial, marketing and biomedical time series in order to uncover patterns to extract valuable information from complex and massive datasets. Clustering of time-series data is widely used for the discovery of interesting patterns in time-series datasets. This task can be applied in finding patterns that frequently appears in the dataset as well as to detect abnormal patterns which happened in datasets surprisingly. We

can define the time series clustering problem as below:

Time-series clustering, given a dataset of n time-series data $D = \{F_1, F_2, ..., F_n\}$; the process of unsupervised partitioning of D into $C = \{C_1, C_2, ..., C_k\}$, in such a way that homogenous time-series are grouped together based on a certain similarity measure, is called time-series clustering. Then, C_i is called a cluster, where $D = \cup_{i=1}^{k} C_i$ and $C_i \cap C_j = \varnothing$ for $i \neq j$.

Many types of statistical data analyses have been proposed from both frequentist view and Bayesian view to identifying similarity patterns in heterogeneous observations.

6.2 Hierarchical Clustering Based on Tapered Test Statistics

A wide variety of clustering methods from a frequentist perspective have been proposed. Mac Queen (1967) proposed the K-Means algorithm for producing clusters that separate data into K groups. However, the main drawback of this algorithm is that of a priori fixation of the number of clusters and seeds. Ran Vijay Singh et al. (2011) presented a modified k-means algorithm based on the sensitivity of the initial centre of clusters. Kaufmann and Rousseeuw (1990) introduced the agglomerative hierarchical clustering. It starts by placing each object in its cluster and then merges these atomic clusters into a more massive cluster until certain termination conditions are satisfied. It uses the Single-Link method and the dissimilarity matrix. Kaufmann and Rousseeuw (1990) also introduced divisive hierarchical clustering. It does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.

In agglomerative (bottom-up) clustering method, we assign each observation to its cluster, then compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat those steps until there is only a single cluster left. Traditional hierarchical clustering used Euclidean distance as the distance measure. However, in the case of time series clustering, such distance measure may lose power due to the unique property of time series. Inspired by our previous work, we apply the tapered test statistics as the distance measure and run the preliminary results on personal weather station data. The related algorithm is shown below.

Algorithm 2 Clustering Algorithm

procedure HAC$_{TS}$($\{X_1\}, ..., \{X_n\}$) ▷ TIME SERIES INPUT
  $c_i = \{\{X_i\}\}$
  $C = \{c_1, ..., c_n\}$
  while $C.size > 1$ **do** ▷ REPEAT UNTIL ALL ITEMS ARE CLUSTERED INTO A SINGLE CLUSTER OF SIZE N
    $(c_{min1}, c_{min2}) = \min$ TAPERED(c_i, c_j) FOR ALL c_i,c_j IN C
    REMOVE c_{min1} AND c_{min2} FROM C
    ADD $\{c_{min1}, c_{min2}\}$ TO C
$=0$ **return** C ▷ THE CLUSTERED RESULT IS C

We apply the proposed algorithm to our dataset. In the first step, we select the daily maximum temperature as the variable of interest. We extract time-series data of the year 2017 and 2018, with 689 timestamps in total. The comparison of each personal weather station concerning NCDC data is shown in Figure 3.12. Based on the exploratory data analysis, PWS KVANORFO2, KVANORFO66 and KVANORFO42 seem to be consistent with official data, while part of KVANORFO20 deviates a lot, KVANORFO72 and KVANORFO305 seem to be inconsistent with the ground truth. In the next step, we carry out time-series clustering method to make clusters and uncover the patterns in each cluster. When we set the number of clusters k equal to

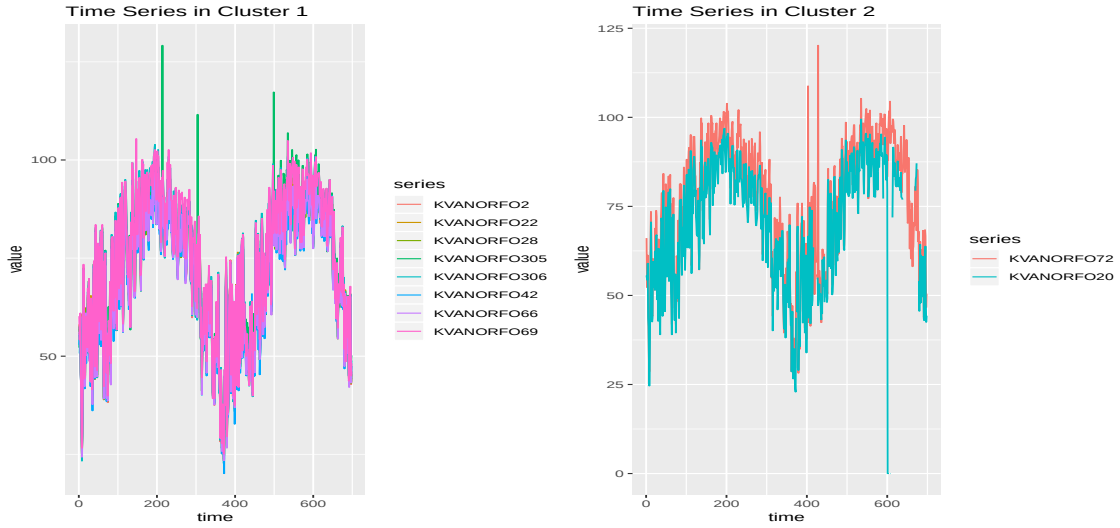

Figure 6.1: Maximum Temperature Hierarchical Clustering for k=2

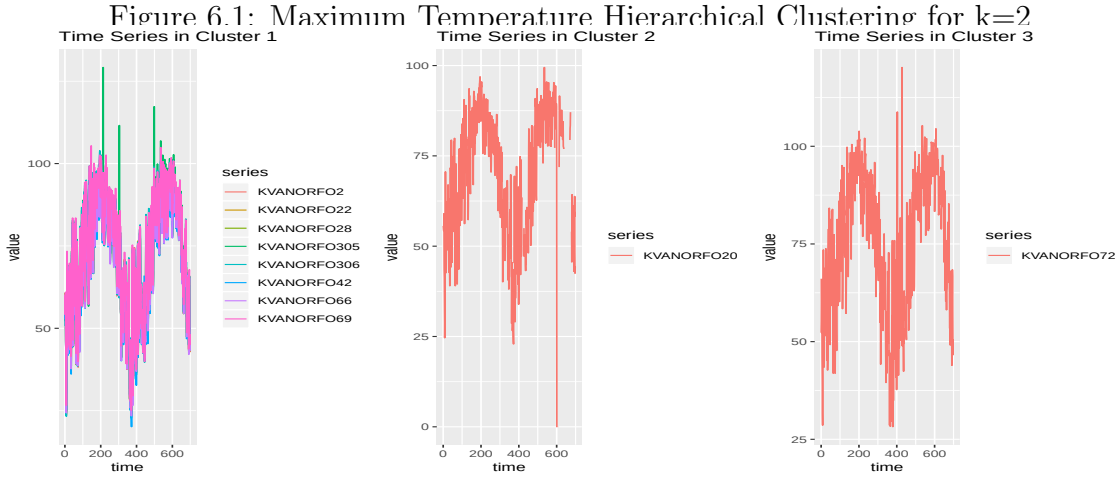

Figure 6.2: Maximum Temperature Hierarchical Clustering for k=3

2, the result is shown in Figure 6.1. Figure 6.2 and Figure 6.3 is the results when we set $k = 3, 4$ separately. All the plots give us very similar conclusions, KVANORFO72, KVANORFO20 are very different from the other personal weather stations time series patterns, and KVANORFO42 and KVANORFO305 look very suspicious when $k = 4$, they are also separated from the clusters of the majority of time series data.

Except for the maximum of daily temperature, we also investigate into the minimum of daily temperature, and Figure 3.12 illustrates that station KVANORFO2 deviates dramatically from the underlying truth at multiple time points. The clus-

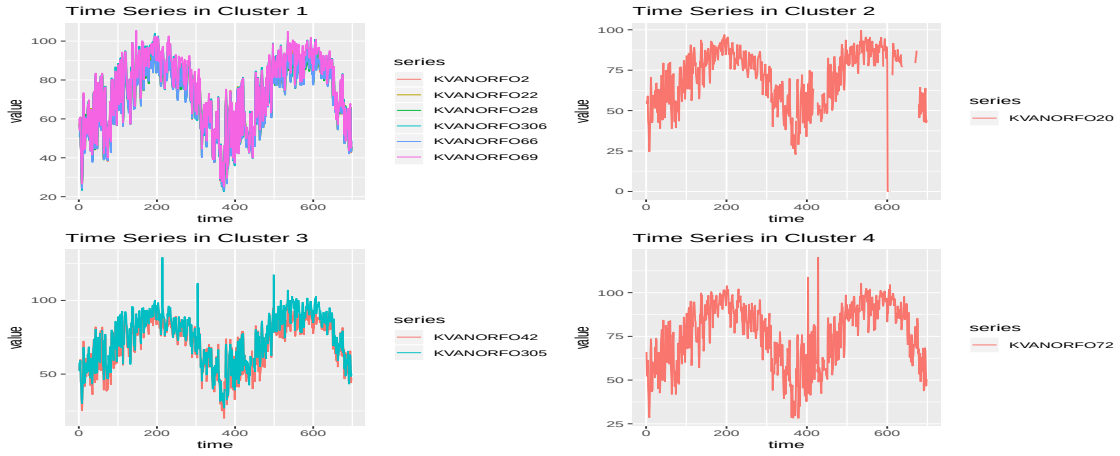

Figure 6.3: Maximum Temperature Hierarchical Clustering for k=4

tering results consolidate the finding that KVANORFO2 is abnormal concerning the minimum daily rainfall report.

Daily precipitation is also modelled by our clustering method. Unlike temperature data, the rainfall data has unique properties: it has a large portion of zeros, and its distribution does not follow the normal distribution. As shown in Figure 1.3, it is hard to see the general trend and seasonality from the time series plot, when clustering method is applied, station KVANORFO72 and KVANORFO306 seem to always in the unique abnormal clusters while the remaining stations tend to distribute equally among the other clusters.

To sum up, stations KVANORFO72, KVANORFO20, KVANORFO306 and KVANORFO2 seem to be untrustworthy, especially station KVANORFO72 is abnormal with regard to minimum daily temperature, maximum daily temperature and daily precipitation.

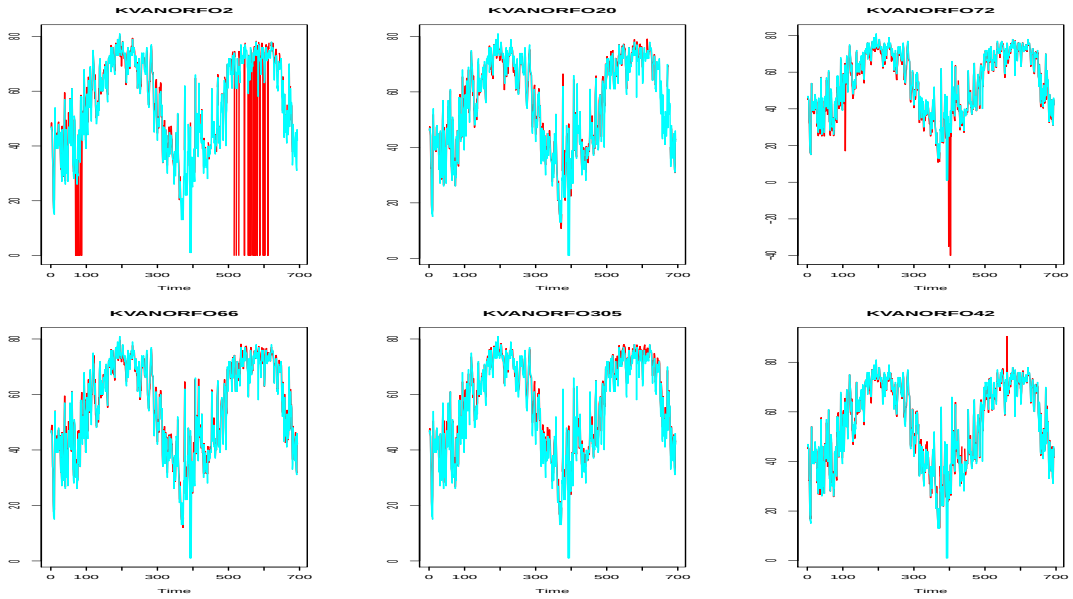

Figure 6.4: Time Series Plot of PWS vs NCDC for Minimum Temperature

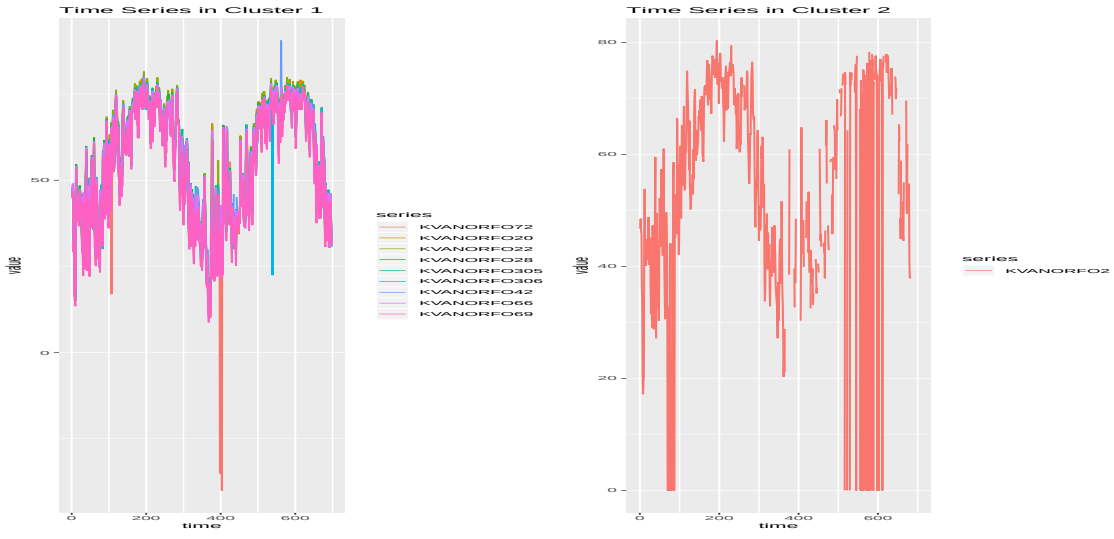

Figure 6.5: Minimum Temperature Hierarchical Clustering for k=2

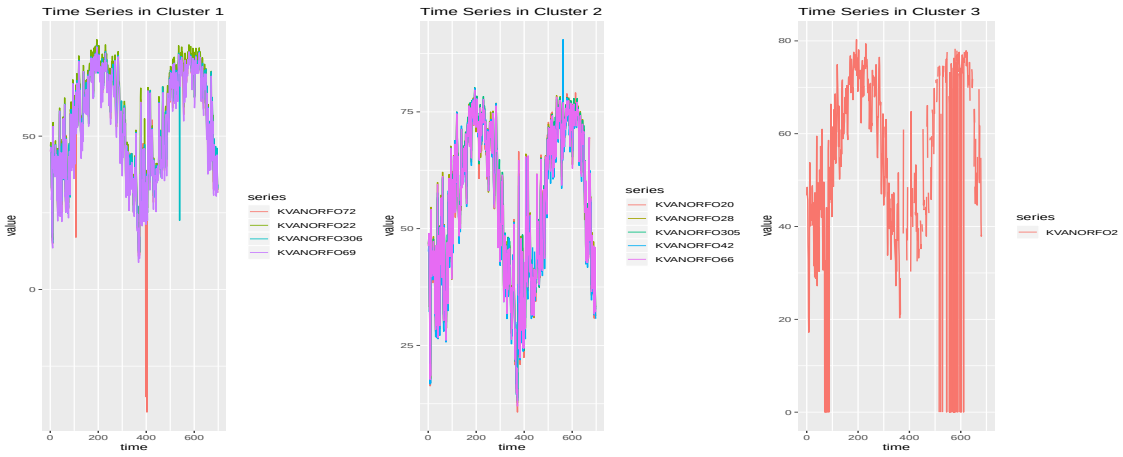

Figure 6.6: Minimum Temperature Hierarchical Clustering for k=3

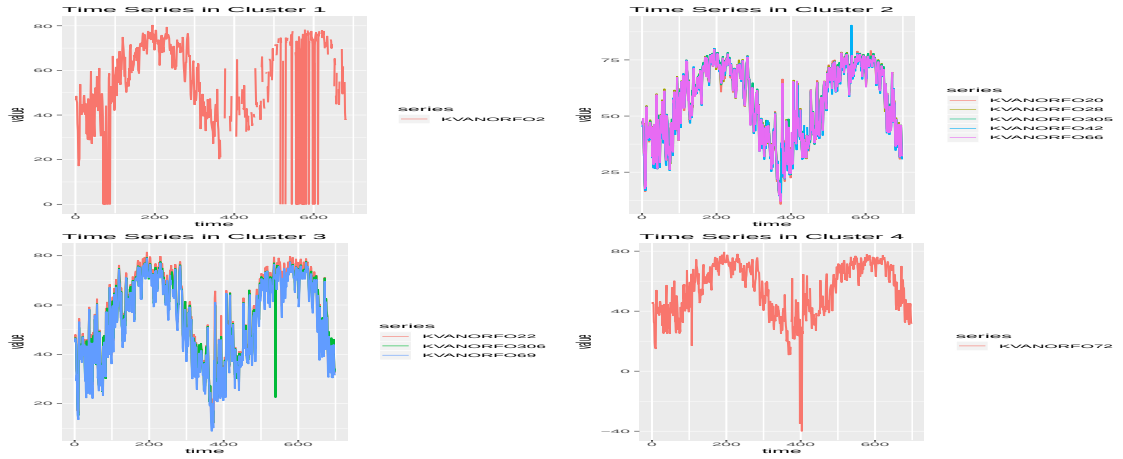

Figure 6.7: Minimum Temperature Hierarchical Clustering for k=4

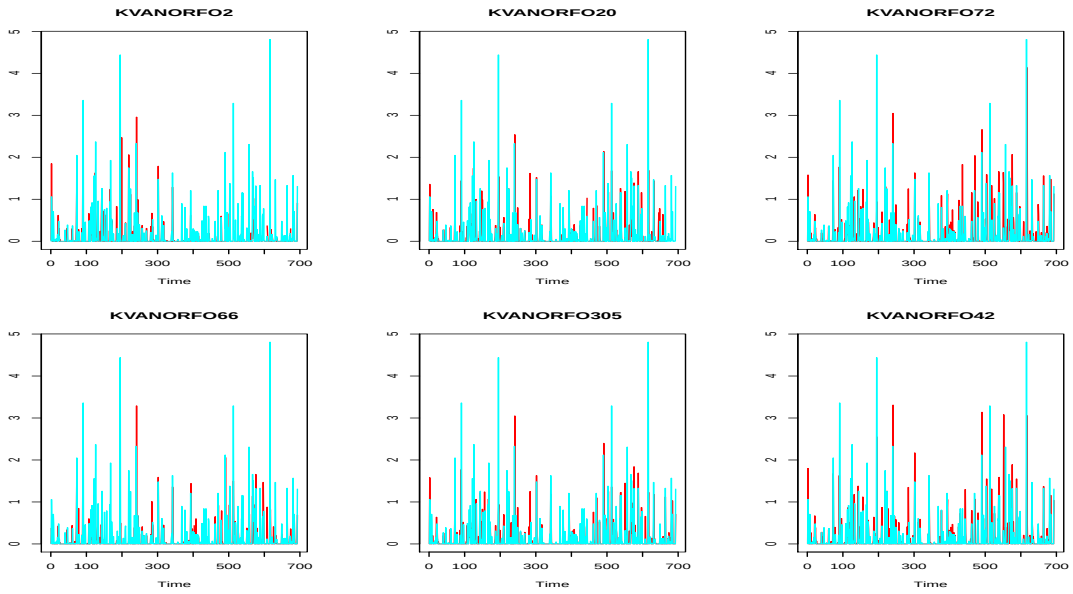

Figure 6.8: Time Series Plot of PWS vs NCDC for Rainfall

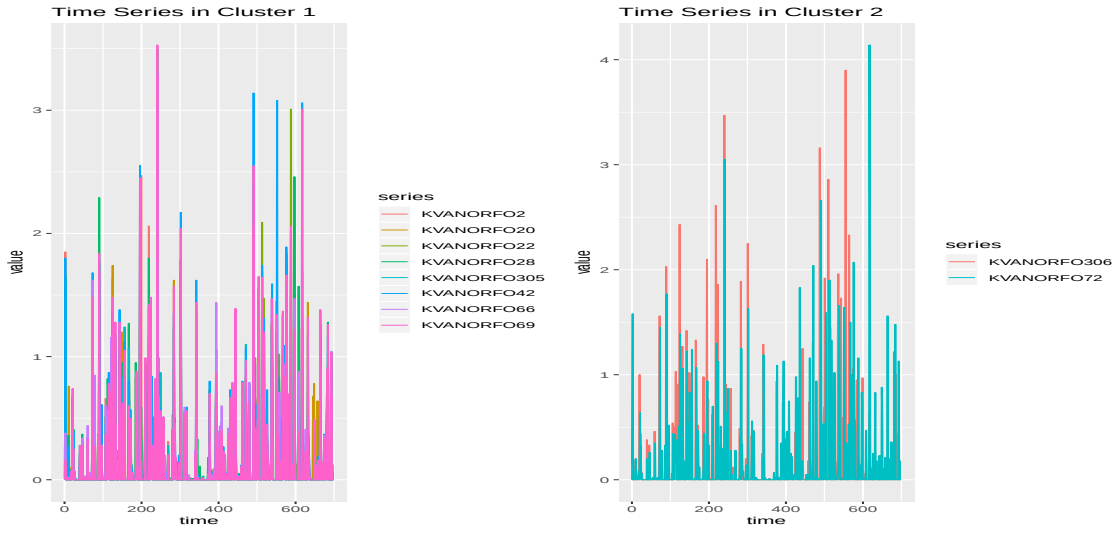

Figure 6.9: Rainfall Hierarchical Clustering for k=2

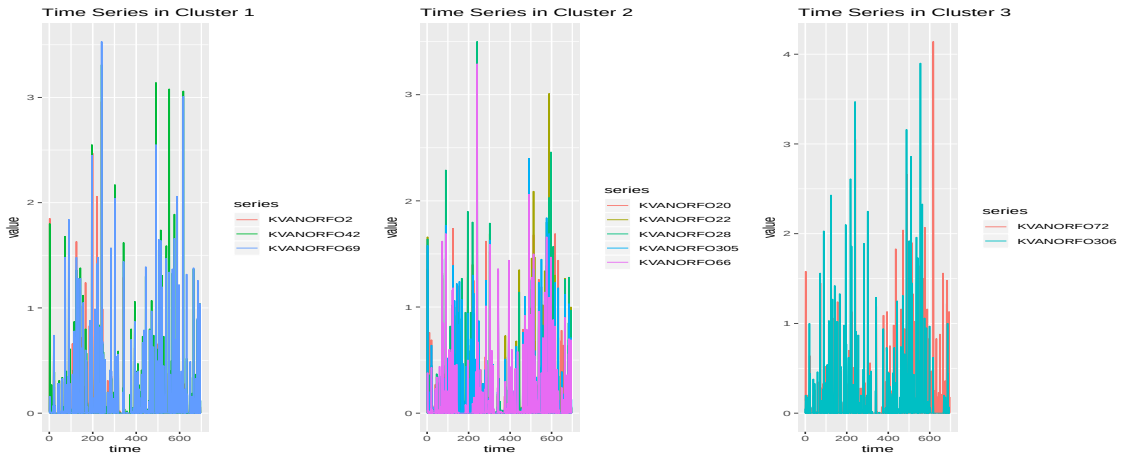

Figure 6.10: Rainfall Hierarchical Clustering for k=3

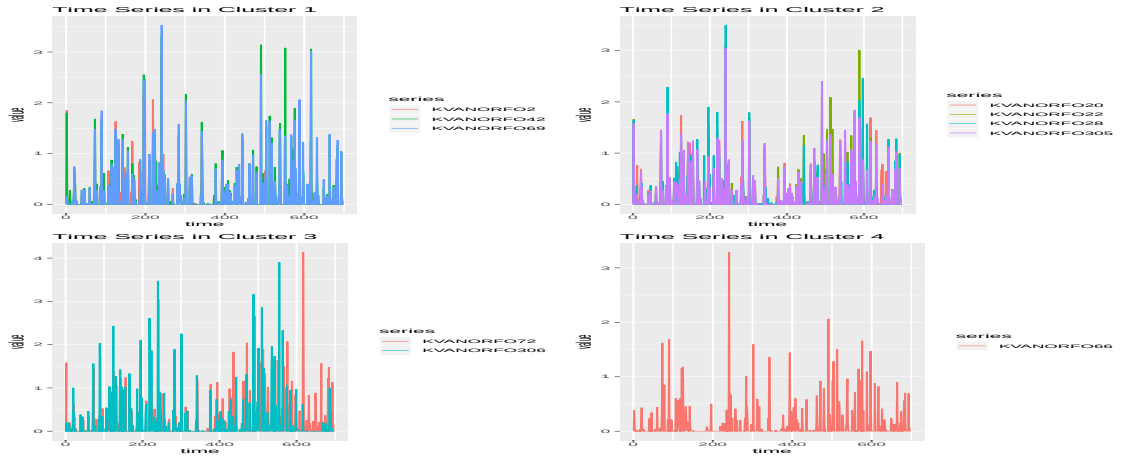

Figure 6.11: Rainfall Hierarchical Clustering for k=4

Chapter 7

Demo on Personal Weather Station

In this chapter, we will demo the entire toolbox applied to PWSs in Norfolk area. As discussed earlier in the thesis, our methods can be used for consistently producing real time Bayes Factors to achieve standardized evaluation of PWSs at large scales. Hypothesis testing framework is dependent on the assumption that we can retrieve NCDC data and use the interpolated data to account for spatial difference as ground truth. This approach is powerful since reliable data sources are introduced as a benchmark, and we can derive Bayes Factors to quantify our 'belief' on whether a PWS is abnormal or not. However, the shortcoming is also straightforward since we mainly rely on NCDC to reflect whether a PWS is reliable. Clustering-based abnormal detection, on the other hand, with the advantage of avoiding introducing outside data sources, can be used as a complementary approach for the abnormal detection purpose. Without other data sources, we only count on PWS API to retrieve all the PWSs within the similar region and make comparisons among them, with the belief that if abnormality exists, they should be clustered apart from the majority ones. We can use the visualization of cluster patterns to verify our conclusion. This method also has the drawback that geographical difference is not considered, unlike

in the first approach, we can handle this concern by interpolation. Fundamentally, both methods share the same theoretical support built from the proposed tapered statistics with kernel transformation.

The entire toolbox works in this fashion: first of all, historical time series data of PWSs of our interest and surrounding NCDC stations within the same region is collected, along with their geographical information for the data interpolation purpose. Once the data has been collected, they fall into two pipelines:

1. With NCDC data as part of input, three hypothesis testing procedures for temperature, precipitation and multiple weather measurements are implemented, and Bayes Factors are calculated correspondingly. For hypothesis testing on daily maximum temperature and daily precipitation, the null hypothesis is that the weather measurement time series of PWS and NCDC follow the same distribution, and the alternative hypothesis is there is dissimilarity between them. We apply the tapered test statistics with Gaussian kernel of optimal bandwidth, and derive the logarithm of Bayes factors. If $\log BF < 0$, it indicates that the evidence is against H_0, showing the PWS is abnormal. If $0 < \log BF < 1$, then the evidence for H_0 is 'not worth a bare mention', we do not hold strong belief on the trustworthiness of PWS. If $1 < \log BF < 3$, we are 'positive' that there exists evidence for H_0. If $3 < \log BF < 5$, 'strong' evidence supports H_0. If $\log BF > 5$, then we have 'very strong' belief on the null hypothesis that the PWS comes from the same distribution as interpolated NCDC, thus the PWS is trustworthy.

2. Without outer data sources like NCDC, clustering based abnormal detection is applied at the same time solely based on surrounding PWSs. We check the pattern of each cluster and identify clustered apart from the majority PWSs and label them as suspicious.

In general, the workflow pipeline generates multiple metrics for different ap-

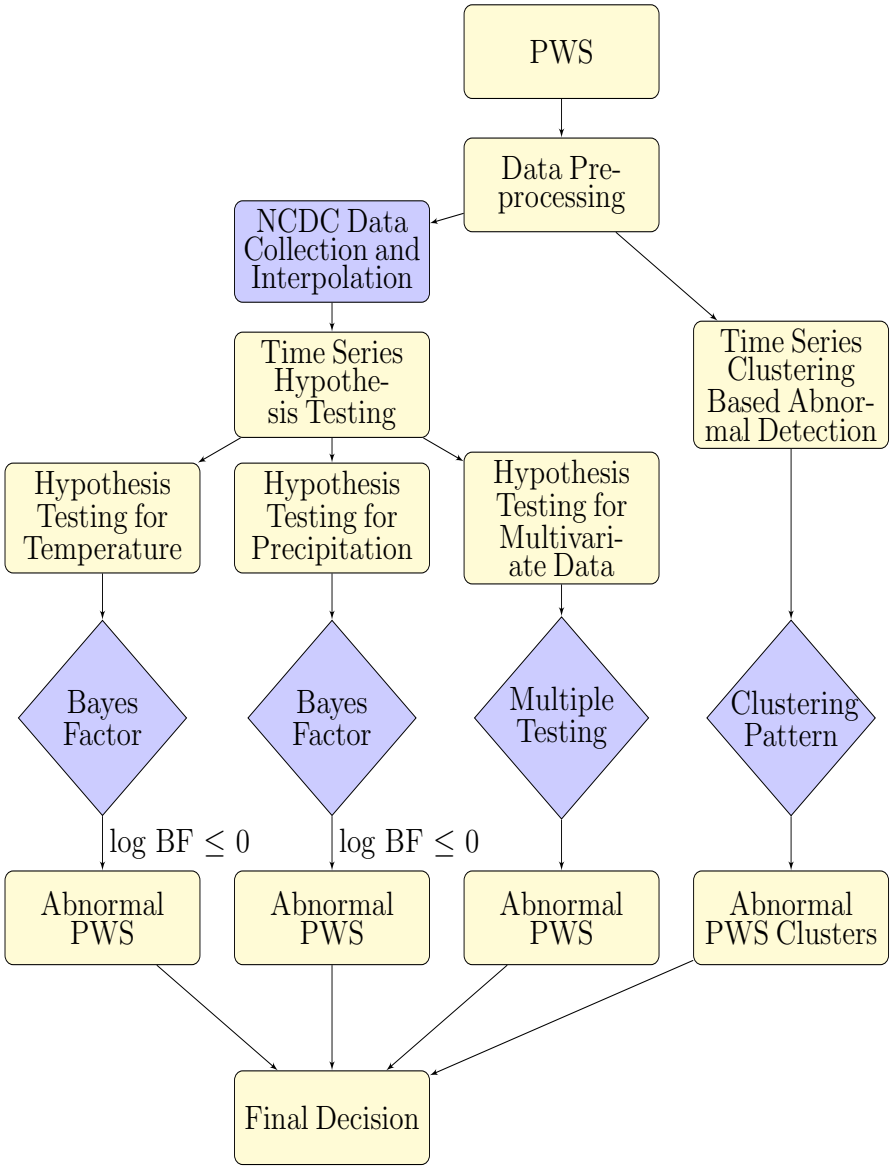

Figure 7.1: Personal Weather Station (Overview of the workflow: Temperature and precipitation data from both PWS and NCDC in Norfolk area are collected from the network jointly; the data is then preprocessed to get interpolated with the NCDC data; the system computes the Bayes factor based on temperature and rainfall data for each PWS individually; the system conduct the clustering-based time series abnormal detection for all the neighboring PWSs jointly.)

proaches. Even though we offer comprehensive solutions, it might be a little sophisticated to draw a consensus conclusion. The guideline is analysts can make the decision based on the practical problems. For example, if they are interested in flood monitoring, then they can emphasize the Bayes Factor generated by the precipitation testing problem. If they care more about the overall performance, they can use the combination of the three metrics, to decide by majority vote. They can always check the density of NCDC in the surrounding area if the number of NCDCs is relatively small, then they can conduct clustering-based methods to capture the abnormal ones. With methods proposed, we come up with well-defined tools to measure the trustworthiness of the crowdsourced data in real-time. The workflow is shown in Figure 7.1.

To make the workflow more directly perceived, I'll mainly use two PWSs, KVANORFO22 and KVANORFO72 to demo the flowchart of our methods proposed in this thesis, and also present the overall results for 11 stations in Norfolk.

7.1 Demo on Temporal Data Preprocessing

With accessed data from weather underground API published by PWS at the frequency of every 15 minutes and daily historical data released by NCDC in the Norfolk area, they will fall into abnormal detection toolbox. To begin with, we carry out data preprocessing steps on daily maximum temperature. It is noteworthy that KVANORFO72 presents several enormous and small values at specific timestamps which makes it quite suspicious and unstable. In our analysis of comparing two time series, we want to eliminate the random effect caused by the outliers in order to get a more convincing result based on the majority of published data points. Figure 7.2 is the outlier detection result for KVANORFO72 daily maximum temperature. The model uncovers during February 2018, December 2018 and January 2019, sev-

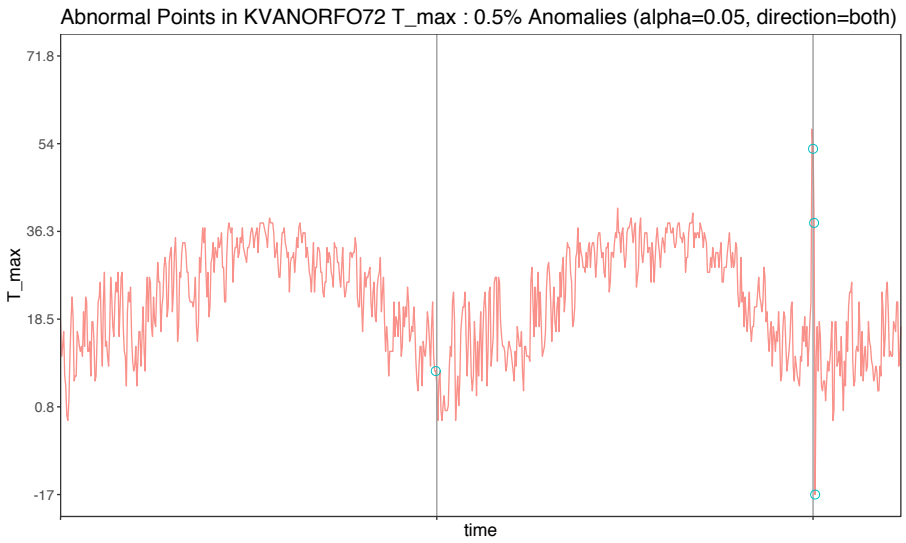

Figure 7.2: Outlier Detection on KVANORFO72

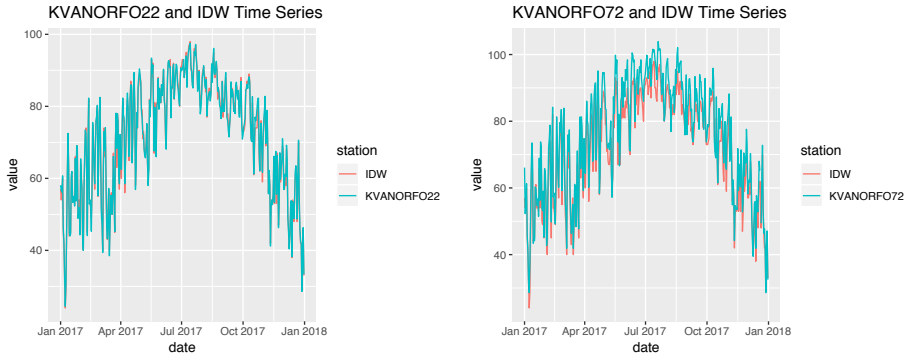

Figure 7.3: Time Series Plot of PWS vs IDW for PWS KVANORFO22 and KVANORFO72

eral suspiciously high and low spikes have been captured in abnormal detection and eliminated in the further analysis.

7.2 Demo on Time Series Hypothesis Testing

In the demo of time series hypothesis testing, besides PWS temporal data, the time series data of surrounding NCDCs is also needed. With the practical concern that PWS and neighboring NCDC does not necessarily share the same locations, sources

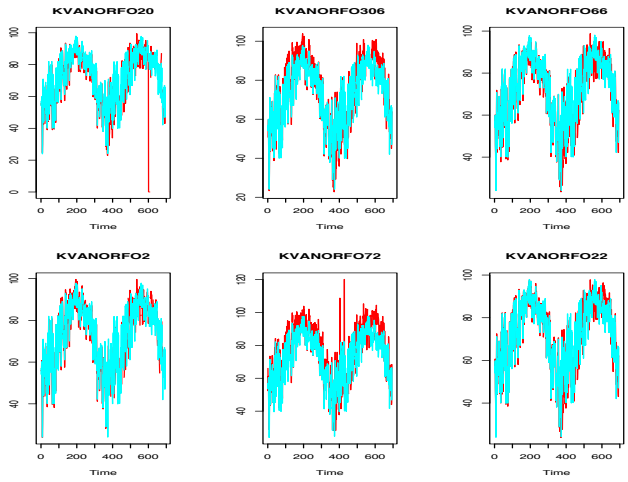

Figure 7.4: Time Series Plot for Daily Maximum Temperature of PWS vs NCDC

of NCDC data must be interpolated based on their geographical information. As shown in the left pipeline, to begin with, we interpolate the surrounding NCDCs to the same longitude and latitude of PWS of interest. The time series plots of both PWS stations compared with NCDC are shown in Figure 7.3.

With both preprocessed PWS and NCDC temporal data, we move on to carry out the time-series hypothesis testing framework.

7.2.1 Hypothesis Testing on Temperature

The next step is to conduct hypothesis testing on the daily maximum temperature, with pre-chosen optimal bandwidth under the Bayesian setup, Bayes Factors are calculated accordingly as shown in Table 7.1.

KVANORFO22 and KVANORFO72, with outliers detected and eliminated, have logarithm of Bayes factors around 43.7 and 47.9, indicating both of them have solid evidence that PWSs come from the same underlying distribution as interpolated NCDC, in other words, KVANORFO22 and KVANORFO72 are trustworthy concerning daily maximum temperature measurements.

Table 7.1: log BF for Daily Maximum Temperature of PWS in Norfolk

PWS	log BF
KVANORFO2	43.74921
KVANORFO20	-8.6301
KVANORFO72	47.9050
KVANORFO305	49.6717
KVANORFO306	-203.2970
KVANORFO22	43.7492
KVANORFO28	49.04882
KVANORFO42	2.6867
KVANORFO66	-17.7271
KVANORFO69	46.3411

Table 7.2: log BF for Daily Precipitation of PWS in Norfolk

PWS	log BF
KVANORFO2	-49.06801
KVANORFO20	26.53325
KVANORFO22	-44.09912
KVANORFO28	-22.98318
KVANORFO42	-338.7884
KVANORFO66	-12.02769
KVANORFO69	19.94753
KVANORFO72	-1326.598
KVANORFO305	21.78035
KVANORFO306	20.64477

7.2.2 Hypothesis Testing on Precipitation

Similarly, for hypothesis testing on daily precipitation, under the two-step setup, in the first step, hypothesis testing on precipitation is conducted. If the null hypothesis is supported, we will conduct time-series hypothesis testing on the daily maximum temperature data, with pre-chosen optimal bandwidth under the Bayesian setup, Bayes Factors are calculated as shown in Table 7.2.

KVANORFO22 and KVANORFO72 have logarithm of Bayes factors around -44.1

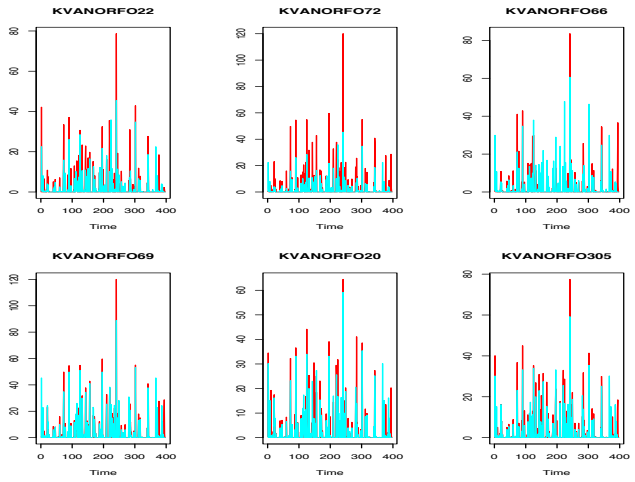

Figure 7.5: Time Series Plot for Daily Precipitation of PWS vs NCDC

and -132.6, which implies that both stations seem to follow different distribution from NCDC data concerning daily precipitation. Visualization from Figure 7.5 shows that there indeed exist significant gaps between PWS and NCDC on the top part, and on the bottom part, such as for KVANORFO20, KVANORFO69 and KVANORFO305, PWSs appear consistent with NCDC rainfall data.

7.2.3 Hypothesis Testing on Multivariate Time Series

When we take into consideration of multiple measurements of maximum temperature, minimum temperature and average temperature jointly, the Bayesian approach is adopted to correct multiplicity, both KVANORFO22 and KVANORFO72 are labelled as trustworthy with regard to multiple weather measurements on maximum temperature, minimum temperature and average temperature.

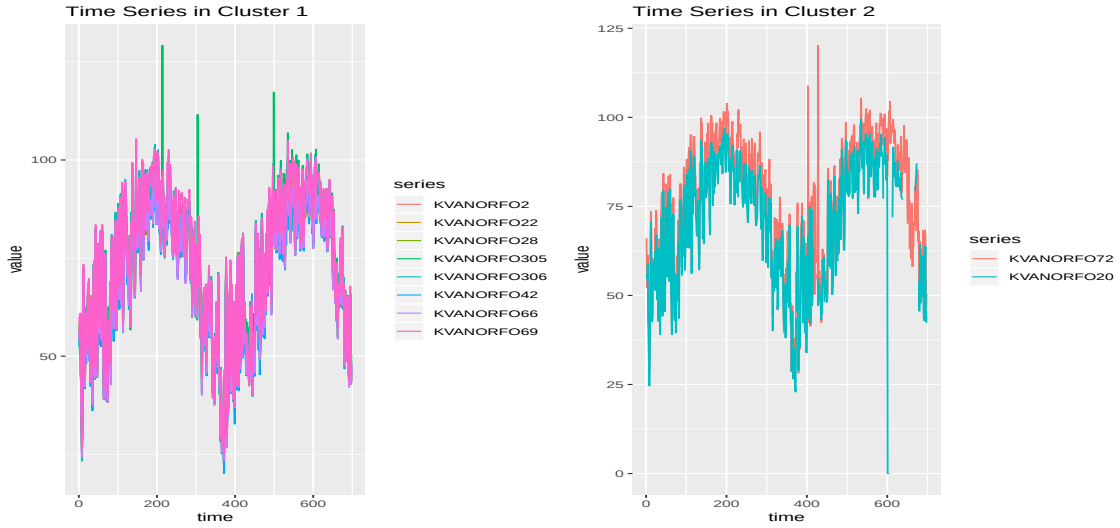

Figure 7.6: Maximum Temperature Hierarchical Clustering for k=2

7.3 Demo on Clustering-based Abnormal Detection

Besides abnormal detection under the hypothesis testing framework, we also apply time series clustering based abnormal detection on daily maximum temperature for all PWSs in the Norfolk area. Given a predefined number of clusters k, for instance, with $k = 2$, clusters generated are shown in Figure 7.6, with $k = 3$, clusters generated are shown in Figure 7.7, we find out that regardless of the setting of the number of clusters k, KVANORFO72 either falls into a separate cluster by itself, or isolated from the majority of clusters, while KVANORFO22 is always grouped with the majority of stations in the same cluster. So based on those patterns, we can conclude that KVANORFO72 exists abnormal behaviors, while KVANORFO22 is trustworthy.

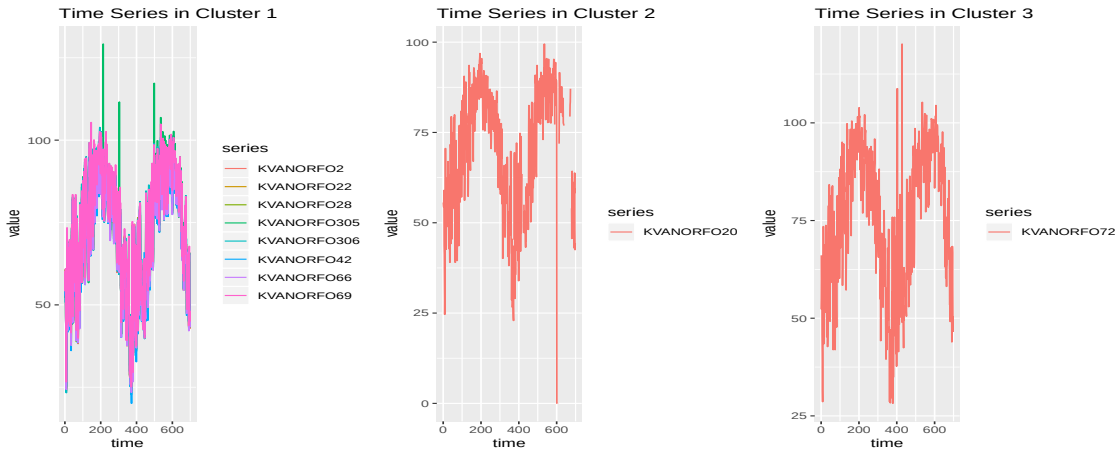

Figure 7.7: Maximum Temperature Hierarchical Clustering for k=3

7.4 Demo on Final Decisions

With methods proposed, we can run the entire toolbox and get a decision for every aspect. For PWS KVANORFO72, when we carry out hypothesis testing on daily maximum temperature, the Bayes factor is 47.9, showing strong evidence to support the null hypothesis that it comes from the same distribution as interpolated NCDC. However, when we conduct the clustering-based abnormal detection, it has been clustered apart from the majority of the PWSs in Norfolk, the reasons can be that for clustering, the geographical information of PWs have not been taken care. Thus the results are not as convincing as hypothesis testing. We would refer to the testing problem to conclude that PWS KVANORFO72 provides reliable measurement on daily maximum temperature, but we should keep in mind to eliminate the few extremely values in the analysis. For PWS KVANORFO22, with Bayes factor equal to 43.7 in the hypothesis testing and also clustered along with the majority of surrounding PWSs in clustering methods, we can safely conclude that KVANORFO22 is reliable concerning daily maximum temperature. Both KVANORFO72 and KVANORFO22 have negative Bayes factors for daily precipitation, meaning we can not trust both stations for reliable daily precipitation measurements.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

Motivated by the application of abnormal detection for personal weather stations (PWS), in our research, we propose comprehensive methods to effectively identify the 'untrustworthy' personal weather stations (PWS) for multiple weather measurements. Periodogram-based tapered test under rate-of-testing framework and clustering-based abnormal detection methods have been proposed.

In Chapter 2, we lay out the preprocessing steps for the weather measurement data, since PWSs cover the personal residences of finer granularity than NCDC, the observations of NCDC stations need to be interpolated at gridded locations using a technique such as Inverse Distance Weighting (IDW) to take care of the geographical difference. Based on the findings that the very few very distinct outliers can significantly influence the outcome of abnormal detection, even though the rest part seems consistent with the NCDC time series data, we apply outlier detection as the preprocessing step to avoid signifying an anomaly.

In Chapter 3, the first approach based on the PWS-NCDC comparison is proposed.

Fundamentally, we test PWS against interpolated NCDC stations through hypothesis testing procedure. We develop the theory for testing periodograms by smoothing and introduce Bayesian testing framework to time series setting, and we come up with an end-to-end framework for hypothesis testing problems in the time series context. This approach also handles the challenge that existing testing procedures for assessing whether two stationary time series have the same dynamics lose power for the periodogram-based test, for the reason that the periodogram is not a consistent estimator of the spectral density. It also brings in the issue of high-dimensionality, which motivates us to incorporate tapered weights, to emphasize low-frequency periodic shapes over high-frequency periodic shapes as it is critical for avoiding accumulating stochastic errors for the high-dimensional data. Theoretical theorems and proofs under each model are explained in details, and comprehensive simulation studies have shown that the Gaussian kernel smoothing model with "optimal bandwidth" is superior and corresponding tapering test procedures are recommended.

In Chapter 4, we deal with the challenge brought by precipitation data because of the unique property of inflated with zero values, which makes it difficult to be converted into a stationary time series. We propose the two-stage approach where in the first stage, we reduce the data to indicators of the presence or absence of precipitation and carry out analysis for binary time series. On condition that the binary time series can pass the test, we carry forward to the next stage of adapting the methods from Chapter 3 for testing non-zero part of rainfall data.

In Chapter 5, we extend the idea of Chapter 3 to the multivariate time series context. We use the optimal m based on the rate-of-testing theory as the guideline of the smoothing methods with simulation studies showing its powerful performance under multivariate time series settings.

In Chapter 6, we develop the clustering-based time series abnormal detection

methods without introducing external data sources. We apply the proposed tapered test statistic to hierarchical clustering as modified dissimilarity measure and conduct time-series clustering model to all the PWS in Norfolk area. Our model tends to classify the similar 'normal' PWS into the same cluster, and we can detect the untrustworthy ones by investigating into the pattern of the remaining clusters.

In chapter 7, the demo of the entire toolbox is presented. We use KVANORFO22 and KVANORFO72 to demo the flowchart of the methods proposed in this thesis and present the results for all 11 stations in Norfolk. We illustrate end-to-end analysis and the decision making process for practical problems.

With the many proposed methods, they have pros and cons and can be applied to various applications. The methodology of hypothesis testing on stationary time series such as temperature data lays out the foundation of the thesis. Temperature data, after the preprocessing steps of interpolation and outlier removal, can be converted to stationary time series. Thus we can use periodogram-based methods with tapering technique to conduct the hypothesis testing framework. This approach has a solid theoretical foundation with consistent and stationary input, making the final decision more convincing. The method developed on the hypothesis testing on precipitation data can be more relevant to the problem of our concern since it directly applies to the rainfall data and can give us an informational conclusion on the flood monitoring. However, it also has the drawback that to avoid the issue of zero inflation, the time series with zero values are either removed or smoothed in the processing process, resulting in the loss of information. Hypothesis testing on multivariate data is motivated by the problem that weather stations report multiple weather measurements such as minimum temperature, maximum temperature and average temperature, and this approach can compare multiple weather measurements jointly and test whether their underlying spectral matrices are the same. Bayesian approach to the correction

of multiplicity is proposed to handle the multiple testing problem. For testing problems, Bayes Factors (BF) are calculated to 'quantify' our belief for each scenario. For those classified as 'alternative hypothesis' time-series data, it will be labelled as 'untrustworthy' ones. Time-series clustering is a widely used approach for the abnormal detection since clustering method can reveal similarities among the PWSs. If abnormality exists, they should be clustered apart from the majority of the data, making it practical to check the credibility of the PWSs within the same district jointly. With our proposed distance measure of tapered test statistics, our approach can handle the temporal dependencies and seasonal variation of time series data. This method can avoid bringing external information, but it also has the drawback that we have to check each cluster manually and make decisions based on human judgement. To sum up, we propose multiple solutions for the abnormal detection in the PWSs, hypothesis testing problems output Bayes factors which can 'quantify' our belief of whether or not to support the null hypothesis for each PWS. However, this may bring in the issue that with multiple Bayes factors produced by different testing procedures, it is likely that one PWS might have contradictory results. It is crucial to identify the problem we are most concerned of, for example, if we care most about the flood monitoring, then we can rely on the Bayes factor generated by the testing on precipitation data. If we care about the overall performance of the PWS, we can rely on the Bayes Factor of the hypothesis testing of multivariate time series or the majority vote of Bayes factors generated by different measurements. For those PWSs classified as 'alternative hypothesis' time-series data, it will be labelled as 'untrustworthy' ones. The clustering method can be applied for the situation where we external data sources are unavailable, to check the pattern of PWSs and find out abnormal clusters.

8.2 Future Work

Hypothesis testing for time series has enormous potential for a variety of scientific studies and industrial applications. In particular, the approaches proposed in our work has favourable computation complexity and theoretical guarantee. For future work, we will calibrate on the following aspects:

1. The rate-of-testing theory provides powerful toolsets for calibrating the modelling procedures for hypothesis testing. Exploring the richer families (such as the Tweedie family) with our proposed framework may lead to more advantageous approaches, which we leave to the future work.

2. From the application perspective, with the increasing popularity of PWS, abnormal detection will play more critical parts in the crowdsourcing weather station studies. Future work may extend the current analysis by including more stations from geologically-different locations. In particular, we are interested in Chapel Hill and Houston. In Chaple Hill, there are 108 PWS, and in Houston, there are 428 PWS. One of the advantages of Huston rainfall data is that it does not have the snow season and can avoid dealing with such special cases.

Chapter 9

References

[1] Jianqing Fan. Test of significance based on wavelet thresholding and neyman's truncation. Journal of the American Statistical Association, 91(434):pp. 674–688, 1996.

[2] Kewei Lu and Linyuan Li. On fan's adaptive neyman tests for comparing two spectral densities. Journal of Statistical Computation and Simulation, 83(9):1585–1601, 2013.

[3] Robert Lund, Hany Bassily, and Brani Vidakovic. Testing equality of stationary autocovariances. Journal of Time Series Analysis, 30(3):332–348, 2009.

[4] Dan J. Spitzner. A powerful test based on tapering for use in functional data analysis. Electronic Journal of Statistics, 2:939–962, 2008.

[5] Brillinger, David R. Time series: data analysis and theory. Society for Industrial and Applied Mathematics, 2001.

[6] Coates, D. S., and P. J. Diggle. "Tests for comparing two estimated spectral densities." Journal of Time Series Analysis 7.1 (1986): 7-20.

[7] Diggle, Peter J., and Nicholas I. Fisher. "Nonparametric comparison of cumulative periodograms." Applied Statistics (1991): 423-434.

[8] Lund, Robert, Hany Bassily, and Brani Vidakovic. "Testing equality of stationary autocovariances." Journal of Time Series Analysis 30.3 (2009): 332-348.

[9] Priestley, Maurice Bertram. "Spectral analysis and time series." (1981).

[10] Shumway, Robert H., and David S. Stoffer. Time series analysis and its applications: with R examples. Springer Science Business Media, 2010.

[11] Spitzner, Dan J. "An asymptotic viewpoint on high-dimensional Bayesian testing." Bayesian Analysis 3.1 (2008): 121-160.

[12] Spitzner, Dan J. "Testing in functional data analysis using quadratic forms." arXiv preprint arXiv:0801.2224 (2008).

[13] Spokoiny, Vladimir G. "Adaptive hypothesis testing using wavelets." The Annals of Statistics 24.6 (1996): 2477-2498.

[14] Brockwell, Peter J., and Richard A. Davis. Time series: theory and methods. Springer Science Business Media, 2013.

[15] Chenyi Pan, Bayesian tapering test for comparing two spectral densities with application to EEG data, 2016

[16] Albert, James H., and Siddhartha Chib. "Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts." Journal of Business Economic Statistics 11.1 (1993): 1-15.

[17] Barry, Daniel, and John A. Hartigan. "Product partition models for change point problems." The Annals of Statistics (1992): 260-279.

[18] Barry, Daniel, and John A. Hartigan. "A Bayesian analysis for change point problems." Journal of the American Statistical Association 88.421 (1993): 309-319.

[19] Crowley, Evelyn M. "Product partition models for normal means." Journal of the American Statistical Association 92.437 (1997): 192-198.

[20] Dahl, David B. "Modal clustering in a class of product partition models." Bayesian Analysis 4.2 (2009): 243-264.

[21] Quintana, Fernando A., and Pilar L. Iglesias. "Bayesian clustering and product partition models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65.2 (2003): 557-574.

[22] Loschi, Rosangela Helena, et al. "A Gibbs sampling scheme to the product partition

model: an application to change-point problems." Computers Operations Research 30.3 (2003): 463-482.

[23] Brockwell, P. J., Davis, R. A., Fienberg, S. E. (1991). Time Series: Theory and Methods: Theory and Methods. Springer Science Business Media.

[24] Johnson, R. A., Wichern, D. W. (2002). Applied multivariate statistical analysis (Vol. 5, No. 8). Upper Saddle River, NJ: Prentice hall.

[25] Spitzner, D. J. (2011) Neutral-data comparisons for Bayesian testing, Bayesian Analysis, 6:603-638.

[26] Womack, A. J., Fuentes, C., Taylor-Rodriguez, D. (2015). Model space priors for objective sparse Bayesian regression. arXiv preprint arXiv:1511.04745.

[27] Scott, J. G., Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics, 38(5), 2587-2619.

[28] Berger, J. O., Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association, 91(433), 109-122.

[29] Chen, A. B., Behl, M., Goodall, J. L. (2018, November). Trust me, my neighbors say it's raining outside: ensuring data trustworthiness for crowdsourced weather stations. In Proceedings of the 5th Conference on Systems for Built Environments (pp. 25-28). ACM.

[30] Ravishanker, Nalini, J. R. M. Hosking, and Jaydip Mukhopadhyay. "Spectrum-based comparison of stationary multivariate time series." Methodology and Computing in Applied Probability 12.4 (2010): 749-762.

[31] Sellke, T., Bayarri, M., Berger, J. O. (2001). Calibration of values for testing precise null hypotheses. The American Statistician, 55 , 62–71.

[32] Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81 , 826–831.

[33] Stephens, M., Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. Nature Reviews Genetics, 10 , 681–690.

[34] Storey, J. D., Tibshirani, R. (2003). Statistical significance for genomewide studies.

Proceedings of the National Academy of Sciences, 100 , 9440–9445.

[35] Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics, 5 , 99–114.

[36] Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. Journal of the American Statistical Association, 92 , 299–306.

[37] Westfall, P. H., Johnson, W. O., Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. Biometrika, 84 , 419–427.

Chapter 10

Proof

Appendix 1

Proof of Theorem 3.3.1. Denote by P_0 and P_1 the respective distributions of the data under H_0 and H_1. Suppose there is a scalar series Z_n that is asymptotically constant such that

$$P_0[\frac{Q_n - E_0[Q_n]}{\sqrt{V_0[Q_n]}} > Z_n] = \alpha$$

where α is the size of the test. The power of the test is:

$$\begin{aligned} P_1[Q_n > E_0[Q_n] + z_n\sqrt{V_0[Q_n]}] &= P_1[\frac{Q_n - E_1[Q_n]}{\sqrt{V_1[Q_n]}} > \frac{E_0[Q_n] + z_n\sqrt{V_0[Q_n]} - E_1[Q_n]}{\sqrt{V_1[Q_n]}}] \\ &= P_1[\frac{Q_n - E_1[Q_n]}{\sqrt{V_1[Q_n]}} > \frac{z_n - (E_1[Q_n] - E_0[Q_n])/\sqrt{V_0[Q_n]}}{\sqrt{1 + \frac{(E_1[Q_n]-E_0[Q_n])^2}{V_0[Q_n]}\frac{V_1[Q_n]-V_0[Q_n]}{(E_1[Q_n]-E_0[Q_n])^2}}}] \end{aligned}$$

We can also rewrite this as:

$$P_1[\frac{Q_n - E_1[Q_n]}{\sqrt{V_1[Q_n]}} > \frac{z_n - A_n}{\sqrt{1 + A_n^2/B_n^2}}]$$

where

$$A_n = \frac{E_1[Q_n] - E_0[Q_n]}{\sqrt{V_0[Q_n]}}$$

and

$$B_n = \frac{E_1[Q_n] - E_0[Q_n]}{\sqrt{V_1[Q_n] - V_0[Q_n]}}$$

So if the rate of testing criteria holds, then it implies that

$$\inf_{\boldsymbol{\theta} \in H_1(\delta_n/\delta_n^\star; s, M)} A_n \to \infty$$

and

$$\inf_{\boldsymbol{\theta} \in H_1(\delta_n/\delta_n^\star; s, M)} B_n \to \infty$$

□

Proof of Theorem 3.3.2. For the mean and variance of the $\bar{Q}_n$:

$$\begin{aligned} E(\bar{Q}_n) &= E(\sum_{j=1}^{p_n} \frac{1}{j^{1/2}} \bar{Y}_j^2) \\ &= \sum_{j=1}^{p_n} \frac{1}{j^{1/2}} E(\bar{Y}_j^2) \\ &= \sum_{j=1}^{p_n} \frac{1}{j^{1/2}} [Var(\bar{Y}_j) + E\bar{Y}_j)^2] \\ &= \sum_{j=1}^{p_n} \frac{1}{j^{1/2}} [\frac{\sigma^2}{m_n} + \bar{\theta}_i^2] \end{aligned}$$

$$\begin{aligned} V(\bar{Q}_n) &= V(\sum_{j=1}^{p_n} \frac{1}{j^{1/2}} \bar{Y}_j^2) \\ &= \sum_{j=1}^{p_n} \frac{1}{j} V(\bar{Y}_j^2) = \sum_{j=1}^{p_n} \frac{1}{j} [E(\bar{Y}_j^4) - E(\bar{Y}_j^2)^2] \\ &= \sum_{j=1}^{p_n} \frac{1}{j} [(\bar{\theta}_i^4 + 6\bar{\theta}_i^2 \frac{\sigma^2}{m_n} + 3\frac{\sigma^4}{m_n^2}) - (\frac{\sigma^2}{m_n} + \bar{\theta}_i^2)^2] \\ &= \sum_{j=1}^{p_n} \frac{1}{j} [4\bar{\theta}_i^2 \frac{\sigma^2}{m_n} + 2\frac{\sigma^4}{m_n^2}] \end{aligned}$$

Similarly, we have the constraint that

$$\theta = (\theta_j, j = 1, 2, ...) \in \mathbb{B}_{s,M}$$

$$\mathbb{B}_{s,M} = (\theta_1, \theta_2, ...) : \sqrt{\sum_{j=1}^{\infty} j^{2s} \theta_j^2} \leqslant M$$

$$H_1(\delta; s, M) = \theta \in \mathbb{B}_{s,M} : \sqrt{\sum_{j=1}^{\infty} \theta_j^2} \geqslant \delta$$

According to Theorem 1, if we want the rate of testing criteria holds, then the following constraint should be satisfied:

$$\frac{\sum_{j=1}^{p} w_j \bar{\theta}_j^2}{\sqrt{\sum_{j=1}^{p_n} \frac{1}{j} \bar{\theta}_i^2 \frac{\sigma^2}{m_n}}} \to \infty \tag{10.1}$$

as well as

$$\frac{\sum_{j=1}^{p} w_j \bar{\theta}_j{}^2}{\sqrt{\sum_{j=1}^{p_n} \frac{1}{j} \frac{\sigma^4}{m_n^2}}} \to \infty \tag{10.2}$$

For the numerator part of both constraints, since

$$\begin{aligned}
\sum_{j=1}^{p} w_j \bar{\theta}_j^2 &= \sum_{j=1}^{\infty} w_j \bar{\theta}_j{}^2 - \sum_{j=p+1}^{\infty} w_j \bar{\theta}_j{}^2 \\
&= \sum_{j=1}^{\infty} w_j || \frac{m_n}{n} \sum_{i \in C_j} \theta_i ||^2 - \sum_{j=p+1}^{\infty} w_j || \frac{m_n}{n} \sum_{i \in C_j} \theta_i ||^2 \\
&= \frac{m_n^2}{n^2} [\sum_{j=1}^{\infty} w_j || \sum_{i \in C_j} \theta_i ||^2 - \sum_{j=p+1}^{\infty} w_j || \sum_{i \in C_j} \theta_i ||^2]
\end{aligned}$$

The lower bound of $\sum_{j=1}^{\infty} \frac{1}{j^{1/2}} || \sum_{i \in C_j} \theta_i ||^2$ subject to constraint $\theta \in \mathbb{B}_{s,M}$ is achieved by

$$\theta_j = \begin{cases} \delta_n / \delta_n^* & j = \hat{j} \\ 0 & \text{o.w.} \end{cases}$$

where $\hat{j}$ is the largest index between 1 and ∞ that satisfies $\hat{j} \leqslant (\frac{\delta_n^* M}{\delta_n})^{1/s}$.

So

$$\sum_{j=1}^{\infty}\frac{1}{j^{1/2}}||\sum_{i\in C_j}\theta_i||^2 \geqslant \hat{j}^{-1/2}(\delta_n/\delta_n^*)^2 = (\delta_n/\delta_n^*)^{\frac{4s+1}{2s}}M^{-\frac{1}{2s}}$$

and

$$\sum_{j=p+1}^{\infty}\frac{1}{j^{1/2}}||\sum_{i\in C_j}\theta_i||^2 \leqslant p^{-(2s+1/2)}\sum_{j=p+1}^{\infty}j^{2s}||\sum_{i\in C_j}\theta_i||^2 \leqslant p^{-(2s+1/2)}\frac{m_n}{n}M^2$$

So

$$inf_{\theta\in H_1}\sum_{j=1}^{p}\frac{1}{j^{1/2}}\bar{\theta}_j^2 \geqslant \frac{m_n^2}{n^2}(\delta_n/\delta_n^*)^{\frac{4s+1}{2s}}M^{-\frac{1}{2s}} - \frac{m_n}{n}p^{-\frac{4s+1}{2}}M^2$$

Which leads constraint (10.1) and (10.2) to the following conditions:

$$limsup_{n\to\infty}\frac{\frac{m_n^2}{n^2}(\frac{\delta_n}{\delta_n^*})^{\frac{4s+1}{2s}}M^{-\frac{1}{2s}}}{\frac{1}{m_n}\sigma^2\sqrt{\sum_{j=1}^{p}\frac{1}{j}}} = \infty \tag{10.3}$$

$$limsup_{n\to\infty}\frac{\frac{m_n}{n}p^{-\frac{4s+1}{2}}M^2}{\frac{1}{m_n}\sigma^2\sqrt{\sum_{j=1}^{p}\frac{1}{j}}} \leqslant \infty \tag{10.4}$$

and

$$\frac{\sum_{j=1}^{p}\frac{1}{j^{1/2}}\bar{\theta}_j^2}{\frac{1}{\sqrt{m_n}}\sigma\sqrt{\sum_{j=1}^{p_n}\frac{1}{j}\bar{\theta}_i^2}} \to \infty \tag{10.5}$$

For the condition of (10.1), since the left part

$$\sqrt{m_n}\frac{\sum_{j=1}^{p}\frac{1}{j^{1/2}}\bar{\theta}_j^2}{\sigma\sqrt{\sum_{j=1}^{p_n}\frac{1}{j}\bar{\theta}_i^2}} > \sqrt{m_n}\frac{\sum_{j=1}^{p}\frac{1}{j^{1/2}}\bar{\theta}_j^2}{\sigma\sqrt{\sum_{j=1}^{p_n}\frac{1}{j^{1/2}}\bar{\theta}_i^2}} = \sqrt{m_n}\sqrt{\sum_{j=1}^{p}\frac{1}{j^{1/2}}\bar{\theta}_j^2} \to \infty$$

which can be implied from condition (10.2).

Since we have the property that

$$\sum_{j=1}^{p_n} \frac{1}{j} \asymp \log p_n \tag{10.6}$$

By taking square of equation (10.17), it is equivalent to:

$$limsup_{n\to\infty} \frac{\frac{m_n^4}{n^2} p^{-4s+1} M^4}{\sigma^4 logp} \leqslant \infty$$

which leads to the conclusion that

$$m_n \asymp (n^2 p^{4s+1} logp)^{1/4} \tag{10.7}$$

Note that since we have the relationship that p=[n/2], so the equation (10.7) is equivalent to that

$$m_n \asymp (p^{s+3/4} logp^{1/4}) \tag{10.8}$$

Similarly, equation (10.3) leads to:

$$limsup_{n\to\infty} (\frac{\delta_n}{\delta_n^*})^{\frac{4s+1}{2s}} \frac{m_n^3}{n^2} \frac{1}{logp} = \infty \tag{10.9}$$

Equation (10.9) implies that

$$\delta_n \asymp (\frac{n^2 logp}{m_n^3})^{\frac{2s}{4s+1}}$$

By the definition of m_n, we have

$$\delta_n \asymp (p^{1/4-3s} \log p^{1/4})^{\frac{2s}{4s+1}} \tag{10.10}$$

□

Proof of Theorem 3.3.3. For the mean and variance of the $\tilde{Q_n}$:

$$\begin{aligned} E(\tilde{Q_n}) &= E(\sum_{j=1}^{p_n} \frac{1}{j^{1/2}} \tilde{Y_j}^2) \\ &= \sum_{j=1}^{p_n} \frac{1}{j^{1/2}} E(\tilde{Y_j}^2) \\ &= \sum_{j=1}^{p_n} \frac{1}{j^{1/2}} [Var(\tilde{Y_j}) + E(\tilde{Y_j})^2] \\ &= \sum_{j=1}^{p_n} \frac{1}{j^{1/2}} [V_{n,j} + \theta_i^2] \end{aligned}$$

$$\begin{aligned} V(\tilde{Q_n}) &= V(\sum_{j=1}^{p_n} \frac{1}{j^{1/2}} \tilde{Y_j}^2) \\ &= \sum_{j=1}^{p_n} \frac{1}{j} V(\tilde{Y_j}^2) = \sum_{j=1}^{p_n} \frac{1}{j} [E(\tilde{Y_j}^4) - E(\tilde{Y_j}^2)^2] \\ &= \sum_{j=1}^{p_n} \frac{1}{j} [(\theta_i^4 + 6\theta_i^2 V_{n,j} + 3V_{n,j}^2) - (V_{n,j} + \theta_i^2)^2] \\ &= \sum_{j=1}^{p_n} \frac{1}{j} [4\theta_i^2 V_{n,j} + 2V_{n,j}^2] \end{aligned}$$

The parameter is restricted to the following constraints that

$$\theta = (\theta_j, j = 1, 2, ...) \in \mathbb{B}_{s,M}$$

$$\mathbb{B}_{s,M} = (\theta_1, \theta_2, ...) : \sqrt{\sum_{j=1}^{\infty} j^{2s}\theta_j^2} \leqslant M$$

$$H_1(\delta; s, M) = \theta \in \mathbb{B}_{s,M} : \sqrt{\sum_{j=1}^{\infty} \theta_j^2} \geqslant \delta$$

According to Theorem1, if we want the rate of testing criteria holds, then the following constraint should be satisfied:

$$\frac{\sum_{j=1}^{p} w_j\theta_j^2}{\sqrt{\sum_{j=1}^{p_n} \frac{1}{j}\theta_i^2 V_{n,j}}} \to \infty \tag{10.11}$$

as well as

$$\frac{\sum_{j=1}^{p} w_j\theta_j^2}{\sqrt{\sum_{j=1}^{p_n} \frac{1}{j} V_{n,j}^2}} \to \infty \tag{10.12}$$

For the numerator part of both constraints, since

$$\sum_{j=1}^{p} w_j\theta_j^2 \quad = \sum_{j=1}^{\infty} w_j\theta_j^2 - \sum_{j=p+1}^{\infty} w_j\theta_j^2$$

The lower bound of $\sum_{j=1}^{\infty} \frac{1}{j^{1/2}}\theta_i^2$ subject to constraint $\theta \in \mathbb{B}_{s,M}$ is achieved by

$$\theta_j = \begin{cases} \delta_n/\delta_n^* & j = \hat{j} \\ 0 & \text{o.w.} \end{cases}$$

where $\hat{j}$ is the largest index between 1 and ∞ that satisfies $\hat{j} \leqslant (\frac{\delta_n^* M}{\delta_n})^{1/s}$.

So

$$\begin{aligned}\sum_{j=1}^{\infty}\frac{1}{j^{1/2}}\theta_j^2 &\geqslant \hat{j}^{-1/2}(\delta_n/\delta_n^*)^2 \\ &= (\delta_n/\delta_n^*)^{\frac{4s+1}{2s}}M^{-\frac{1}{2s}}\end{aligned} \tag{10.13}$$

and

$$\begin{aligned}\sum_{j=p+1}^{\infty}\frac{1}{j^{1/2}}\theta_j^2 &= \sum_{j=p+1}^{\infty}\frac{1}{j^{1/2}}j^{-2s}j^{2s}\theta_j^2 \\ &\leqslant p^{-(2s+1/2)}\sum_{j=p+1}^{\infty}j^{2s}\theta_j^2 \\ &\leqslant p^{-(2s+1/2)}M^2\end{aligned} \tag{10.14}$$

So

$$inf_{\theta\in H_1}\sum_{j=1}^{p}\frac{1}{j^{1/2}}\tilde{\theta_j}^2 \geqslant (\delta_n/\delta_n^*)^{\frac{4s+1}{2s}}M^{-\frac{1}{2s}} - p^{-\frac{4s+1}{2}}M^2 \tag{10.15}$$

Which leads constraint (10.6) to the following conditions:

$$limsup_{n\to\infty}\frac{(\frac{\delta_n}{\delta_n^*})^{\frac{4s+1}{2s}}M^{-\frac{1}{2s}}}{V_{n,j}\sqrt{\sum_{j=1}^{p}\frac{1}{j}}} = \infty \tag{10.16}$$

$$limsup_{n\to\infty}\frac{p^{-\frac{4s+1}{2}}M^2}{V_{n,j}\sqrt{\sum_{j=1}^{p}\frac{1}{j}}} \leqslant \infty \tag{10.17}$$

For the condition of (10.5), it can be implied from condition (10.6).

Since we have the property that

$$\sum_{j=1}^{p_n}\frac{1}{j} \asymp \log p_n \tag{10.18}$$

By taking square of equation (10.17), it is equivalent to:

$$limsup_{n\to\infty}\frac{p^{-(4s+1)}M^4}{V_{n,j}^2\log p}\leqslant\infty \tag{10.19}$$

which leads to:

$$V_{n,j}^2\asymp p^{-(4s+1)} \tag{10.20}$$

Since we consider the case for the uniform kernel smoothing, the kernel function $K(u)$ is defined as:

$$K(u)=\begin{cases}1/2 & |u|\leqslant 1\\ 0 & \text{o.w.}\end{cases}$$

When we investigate the definition of $C_{j,k}$, it is represented in the form of $K(\frac{w_k-w_j}{b_n})$, which can be expressed as:

$$K(\frac{w_k-w_j}{b_n})=\frac{1}{2}1_{|w_k-w_j|\leqslant b_n}=\begin{cases}1 & |w_k-w_j|\leqslant 1\\ 0 & \text{o.w.}\end{cases}$$

Thus, $V_{n,j}$ can be expressed as:

$$
\begin{aligned}
V_{n,j} &= \sigma^2 \sum_{k=1}^{[n_1/2]-1} C_{j,k}^2 \\
&= \sigma^2 \sum_{k=1}^{p} [K(\frac{w_k - w_j}{b_n}) / \sum_{k=1}^{p} K(\frac{w_k - w_j}{b_n})]^2 \\
&= \sigma^2 \sum_{k=1}^{p} [\frac{1}{2} 1_{|w_k - w_j| \leqslant b_n}]^2 / [\frac{1}{2} \sum_{k=1}^{p} 1_{|w_k - w_j| \leqslant b_n}]^2 \\
&= \sigma^2 \sum_{k=1}^{p} [1_{|w_k - w_j| \leqslant b_n}]^2 / [\sum_{k=1}^{p} 1_{|w_k - w_j| \leqslant b_n}]^2 \\
&= \sigma^2 \sum_{k=1}^{p} 1_{|w_k - w_j| \leqslant b_n} / [\sum_{k=1}^{p} 1_{|w_k - w_j| \leqslant b_n}]^2 \\
&= \sigma^2 \frac{nb_n/2\pi}{(nb_n/2\pi)^2} \\
&= \sigma^2 2\pi / nb_n (10.21)
\end{aligned}
$$

Thus, according to equation (10.25), we have $2\pi/nb_n \asymp p^{-(4s+1)/2}$, which leads to:

$$b_n \asymp p^{\frac{4s+1}{2}} \tag{10.22}$$

Similarly, equation (10.16) leads to:

$$\limsup_{n \to \infty} (\frac{\delta_n}{\delta_n^*})^{\frac{4s+1}{2s}} \frac{1}{V_{n,j}^2 logp} = \infty \tag{10.23}$$

Equation (10.33) implies that

$$\delta_n \asymp (p^{(4s+1)} logp)^{\frac{2s}{4s+1}} \tag{10.24}$$

□

Proof of Theorem 3.3.4. Based on Theorem 3, we have the conclusion that in order to achieve rates of testing, we would have:

$$V_{n,j}^2 \asymp p^{-(4s+1)} \tag{10.25}$$

Since we consider the case for the Gaussian kernel smoothing, the kernel function $K(u)$ is defined as:

$$K(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{10.26}$$

So we have the expression for kernel as:

$$K(\frac{w_k - w_j}{b_n}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{w_k - w_j}{b_n})^2}$$

Thus, $V_{n,j}$ can be expressed as:

$$\begin{aligned}
V_{n,j} &= \sigma^2 \sum_{k=1}^{[n_1/2]-1} C_{j,k}^2 && (10.27)\\
&= \sigma^2 \sum_{k=1}^{p} [K(\frac{w_k - w_j}{b_n}) / \sum_{k=1}^{p} K(\frac{w_k - w_j}{b_n})]^2 && (10.28)\\
&= \sigma^2 \sum_{k=1}^{p} \frac{1}{2\pi} e^{-(\frac{w_k - w_j}{b_n})^2} / [\sum_{k=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{w_k - w_j}{b_n})^2}]^2 && (10.29)\\
&= \sigma^2 \sum_{k=1}^{p} e^{-(\frac{w_k - w_j}{b_n})^2} / [\sum_{k=1}^{p} e^{-\frac{1}{2}(\frac{w_k - w_j}{b_n})^2}]^2 && (10.30)\\
& && (10.31)
\end{aligned}$$

For the term $e^{-(\frac{w_k - w_j}{b_n})^2}$, it falls within the range that $e^{-\frac{n^2}{b^2}} \leqslant e^{-(\frac{w_k - w_j}{b_n})^2} \leqslant e^{-\frac{1}{b^2}}$

So equation (10.27) would falls in the range that

$$V_{n,j} \geqslant \sigma^2 \frac{1}{n} e^{\frac{n^2 - 1}{b_n}}$$

and

$$V_{n,j} \leqslant \sigma^2 \frac{1}{n} e^{\frac{n^2-1}{b_n}}$$

Thus, we have $\frac{1}{n} e^{\frac{n^2-1}{b_n}} \asymp p^{-(4s+1)/2}$, which leads to:

$$b_n \asymp \frac{n^2-1}{\log p^{4s}} \tag{10.32}$$

Similarly, equation (10.16) leads to:

$$\limsup_{n\to\infty} (\frac{\delta_n}{\delta_n^*})^{\frac{4s+1}{2s}} \frac{1}{V_{n,j}^2 logp} = \infty \tag{10.33}$$

Equation (10.33) implies that

$$\delta_n \asymp p^{4s} logp^{4s} \tag{10.34}$$

□