Exploring The Applicability of Functional Data Analysis (FDA) for dimension reduction in bigdata applications

Mapitsi Roseline Rangata

WITS
UNIVERSITY

Supervised by:
Prof Sonali Das
Prof Montaz Ali
A dissertation submitted for the degree of
Master of Science to the Faculty of Science, University of the Witwatersrand, Johannesburg.

Declaration

I _____ declare that this thesis is my own, unaided work. It is being submitted for the Degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Signature _____

_____ day of _____ 20 _____ at _____.

Acknowledgments

I would like to extend my sincere gratitude and appreciation to my supervisors Prof Sonali Das of the University of Pretoria, Department of Business Management Sciences and University of the Witwatersrand and Prof Montaz Ali of University of the Witwatersrand, School of Computer Science and Applied Mathematics for the constant support, advice and guidance through the course of this dissertation. I would also like to thank the Council of Scientific and Industrial Research (CSIR), Next Generation Enterprises and Institutions cluster for granting me the opportunity in Master's studentship and providing us with data. Lastly I would like to extend my gratitude to my wonderful God, my mother, best friend 'Remember', my fiancé and my siblings for their love and support.

Abstract

In this thesis we use Functional Data Analysis (FDA) methods to explore space and time variations in climate patterns from historical data of 16 locations spread across South Africa. Specifically, we focus on maximum monthly temperature data for the period 1965 to 2010 with 5 years of intervals. We explore this data with methods of aligning curves, the phase-plane analysis, functional Principal Component Analysis (fPCA), and functional clustering. The purpose of this exercise is to investigate within the FDA framework how the temperature patterns may have changed over time and space for the data analysed. To the best of our knowledge this is the first attempt to understand maximum temperature over 40+ years from South Africa using FDA methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The term 'bigdata' is increasingly being heard in contemporary research primarily because of the ease in which modern sensors are able to capture huge volumes of data at a very rapid rate. Analysing such huge volumes of data requires the ability to store such data and then the application of novel methods to infer in an effective and efficient way. In this thesis our efforts are focused on using the Functional Data Analysis (FDA) approach, which is a novel statistical technique, to reduce high-dimensional data problems to manageable forms, and then use various FDA techniques for inference purpose. FDA analyses data as a continuous process observed at finite time points. One of the most appealing properties of the functional approach is that interesting features such as timing (start, end, length of regime) correspond to derivatives of the smooth functional model. As such, an initial step to using FDA methods is to convert discrete data into a continuum that can be represented as smooth curves or functions [Wang et al., 2015]. When the discrete data is converted into a functional form, it in essence becomes infinite dimensional, in the sense that once the appropriate function is fitted, then for any point on the range (that may not be any of the original points of observation) we can get the corresponding value in the range using the fitted function.

A common question that one may ask is that why use FDA over other methods. Multivariate approaches tend to ignore information about the smooth functional behaviour of the generating process underlying the data [Ullah and Finch, 2013]. But FDA has the advantage of producing models that can be described by continuous smooth dynamics which will help carry out accurate parameter estimation [Ullah and Finch, 2013]. It also reduces data noise through the process of curve smoothing and the other advantage of FDA approach over multivariate is that FDA can extract information in the function and its derivatives as some of the variation in a curve can be explained at the level of certain variation [Allen, 2011]. In FDA there is no concern about the correlations between repeated measures as it treats curves as a single entity. It also makes no parametric assumption about time effects [Wang et al., 2015].

Ullah and Finch [2013] has indicated that FDA is increasily being applied in various field of study such medicine, biomechanics, education, ecology and others that are mentioned in [Ullah and Finch, 2013]. This shows a recognition of FDA tools or features. For example

FDA applications are generally in time-series data, such as economic or environmental data. For economic data, one can regard a year as the window over which the behaviour of the annual curves of the economic indicator can be analysed. Similarly in our case of application, for environmental data, say daily-hourly temperature, one can analyse the behaviour of the daily temperature curves.

In this thesis we explore the methods of FDA to investigate space and time variation of monthly maximum temperature data of 16 locations spread across South Africa in line with the context of climate change particularly in the Southeren Hemisphere. The purpose of this study is to see if we can detect change in the maximum temperature across 16 locations of South Africa over 40+ years. Our focus will be to explore appropriate techniques for converting discrete data on a continuum into a functional form, especially with regards issues related to polynomial smoothing and splines, and alignment. Thereafter, we will study the variability within a set of smoothed curves with regards to both amplitude and phase, and also the first and second derivatives [Jacques and Preda, 2014]. Finally, we will explore the FDA tools of functional Principal Component Analysis (fPCA) which is a counterpart of the traditional statistical PCA (Principal Component Analysis) technique used as a dimension reduction tool. We will also demonstrate the usability of clustering temperature curves using the functional clustering.

In Chapter 2 of the thesis we introduce the FDA, we first demonstrate how raw discrete data is converted into functional data, the derivatives of smooth functions. In Chapter 3 we introduce the various techniques for aligning the curves. In Chapter 4 we introduce the application of traditional PCA and fPCA. In Chapter 5 we introduce clustering and functional clustering algorithm such as *K-means* and *Hierarchical* algorithm to classify data. The results of application of FDA are presented in Chapter 6 (exploratory data analysis), where we first explain the data to be used for analysis, fit the data, align the curves, investigate the derivatives in terms of phase-plane plots and perform fPCA and functional clustering. In Chapter 7 we explain the software used for the analysis. In Chapter 8 we present future extension from this thesis and conclude in Chapter 9.

Chapter 2

Functional Data Analysis (FDA)

Functional Data Analysis (FDA) is the analysis of data in a form of curves [Wang et al., 2015]. FDA methods thus essentially start with transforming discrete data into functions by smoothing them using basis functions over a specific continuum. Once the appropriate smooth function is identified, one can then proceed to registering curves (alignment) as described in this Chapter, Section 3. For example, as temperature patterns change every year, winter may come early in certain years and late in other years, we use the warping functions described in Section 3.1 for the registration process. Registration of curves will then allow temperature measurements over the year to speed up and slow down within each year so that the seasons will change at the same time across all years, for instance to get better average estimates [Ramsay and Silverman, 2007].

In FDA data is observed discretely overtime on the time grid of either sparse or dense. With densely observed curves a standard approach is to form a grid time point and sampling the curve at each point which will result in high finite dimensional representation of each curve [James, 2010] and there are standard methodologies in FDA that can only be applied over densely observed data points or cannot be applied over sparsely observed data points. For example, James [2010] they have used "growth data" which consists of measurements of spinal bone mineral density for 280 males and females taken at various ages the use of basis functions tend to fail to fit the curves for extremely sparse data. There is also a number of authors that have applied FDA methods on densely observed curves, for instance Mangisa et al. [2019] a recent publication "growth curve" have used functional regression models for South African economic indicators which outperforms both the concurrent model and the historical model by far. This shows that there is a lot of scope to revisit traditional time series solutions using functional growth curve. In light with FDA approach the objectives of the methodology is the same as any standard statistical analysis, that is to investigate patterns of variability in the data, build models, estimate summary statistics [Ramsay and Silverman, 2007] and aid in the process of inference.

2.1 Applying FDA

FDA transforms discrete data into functional data of the observed values which can include noise. The difference between discrete and functional data is that discrete data is defined as data measured at a certain times and can have finite values. Functional data on the other hand are data derived from a set of observations from a continuous underlying process $X(t_j)$ at time t_j. We denote by $y(t_j)$ observation with noise $\epsilon(t_j)$ in the process $X(t_j)$, where $j = 1, ...n$ [Levitin et al., 2007]. Therefore a single functional observation $y(t_j)$ is derived from of n pairs $(t_j, y(t_j))$ as follows:

$$y(t_j) = x(t_j) + \epsilon(t_j), \tag{2.1}$$

where $y(t_j)$ is a smooth function and the smoothness depends on the choice of basis function K, $x(t_j)$ is the observation and $\epsilon(t_j)$ is a measurement error at observed time t_j. Next we fit a curve, or smooth the data, by transforming discrete data into functional data using a smoothing process as follows:

$$x(t_j) = \sum_{i=1}^{K} \phi_i(t_j) c_i, \tag{2.2}$$

where ϕ_i are basis functions and c_i are the coefficients associated with K basis function [Olorunmaiye and Awoola, 2016].

2.2 Smoothing

Smoothing is the initial step to using FDA methods by converting discrete data $y(t_j)$ into a smooth functional observation $x(t_j)$ [Ramsay et al., 2009]. There are different types of smoothing techniques knowns as Fourier Series and B-spline which are represented through Basis Functions. The choice of a smoothing techinque is mainly dependent on the underlying behaviour of data to be analysed. For example Fourier Series is mostly used to fit periodic data and B-Splines are mostly used to fit non-periodic data [Ramsay and Silverman, 2007]. Smoothing techniques and smoothing penalties such as smoothing with least squares estimates and smoothing with roughness penalty are discussed in detail below.

2.2.1 Smoothing with Basis Functions

A basis function is a set of known functions ϕ_k that are independent of each other which can approximate a function by a weighted sum of large number, K, of these functions. It can be represented by:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t), \tag{2.3}$$

where c is the vector of length K of the coefficient c_k and ϕ is the functional vector of the basis function ϕ_k. The amount of smoothness of the function is determined by number of basis function K [Ramsay and Li, 1998].

2.2.2 Smoothing with Fourier Series

Fourier series is mostly used to fit periodic data [King, 2014] which can be represented by one constant function and pairs of sine and cosine functions as follows:

$$\phi_1(t) = 1,$$
$$\phi_2(t) = sin(tw),$$
$$\phi_3(t) = cos(tw),$$
$$\phi_4(t) = sin(2tw),$$
$$\phi_5(t) = cos(2tw),$$
$$\dots$$
$$\phi_k(t) = sin\Big(\frac{k}{2}tw\Big),$$
$$\phi_{k+1}(t) = cos\Big(\frac{k}{2}tw\Big),$$

we assume that k is an even number.
where $\phi(t)$ is a basis function, $w = 2\pi/T$ and T the period of the function [Ramsay et al., 2009]. They are defined by the number of K and the period T. As such, for a Fourier series, we need to have the value to T and value for k, which has to be an even number.

2.2.3 Smoothing with B-splines

B-splines are polynomials joined together at interval endpoints, known as knots, and they are also defined by order and degree of the polynomial, where the order of a polynomial is one higher than its degree (e.g. a cubic polynomial of degree 3 is of order 4 spline) [Levitin et al., 2007]. Thus number of basis functions is equal to the number of knots plus order of spline [Ramsay et al., 2009].

2.2.4 Smoothing with Least Squares

Smoothing with least squares converts discrete data into functional form, using the basis expansion:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = c^T \phi, \tag{2.4}$$

vector c of length K contains the coefficients c_k and we define n by K matrix Φ (n is the number of observation and K is basis functions) as containing the values $\phi_k(t_j)$ and we smooth data using either Ordinary Least Squares or Weighted Least Squares [Ramsay and Li, 1998].

2.2.4.1 Ordinary Least Squares

Ordinary Least Squares minimizes the mean squared error of the residual ϵ_j, it is mostly used in situations where we assume that the residuals ϵ_j of the curve has mean zero and constant variance. We first obtain the coefficients of the expansion c_k by minimizing the least squares criterion:

$$SMSSE(y|c) = \sum_{j=1}^{n}[y_(t_j) - \sum_{k}^{K} c_k\phi_k(t_j)]^2, \tag{2.5}$$

in matrix term is expressed as:

$$SMSSE(y|c) = (y - \Phi c)^T(y - \Phi c) = ||y - \Phi c||^2, \tag{2.6}$$

Taking the derivative of criterion $SMSSE(Y|C)$ with respect to c we get:

$$2\Phi\Phi^T c - 2\Phi^T y = 0, \tag{2.7}$$

Solving c provides the estimate $\hat{c}$ that minimizes the least square solution:

$$\hat{c} = (\Phi^T\Phi)^{-1}\Phi^T y, \tag{2.8}$$

and the vector $\hat{y}$ of fitted values is

$$\hat{y} = \Phi\hat{c} = \Phi(\Phi^T\Phi)^{-1}\Phi^T y, \tag{2.9}$$

[Ramsay and Li, 1998].

2.2.4.2 Weighted Least Squares

Weighted Least Squares minimizes the weighted residual ϵ_j. It is mostly used to standardise data. We can extend the least square criterion to bring in a differential weighting of residuals as follows :

$$SMSSE(y|c) = (y - \Phi c)^T W(y - \Phi c), \tag{2.10}$$

where W is a symmetric positive matrix, and if the variance-covariance matrix $\sum_e$ for the residual ϵ_j is known then

$$W = \sum_e^{-1}.$$

Therefore the weighted least square estimate $\hat{c}$ of the coefficient vector c is

$$\hat{c} = (\Phi^T W\Phi)^{-1}\Phi^T W y, \tag{2.11}$$

[Ramsay and Li, 1998].

2.2.5 Smoothing with Roughness Penalty

The roughness of a function is measured by the square of the second derivative $[D^2x(t)]^2$ of a function at t defined by:

$$PEN_2(x) = \int [D^2x(s)]^2 ds. \tag{2.12}$$

Smoothing with roughness penalty allow control over smoothness [Ramsay and Silverman, 2005]. The penalty $PEN_2(x)$ provide smoothing because if a function is highly variable, the square of second derivative $[D^2x(s)]^2$ is large. The second derivative of roughness of a function is defined by:

$$PEN_4(x) = \int [D^4x(s)]^2 ds, \tag{2.13}$$

where x is a height function. From the example of growth acceleration in Ramsay et al. [2009] page 62, they have indicated that the estimated acceleration curve can be rougher than the true curve and since they are the second derivatives. The generalized roughness penalty is denoted by:

$$PEN_m(x) = \int [D^mx(s)]^2 ds, \tag{2.14}$$

where m is the arbitrary order of the derivative [Ramsay, 2006] page 84. Thus smoothing with the roughness penalty ensure that the function is smooth and it also gives better results. For more on roughness penalty the reader may refer Ramsay et al. [2009], page 62.

2.2.6 Derivatives and Phase-Plane Plots

Once the smoothed curves are obtained, the derivatives can be obtained as follows:

$$D^mx(t) = c_1D^m\phi_1(t) + ... + c_kD^m\phi_k(t)$$

where D^m denotes the m^{th} derivative operator [Ramsay, 2006]. As some of the variation in a curve can be explained at the level of certain derivatives, we use phase-plane plots to visualise velocity against acceleration [Hall et al., 2009]. Phase-plane plot is a graphical representation of derivatives, specifically acceleration versus velocity. It is a way of learning how derivatives relate to each other through the energy transfer between kinetic and potential energy within the system. Ramsay et al. [2009], page 32 has indicated that the properties that one should look for in a phase-plane plot are namely:

- The size of cycle
- The magnitude of the radius: the larger it is, the higher the energy occur in the event
- The horizontal location of the center: if it is in the right-hand side, there is net positive velocity, and if it is in the left-hand side, there is net negative velocity

- The vertical location of the center: if it is above zero, there is net velocity increase, if it is below zero , there is velocity decrease
- Changes in the shapes of the cycles in periods (yearly or monthly).

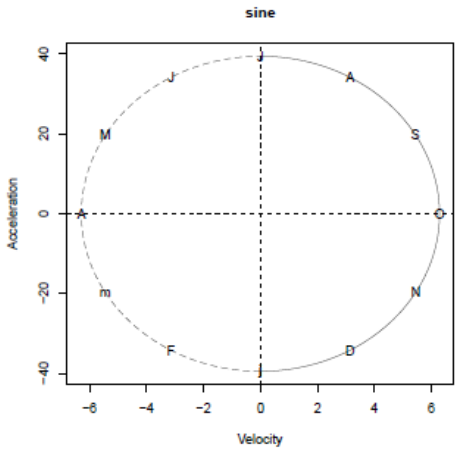


Figure 2.1: Phase-plane plot of a sine function [Ramsay et al., 2009], page 32.

In Figure 2.1 is an example of a phase-plane plot of the function $\sin(2\pi t)$ from Ramsay et al. [2009] used to describe the energy transfer. On the Figure 2.1 we have the months labelled starting with January (lowercase j). They are described as a harmonic process that can be compared to the vertical position of the end of the suspended spring [Keser, 2015]. The spring start again at $t = 0$, where the spring oscillates due to an exchange between kinetic and potential energy. When the spring passes through 0, the velocity is at its highest but the acceleration is zero [Keser, 2015]. This may show that potential energy is associated with high acceleration and kinetic energy with high velocity.

In this Chapter we introduced FDA and presented the smoothing techniques as tools to deal with curves with limited functional parameters rather than using the entire data (which can be of huge volume), which in turn reduces the amount of data one needs to handle everytime for inferential statistics. As a result smoothed curves may exhibit both amplitude and phase variation that may need to be registered to be able to extract important features. In the next Chapter we introduce and discuss methods that are used to align the curves or functions.

Chapter 3

Curve Registration

When we have our observations in functional form, we often see that the functional observations have variation both in phase (x axis) and amplitude (y axis). Curve registration is a non-linear method that is used to remove variation in functional observations. It is also known as time warping or curve alignment. For example most of the curves in functional space, like human growth curves from [Ramsay, 2006] page 123, exhibit these two type of variability. If the presence of variation is ignored the mean functions will not be of good use in terms of interpretation because if the important features in a curves vary from curve to curve it will dismally ignore these feature [Wang et al., 2015]. Therefore the representation of the sample curves will fail. Figure 2 below shows a representation of phase and amplitude variation in curves.

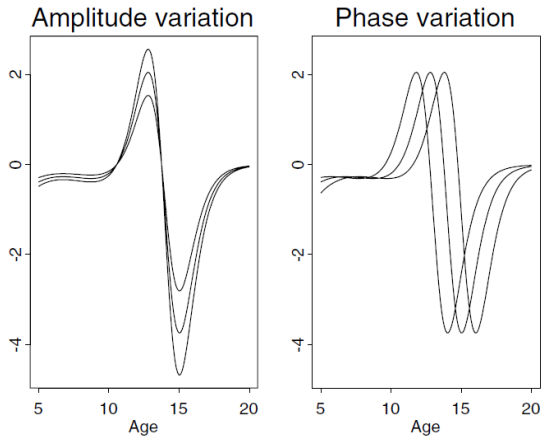


Figure 3.1: The left panel shows three height acceleration curves of human growth data varying in amplitude and the right panel shows three curve of phase variation [Ramsay et al., 2009], page 123.

One of the challenges in curve registration is the identification of both amplitude and phase variation, which can help to identify what caused each type of variance. For example from

Ramsay et al. [2009] page 123 in Figure 3.2 below they have indicated that after eight years of age the peaks in the curves exhibit phase variation. Phase variation is a phenomenon in FDA that often comes with deformation in curves that may lead to increases in data variance, and can cause blury data structure which may lead to principal components pulling out of shape. For example puberty for girls usually take place at the age of 11.7 years Ramsay et al. [2009] page 117, from [Marron et al., 2015] factors such as hormones and physiology can shift the age forward and backward. It is very difficult to remove phase variation from amplitude variation and the problem that comes with ignoring the phase variation can result to failure in most common or known data analysis as the presence of phase variation increases variance in data [Marron et al., 2015].

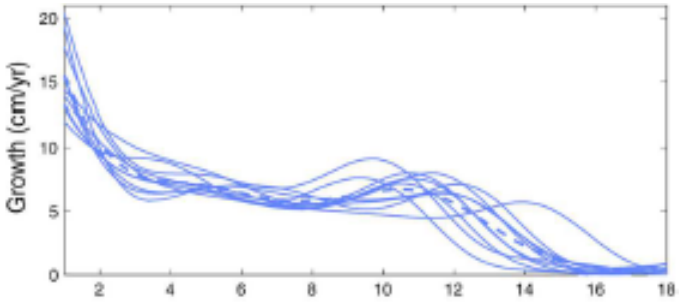


Figure 3.2: The first derivative of height of growth data [Marron et al., 2015].

One of the steps that can be implemented to align curves is removal of amplitude variation as it is relatively easy to remove amplitude variation than phase variation. Phase variation contain interesting information in the timing of the curves of important peaks. In Figure 3.3 the top panel represent unregistered curves and the bottom panel phase registered curves. We observe that registered curves helps us to carry out a better averaging estimate, for example from Ramsay et al. [2009] page 118 they have explained that registering growth curves helps pubertal spurt of all girls to occur roughly at the same time to get a better or acurate average. Both phase and amplitude variation can be removed using various curve registration methods known as *landmark* registration, *continuous* registration, and *shift* registration. These methods use the time warping function h to transform the domain time for each curve [Marron et al., 2015].

3.1 Methods of Curve Registration

Time-Warping Function h

Time-warping is a technique that transforms the domain (e.g., time) for each curve for aligning certain features of interest so that for any given landmark. The properties of time-warping function include:

- They must always be strictly increasing, as time cannot go backward

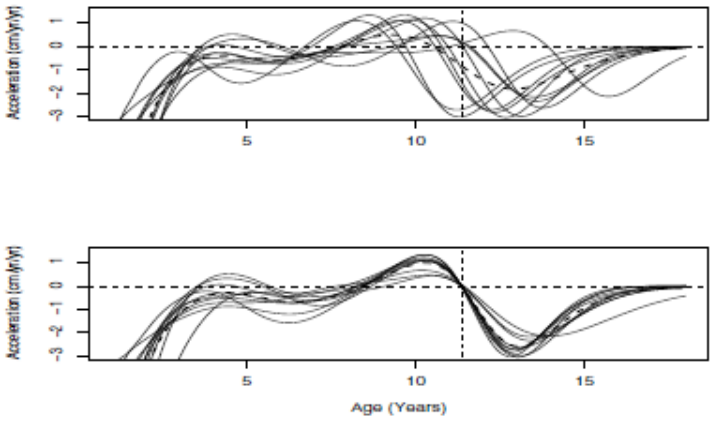


Figure 3.3: Top panel is the unregistered of the second derivative of height of the growth data and bottom panel is registered [Ramsay et al., 2009], page 118 .

- Assuming that the curves are observed over an interval $[0, T]$, the time warping functions must satisfy the constraints $h(0) = 0$ and $h(T) = T$.

The time-warping function transforms time by changing the initial starting time to an observed starting time and the initial ending time to the observed time at the same time [Zhong, 2008]. For example: given $h_i(t)$ is our time-warping function for the i^{th} subject over the interval $[0, T_i]$, the values of $h_i(t)$ satisfy $h(0) = 0$ and $h_i(T_0) = T_i$, where T_0 is the initial ending time and T_i is the observed ending time. If the observed process is delayed, then $h(t) > t$ (we speed up time), and if the observed process is ahead of time then $h(t) < t$ (we slow down time). For example, Ramsay et al. [2009] page 119 have estimated a time-warping function $h_i(t)$ for the growth data that transforms growth time t to clock time for child i, if we observe that $h_i(t) < t$ we may say that the girl is growing faster than average clock time. Below we present details about three popular registration techniques.

3.1.1 Landmark Registration

A landmark is a feature that can be identified at zero crossing of curves. The *landmark* registration aligns the curves by first identifying minima, maxima or crossing of zero in a curve. It uses the time-warping function $h_i(t)$ to transform time by calculating the registered function values of the curve

$$x_i^*(t) = x[h_i(t)] \tag{3.1}$$

where x_i^* denotes the function to be registered and x_i denotes the initial function of curve i, so that the landmarks are the same for all curves. *Landmark* registration is usually a preferable first method to use in registration as landmarks or minima or maxima are mostly always visible in curves but when the landmarks are not visible we will need a more refined registration method such as continuous registration [Ramsay, 2006].

3.1.2 Continuous Registration

Continuous registration is a method or technique that aligns the entire curve rather than using features like landmarks to align the curves [Ramsay, 2006] *Continuous* registration can perform a complete curve registration. It uses the time-warping function $h_i(t)$ to minimize amplitude variation. In Figure 3.4 is an example of curve x_1 and x_2 that only differ in amplitude and not in phase variation. The curve to be registered to the dashed line curve is the dotted curve and the solid line is the registered curve. The left panel is the time-warping function $h(t) = 0$ [Ramsay, 2006], page 139.

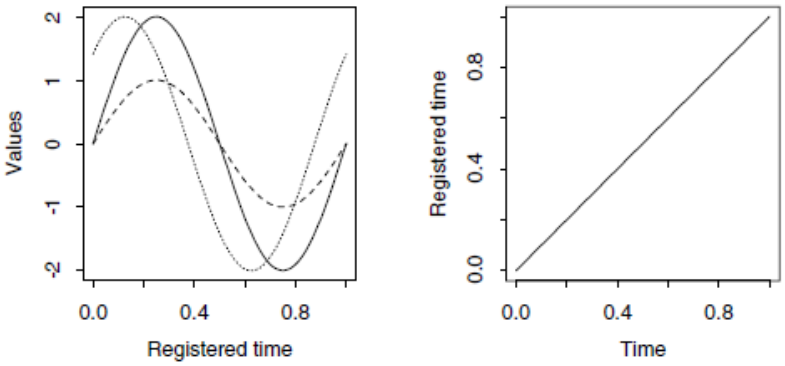


Figure 3.4: Artificial temperature results of registration [Ramsay, 2006], page 139.

3.1.3 Shift Registration

Shift registration is the registration technique that only require a shift in the time scale, in Figure 3.5 is an example of a curve that only require a shift in the time scale to align it. From Ramsay [2006] page 130 they state that Edmonton station have earlier seasons than marine station St.John's because of the capacity of oceans to store heat and to release it slowly therefore the stations's weather will have to be shifted three weeks to align the two. For example we let τ to be the interval over which the functions are to be registered and assume that each function x of curve i exits at time t in the region of τ. Then

$$x_i^*(t) = x_i(t + \delta_i), \tag{3.2}$$

where x^* denote the unregistered curve i. We first estimate the shift parameter δ_i using the global registration criterion in equation (3.3) to properly align the curve, then calculate the unregistered curve using equation (3.2) [Ramsay and Li, 1998]. The shift parameter δ_i is

estimated using the global registration criterion as follows:

$$REGSSE = \sum_{i=1}^{N} \int_{\tau} [x_i(t + \delta_i) - \hat{\mu(t)}]^2 ds = \sum_{i=1}^{N} \int_{\tau} [x_i^*(t) - \hat{\mu(t)}]^2 ds, \tag{3.3}$$

where $\hat{\mu(t)}$ is the estimated mean.

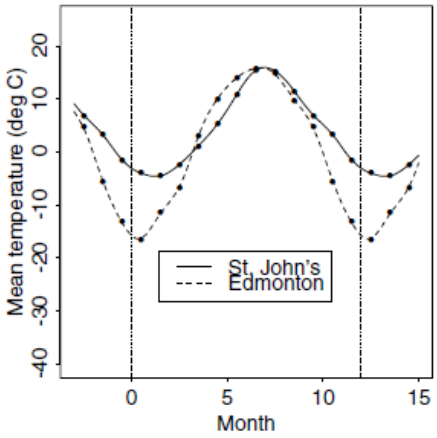


Figure 3.5: Temperature records for St.John's and Edmonton weather station where the seasons have phase variation [Ramsay, 2006], page 130.

In the next 'Chapter' 4 we introduce techniques: the traditional PCA and functional PCA and its examples used to reduce the dimension of the data.

Chapter 4

Functional Principal Component Analysis (fPCA)

fPCA is a linear approach that is used in multivariate kind of data, to understand the dominant factors that affect variabilty, PCA and fPCA are dimension reduction techniques that reduces the number of variables in data while still preserving all the information related to the variability inherent in the data. We briefly present the traditional PCA followed by the details of fPCA.

PCA is a dimension reduction technique which intend to find the principal components using eigenvalues and eigenvectors to reduce the dimension of a set of data. It is a method of identifying principal directions of variability in a data set [Rajaraman and Ullman, 2010]. Eigenvalues show how data is spread out in the direction of an eigenvector, while eigenvector is a direction. Suppose we have an $n \times m$ matrix A. A scalar λ is called an eigenvalue of A if there is a $m \times 1$ non-zero vector X such that $AX = X\lambda$. X is called an eigenvector of A corresponding to the eigen value λ [Rajaraman and Ullman, 2010]. We demostrate the methods with a toy example. Suppose we have a 2-dimension data set and we want to reduce the dimension using PCA. Specifically, let our data be (2,3),(3,2),(4,5),(5,4) [Rajaraman and Ullman, 2010].

- We first convert the data set into a matrix B as follows:
  $B = \begin{bmatrix} 2 & 3 \\ 3 & 2 \\ 4 & 5 \\ 5 & 4 \end{bmatrix}$, with $B^T B = \begin{bmatrix} 54 & 52 \\ 52 & 54 \end{bmatrix}$

- Next we calculate for λ by solving the characteristic equation $det(B^T B - I\lambda) = 0$. This gives $\lambda = 106$ or $\lambda = 2$.

  Hence, $\lambda = 106$ is the 1st most important principal eigenvalue.

- To calculate eigenvectors corresponding to eigenvalue $\lambda = 106$ and $\lambda = 2$ we proceed as follows:

We first start with calculating eigenvector corresponding to $\lambda = 106$. Let $X = \begin{bmatrix} a \\ b \end{bmatrix}$ be the unknown vector. If we suppose $B^T BX = X\lambda$, then

$$\begin{bmatrix} 54 & 52 \\ 52 & 54 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 106 \begin{bmatrix} a \\ b \end{bmatrix}$$

Solving the equations gives us the eigenvector X corresponding to $\lambda = 106$ as $X = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Since it is easier to find standardised eigenvector, subject to the constraint $||\mathrm{X}||^2 = 1$, we can calculate the distance of eigenvector X as follows: $||\mathrm{X}|| = \sqrt{1^2 + 1^2} = \sqrt{2}$, which results in the standardised eigenvector X being:

$X = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

we get the standardised eigenvector X corresponding to $\lambda = 2$, as $X = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

- Next we form a matrix E of the set of standardised eigenvectors in the descending order of magnitude of the eigen values as follows:

  $E = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$.

- Since our toy example is of 2-dimension, the only reduction in dimension that can be achieved is reducing it to dimension one. We next calculate BX to get a new set of dimension as follows:

  $$BX = \begin{bmatrix} 2 & 3 \\ 3 & 2 \\ 4 & 5 \\ 5 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 5 \\ 5 \\ 7 \\ 7 \end{bmatrix}.$$

  This transformation reduces the original matrix B in dimension from 2-dimension to 1-dimension, with the reduced dimension vector being $\frac{1}{\sqrt{2}}(5, 5, 7, 7)^T$.

fPCA is a dimension reduction tool for multivariate data that has been extended to functional space [Wang et al., 2015]. fPCA is similar to a standard Principal Component Analysis (PCA), with the primary difference being that PCA does not account for smoothness and continuity while fPCA do account for smoothness and continuity [Hadjipantelis and Müller, 2018]. fPCA transforms or reduces high dimensional dataset to a low-dimensional dataset, which contains a set of uncorrelated components that summaries features that represent the original dataset. In short fPCA captures the main modes of variability of data [Cardot et al., 2008]. By using eigenvalues and the eigenfunctions to reduce the dimension of a dataset. Eigenfunctions in a dataset represent rotation of the original functional data of a diagonal

covariance matrix of the data. One of the challenges in fPCA application is the selection of the number of components to retain or reject in fPCA [Hadjipantelis and Müller, 2018].

To explain the effect of smoothing on functional principal component, we refer to criminology data example from Ramsay and Silverman [2007] page 25. In Figure 4.1 The left panels consists of three unsmoothed functional principal components and the right panels consists of smoothed functional principal components. Where the dashed line denote the overall mean curve. Note, Smoothing removes roughness in the raw functional principal component curves [Ramsay and Silverman, 2007]. The smoothness can affect application of fPCA, thus smoothing can be done directly on the eigendecomposition step in fPCA or Smoothing can be done directly on raw data [Hadjipantelis and Müller, 2018]. Hence for this thesis we will pursue smoothed and registered curves.

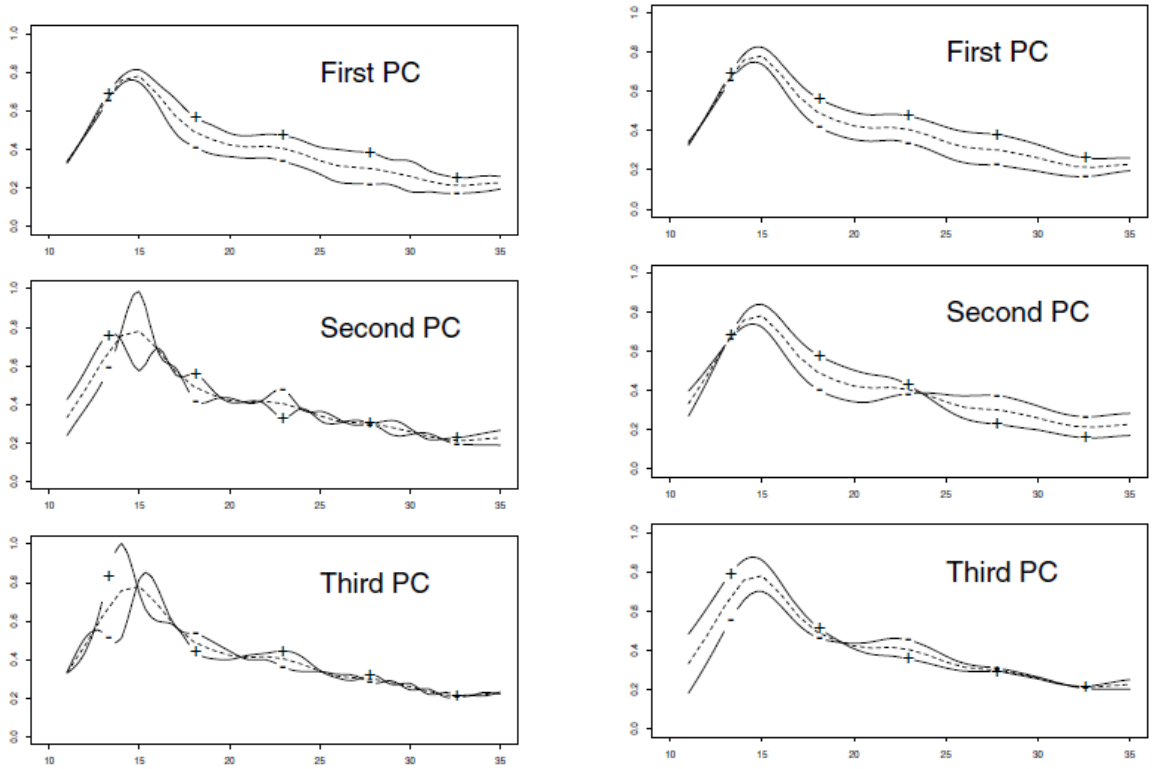


Figure 4.1: The left panel is the three unsmoothed functional principal components and on the right panel we have smoothed functional principal components of criminology data [Ramsay and Silverman, 2007], page 25.

fPCA express the eigenequation to matrix form of each function x_i of the basis function ϕ_k through the estimation of covariance operators, eigenvalues and their corresponding eigenfunction [Hall and Hosseini-Nasab, 2006]. Let us suppose that each function has basis func-

tions as follows:

$$x_i(t) = \sum_{i=1}^{K} c_i \phi_k(t). \tag{4.1}$$

where x have components $x_1, ..., x_N$, and the vector-valued function ϕ have components $\phi_1, ..., \phi_k$, and c is the $n \times k$ (n is the number of observations and k is number of basis functions) dimension coefficient matrix,

Functional eigenequation is defined as follows:

$$\int v(s,t)\xi(t)dt = \rho\xi(s). \tag{4.2}$$

where ρ is an eigenvalue, $\xi(s)$ is an eigenfunction and $v(s,t)$ is the sample variance-covariance function which is defined by:

$$v(s,t) = n^{-1} \sum_{i=1}^{n} x_i(s)x_i(t), \tag{4.3}$$

Solving the eigenvalues and eigenfunctions one proceeds as in Ramsay [2006]:

- Suppose the observed functions are expanded as a vector $\phi(t)$ of K basis functions:

$$x(t) = c^T\phi(t), \tag{4.4}$$

$$x(s) = c^T\phi(s), \tag{4.5}$$

- Substituting equation (4.4) and (4.5) into $v(s,t)$ in equation(4.3) we get:

$$v(s,t) = n^{-1}\phi^T(s)cc^T\phi(t), \tag{4.6}$$

- Now suppose the eigenfunctions are defined as follows:

$$\xi(s) = \phi^T(s)b, \tag{4.7}$$

$$\xi(t) = \phi^T(t)b, \tag{4.8}$$

  where b is kx1 vector.

- Substituting equations (4.6), (4.7) and (4.8) in the eigen equation (4.2) we get:

$$n^{-1}\phi^T(s)c^Tc \int \phi(t)\phi^T(t)dtb = \rho\phi^T(s)b. \tag{4.9}$$

- Then we define the matrix Z of order K as follows:

$$Z = \int \phi(t)\phi^T(t)dt \tag{4.10}$$

to get the eigen equation :

$$n^{-1}\phi^T(s)c^T cZb = \rho\phi^T(s)b \implies n^{-1}c^T cZb = \rho b \tag{4.11}$$

- If we can define $U = Z^{1/2}b$ and solve for b we get:

$$b = Z^{-1/2}u \tag{4.12}$$

and substituting equation (4.12) in equation (4.11) we get:
$n^{-1}Z^{1/2}c^T cZ^{1/2}Z^{-1/2}u = \rho Z^{-1/2}u$

$\implies \rho = n^{-1}Z^{1/2}c^T cZ^{1/2}$.

Once we have the fPCA, we can identify the dominant direction of the variability. In the next Chapter we proceed to introduce and explore traditional clustering and functional clustering techniques.

Chapter 5

Functional Clustering

5.1 Traditional Clustering

Clustering is an unsupervised learning process which groups a set of data into clusters such that data objects in the same cluster are more similar than those in other clusters [Krishna and Murty, 1999]. There are various types of algorithms for clustering, however in this thesis we will focus on *K-means* and *Hierarchical* clustering.

Clustering algorithms
K-means clustering is a method that groups the objects into a pre-specified number of clusters based on the minimum distance calculated using the euclidean distance method [Krishna and Murty, 1999]. It starts with choosing the number of clusters defined by K, then proceeds to calculate the centroid and distance between objects, and finally groups the objects into one of the K clusters based on minimum distance criteria [Hartigan and Wong, 1979]. The steps to follow when using K-means algorithm to cluster datasets [Hartigan and Wong, 1979] are as follows: Let us consider a 2-dimension datasets $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_n$

(i) We choose number of clusters defined by K.

(ii) For each point x_i we find the nearest centriod $c_j,\ j = 1, ..., K$ using euclidean distance defined by $[(x, y); (a, b)] = \sqrt{(x-a)^2 + (y-b)^2}$. We calculate the distance of all data points.

(iii) We then construct the table containing data with reference to the clusters.

(iv) We display the clusters.

(v) We re-calculate the mean for the new cluster and repeat steps (ii) to (iv).

(vi) If clusters containing the same data points repeat themselves then we stop [Hartigan and Wong, 1979].

Hierarchical clustering is a method that uses either 'agglomerative algorithm' or 'divisive algorithm' to group objects into clusters without initially specifying the number of clusters

[Fernández and Gómez, 2008]. It starts with one cluster and iteratively divides (divisive) or iteratively combines (agglomerative) objects. A visual representation of the algorithm is called a 'dendogram' [Blei, 2008]. The agglomerative algorithm uses single, complete, and average linkage to calculate the distance between the objects. The single linkage is the distance between the closest members of the two clusters; the complete linkage is the distance between the members that are farthest apart; and the average linkage is the distances between all pairs and averages of of all distances. Steps to perform when using agglomerative algorithm to cluster datasets [Blei, 2008] are as follows:

(i) First we calculate the euclidean distance between two points to create the distance matrix which is a square matrix containing the distances between the elements of a set.

(ii) We construct the distance matrix and display it.

(iii) From the distance matrix we select the minimum data point between two sets.

(iv) We construct a cluster in dendrogram format.

(v) Update the distance matrix depending on whether its a single, complete or average linkage.

(vi) We repeat steps (ii) to (v).

(vii) We stop when clusters has merged to form one cluster [Johnson, 1967].

Figure 5.1 below shows an example of a dendrogram.

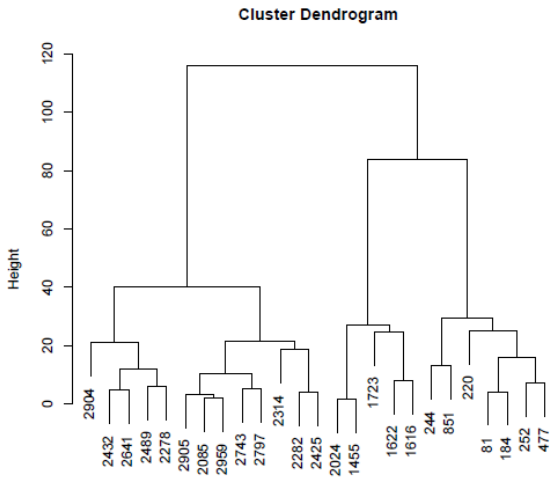


Figure 5.1: Cluster dendrogram of groups that merge at high values relative to the merger values of their subgroups are candidates for natural cluster [Blei, 2008].

In Table 5.1 we briefly explain the advantages and disadvantages of *K-means* and *Hierarchical* clustering.

Table 5.1: Advantages and disadvantages of clustering algorithms [Santini, 2016].

Clustering Algorithm	Advantages	Disadvantages
K-means	Works well with large datasets	Difficult to predict number of clusters (K)
Hierarchical	Does not work well with large datasets	Easy to make a choice of the number of clusters

5.2 Functional Clustering

The focus of functional clustering is to identify representative curve patterns which are likely generated from the similar process [Zhang and Telesca, 2014]. *K-means* and *Hierarchical* functional are the two popular algorithms that we will focus on in this thesis clustering. Clustering functional data can be a challenge because functional data has infinite dimension. Different approaches have been proposed but one of the most popular approach is reducing the infinite dimensional problem to finite. Therefore there can be two steps that we can use to approach for clustering

(i) First we can reduce the dimension of data using fPCA

(ii) Second we can apply K-means and/or hierarchical clustering.
We present below some details on *K-means* and *Hierarchical* clustering and their counterparts.

K-means functional clustering.
Given a set of functional data $\{X_i(t); i = 1, ..., n\}$, *K-means* finds a set of cluster means denoted by $\{\mu^c; c = 1, ..., L\}$ by minimising the sum of the squared distance between $\{X_i\}$ and the cluster centers $\{C_i; i = 1, ..., n\}$ with their cluster labels $C_i; i = 1, ..., n$, for functional distance d. Here L is the number of clusters of interest $\leq n$. Equation (5.1) denotes objective function to be minimised:

$$\frac{1}{n}\sum_{i=1}^{n} d^2(X_i, \mu_n^c), \tag{5.1}$$

where $\mu_n^c = \sum_{i=1}^n X_i(t) 1_{\{C_i=c\}}/N_c$, and $N_c = \sum_{i=1}^n 1_{\{C_i=c\}}$.
This implies $\mu_n^c = \sum_{i=1}^n X_i(t) N_c$
Minimising the equation above converts an infinite-dimensional functional data into a lower dimensional space [Wang et al., 2015].
Hierarchical functional clustering is no different from the regular *hierarchical* clustering and uses either the 'agglomerative algorithm' or the 'divisive algorithm' to group curves

into clusters. The agglomerative algorithm in functional space starts by calculating the distance between the curves, then calculating the distance or dissimilarity matrix, then finally proceeds to apply agglomerative criterion (single, complete, or average linkage).
Suppose that we have a sample of curves $X_1(t), ..., X_n(t)$ defined for $t \in T = [a, b]$. We assume that the functions can be represented using some basis functions as follows:

$$X_i(t) = \sum_{j=1}^{K} a_{ij}\phi_j(t) \tag{5.2}$$

where $\mathbf{a}_i$ and $\mathbf{a}_j$ are the vectors of the basis coefficient for the i^{th} and j^{th} curves, and $\phi(t)$ is the vector of the basis function. We next calculate the distance between curves $X_i(t)$ and $X_j(t)$ as follows:

$$\begin{aligned} d_{ij} &= \sqrt{\int_{[a,b]} (x_i(t) - x_j(t))^2 dt} \\ &= \sqrt{\int_{[a,b]} (\mathbf{a}_i - \mathbf{a}_j)^T \phi(t)\phi(t)^T (\mathbf{a}_i - \mathbf{a}_j) dt} \\ &= \sqrt{(\mathbf{a}_i - \mathbf{a}_j)^T P (\mathbf{a}_i - \mathbf{a}_j)} \end{aligned} \tag{5.3}$$

where is $P = \int_{[a,b]} \phi(t)\phi(t)^T dt$, and P is the identity matrix. We then calculate the distance matrix to apply the agglomerative algorithm [Giraldo et al., 2012].

Once we have the clustering algorithm, we can apply it on smoothed registered data. In this thesis we will be focusing on the application of functional clustering algorithms (*K-means* and *hierarchical* clustering) on registered curves, specifically in this thesis our focus is to understand temperature patterns across South Africa for which we will be using both fPCA and functional clustering algorithms.

In the next Chapter we present exploratory data analysis on monthly maximum temperature data.

Chapter 6

Exploratory Data Analysis of Monthly Maximum Temperature Using FDA Methods

The African continent has been established as vulnerable to climate change [Kusangaya et al., 2014]. Within the African continent, the Southern African region is established as one of the most vulnerable regions in Africa [Kusangaya et al., 2014]. Vulnerability is defined as an amount by which a system may be harmed by climate change [Kiker, 2000]. Ziervogel et al. [2014] have indicated that climate change is one of the concern in South Africa with experiences of an increase in mean annual temperature and extreme rainfall events. As a result it is mentioned that it poses a threat to South Africa's water resources, food security, health, infrastructure, and biodiversity [Ziervogel et al., 2014].

For example from Ziervogel et al. [2014] in South Africa climate change was identified through the case study of a short term insurance, with insurance claims in the Western Cape province caused by extreme weather events such as those fires caused by periods of drought, flooding, and storms and global warming is related to such extreme events [Dosio, 2017]. Coastal areas have also been identified as vulnerable to global warming [Lakhraj-Govender and Grab, 2018], in which it is mentioned from Lakhraj-Govender and Grab [2018] that the Southeren African Mozambique regions experiences daytime warming and nighttime cooling from a period of 1939-1992, while northern part of Eastern Africa indicate the opposite of Mozambique (dayttime cooling and nighttime warming) at the same time.

Komen et al. [2015] have shown that there is a strong correlation between an increase in temperature and malaria cases in the Limpopo Province in South Africa. Tshiala et al. [2011] have also shown that temperature changes may lead to changing of patterns in rainfall, soil moisture and an increase of droughts and floods in the Southeren Africa. From the a formation studies it is evident that an increase in temperature is a concern in Africa resulting in extreme weather events which may affect human health and well being. In this thesis we identified to work with temperature data for South Africa.

Even though there is research on temperature related to South Africa, we have not found any work that involve FDA methods. Our contribution is thus analysis of monthly maximum temperature using FDA methods and to get insights into variability in patterns. It is evident that there is a need or gap for continuous research in different aspects of climate in Africa and hence this thesis aims to contribute with the space, by focusing on monthly maximum temperature patterns of 16 locations over the duration of 10 years (with 5 years intervals). Below we present details about data that was subject to FDA application as part of the scope in this thesis.

For our analysis weather data was acquired from the Natural Resource and the Environment (NRE) operating unit of the Council of Science and Industrial Research (CSIR). The original data contain, temperature , precipitation, and relative humidity in the form of NETCDF (Network Common Data Form) in kelvin. In this thesis we will only focus on monthly maximum temperature from the data which is first converted from the NETCDF file into a .CSV file (Comma-Seperated Values), the reader may refer to Appendix A for code in R for converting files. Then converting the original temperature unit from Kelvin (K) to Celsius (C). The unit of temperature is usually measured in Kelvin or Celsius, with the magnitude of C equal to K with the relationship between C and K as follows $T_{(C)} = T_{(K)} - 273.15$, where T is temperature [Preston-Thomas, 1990].

We have organised the monthly maximum temperature data by extracting temperature data from 16 cities in South Africa and converted the data to series of annual curves for each location. The maximum temperature has a dimension of 160 rows and 15 columns of which column 1 is Longitude, column 2 is Latitude, column 3 is Year and colum 4 to 15 is the months of the year. Note the time spam of our data is from 1960-2010, and we analysed data of every 5 year interval specifically 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010. We selected 16 locations of major cities spread across South Africa from which we acquired maximum monthly temperature data. We visualise the 16 cities on a map in Figure 6.1 and a table of reference in Table 9.1 geographical locations of the cities in Appendix A.

6.1 Fitting of Monthly Maximum Temperature Data

For fitting monthly maximum temperature data of all 16 locations together and each location separately, we constructed a code in R (*fda* library) using B-spline functions explained in Section 2.2.3 with number of basis function $K = 5$ which is determined by: number of basis function = order of spline + number of interior points. We chose B-spline basis functions as our basis function to smooth the data because we want a basis system that will allow us to have control of the smoothness of our functions and take into account the general temperature of the Southern hemisphere, where high temperature prevail around November to February and low temperature around June to August. In Figure 6.2 below we visualise the results of raw and smoothed data of 160 curves which are the annual curves of 16 locations across South Africa of each year from 1965 to 2010 with 5 years of intervals. were the curves are identified by location and year, each location consists of 10 curves of the period from 1965

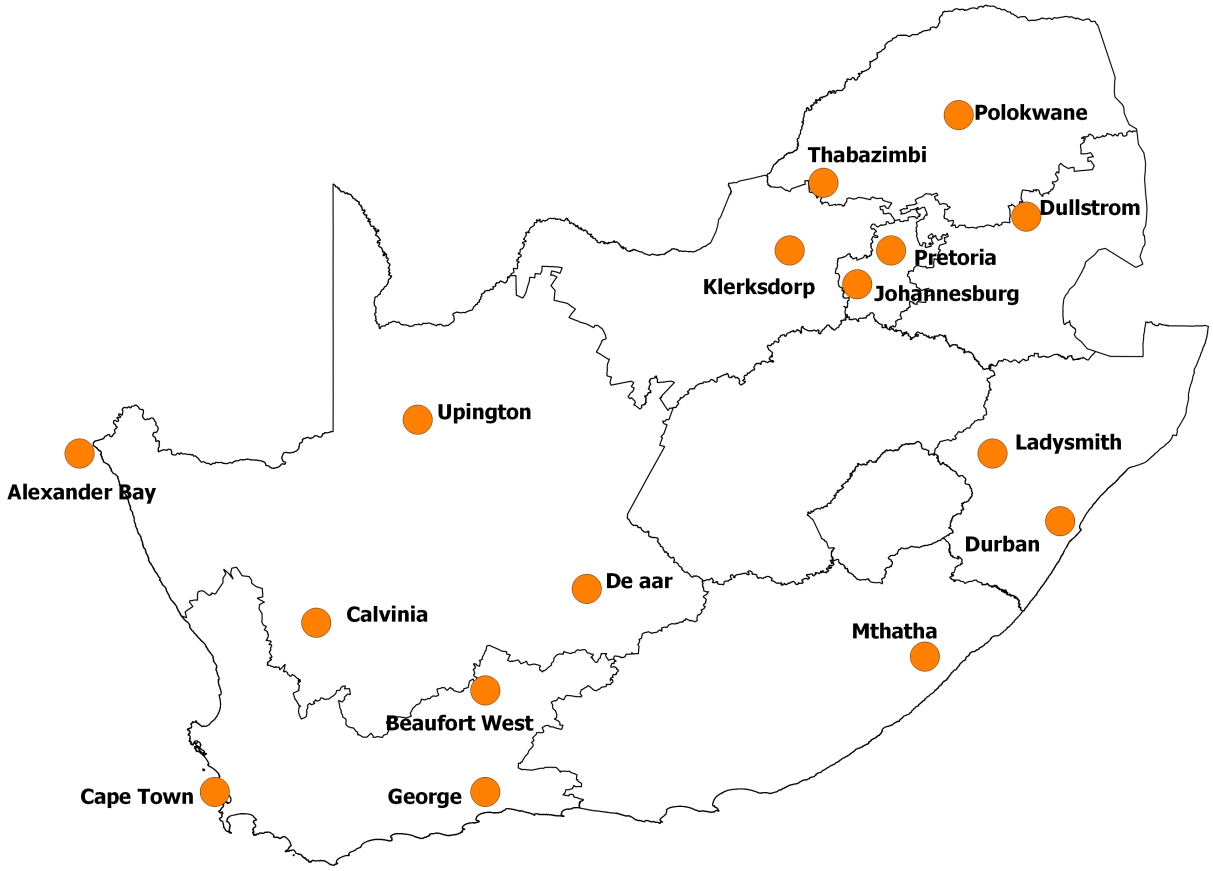


Figure 6.1: Map of South Africa with the close locations to the original latitude and longitude of monthly maximum temperature data.

to 2010 with 5 years intervals.

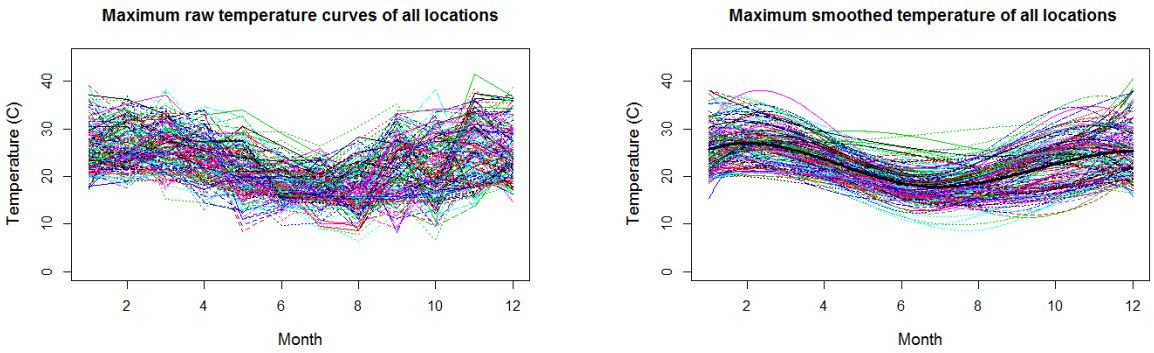


Figure 6.2: Raw and smoothed 160 monthly maximum temperature curves in all 16 locations across South Africa from 1965-2010 with 5 years of intervals.

6.2 Curve Registration of Monthly Maximum Temperature

The curves in functional data may differ in amplitude and may also exhibit phase variation, for example were peaks and valleys at the same locations of two or more curves occuring at the same time. We align the curves with the target function so that the peaks, valleys and crossings occur at the same argument as those of the target. We used *continuous* registration explained in Section 3.1.2 to register our maximum monthly temperature curves rather than using the method of identifying landmarks (*landmark* registration), which in our case will be difficult to identify as we are using large datasets. Moreover, to align curve $x_1(t)$ and $x_n(t)$ defined on the domain [0,12] we estimate time warping functions explained in Section 3.1 that transforms the time values to align the common features in the curves and smooth the data which we have already done in Section 6.1.

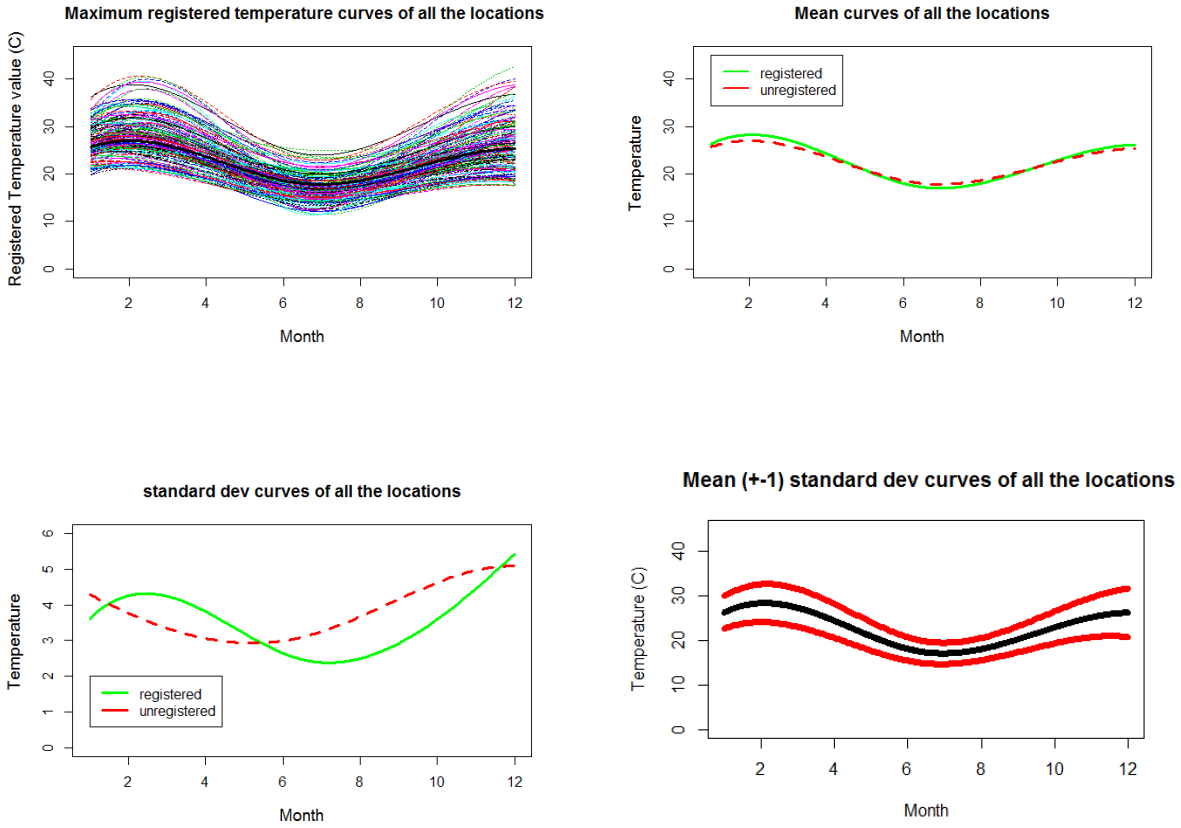


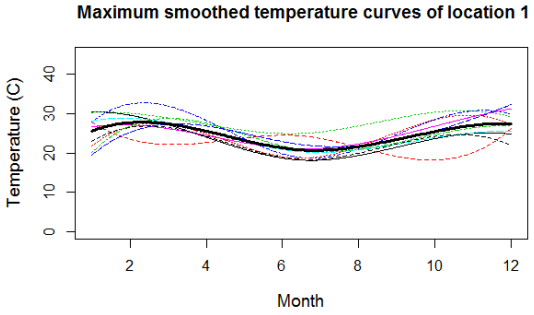
Figure 6.3: Top panel is monthly maximum 160 registered temperature curves of 16 locations across South Africa from a period of 1965 to 2010 with 5 years intervals , its standard deviation curves and bottom panel is mean for both unregistered and registered maximum monthly temperature curves, mean(+-) standard deviation curves.

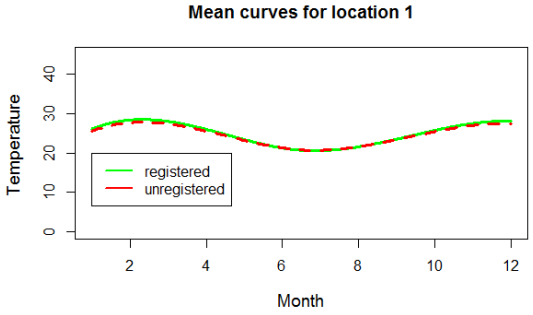
Therefore we will use our smoothed data to define our target function as the mean of temperature curves across locations. and we used the warping function and "register.fd" function in R (*fda* library) to register our curves. In Figure 6.3 below we visualise registered curves of all l6 locations together with their mean and standard deviation. From Figure 6.3 (top left panel) graph, the black solid line represent the mean of maximum monthly temperature curves of all 16 locations across South Africa. We see from the Figure in the top left panel that the peaks (higher temperature) occurs around February to April and again from October to December, while the troughs occur around July. The standard deviation function plot (bottom right) in Figure 6.3 suggests that the registered curves have the highest variability in temperature in summer months (from February to April and October to December) and have lower variability in winter months (from May to August). While the unregistered curves have the greatest variability increasing from July to October, and decreasing from February to June.

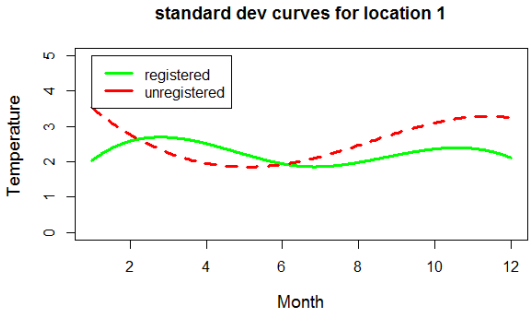
6.2.1 Curve Registration of Monthly Maximum Temperature of each Location Separately

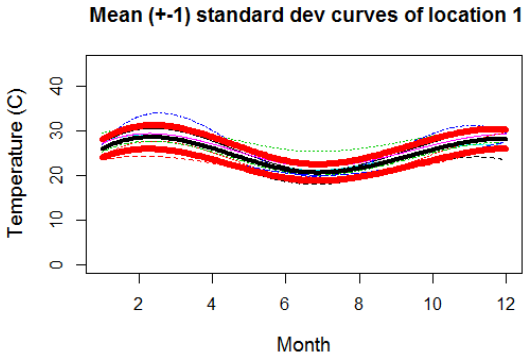
In this section we visualise curve registration plots of each location with their mean, standard deviation and mean (+-1) standard deviation separately. Each Figure of a location contain 10 annual curves of each year from 1965 to 2010 with 5 years of intervals.

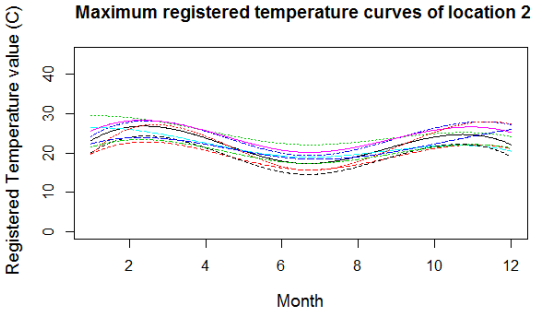
The standard deviation function plot of location 1 to 7 of the registered curves suggest that there were greater variability in summer months and lower variability in winter months. and in location 8,11,14 and 16 the registered curve of standard deviation is slightly decreasing with greater variability in January, it suggests that the temperature values are similar throughout the years. While in location 9,12,13 and 15 there is highest variability from May to August, January and December and the lower in February, thus Calvinia, De Aar, Klerksdorp, Johannesburg, and Pretoria had warmer winter in the analysed years. Location 10 had increasing temperature values from October to December which suggests the there were similar temperature values from January to August expect October to December.

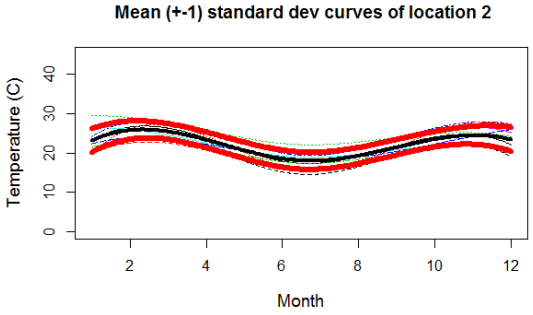


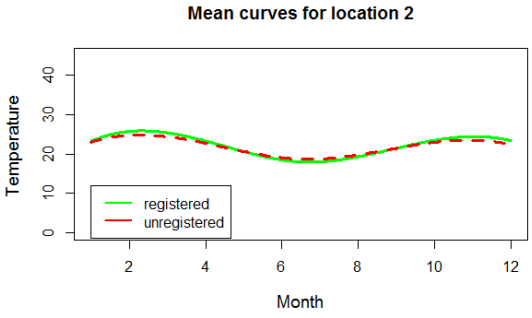


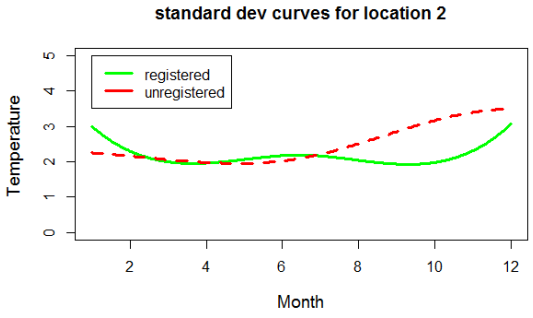

standard dev curves for location 1
registered
unregistered
Temperature
Month

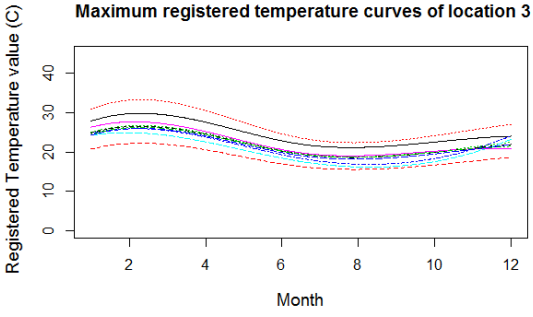

Mean (+-1) standard dev curves of location 1
Temperature (C)
Month

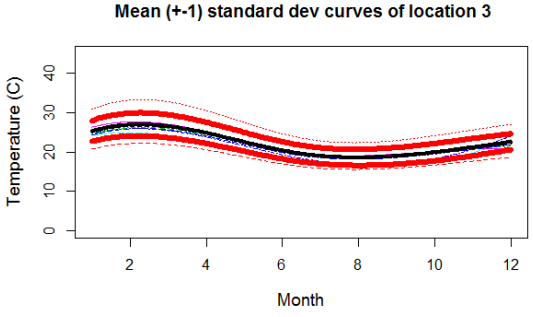

Maximum registered temperature curves of location 2
Registered Temperature value (C)
Month

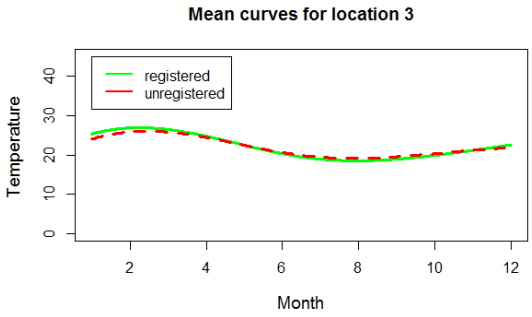

Mean (+-1) standard dev curves of location 2
Temperature (C)
Month

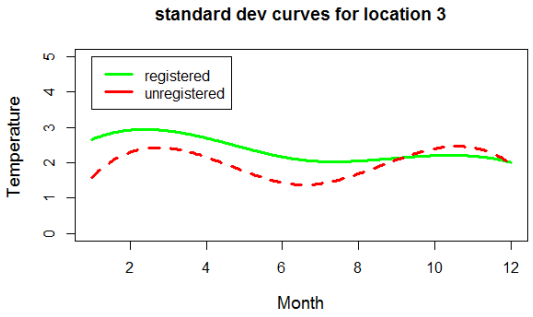

Mean curves for location 2
registered
unregistered
Temperature
Month

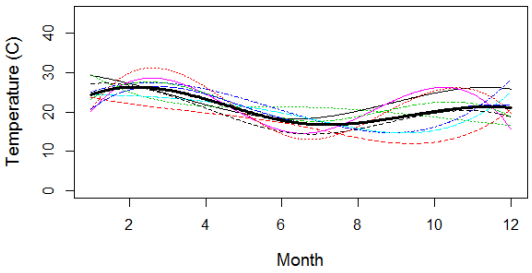

standard dev curves for location 2
registered
unregistered
Temperature
Month

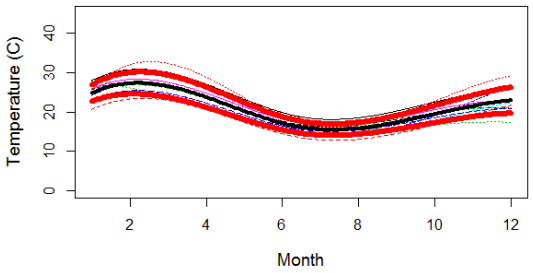

Maximum registered temperature curves of location 3
Registered Temperature value (C)
Month

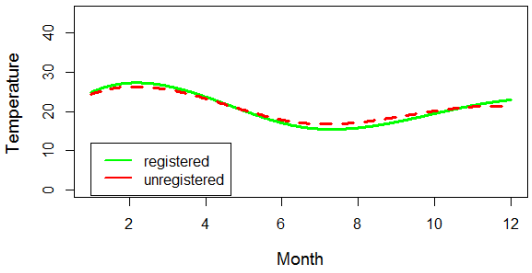

Mean (+-1) standard dev curves of location 3
Temperature (C)
Month

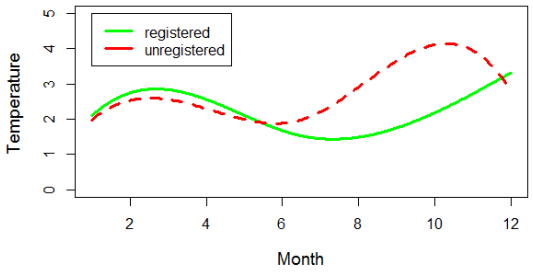

Mean curves for location 3
registered
unregistered
Temperature
Month

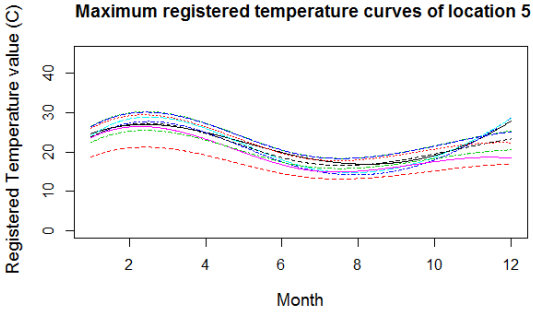

standard dev curves for location 3
registered
unregistered
Temperature
Month

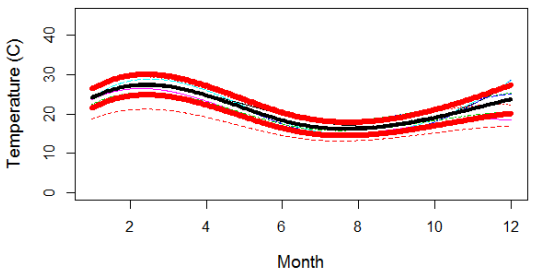

Maximum smoothed temperature curves of location 4
Temperature (C)
Month

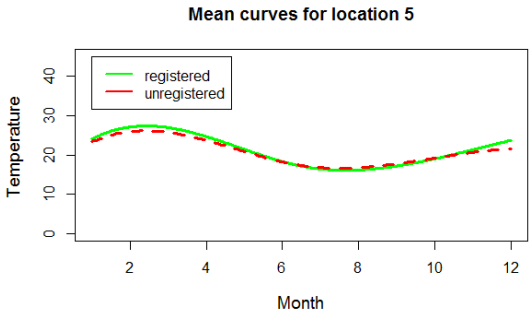

Mean (+-1) standard dev curves of location 4
Temperature (C)
Month

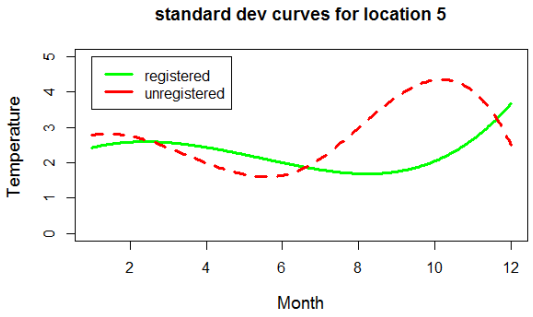

Mean curves for location 4
registered
unregistered
Temperature
Month

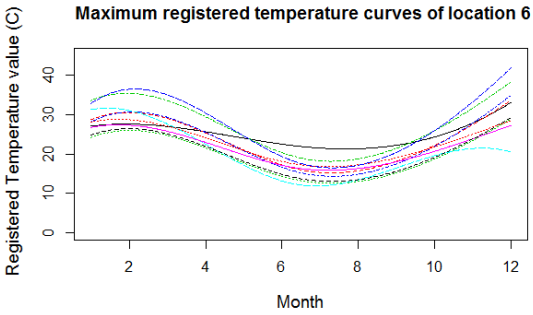

standard dev curves for location 4
registered
unregistered
Temperature
Month

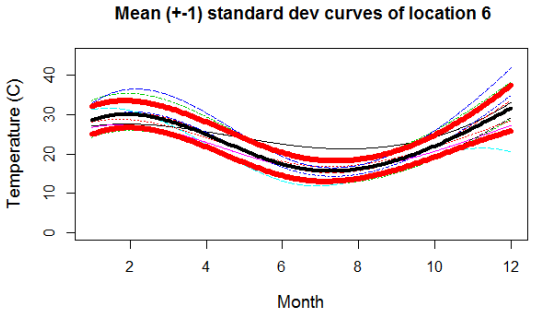

Maximum registered temperature curves of location 5
Registered Temperature value (C)
Month

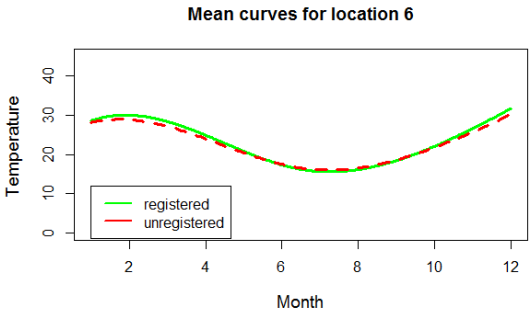

Mean (+-1) standard dev curves of location 5
Temperature (C)
Month

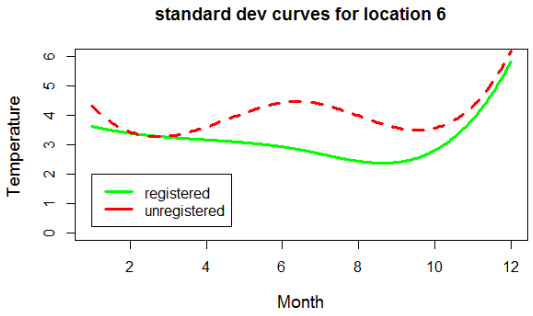

Mean curves for location 5
registered
unregistered
Temperature
Month

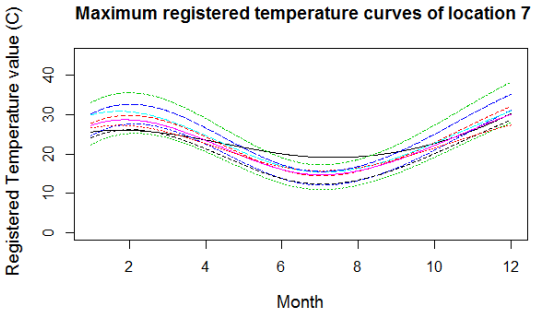

standard dev curves for location 5
registered
unregistered
Temperature
Month

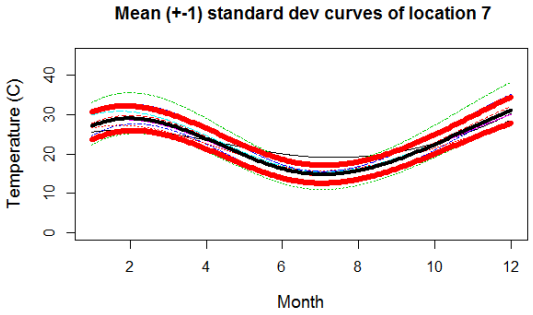

Maximum registered temperature curves of location 6
Registered Temperature value (C)
Month

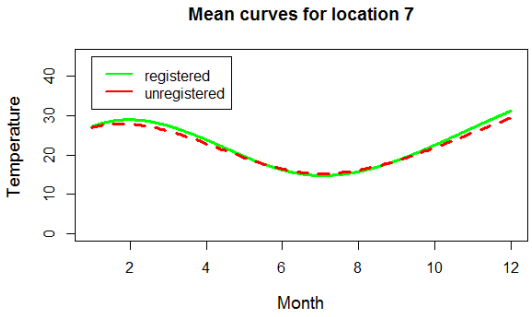

Mean (+-1) standard dev curves of location 6
Temperature (C)
Month

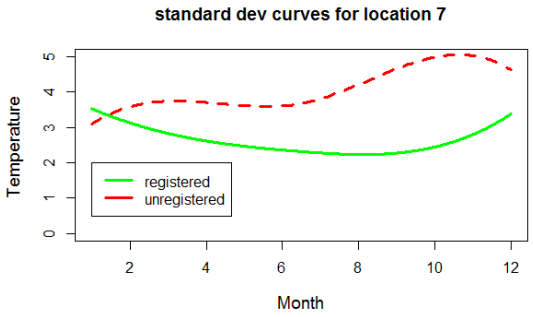

Mean curves for location 6
registered
unregistered
Temperature
Month

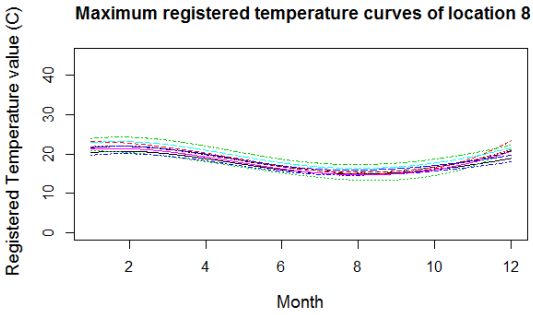

standard dev curves for location 6
registered
unregistered
Temperature
Month

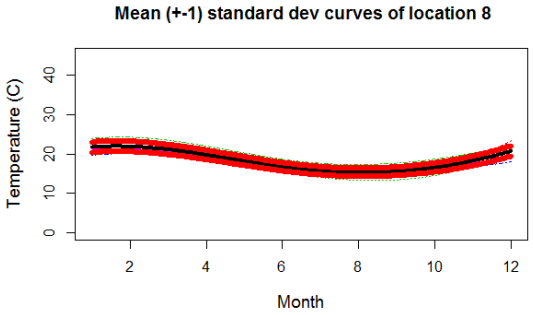

Maximum registered temperature curves of location 7
Registered Temperature value (C)
Month

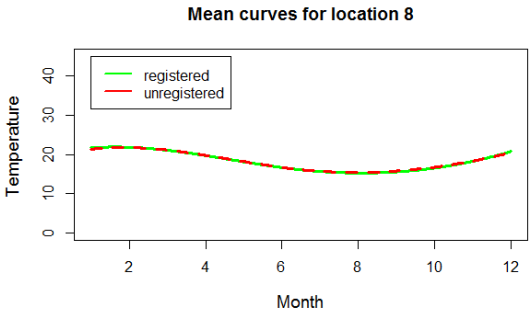

Mean (+-1) standard dev curves of location 7
Temperature (C)
Month

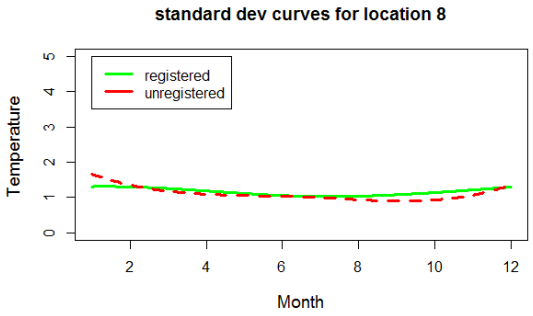

Mean curves for location 7
registered
unregistered
Temperature
Month

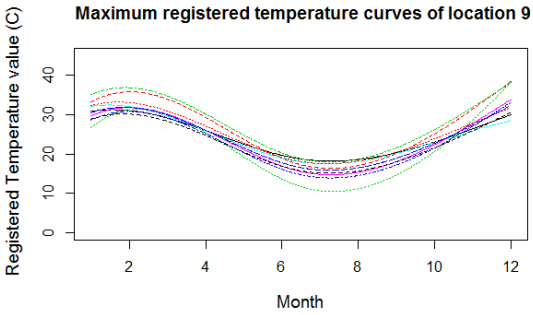

standard dev curves for location 7
registered
unregistered
Temperature
Month

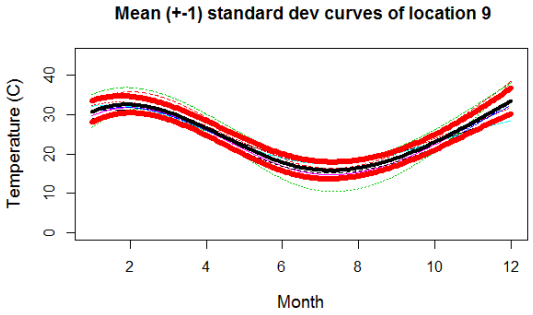

Maximum registered temperature curves of location 8
Registered Temperature value (C)
Month

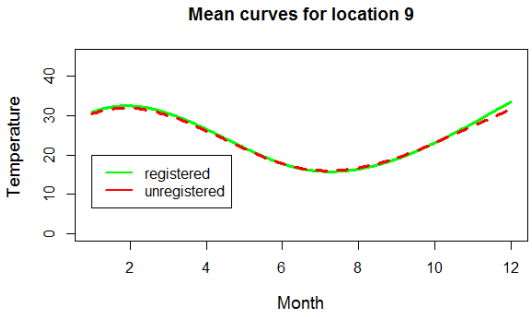

Mean (+-1) standard dev curves of location 8
Temperature (C)
Month

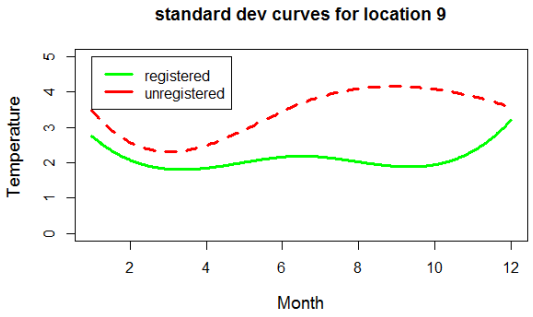

Mean curves for location 8
registered
unregistered
Temperature
Month

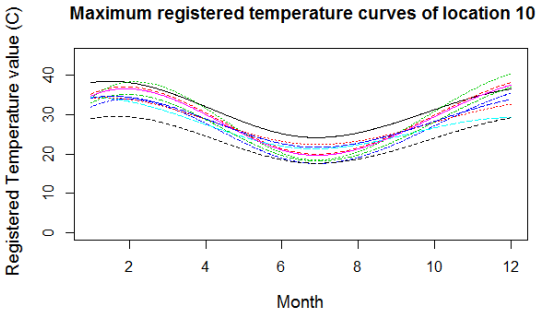

standard dev curves for location 8
registered
unregistered
Temperature
Month

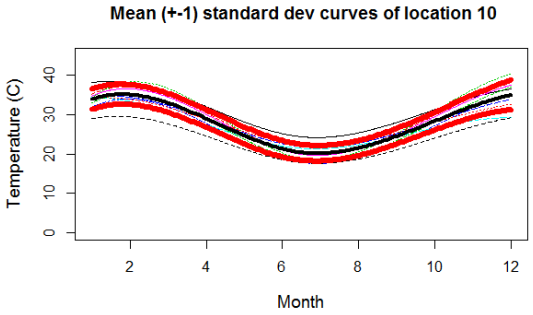

Maximum registered temperature curves of location 9
Registered Temperature value (C)
Month

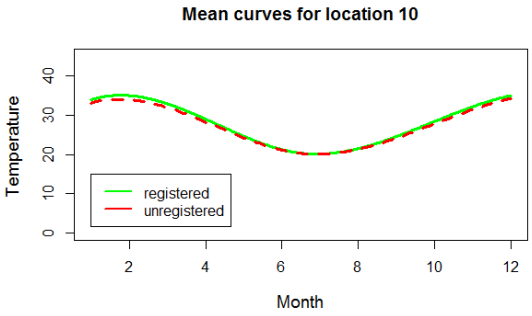

Mean (+-1) standard dev curves of location 9
Temperature (C)
Month

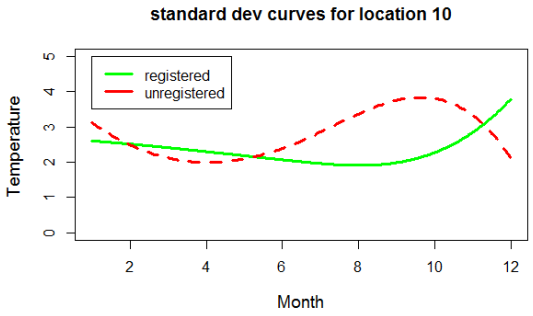

Mean curves for location 9
Temperature
Month
registered
unregistered

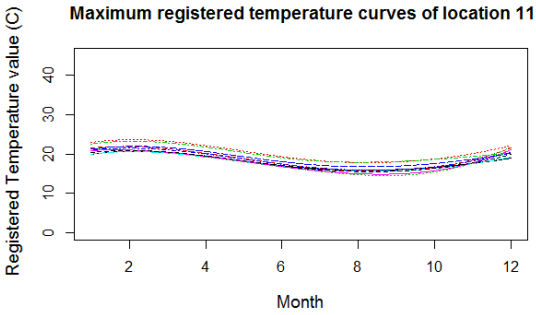

standard dev curves for location 9
Temperature
Month
registered
unregistered

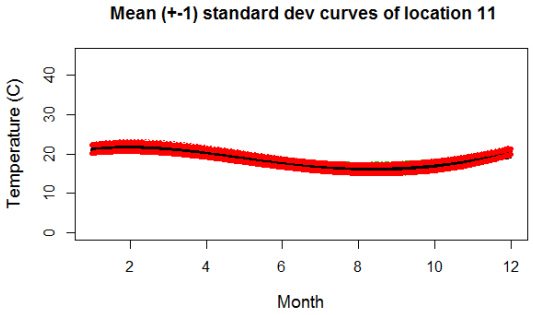

Maximum registered temperature curves of location 10
Registered Temperature value (C)
Month

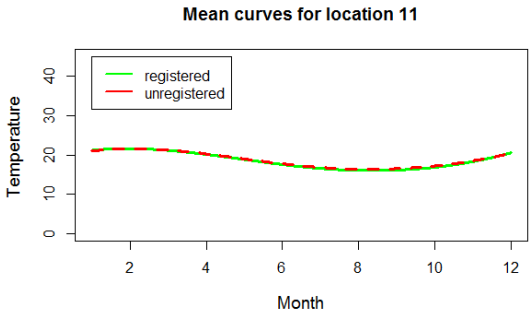

Mean (+-1) standard dev curves of location 10
Temperature (C)
Month

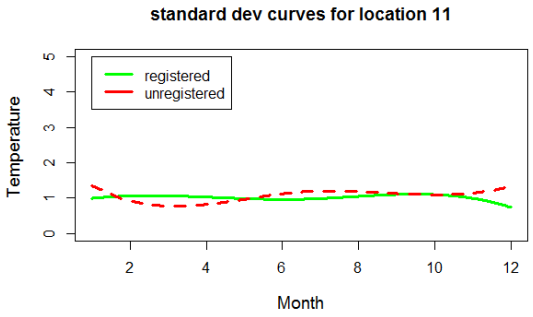

Mean curves for location 10
Temperature
Month
registered
unregistered

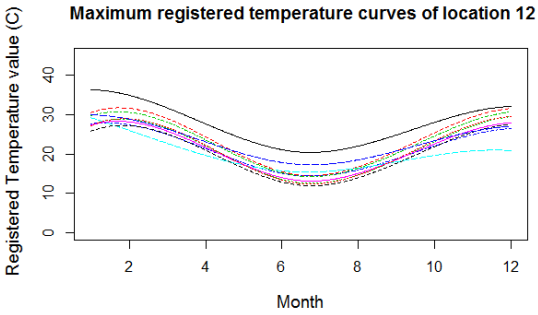

standard dev curves for location 10
Temperature
Month
registered
unregistered

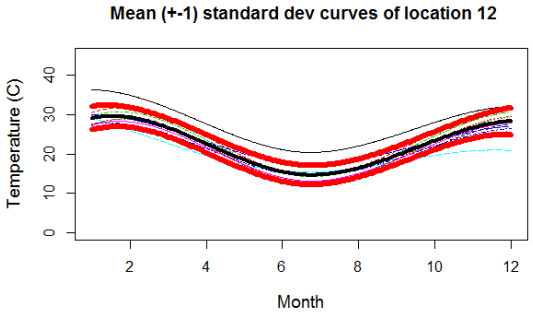

Maximum registered temperature curves of location 11
Registered Temperature value (C)
Month

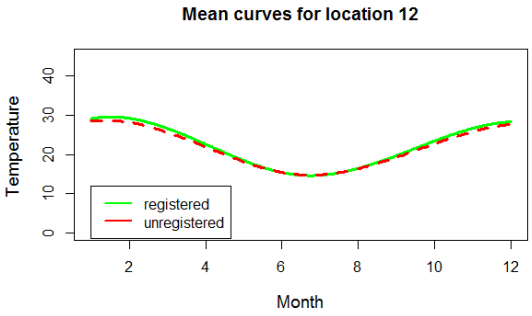

Mean (+-1) standard dev curves of location 11
Temperature (C)
Month

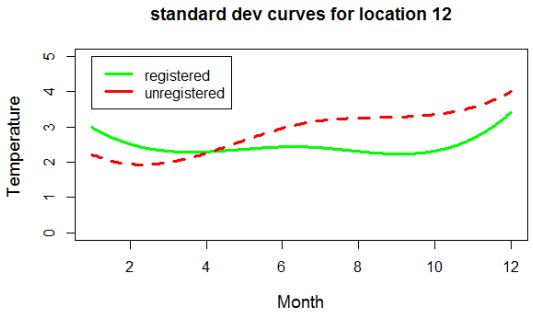

Mean curves for location 11
registered
unregistered
Temperature
Month

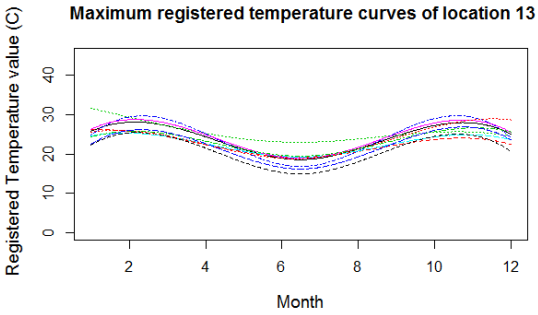

standard dev curves for location 11
registered
unregistered
Temperature
Month

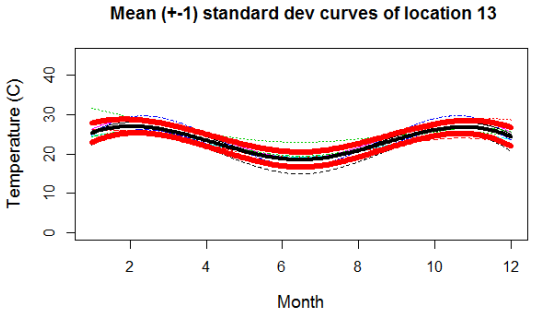

Maximum registered temperature curves of location 12
Registered Temperature value (C)
Month

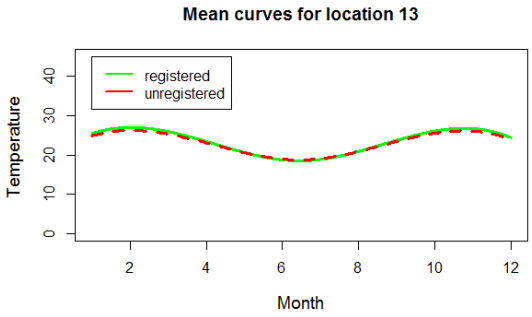

Mean (+-1) standard dev curves of location 12
Temperature (C)
Month

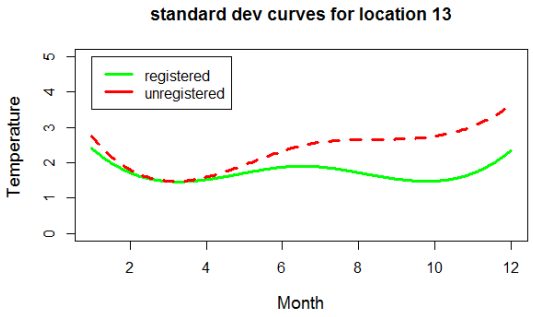

Mean curves for location 12
registered
unregistered
Temperature
Month

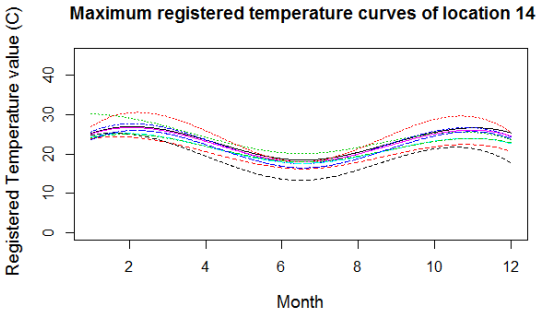

standard dev curves for location 12
registered
unregistered
Temperature
Month

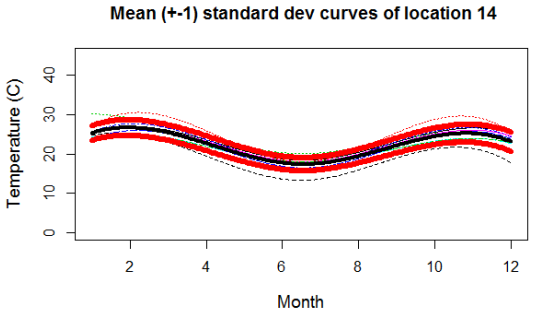

Maximum registered temperature curves of location 13
Registered Temperature value (C)
Month

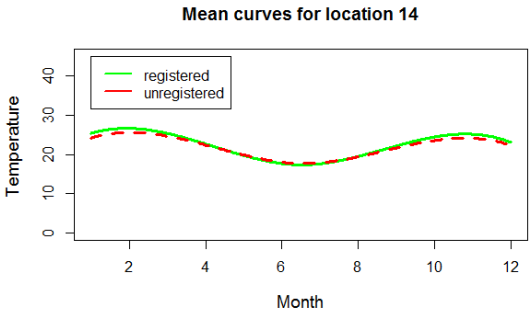

Mean (+-1) standard dev curves of location 13
Temperature (C)
Month

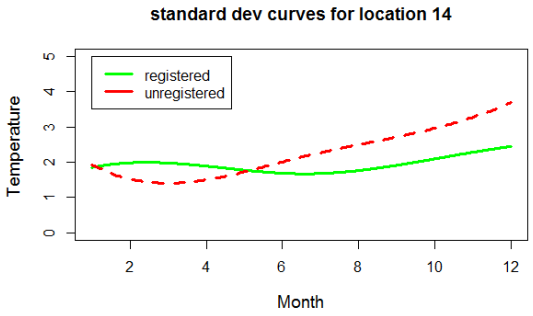

Mean curves for location 13
registered
unregistered
Temperature
Month

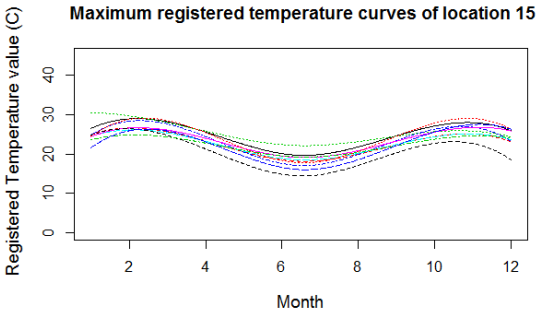

standard dev curves for location 13
registered
unregistered
Temperature
Month

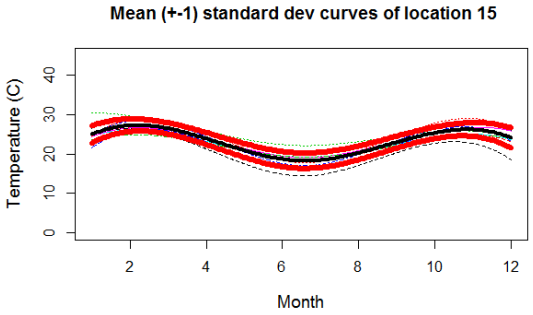

Maximum registered temperature curves of location 14
Registered Temperature value (C)
Month

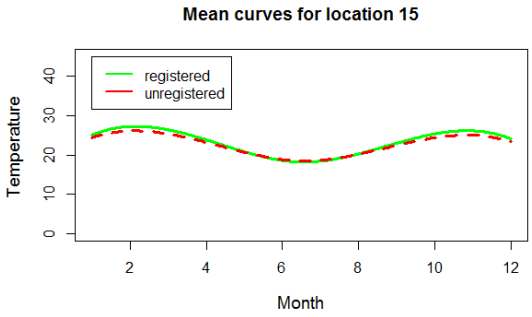

Mean (+-1) standard dev curves of location 14
Temperature (C)
Month

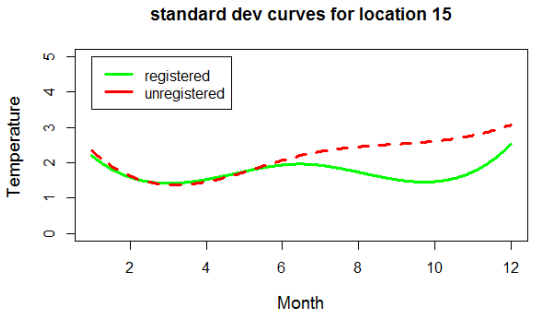

Mean curves for location 14
registered
unregistered
Temperature
Month

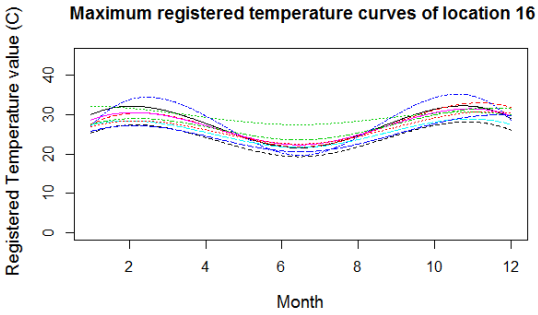

standard dev curves for location 14
registered
unregistered
Temperature
Month

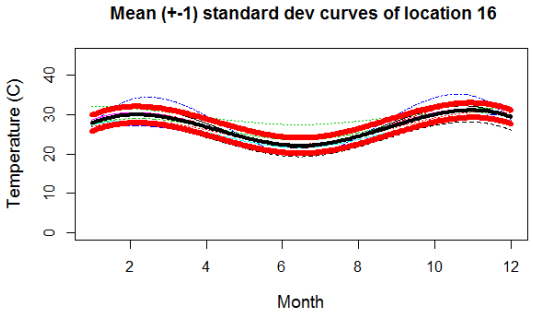

Maximum registered temperature curves of location 15
Registered Temperature value (C)
Month

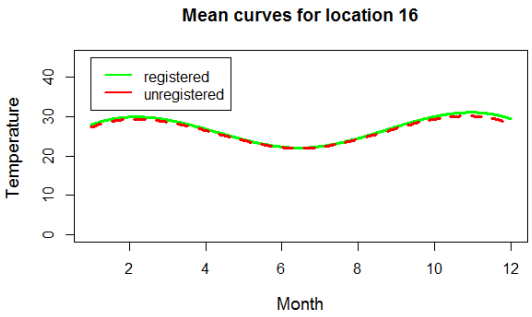

Mean (+-1) standard dev curves of location 15
Temperature (C)
Month

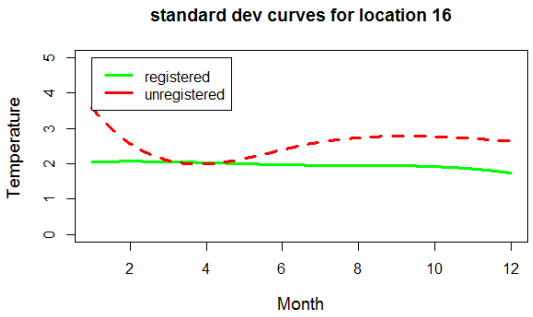

Mean curves for location 15
registered
unregistered
Temperature
Month


standard dev curves for location 15
registered
unregistered
Temperature
Month


Maximum registered temperature curves of location 16
Registered Temperature value (C)
Month


Mean (+-1) standard dev curves of location 16
Temperature (C)
Month


Mean curves for location 16
registered
unregistered
Temperature
Month


standard dev curves for location 16
registered
unregistered
Temperature
Month

6.3 functional Principal Component Analysis (fPCA) of Monthly Maximum Temperature

We performed fPCA in the functional data analysis (*fda* package) in R using the registered monthly maximum temperature data of all 16 locations and consider the first two harmonics functions or functional principal component to explain variation in the data. We observe that the first two fpc's explain 98.7 percent of the variation. We rotate the functional principal components using the VARIMAX rotation algorithm to disclose more relevant components of variation. From the plots in Figure 6.4 we observe that both PCA function 1 and PCA function 2 have valley in June to July which indicate lower temperatures at those months and they both have peaks in the month of February to March and again in November and December. Therefore PCA function 1 explain summer variation while PCA function 2 explain winter variation.

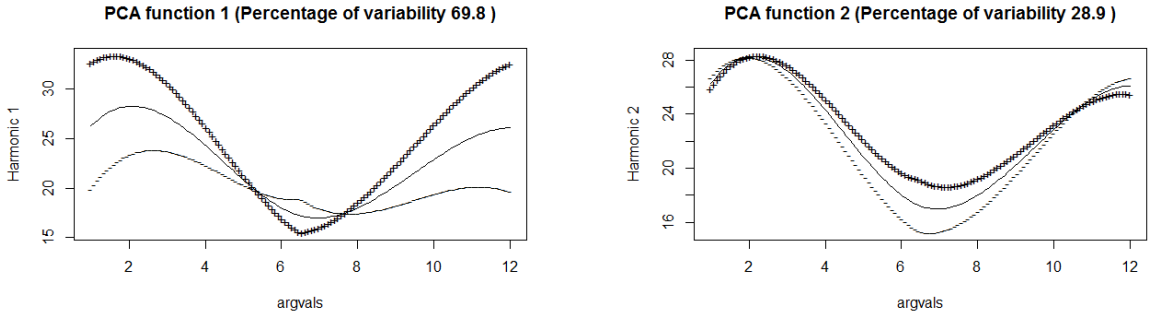


Figure 6.4: The 2 principal components and the percentage they capture.

The solid line in each curve represents the mean of temperature curve, while the curves labelled with a'+' or a '-' indicate what happens when one standard deviation of the harmonic or PC is added or subtracted from the mean.

6.4 Phase-plane Plots of Monthly Maximum Temperature

We obtain phase-plane plots of monthly maximum temperature data of each year of all the locations using the registered data in section 6.2. The number 1 to 12 on the graphs represent the months January to December. From our plots we observe that in 1980 and 1985 we have similar weather patterns, and also in the year 1990 and 2005. In 1965 positive potential energy or acceleration is the greatest around August to December, this shows that the temperature is steadily increasing. While in March there is zero acceleration thus temperature is decreasing. In 1975 from January to May the temperature is decreasing with zero velocity in June this shows that June had lower temperature and also we observe that the temperature

is increasing from July to December with high acceleration from september to December. In 1980 and 1985 the phase-plane plots suggests that we had similar weather patterns, steadily increasing from January to April and decreased in May, November and December with larger kinetic energy. In both the year 2000 and 2005 there is a larger potential energy in January and November this suggests that there is higher temperature. In 1995 and 2010 there is a larger kinetic energy in May and September and a larger potential energy in December 2010 with acceleration near zero. This shows that there was a higher temperature in December 2010, February 1995, with lower temperature in May and September.

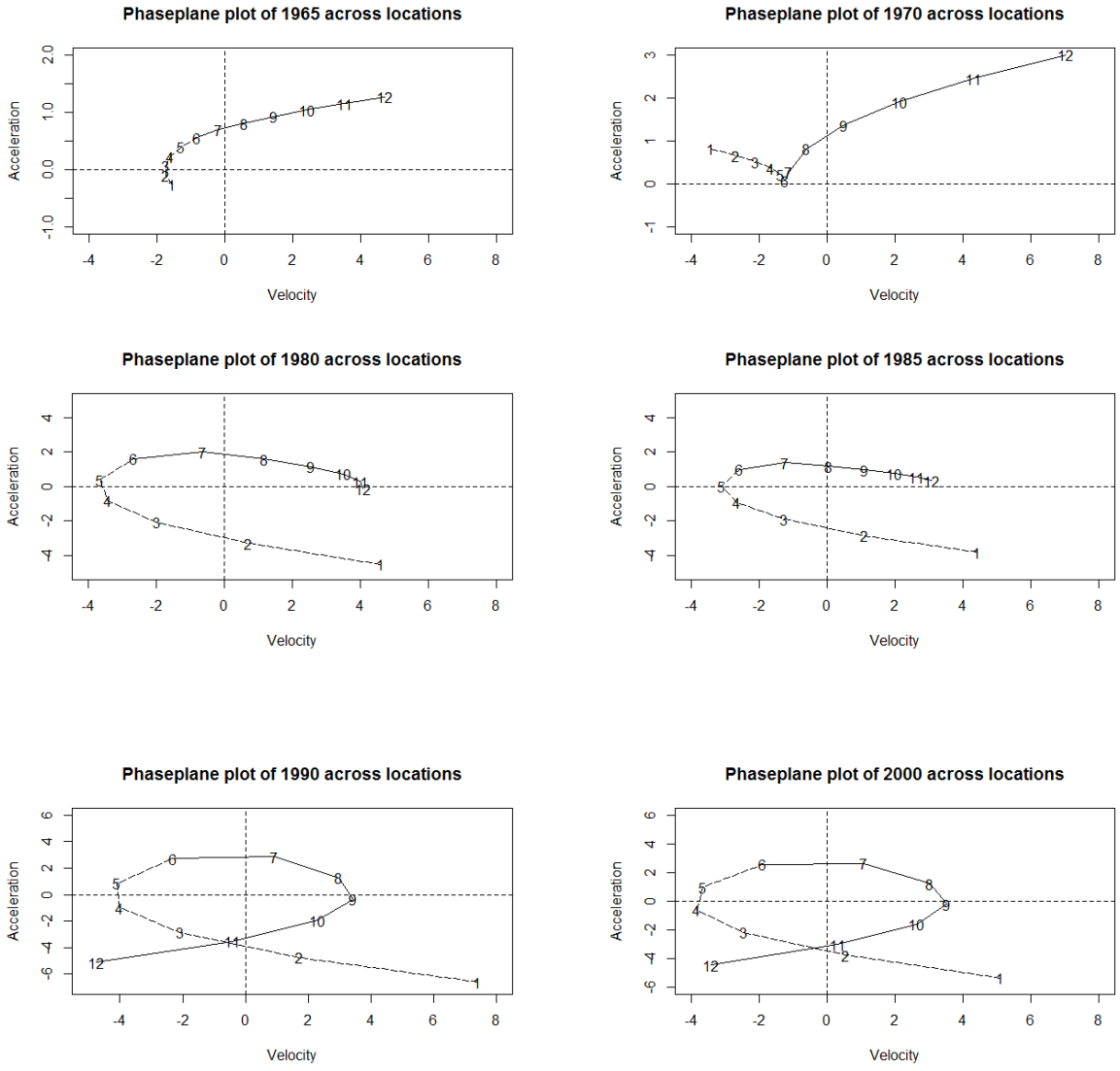


Figure 6.5: Phase-plane plots of the year 1965, 1970, 1980, 1985, 1990 and 2000.

In 2005 we observe that there is larger kinetic energy from March to June, this suggests that there were lower temperature values and we also observe that there is a larger potential energy in July, August, October, November and December. It suggests that there were higher temperature values. In 1975 from April to July phase plane plots suggests that there were lower temperature values (larger kinetic energy) and in August, September, November and December there were higher temperature values (larger potential energy). On the phase-plane of 1995 temperature is steadily increasing from January to December and in 2010 it shows that there is positive velocity or kinetic energy with zero acceleration or potential energy, this suggests that the temperature is increasing at that point and decreasing with higher acceleration and velocity at zero in 1995 but near zero in 2010.

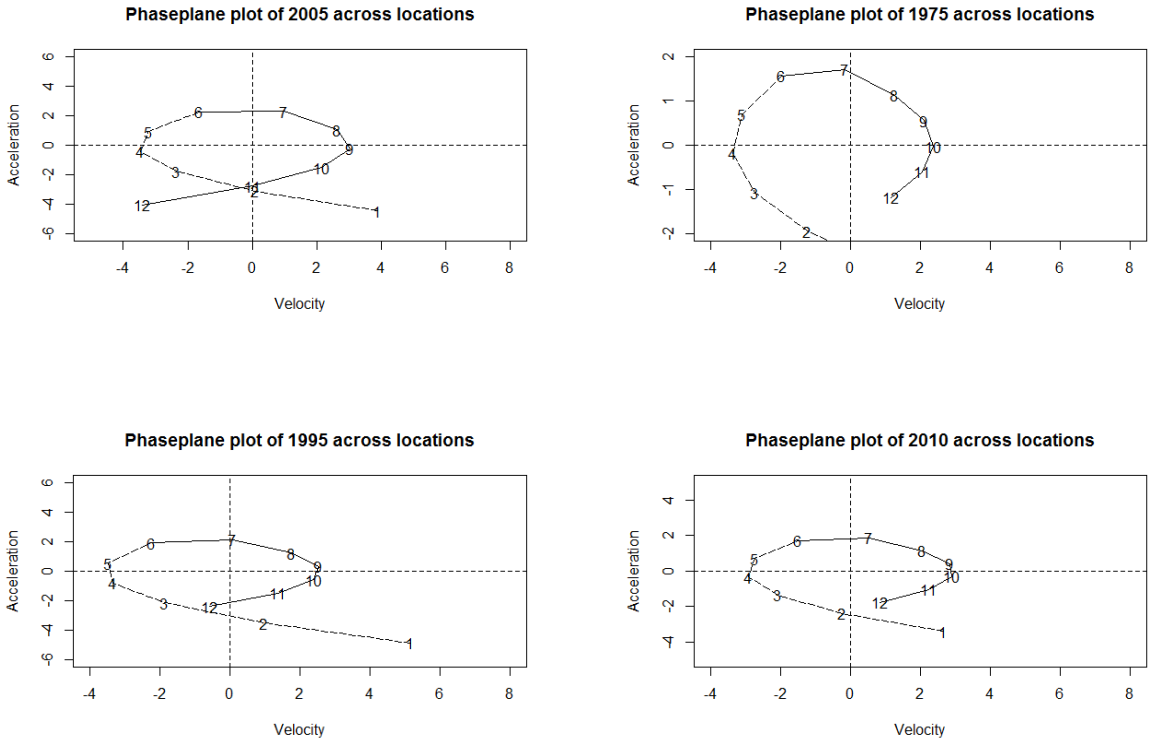


Figure 6.6: Phase-plane plots of the year 1975, 1995, 2005 and 2010.

6.5 Functional Clustering of Monthly Maximum Temperature

We use the registered monthly maximum temperature curves registered in Section 6.2. We group the curves in a way that curves that are similar to each other are grouped in one cluster. We chose the number $K = 3$ (number of clusters) using *K-means* and *Hierarchical* algorithm to cluster the data. In the Figure 6.7 below we present monthly maximum temperature of 16

locations. This Figure shows 160 temperature curves colour coded by cluster and each curve is identified by both location and year. In cluster 1 there are two locations which result in 20 curves, there are five locations in cluster 2 which result in 50 curves and there are nine locations in cluster 3 which result in 90 curves across 1965 to 2010 with 5 years of intervals. From the Figure 6.7 on the left panel we can see that the clustered curves have a common shape which shows that from the month of February to April and October to December we have higher temperatures and lower from May to September and the lowest from June to August which correspond to the winter season.

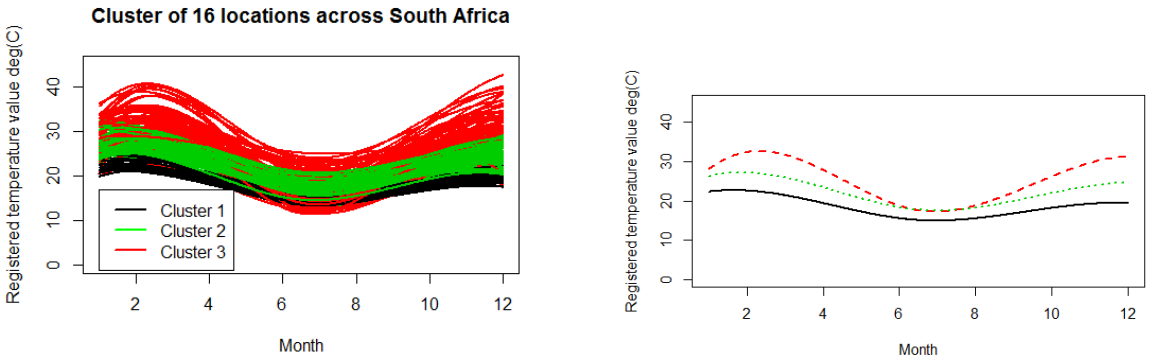


Figure 6.7: Left panel is the registered monthly maximum temperature curves of 16 locations by cluster which consist of 160 curves, 20 curves in cluster 1, 50 curves in cluster 2, and 90 curves in cluster 3. Where each is identified by both location and year and the right panel is mean curve of each cluster.

Looking at the mean curve of each cluster in Figure 6.7 we can clearly see that the highest mean curve is in cluster 3 which indicate that cluster 3 has higher temperature, followed by cluster 2 then cluster 1 which is the lowest among where coastal locations (Cape Town and Alexander Bay) are dominant in the cluster. We clustered the monthly maximum temperature curves across 16 locations for year 1965 to 2010 with 5 years of intervals. There are two locations in cluster 1 (Cape Town and Aleaxander Bay) which shows that both Cape Town and Alexander Bay had similar temperature values/patterns in 1965, nine locations in cluster 2 (Polokwane, Dullstroom, Durban, Ladysmith, Mthata, Beaufort West, Klerksdorp, Johannesburg, Pretoria), and five locations in cluster 3 (George, Calvinia, Upington, De Aar and Thabazimbi). In the year 1970 we see that Cape Town and Alexander Bay had similar temperature patterns/values with Dullstroom, Durban, Ladysmith and Mthatha, and in cluster 2 (Polokwane, Klerksdorp, Johannesburg, Pretoria, Thabazimbi) had similar temperature values. We also observe that the mean curves of each cluster for year 1965 have the same shape with the lowest temperature curves in the coastal areas Alexander Bay and Cape Town. This shows that George, Calvinia, Upington, De Aar and Thabazimbi had warmer temperatures amongst the selected locations. In cluster Figure of 1975 there are three locations in cluster 1 (Calvinia, Upington,and De Aar) which are all located in the Northen Cape Province. Looking at the cluster 1 mean curve in year 1975 it shows that Calvinia,

Upington, and De Aar are much warmer in summer and also in winter. In 1980 Polokwane and Thabazimbi was clustered in cluster 3 with coastal areas such as George, Beaufort West, Calvinia, and Upington which indicate that they had similiar weather patterns. While in 1985 Beaufort West, Ladysmith, Mthatha, and Durban were clustered along together with inland areas, Polokwane, Johannesburg, Pretoria, Dullstroom. Both in 1980 and 1985 Alexander Bay and Cape Town were together in one cluster. We also observe that both the cluster graphs of 1980 and 1985 have the similar shapes therefore one can conclude that in those years we had similar temperature patterns across the locations. In the year 1990 and 1995 there are three locations (Calvinia, Upington,and, De Aar) in cluster 1, if we observe the mean curves of the clusters. It suggets that Calvinia, Upington, and De Aar are much warmer than the rest of locations in these years. In the year 2000 Ladysmith and De Aar were clustered in cluster 2, the mean curve of the cluster suggests that these two locations are much colder than other locations in the winter months (June to August). In 2005 there are two locations in cluster 2 (Alexander Bay and Cape Town) and in 2010 there are four locations (Alexander Bay, Cape Town, Ladysmith and Durban). Therefore if we observe mean of cluster 2 in both years it suggests that Alexander Bay, Cape Town, Ladysmith and Durban are more warmer than other locations.

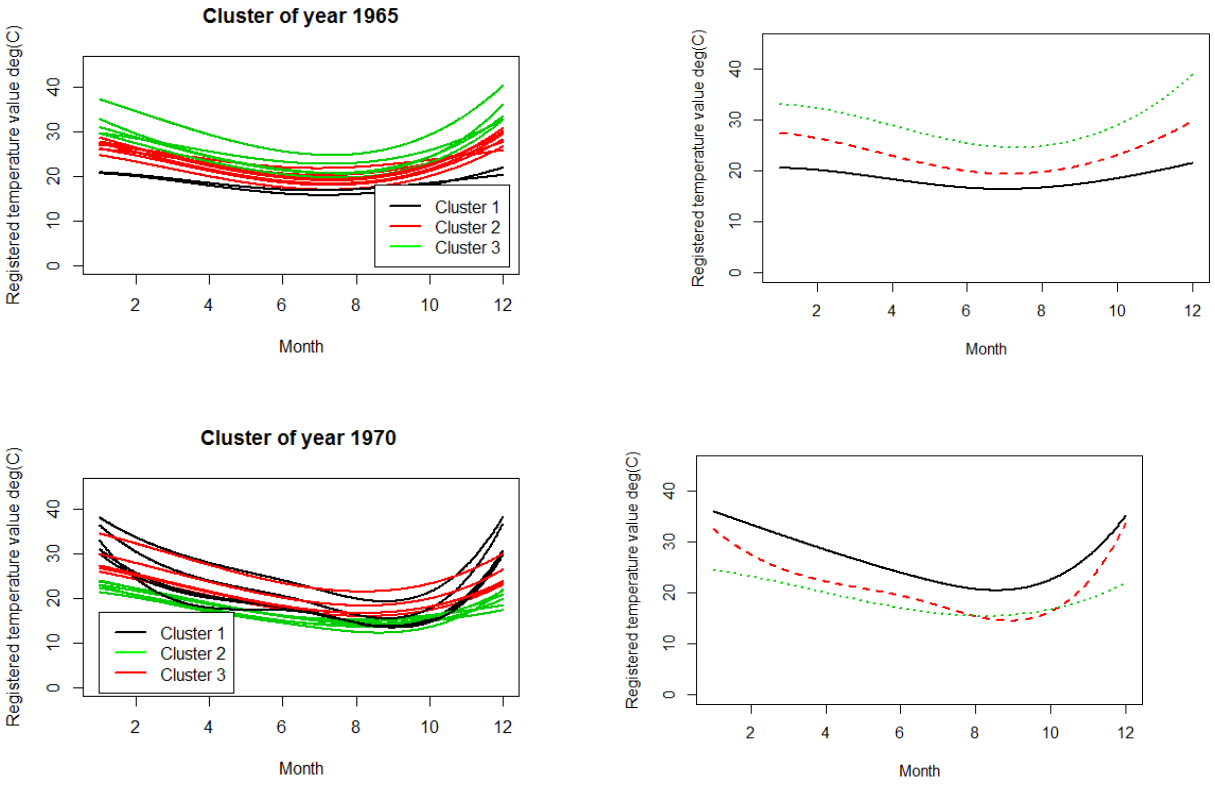


Figure 6.8: *K-means* clustering of monthly maximum temperature curves of all 16 locations in year 1965 and 1970, each year analysed consists of 16 curves identified by location.

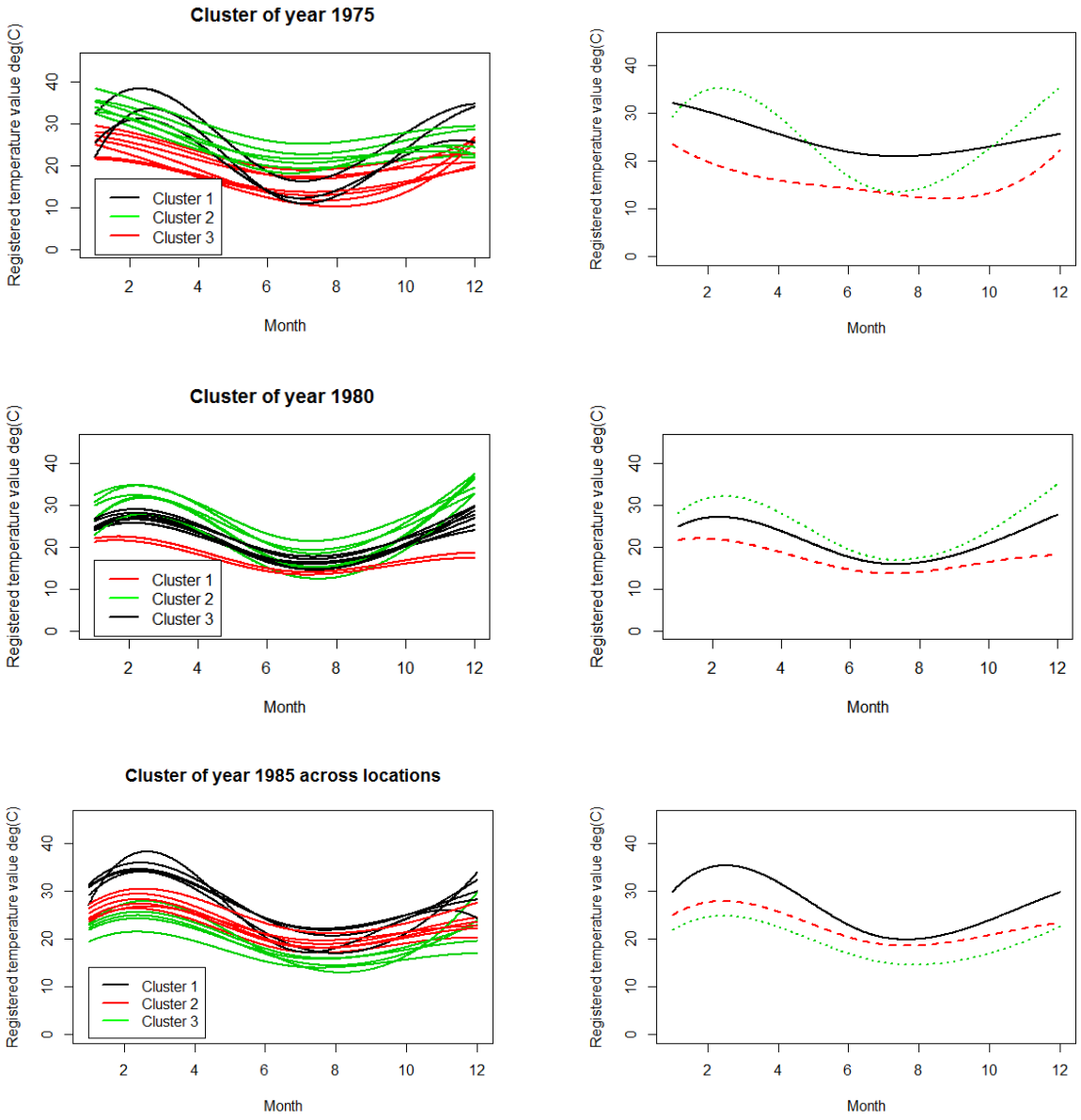


Figure 6.9: *K-means* clustering of monthly maximum temperature curves of all 16 locations in year 1975, 1980, and 1985, each year analysed consists of 16 curves identified by location.

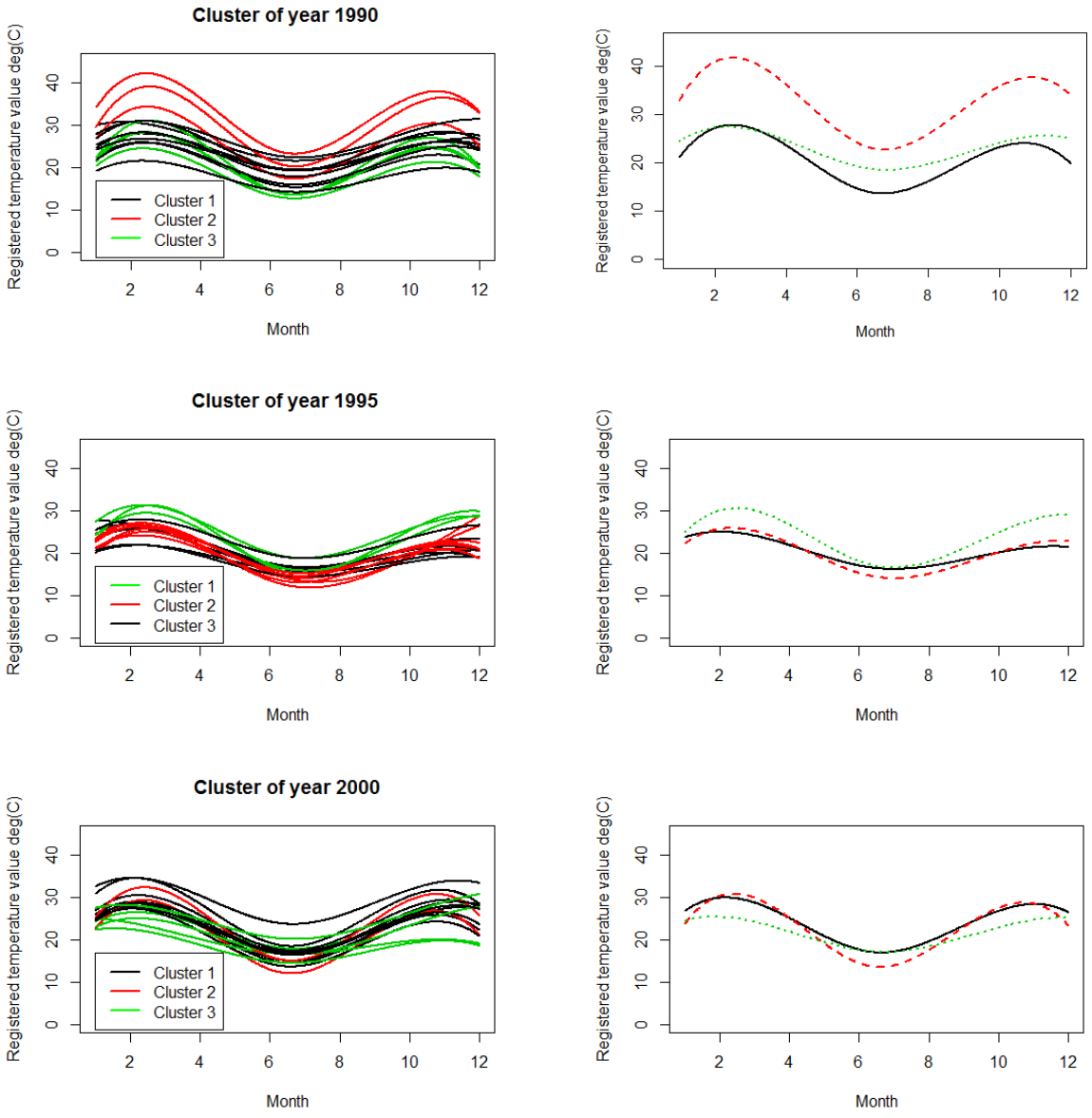


Figure 6.10: *K-means* clustering of monthly maximum temperature curves of all 16 locations in year 1990, 1995, and 2000, each year analysed consists of 16 curves identified by location.

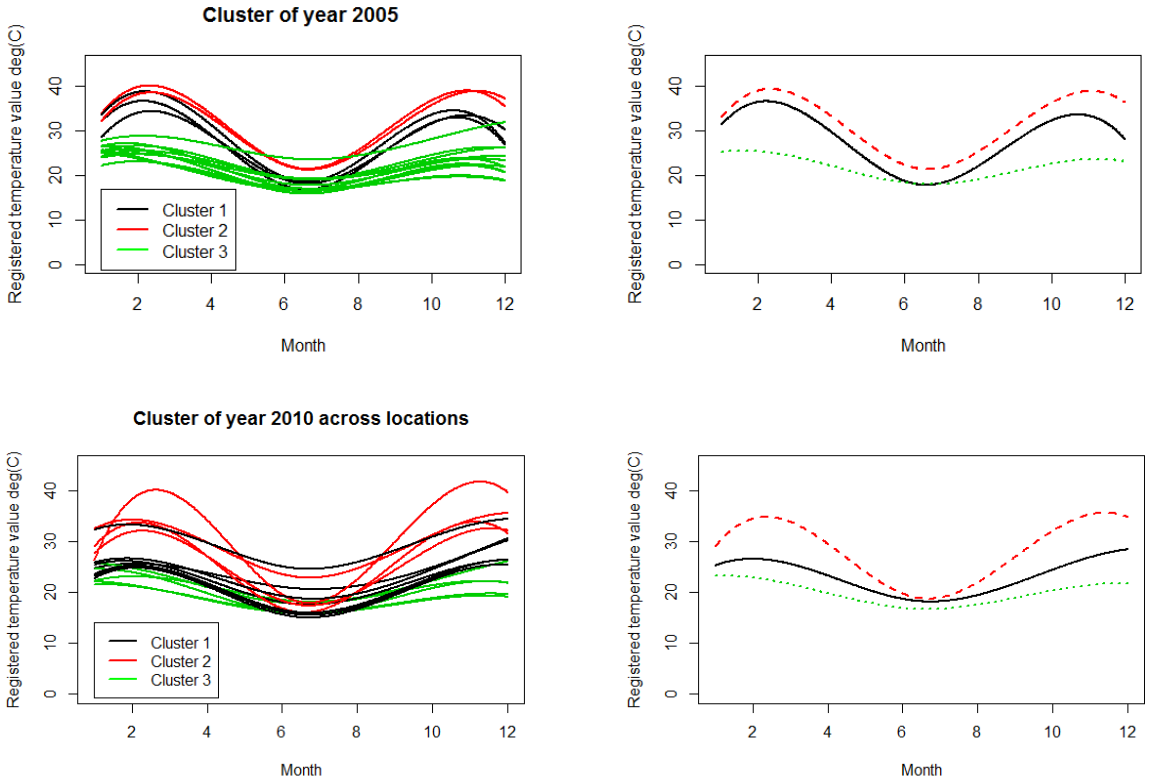


Figure 6.11: *K-means* clustering of monthly maximum temperature curves of all 16 locations in year 2005 and 2010, each year analysed consists of 16 curves identified by location.

We performed *hierarchical* functional cluster analysis on monthly maximum temperature curves of all 16 location in each year analysed. We used complete linkage of agglomerative method. From the dendrogram of year 2000 in Figure 6.13 (middle right panel), there are two clusters one that contain Upington and Thabazimbi and the other contain the remaining locations. Upington and Thabazimbi (coastal and inland area) are seperated from the other locations. This suggest that Upington and Thabazimbi had similar temperature values in 2000 and that there is a big difference between Upington, Thabazimbi and other locations. Looking at the lowest threshold on each dendrogram in year analysed, we can observe that most of the clusters contains the locations that are in the same region.

We also observe that in all years analysed Ladysmith, Mthatha, and Durban are always grouped together in a cluster. This locations fall under Kwa-zulu-natal Province in the coastal area. Alexander Bay and Cape Town are always grouped in one cluster (coastal area). Just as Johannesburg and Pretoria, they are always grouped in one cluster with mostly Klerksdorp or either Dullstroom or Polokwane (which are the inland areas). We also observe that from the dendrograms in year 1965, Klerksdorp and Pretoria have most similar temperature values since the height of the link that join them together is the smallest

amongst others. In 1975, 1985, 1990, 2005, and 2010 Johannesburg and Pretoria have the most similar temperature values.

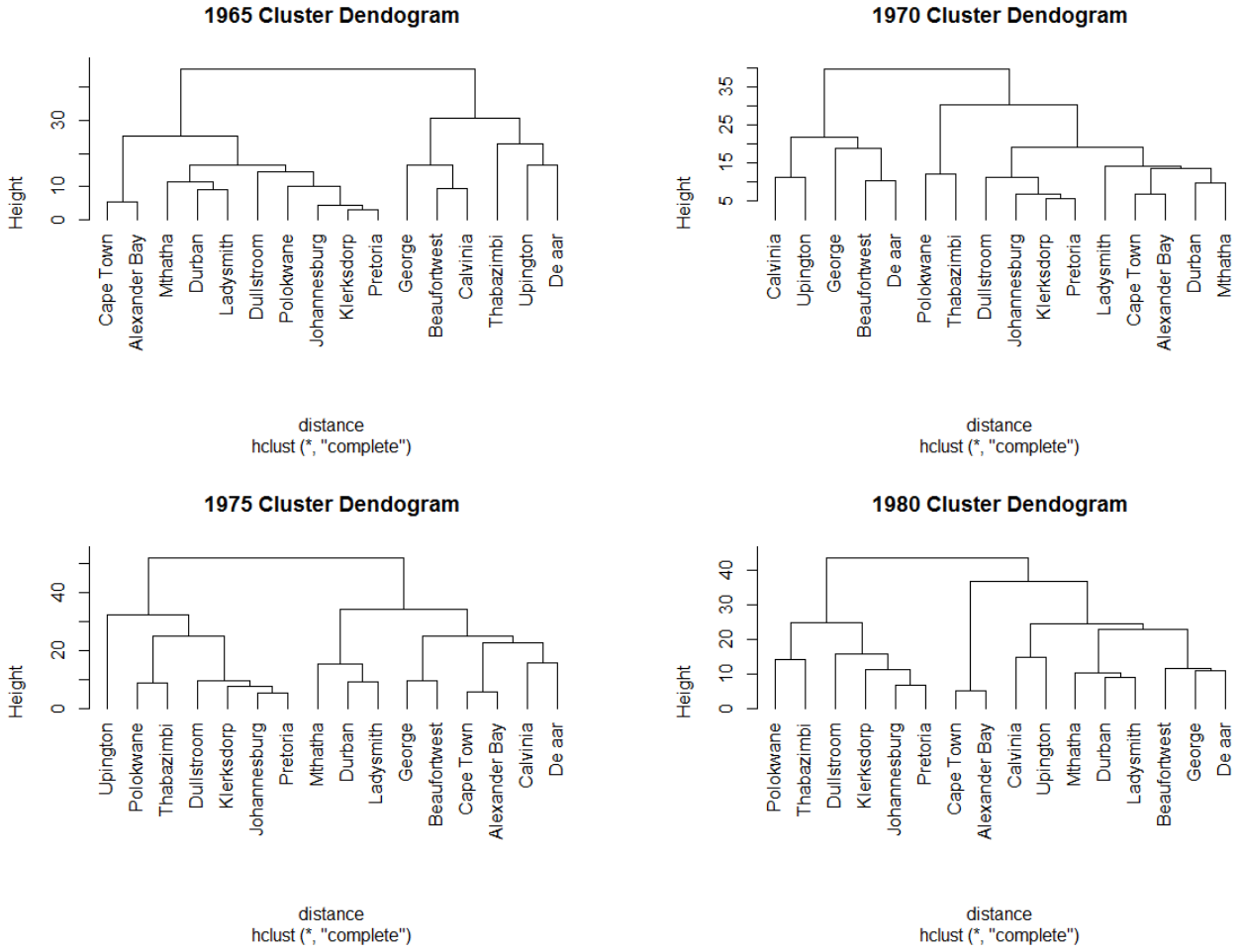


Figure 6.12: *Hierarchical* clustering of maximum monthly curves of all 16 locations spread out across South Africa from the period of 1965 to 1980 with 5 years of intervals.

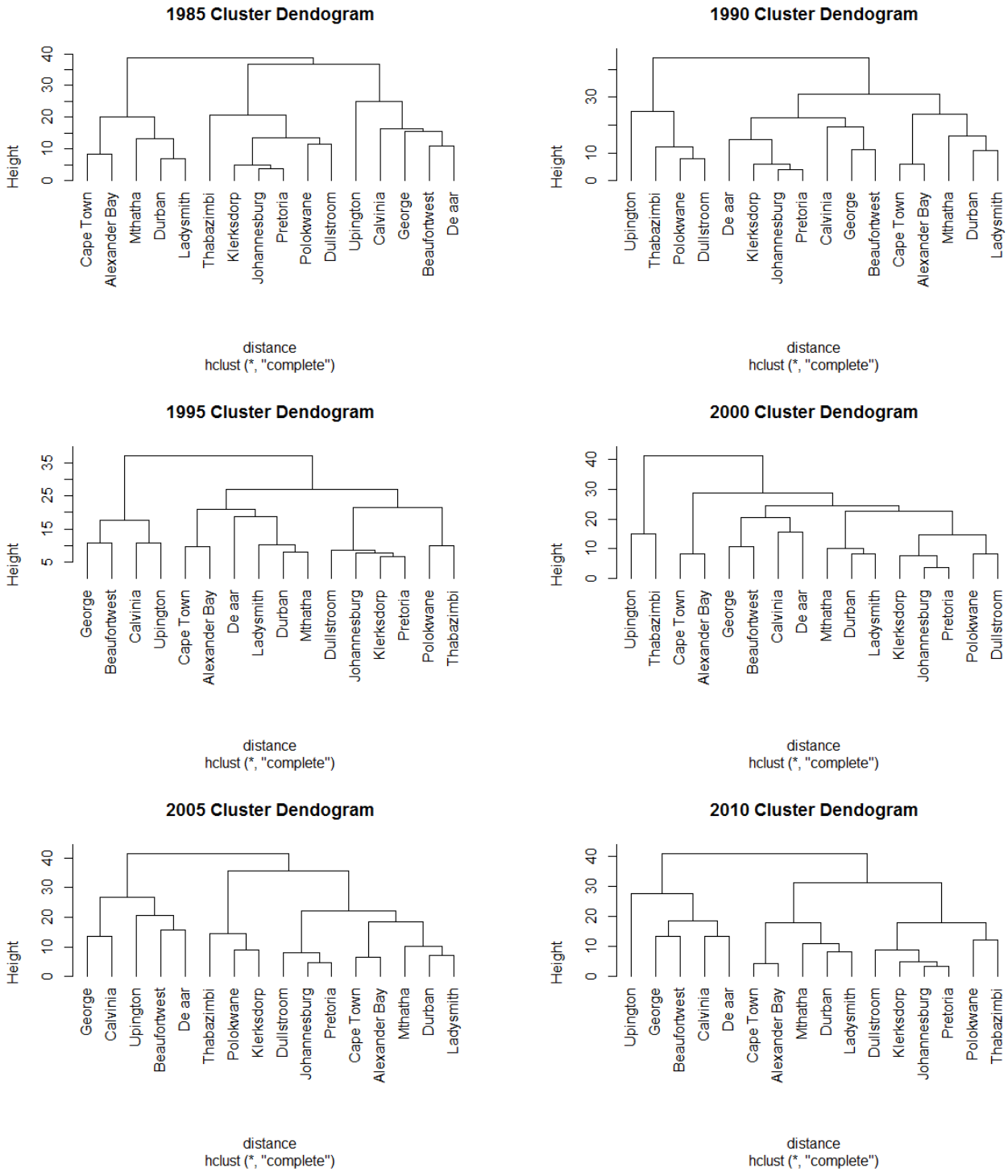


Figure 6.13: *Hierarchical* clustering of maximum monthly curves of all 16 locations spread out across South Africa from the period of 1985 to 2000 with 5 years of intervals.

Table 6.1: Proportion by cluster for all locations

Cluster no:	Proportion (%)
1	0.1667937
2	0.3558855
3	0.4773208

.

Table 6.2: Year proportion by cluster.

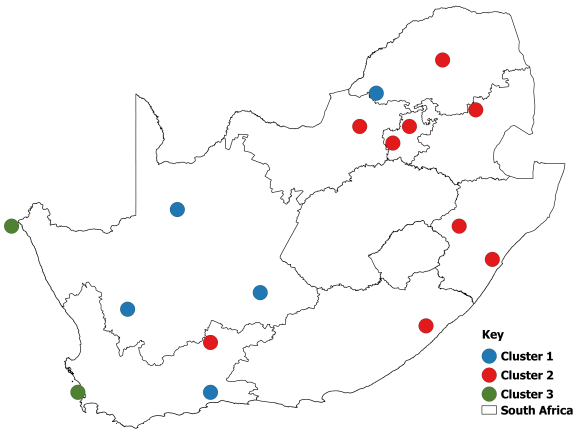
Year	Cluster 1	Cluster 2	Cluster 3
1965	0.1250000	0.3217785	0.5532215
1970	0.3125000	0.3127224	0.3747776
1975	0.1875000	0.4208527	0.3916473
1980	0.1250000	0.4983823	0.3766177
1985	0.2654057	0.4242246	0.3103697
1990	0.3753079	0.2459260	0.3787661
1995	0.1883523	0.3638433	0.4478044
2000	0.5344905	0.1250000	0.3405095
2005	0.1875076	0.1249924	0.6875000
2010	0.2500726	0.2500567	0.4998707

6.5.1 Visualisation of Functional Clusters of Monthly Maximum Temperature

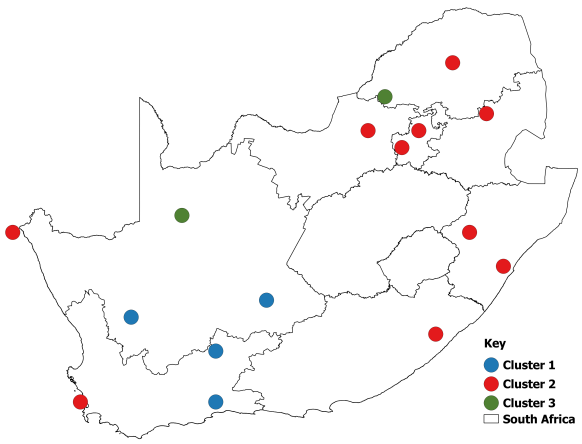
In this section we visualise our cluster results on map in each year of 1965 to 2010 with 5 years of intervals in Figure 6.14, and 6.15 below. In Figure 6.14 (a) we observe that the three locations in Northen Cape are grouped in one cluster and (b) we observe that the locations close to Indian ocean are grouped together with the inland locations expect Thabazimbi. In (c) we observe that all locations close to Indian and Atlantic ocean are grouped in one cluster together with Dullstroom and in Figure 6.15 (b) we observe that the inland locations were grouped together in one cluster, coastal locations in the other, and the Northern cape province locations which also falls in the coastal area in the other one.

Therefore from the visualisation of the map we observe that Cape Town and Alexander Bay are always grouped together in one cluster, and also Johannesburg and Pretoria. This suggests that in all the years analysed Cape Town and Alexander Bay, Pretoria and Johannesburg always had similar maximum temperature patterns and we also observe that in 1990 inland locations were grouped together in one cluster and while the coastal locations were grouped in the other two clusters. This suggests that in the year of 1990 the inland locations

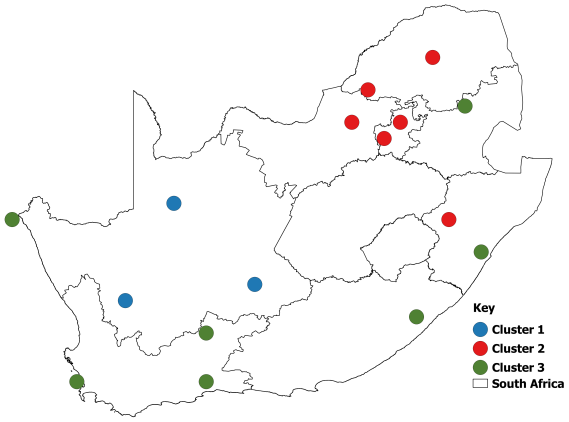
had similar temperature patterns different to of the coastal locations while in other years both some of inland and coastal locations had similar temperature patterns.



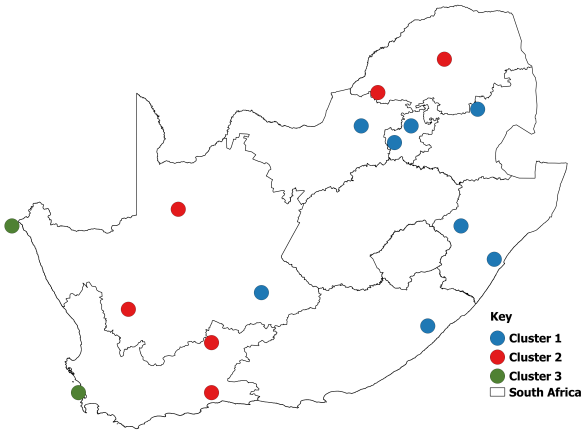
(a) 1965



(b) 1970

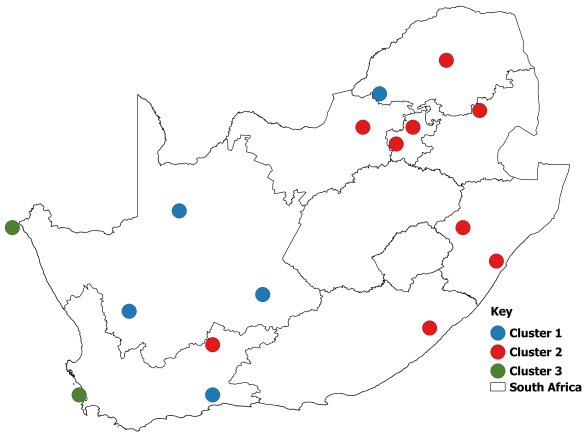


(c) 1975

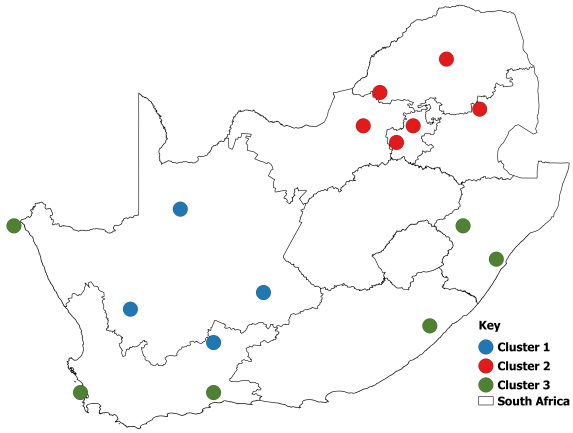


(d) 1980

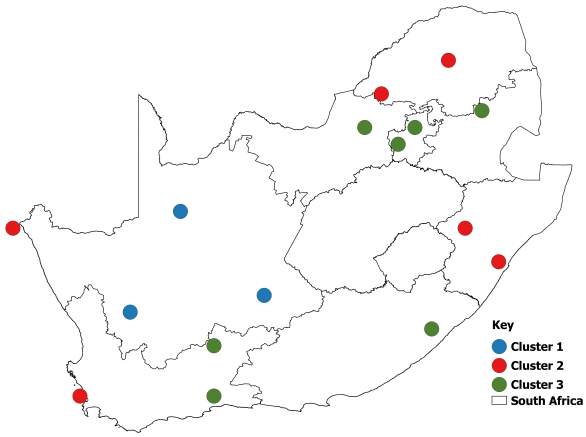
Figure 6.14: visualisation of cluster result on a map from 1965 to 1980 with 5 years of intervals.



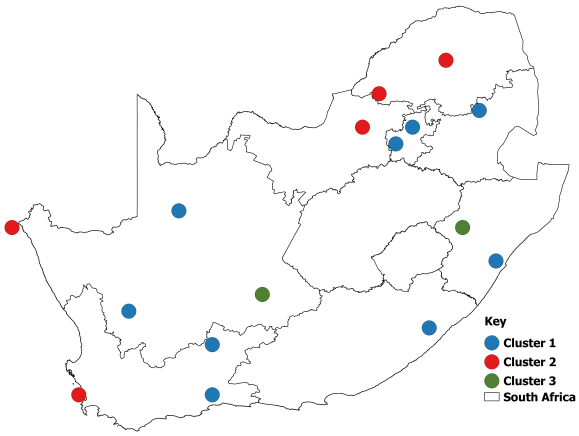
(a) 1985



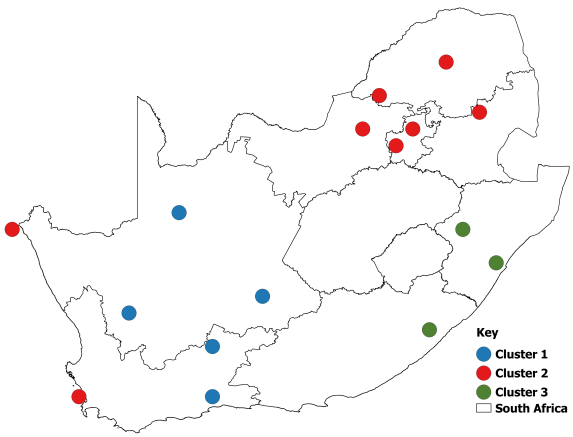
(b) 1990



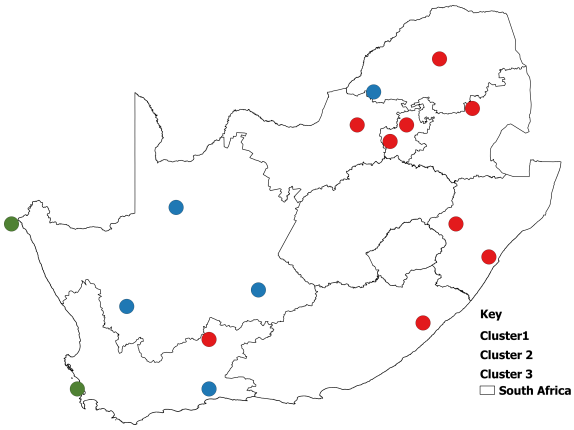
(c) 1995



(d) 2000



(e) 2005



(f) 2010

Figure 6.15: visualisation of cluster result on a map from 1985 to 2010 with 5 years of intervals.

Chapter 7

Software

All of our applications and compilation of this thesis were done under R software using package Sweave (R and Latex) environment and the application of visualising cluster results was performed in QGIS (quantum geographic information system).
Sweave is a package under R that is used to write text and code in the same environment, it works together with Latex [Leisch, 2002] and QGIS is framework for viewing, editing, and analysis of geospatial data.

- We smoothed raw monthly maximum temperature data to functional curves and registered the curves using "*fda*" package
- We applied the methods of FDA: Phase-plane plots and fPCA on registered monthly maximum temperature curves using both *fda* and *fda.usc* package, the reader may refer to Febrero-Bande et al. [2012] on more about the package
- We applied functional clustering algorithms on our registered monthly maximum temperature curves was performed under funFEM package in R. funFEM is an algorithm used to cluster funfional data [Bouveyron, 2015], you can perform either functions k-means or hclust in the package together with fda package.
- The application of visualisation of functional clustering on a map was performed under QGIS.

Chapter 8

Future Scope of the Thesis

Within the scope of this thesis our focus was to understand monthly maximum temperature patterns from 16 locations and to that end we subjected to three methodologies, registration, fPCA, and functional clustering. An extension to the thesis is application of classification for functional registered temperature curves. Below we explain in detail traditional classification and functional classification.

Classification is a supervised learning process that uses training data to train a model using a classification algorithm. There are different types of classification algorithms and the reader is refered to [Leung, 2007] for an indepth insight. we will only focus on the use of *Naïve Bayes* classifier because of its various advantages particularly when applying to large data sets [Leung, 2007].

Naïve Bayes classifier algorithm is easy to understand and implement. It is based on Baye's theorem which comprises of conditional probabilities. Let A and B denote two events, and $P(A \mid B)$ denote the conditional probability of event A occuring given that event B has occured, and similalry one can define $P(B \mid A)$. In general these two conditional probabilities are different, and Baye's theorem gives a relation between $P(A \mid B)$ and $P(B \mid A)$ as follows [Leung, 2007]:

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)}, P(B) \neq 0 \\ P(B \mid A) &= \frac{P(B \cap A)}{P(A)}, P(A) \neq 0 \\ \implies P(A \cap B) &= P(A \mid B)P(B) \text{ or} \\ \implies P(A \cap B) &= P(B \mid A)P(A). \\ \text{Thus, } P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B)}, P(B) \neq 0. \end{aligned} \tag{8.1}$$

Functional classification, similar to standard classification, is also a supervised learning algorithm and the goal is to predict the class label for each of the observed curve [Dai et al.,

2017]. We explain below the functional classification counterpart of the *Naïve Bayes* classifier. Let $X_i(t) = \{X_{1,i}, ..., X_{N,i}\}, i = 1, ..., n$ be the training set with the number of K given by $C_1, ..., C_k$. Let $L(X(t) \mid)\theta_i)$ be the likelihood of observing $X_i(t)$. Then we can predict a new observation function $X(t)$ that can be classified by the predictive probability density function defined by:

$$P(X(t) \mid X_i(t)) = \int L(X(t) \mid \theta_i) P(\theta_i \mid X_i(t)) d\theta_i, \tag{8.2}$$

by assuming the classes have the same prior probability. To solve for $P(X(t) \mid X_i(t)$ for a given class we estimate the distribution of posterior probability density function defined by:

$$P(\theta_i \mid X_i(t)) = \int P(\theta_i, \phi_i \mid X_i(t)) d\phi_i = \int P(\theta_i \mid \phi_i) P(\phi_i \mid X_i(t)) d\phi_i, \tag{8.3}$$

where $P(\theta_i \mid X_i(t)) = \int P(\theta_1, k, ..., \theta_{Nk}, k, \phi_i \mid X_i(t))$, for all $i \neq j \in (1, ..., k)$. Therefore the class with the highest posterior is the outcome of the prediction [Rabaoui et al., 2012].

From the clustering results, a direct extension of this thesis would be to investigate the prediction of a new location in clusters. In this work we only investigated functional clustering of of 16 locations spread across South Africa. Therefore with the extension of classification of functional data, we would like to investigate space and time variation of the new location in South Africa related to other locations.

Chapter 9

Conclusion

In this thesis we have investigated time and space variation for monthly maximum temperature curves using the methods of FDA. Our analysis shows that registering monthly maximum temperature curves of 16 locations spread across South Africa has displayed time variation and revealed the general temperature pattern of the Southern hemisphere, were high temperature occur during month of December to March and low temperature in June to August. The phase-plane plots has offered an advantage to the analysis of registered monthly maximum temperature curves, where the energy is constantly shifting in most years analysed.

The application of fPCA allows us to summarise high dimensional modes of variation in temperature curves without loss of revelant information. functional clustering has revealed that there is three distinct temperature curves (one with higher temperature values, the other with second higher/ average temperature and the other one with lower temperature), where not all locations are in each cluster. The cluster with higher temperature values contain most of inland locations and ones with average and lower and average temperature values have coastal locations. It has also revealed that in most years some locations are always grouped in one cluster while other move around between clusters. This shows that in some years analysed locations can have similar temperature patterns and be different in other years. The application of FDA on temperature data has shown that it is possible to detect change in temperature through space and time variation. In this work we focused only on the application of temperature data but Functional Data Analysis can be applied in areas like medical imaging and other applications. We intend to consider these as our future research endeavour.

Bibliography

Jake Allen. Comparison of time series and functional data analysis for the study of seasonality. 2011.

David M Blei. Hierarchical clustering. 2008.

C Bouveyron. funfem: Clustering in the discriminative functional subspace. *R package version*, 1, 2015.

Hervé Cardot, Mohamed Chaouch, Camelia Goga, and Catherine Labruere. Functional principal components analysis with survey data. In *Functional and Operatorial Statistics*, pages 95–102. Springer, 2008.

Xiongtao Dai, Hans-Georg Müller, and Fang Yao. Optimal bayes classifiers for functional data and density ratios. *Biometrika*, 104(3):545–560, 2017.

Alessandro Dosio. Projection of temperature and heat waves for africa with an ensemble of cordex regional climate models. *Climate Dynamics*, 49(1-2):493–519, 2017.

Manuel Febrero-Bande, M Oviedo de la Fuente, et al. Statistical computing in functional data analysis: The r package fda. usc. *Journal of statistical Software*, 51(4):1–28, 2012.

Alberto Fernández and Sergio Gómez. Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, 25(1):43–65, 2008.

Ramón Giraldo, Pedro Delicado, and Jorge Mateu. Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, 66(4):403–421, 2012.

Pantelis Zenon Hadjipantelis and Hans-Georg Müller. Functional data analysis for big data: A case study on california temperature trends. In *Handbook of Big Data Analytics*, pages 457–483. Springer, 2018.

Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):109–126, 2006.

Peter Hall, Hans-Georg Müller, and Fang Yao. Estimation of functional derivatives. *The Annals of Statistics*, pages 3307–3329, 2009.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Julien Jacques and Cristian Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

Gareth James. Sparseness and functional data analysis. *The Oxford Handbook of Functional Data Analaysis*, pages 298–323, 2010.

Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

İstem Köymen Keser Keser. Döviz kuru ve hisse senedi enerji değişimlerinin faz düzlemleri aracılığıyla incelenmesi. *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, 8(2):36–57, 2015.

GA Kiker. South african county study on climate change. *Synthesis report for the vulnerability and adaptation assessment section. Pretoria: Department of Environmental Affairs and Tourism*, 2000.

Katherine S King. Functional data analysis with application to united states weather data. 2014.

Kibii Komen, Jane Olwoch, Hannes Rautenbach, Joel Botai, and Adetunji Adebayo. Long-run relative importance of temperature as the main driver to malaria transmission in limpopo province, south africa: A simple econometric approach. *EcoHealth*, 12(1):131–143, 2015.

K Krishna and Narasimha M Murty. Genetic k-means algorithm. *IEEE Transactions on Systems Man And Cybernetics-Part B: Cybernetics*, 29(3):433–439, 1999.

Samuel Kusangaya, Michele L Warburton, Emma Archer Van Garderen, and Graham PW Jewitt. Impacts of climate change on water resources in southern africa: A review. *Physics and Chemistry of the Earth, Parts A/B/C*, 67:47–54, 2014.

R Lakhraj-Govender and SW Grab. Temperature trends for coastal and adjacent higher lying interior regions of kwazulu-natal, south africa. *Theoretical and Applied Climatology*, pages 1–9, 2018.

Friedrich Leisch. Sweave, part i: Mixing r and latex. *R News*, 2(3):28–31, 2002.

K Ming Leung. Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

Daniel J Levitin, Regina L Nuzzo, Bradley W Vines, and JO Ramsay. Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne*, 48(3):135, 2007.

Siphumlile Mangisa, Sonali Das, Surajit Ray, and Gary Sharp. Functional regression models for south african economic indicators: a growth curve perspective. *OPEC Energy Review*, 2019.

James Stephen Marron, James O Ramsay, Laura M Sangalli, Anuj Srivastava, et al. Functional data analysis of amplitude and phase variation. *Statistical Science*, 30(4):468–484, 2015.

JA Olorunmaiye and OO Awoola. Models of hourly dry bulb temperature and relative humidity of key selected areas in nigeria for engineering applications. *Nigerian Journal of Technology*, 35(2):360–369, 2016.

H Preston-Thomas. The international temperature scale of 1990 (its-90). *metrologia*, 27(1): 3, 1990.

Asma Rabaoui, Hachem Kadri, and Manuel Davy. Nonparametric bayesian supervised classification of functional data. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3381–3384. IEEE, 2012.

Anand Rajaraman and Jeffrey D Ullman. Jure leskovec. 2010.

James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

James O Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363, 1998.

James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2007.

James O Ramsay, Giles Hooker, and Spencer Graves. *Functional data analysis with R and MATLAB*. Springer Science & Business Media, 2009.

JO Ramsay and BW Silverman. From functional data to smooth functions. *Functional Data Analysis*, pages 37–58, 2005.

Marina Santini. Advantages & disadvantages of k-means and hierarchical clustering (unsupervised learning). *Department of Linguistics and Philology, Uppsala University: http://stp. lingfil. uu. se/˜ santinim/ml/2016/Lect_10/10c_ UnsupervisedMethods. pdf*, 2016.

Milambo Freddy Tshiala, Jane Mukarugwiza Olwoch, and Francois Alwyn Engelbrecht. Analysis of temperature trends over limpopo province, south africa. *Journal of Geography and Geology*, 3(1):13, 2011.

Shahid Ullah and Caroline F Finch. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):43, 2013.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Mueller. Review of functional data analysis. *arXiv preprint arXiv:1507.05135*, 2015.

Yafeng Zhang and Donatello Telesca. Joint clustering and registration of functional data. *arXiv preprint arXiv:1403.7134*, 2014.

Zimin Zhong. *Curve registration in functional data analysis*. Arizona State University, 2008.

Gina Ziervogel, Mark New, Emma Archer van Garderen, Guy Midgley, Anna Taylor, Ralph Hamann, Sabine Stuart-Hill, Jonny Myers, and Michele Warburton. Climate change impacts and adaptation in south africa. *Wiley Interdisciplinary Reviews: Climate Change*, 5(5):605–620, 2014.

Appendix A

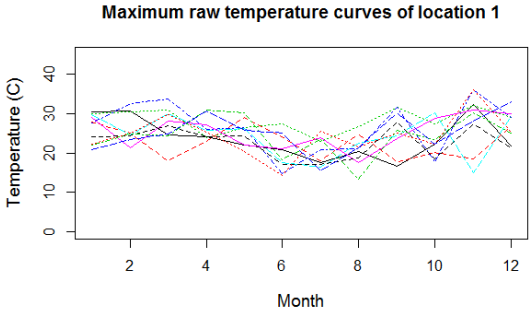
```
#Code for converting nc.df file to
library(raster)
library(ncdf4)
fn1<-("C:/"directory file")
nc<-nc_open(fn1)
dat1<-ncvar_get(nc,attributes(nc$var)$names[3])
nc.brick<-brick(dat1)
write.csv(nc.df,file.choose())
test<-read.csv(file.choose())
```

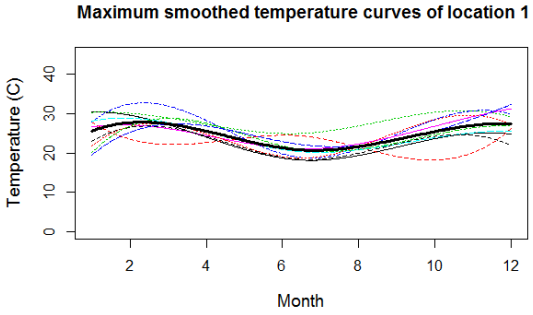
Table 9.1: Table of reference by location

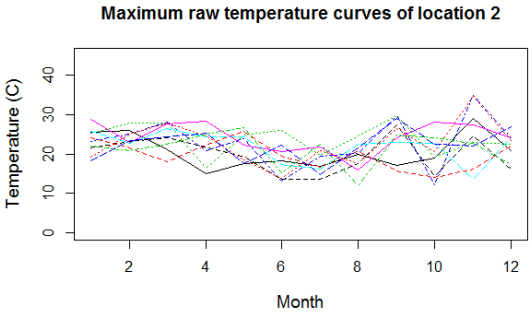
Location	Close Place	Longitude(Lon)	Latitude(Lat)
1	Polokwane	29.25	-23.75
2	Dullstrom	30.25	-25.25
3	Durban	30.75	-29.75
4	Ladysmith	29.75	-28.75
5	Mthatha	28.75	-31.75
6	George	22.25	-33.75
7	Beaufortwest	22.25	-32.25
8	Cape Town	18.25	-33.75
9	Calvinia	19.75	-31.25
10	Upington	21.25	-28.25
11	Alexander Bay	16.25	-28.75
12	De Aar	23.75	-30.75
13	Klerksdorp	26.75	-25.75
14	Johannesburg	27.75	-26.25
15	Pretoria	28.25	-25.75
16	Thabazimbi	27.25	-24.75

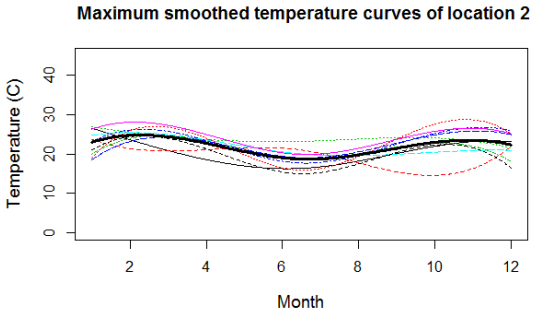
Appendix B

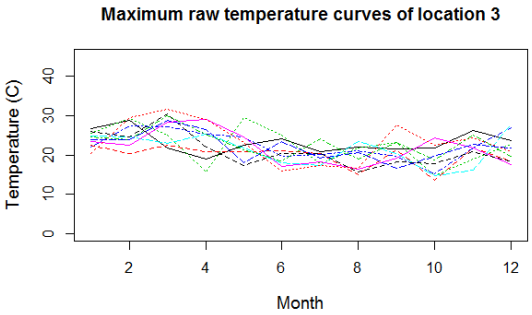
Raw and Smoothed maximum temperature curves of location 1 to 16

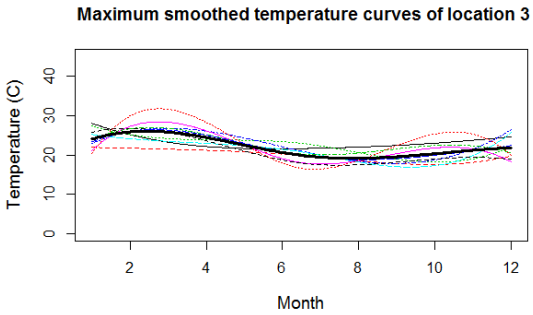


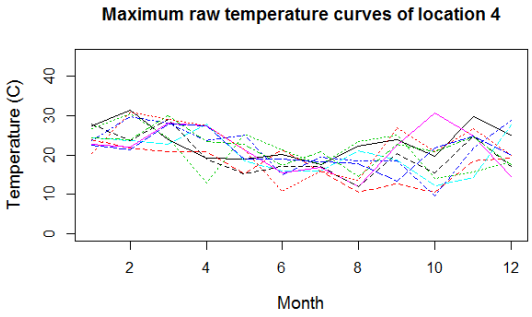


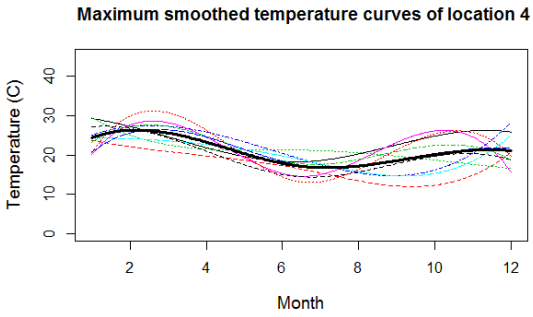


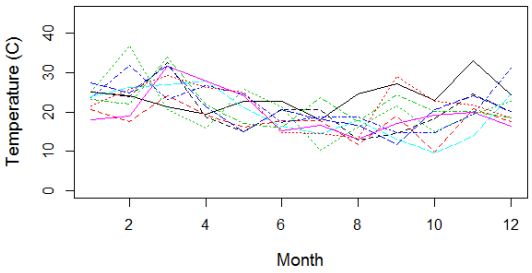


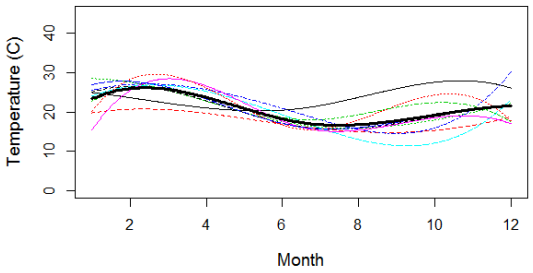


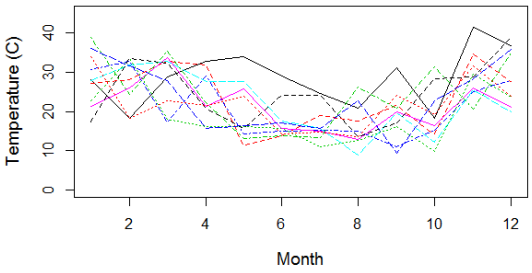


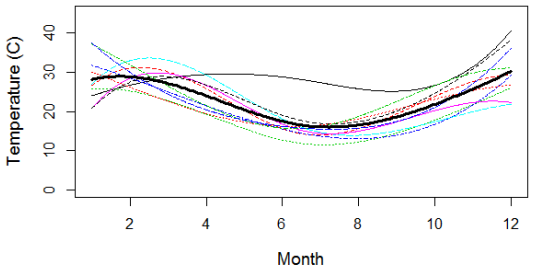

Maximum raw temperature curves of location 4
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

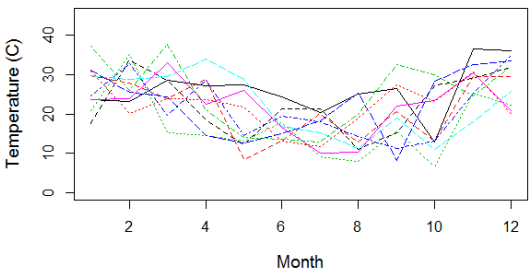

Maximum smoothed temperature curves of location 4
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

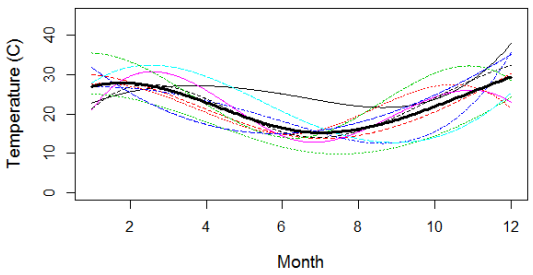

Maximum raw temperature curves of location 5
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

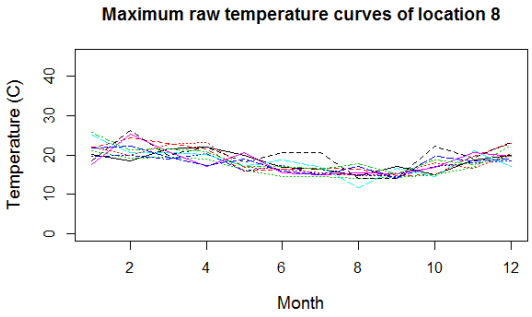

Maximum smoothed temperature curves of location 5
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

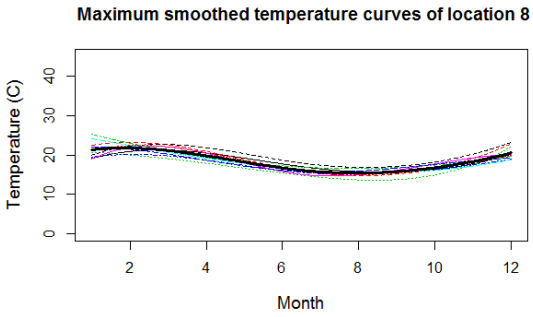

Maximum raw temperature curves of location 6
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

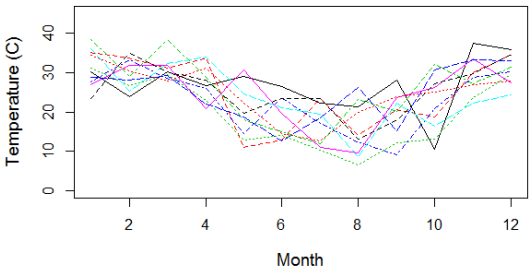

Maximum smoothed temperature curves of location 6
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

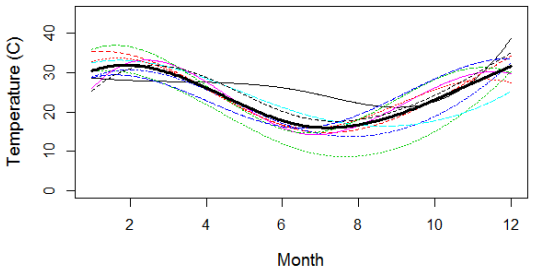

Maximum raw temperature curves of location 7
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

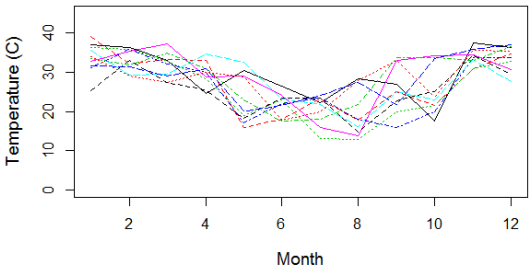

Maximum smoothed temperature curves of location 7
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

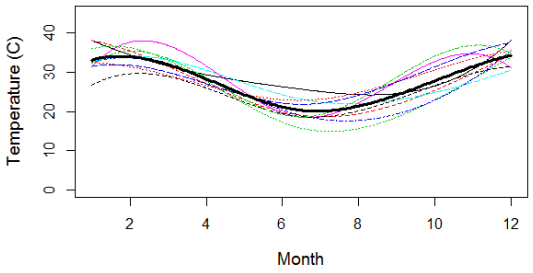

Maximum raw temperature curves of location 8
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

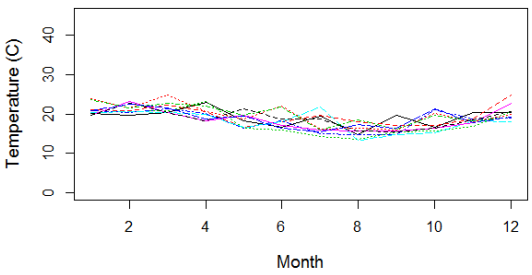

Maximum smoothed temperature curves of location 8
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

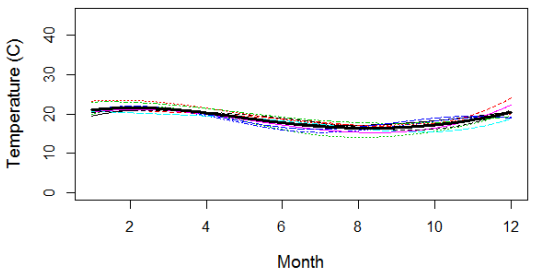

Maximum raw temperature curves of location 9
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

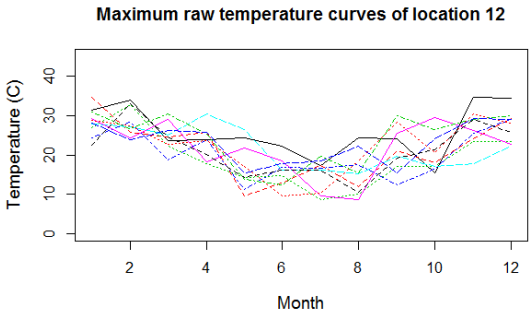

Maximum smoothed temperature curves of location 9
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

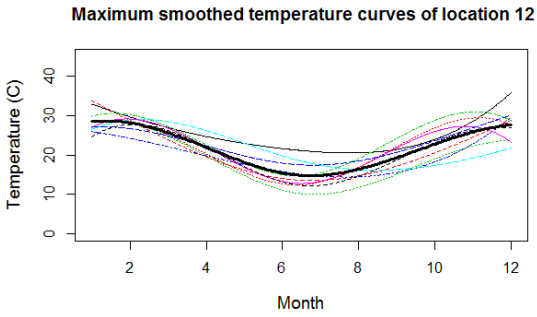

Maximum raw temperature curves of location 10
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

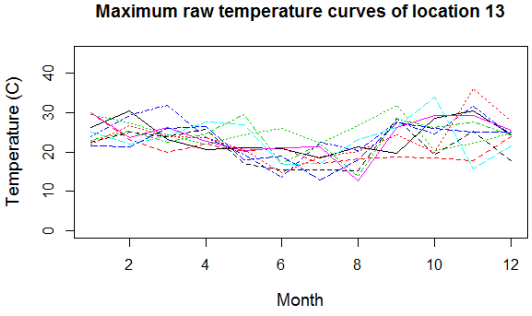

Maximum smoothed temperature curves of location 10
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

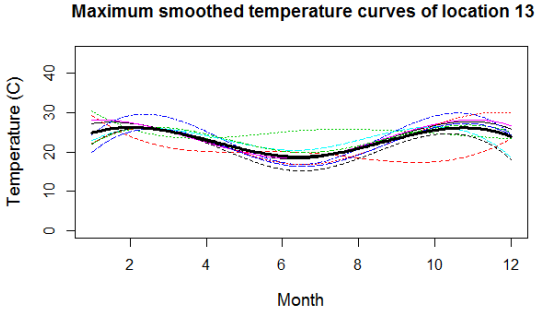

Maximum raw temperature curves of location 11
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

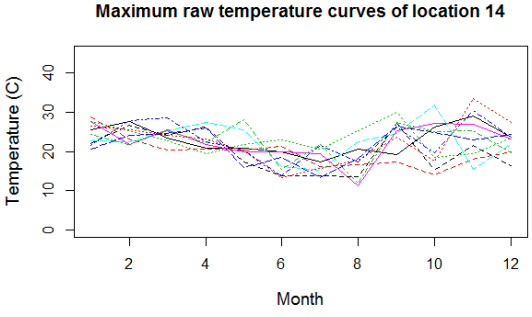

Maximum smoothed temperature curves of location 11
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

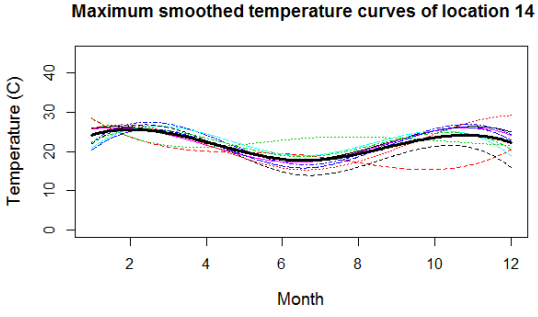

Maximum raw temperature curves of location 12
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

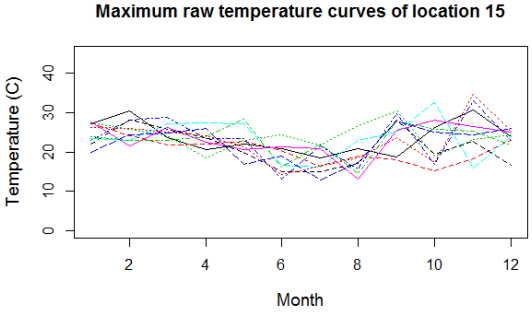

Maximum smoothed temperature curves of location 12
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

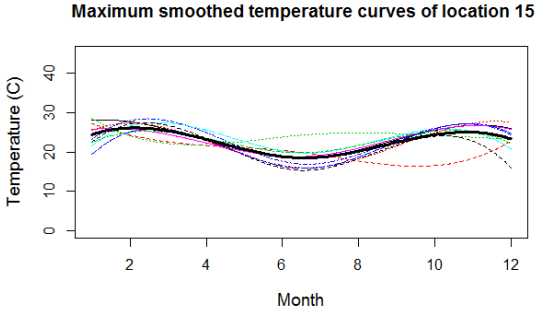

Maximum raw temperature curves of location 13
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

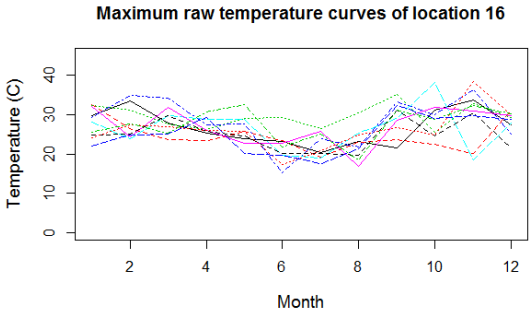

Maximum smoothed temperature curves of location 13
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month

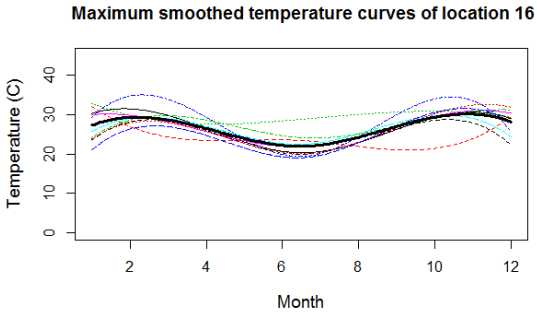

Maximum raw temperature curves of location 14
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month


Maximum smoothed temperature curves of location 14
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month


Maximum raw temperature curves of location 15
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month


Maximum smoothed temperature curves of location 15
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month


Maximum raw temperature curves of location 16
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month


Maximum smoothed temperature curves of location 16
Temperature (C)
0
10
20
30
40
2
4
6
8
10
12
Month