ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

Advances in Decision Sciences

Volume 24
Issue 3
September 2020

Michael McAleer

Editor-in-Chief
University Chair Professor
Asia University, Taiwan

Published by Asia University, Taiwan

ADS@ASIAUNIVERSITY

Analysing Maximum Monthly Temperatures in South Africa for 45 years Using Functional Data Analysis*

Mapitsi Rangata

Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research (CSIR)
Pretoria, South Africa

Sonali Das**

Department of Business Management
University of Pretoria
Pretoria, South Africa

Montaz Ali

Department of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa

Revised: July 2020

* The authors wish to thank a referee for helpful comments and suggestions.

** Correspondence: sonali.das@up.ac.za

Abstract

The paper uses Functional Data Analysis (FDA) to explore space and time variation of monthly maximum temperature data of 16 locations in South Africa for the period 1965 - 2010 at intervals of 5 years. We explore monthly maximum temperature variation by first representing data using the B-spline basis functions. Thereafter registration of the smooth temperature curves was performed. This data was then subjected to analysis using phase-plane plots which revealed the constant shifting of energy over the years analysed. We next applied functional Principal Component Analysis (fPCA) to reduce the dimension of maximum temperature curves by identifying the maximum variation without loss of relevant information, which revealed that the first functional PCA explains mostly summer variation while the second functional PCA explains winter variation. We next explored the functional data using functional clustering using K-means to reveal the spatial location of maximum temperature clusters across the country, which revealed that maximum temperature clusters were not consistent over the 45 years of data analysed, and that the cluster points within a cluster were not necessarily always spatially adjacent. The overall analysis has displayed that maximum temperature clusters have not been static across the country over time. To the best of our knowledge, this the first instance of performing in-depth analysis of maximum temperature data for 16 locations in South Africa using various FDA methods.

Keywords: Functional data analysis, Principle components, South Africa, Temperatures.
JEL: C10; C21; C22; C51; O13

1. Introduction

South Africa has raised concerns within the climate change community by sharing its experiences of frequent extreme events (Ziervogel, 2014). These extreme weather episodes, and consequences thereof, pose acute stress on South Africa's water resources, food security, health, infrastructure and biodiversity, which directly and indirectly affects human well-being, particularly of those who are most vulnerable.

If we focus only on the understanding of temperature in South Africa, there remains a need to obtain better insights into extreme temperature patterns to better understand how annual temperature profiles vary across South Africa. Note, analysing average monthly temperature does little justice as it smooths out the effect of extreme temperatures that occur during the course of the month.

Furthermore, it is important to keep cognisance of the fact that each year, the maximum temperature (and indeed the minimum temperature) does not necessarily occur at the same time in the year, and it is important that before one can analyses variations of a specific weather measurement (such as maximum temperature), one aligns the landmarks (such as annual peaks, valleys) across the years.

The framework under which such analysis can be undertaken is the Functional Data Analysis (FDA) framework (Wang, 2015), in which first the monthly data are smoothed into annual curves, and thereafter the smoothed annual curves are aligned with regards landmark features of interest, allowing investigation of variations in both phase and amplitude across the years. In addition, the FDA framework allows the analysis of other features contained within the data such as identification of the principal components of variations, identification of clusters of curves and the analysis of the rates of change of the curves.

In the analysis, we revisit the understanding of the maximum temperature variations across South Africa using the FDA framework. We specifically focus on maximum monthly temperature data between 1965-2010 from 16 locations spread across the South Africa. Below we present in detail the historic and current state of South African, and indeed of southern African, climate and its impact on the global climate scene with special focus on literature

discussing the impact of extreme temperature variations on the lives of those living in the region, and motivate why the continued understanding of temperature data using emerging statistical tools is hence vital.

We next introduce FDA methodology in greater detail as is used within the scope of this paper. We next present the data that is analysed using various FDA methods and discuss the FDA results thereof. Finally, we present concluding remarks in the context of what the findings may mean for those living in the region.

2. Methodology

Functional Data Analysis (FDA) is the analysis of data that is in the form of curves (Wang, 2015). The objectives of FDA are similar to any standard statistical analysis that is to investigate patterns of variability in the data, to estimate summary statistics, to build models and aid in the process of inference (Ramsay J. O., 2007). FDA methods start with transforming discrete data into functions by smoothing them over a specific continuum (such as a year).

Mathematically, functional data are derived from a set of observations from a continuous underlying process $X(t_j)$, observed at time t_j. We denote by $y(t_j)$ the observed $X(t_j)$ with a noise component $\varepsilon(t_j)$, where $j = 1, \dots, n$ (Levitin, Introduction to functional data analysis, 2007). Therefore, a single functional observation $y(t_j)$ is derived from of n pairs $(t_j, y(t_j))$ as follows:

$$y(t_j) = x(t_j) + \varepsilon(t_j),$$

where $y(t_j)$ is a smooth function, $x(t_j)$ is the observation and $\varepsilon(t_j)$ is a measurement error or noise. To fit a curve to transform the discrete data into functional data curves, a smoothing process is used as follows:

$$x(t_j) = \sum_{k}^{i=1} \varphi_i(t_j) c_i,$$

where $\varphi_i(t_j)$ are basis functions and c_i are the coefficients associated with k basis functions (Olorunmaiye, 2016).

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 a large number k, of such functions. The amount of smoothness of the function is determined by number k (Ramsay J. O., 1998). 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, Introduction to functional data analysis, 2005). Thus, the number of basis functions is equal to the number of knots plus order of the spline (Ramsay J. O., 2009).

Most of the curves in a functional space, like the much cited human growth curve data (Ramsay J. O., 2006), exhibit variability both in terms of amplitude (vertical) as well as phase (horizontal). Phase variation contains interesting information in the timing of the curves' important peaks (or troughs). To investigate phase and amplitude variation, the curves need to be aligned using any of the curve registration methods such as *landmark* registration, *continuous* registration, and *shift* registration, which use the time-warping function to transform the domain for each curve (Marron, 2015).

Time-warping is a technique that transforms the domain (e.g., time) for each curve to align certain features of interest, with the property that they must always be strictly increasing, as time cannot go backward over an interval $[0, T]$. A time-warping function $h(.)$ must satisfy the constraints that $h(0) = 0$ and $h(T) = T$. In this paper we use *continuous* registration to align monthly maximum temperature curves. *Continuou*s registration is a method that uses the entire curve, rather than specified features, to align the curves, and uses the time-warping function $h(.)$ to minimize amplitude variation (Ramsay J. O., 2009).

Derivatives and Phase-Plane Plots

After the smoothing process, the derivatives of the curves can be obtained as follows:

$$D^m x(t) = c_1 D^m \varphi_{1(t)} + \cdots + c_k D^m \varphi_{k(t)},$$

where D^mdenotes the m^{th} derivative operator (Ramsay J. O., 2006). As some of the variation in a curve can be explained at the level of certain derivatives, the use of phase-plane plots to visualise velocity against acceleration provides valuable insights (Hall, 2009), and provide a graphical representation of energy within the system, with the amount of energy in the system being related to the height and width of the ellipse. Specifically, kinetic energy is associated with high velocity and low acceleration, while the potential energy is characterised by high acceleration.

Functional Principal Component Analysis (fPCA)

fPCA is a dimension reduction tool for multivariate data that has been extended to functional space (Wang, 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 does (Hadjipantelis, 2018). fPCA transforms or reduces high dimensional dataset to a low-dimensional dataset which contains a set of uncorrelated components that summarises features that represent the original dataset and captures the main modes of variability in the data (Cardot, 2008) using eigenvalues and the eigenfunctions. One of the challenges in fPCA application is the selection of the number of components to retain or reject (Hadjipantelis, 2018).

Functional Clustering

The purpose of functional clustering is to identify representative curve patterns which are likely generated from the similar process (Zhang, 2014). *K-means* and *Hierarchical* functional clustering are two popular algorithms for functional cluster analyses. Given a set of functional data $\{X_i(t); i = 1, \dots, n\}$, *K-means* finds a set of cluster means denoted by $\{\mu^c; c = 1, \dots, L\}$ by minimising the sum of the squared distance between $\{X_i\}$ and the cluster centres $\{C_i; i = 1, \dots, n\}$ (Wang, 2015).

Hierarchical functional clustering is similar to 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, and finally proceeds to apply agglomerative criterion (single, complete, or average linkage) (Giraldo, 2012).

3. Data

For our analysis, weather data was acquired from the Council of Science and Industrial Research (CSIR). The original data comprised of temperature, precipitation and relative humidity in the form of NETCDF (Network Common Data Form). In this paper we only focus on monthly maximum temperature, which are first converted from the NETCDF file format into a .CSV (Comma-Separated Values) file.

The unit of temperature in the data was in degree Kelvin, which was converted to degree Celsius, with the relationship being $\{T_{(c)} = T_{(k)} - 273.15\}$, where $T$is temperature (Preston-Thomas, 1990), as Celsius measurements are easy to relate with. We then extracted data of selected 16 locations spread across South Africa and limited the data to only comprise of maximum monthly temperature values.

The spatial locations of the 16 locations on a map are presented in Figure 1. 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 column 4-15 are the 12 months of the year. Note, while the time span of the data was from 1960 to 2010, in this paper we analyse data of every 5-year interval, specifically 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005 and 2010.

The summary statistics of the data is presented in the table below, with the location numberings as follows: location 1: Dullstrom, location 2: Durban, location 3: Ladysmith, location 4: Mthatha, location 5: George, location 6: Beaufortwest, location 7: Cape Town, location 8: Calvinia, location 9: Upington, location 10: Alexander Bay, location 11: De Aar, location 12: Klerksdorp, location 13: Johannesburg, location 14: Pretoria, location 15: Thabazimbi, and location 16: Polokwane.

All analysis in this paper are done in the RStudio environment, using the following R packages: *fda*, *FunFEM*, *fda.usc*, and *fdasrvf*. For the spatial visualisation, we used the *QGIS* software.

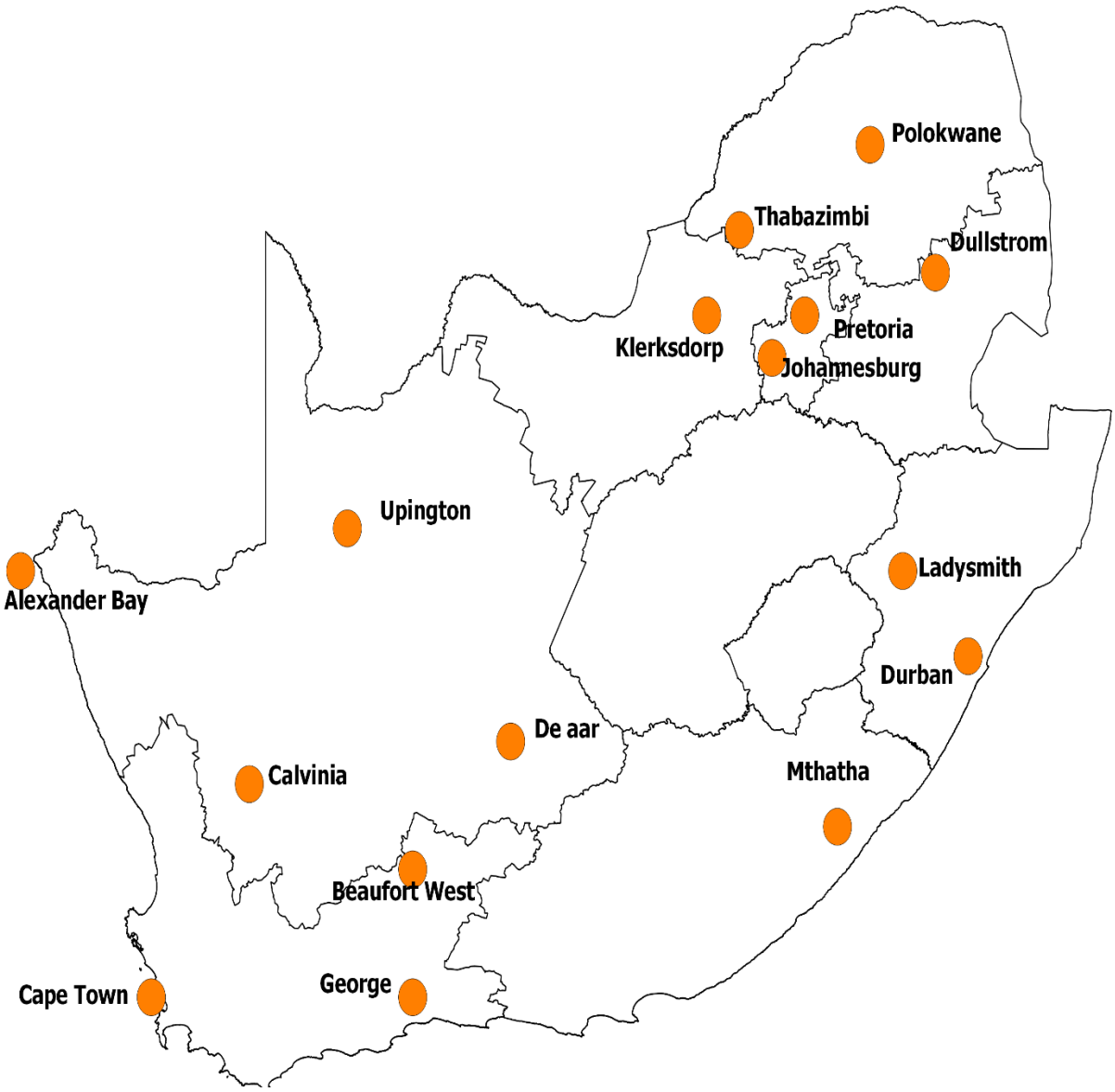

Figure 1

Map of South Africa with lo cations considered in the paper

Table 1

Summary Statistics of Maximum Monthly Temperatures in 16 locations for 1965-2010

Location	Min	1st Quartile	Median	Mean	3rd Quartile	Max
1. Dullstrom	11.93	18.43	22.43	21.93	24.90	34.86
2. Durban	13.61	19.10	21.84	22.06	24.58	31.66
3. Ladysmith	9.59	17.55	21.18	21.05	24.66	31.44
4. Mthatha	9.61	17.34	20.62	20.88	24.29	36.83
5. George	8.82	16.00	22.45	22.91	28.75	41.57
6. Beaufortwest	6.53	15.24	22.84	22.20	28.59	37.80
7. Cape Town	11.68	16.30	18.49	18.45	20.59	26.15
8. Calvinia	6.48	19.30	25.57	24.32	30.26	38.42
9. Upington	12.95	22.59	29.02	27.66	33.18	39.06
10. Alexander Bay	13.39	16.56	19.05	18.90	20.61	24.82
11. De Aar	8.56	17.07	23.05	22.03	27.04	34.68
12. Klerksdorp	12.56	20.09	23.41	23.16	26.23	36.05
13. Johannesburg	11.21	19.30	22.13	21.96	25.26	33.46
14. Pretoria	12.93	19.35	23.05	22.72	25.85	34.78
15. Thabazimbi	15.21	23.48	26.30	26.58	29.83	38.30
16. Polokwane	13.42	21.35	24.82	24.69	28.84	35.98

4. Results

Data of 10 years between 1965 and 2010, with an interval of 5 years, from 16 locations was analysed, first by smoothing each of the 10x16 = 160 curves, then aligning the curves and then exploring the fPCA, phase-plane plots, and functional cluster methods on the registered curves. To smooth monthly maximum temperature data of 10 years from all 16 locations, using B-spline with 3 knots and degree 2. We then aligned the curves with the target function so that the peaks, valleys and crossings occur at the same argument as those of the target, using *continuous* registration.

In Figure 2 we visualise the raw smoothed maximum monthly temperature curves (a), and the corresponding registered smoothed maximum monthly temperature curves (b). In our exploratory analysis investigation of the data using phase-plane plots, we use the average monthly maximum from the registered curves of the 16 locations. The numbers 1 to 12 inside the phase-plane plots represent the months January (1) through December (12). We do so for all the 10 years considered in this paper.

The phase-plane plots reveal that in 1980 and 1985 we had similar weather patterns, as are for the years 1990 and 2005. In 1965 positive potential energy is greatest around August to December, implying steady increase in temperature. In 1975, from January to May, temperature is decreasing with zero velocity in June, with temperature increasing from July to December with high acceleration from September to December.

In 1980 and 1985 the phase-plane plots suggest that we had similar weather patterns, steadily increasing from January to April and decreasing in May, with November and December with larger kinetic energy.

In both 2000 and 2005 there is large kinetic energy in April and September, implying lower temperature in these months, and there is also larger potential energy in January and November. 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.

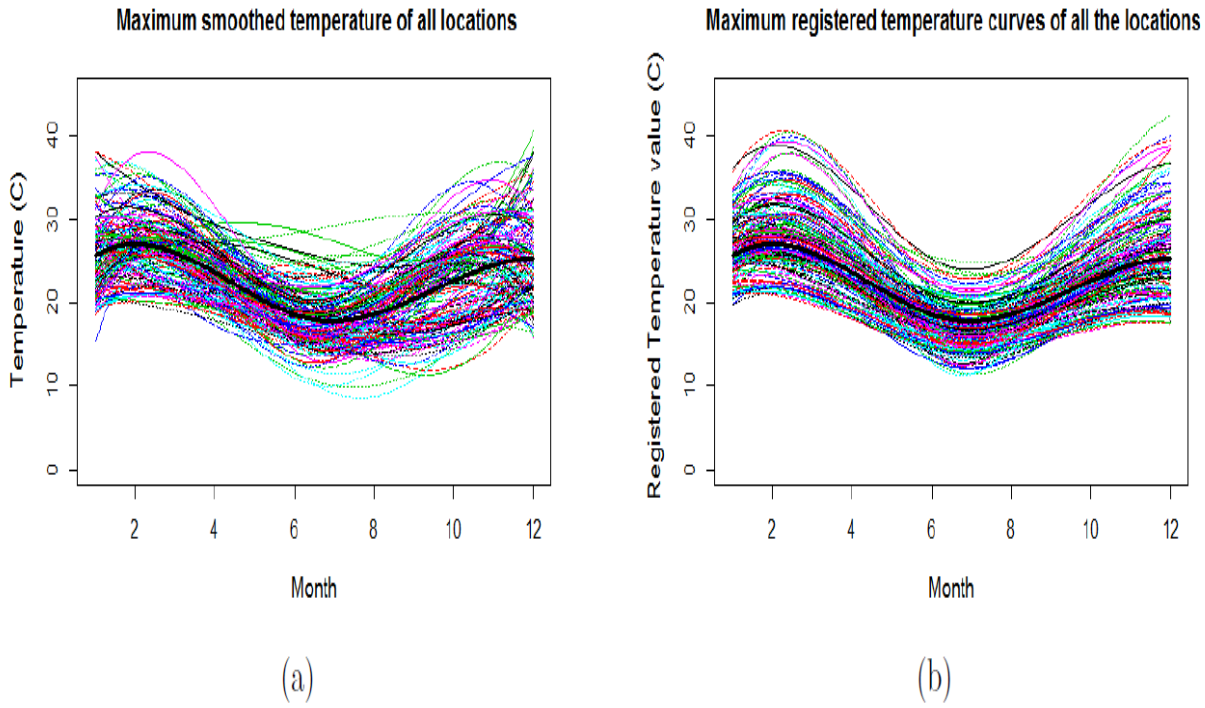

Figure 2

Smoothed and Registered Monthly Maximum Temperatures in 16 locations from 1960-2010

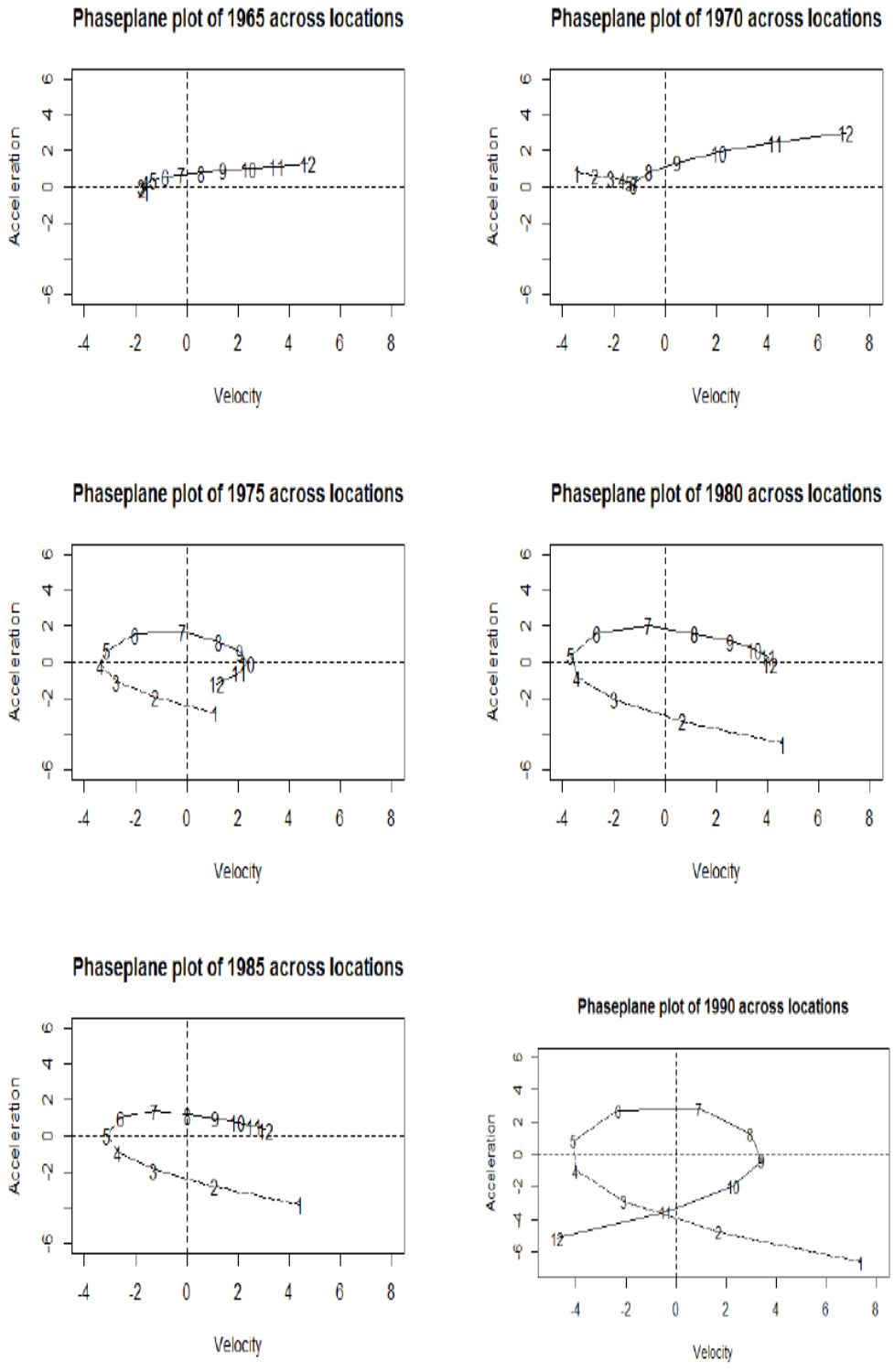

Figure 3

Phase-plane Plots of Average Monthly Maximum Temperatures in all Locations for 1965 and 1990

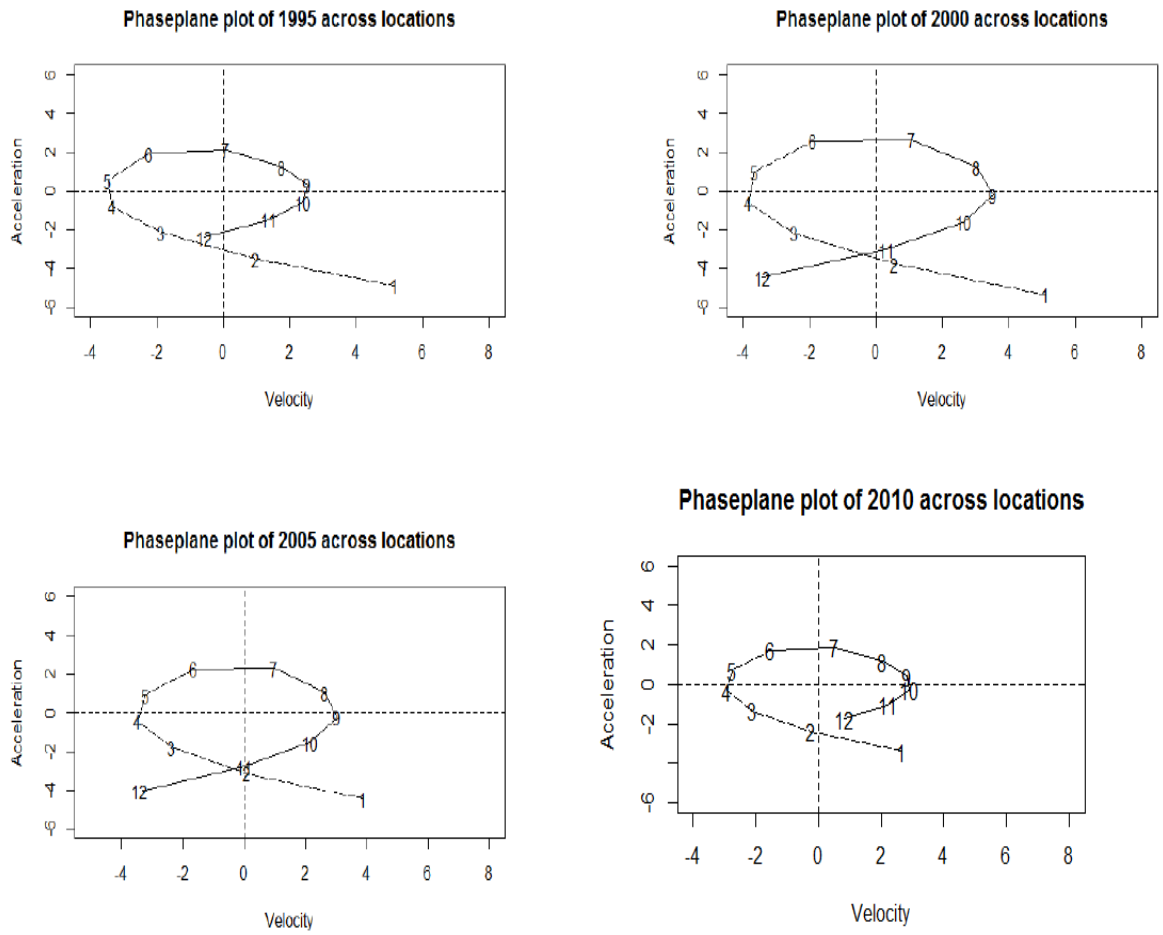

Figure 4

Phase-plane Plots of Average Monthly Maximum Registered Temperatures at All Locations for 1995-2010 with an interval of 5 years

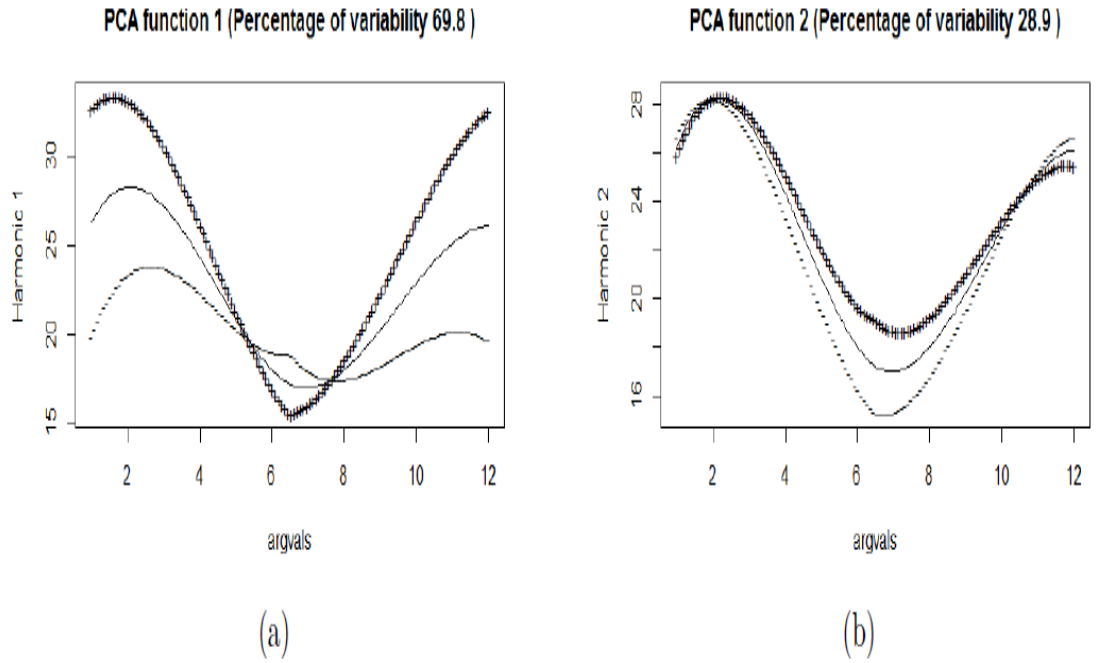

Figure 5

First and Second Functional Principle Components
of Monthly Maximum Temperatures at 16 Locations

We performed fPCA using the *fda* library function in the *fda* package on the registered monthly maximum temperature data for the 16 locations, and consider the first two harmonic functions, or functional principal components, to explain variations in the data. We observe that the first two fPCs explain 98.7% of the total variations in the data as evidenced from Figure 5.

Furthermore, we observe that fPC 1 explains about 70% of the variations (Figure 5(a)), which correspond to the summer variations, while fPC 2 explains about 29% of the variations (Figure 5(b)), which correspond to the winter variations. The solid lines in Figure 5 represent the mean curve, while the curves labelled with a '+' or a '-' indicate the one standard deviation added or subtracted from the mean respectively.

We use the registered monthly maximum temperature curves from the 16 locations and 10 years to group them into clusters that have high inter-cluster variability, and low intra-cluster variability. We chose number of clusters $k = 3$ when using *K-means* and *Hierarchical* algorithm the *funFEM* library function within the *fda* package.

In the Figure 6 we present results from the clustering algorithm. When we investigate the clustering analysis output, we observe that in Cluster 1 (black) two locations feature dominantly, namely Cape Town and Alexander Bay; in cluster 2 (green) five locations are feature dominantly, namely George, Calvinia, Upington, De Aar and Thabazimbi, while in cluster 3 (red), nine locations feature dominantly namely Polokwane, Dullstrom, Durban, Ladysmith, Beaufortwest, Klerksdorp, Johannesburg and Pretoria.

From the Figure 6 (a) we observe 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 temperature from May to September, with the lowest being between June to August, which correspond to the winter season. The mean curve of each cluster in Figure 6 (b) reveal interesting patterns with regards the magnitude and variability within each cluster.

Hierarchical functional cluster results using agglomerative algorithm on the monthly maximum temperature curves of all 16 locations and 10 years are displayed as dendrograms in Figure 8. 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 geographic region.

We also observe that in all the years analysed, Ladysmith, Mthatha and Durban are always grouped together in a cluster, which fall under Kwa-Zulu-Natal Province in the coastal area; Alexander Bay and Cape Town are always grouped in one cluster (coastal area); and Johannesburg and Pretoria are always grouped in one cluster with mostly Klerksdorp, or either Dullstroom or Polokwane (which are the inland areas) in one cluster.

Furthermore, from 1965, Klerksdorp and Pretoria have most similar temperature values since the height of the link that joins them together is the smallest amongst others. When we observe dendrogram of year 2000, we see that there are two clusters, one that contain Upington and Thabazimbi and the other contain the remaining locations.

Table 3 shows the positions of the three clusters by proportion of locations in all 10 years analysed, that there is 48 % of locations in cluster 3, 17 % of locations in cluster 1, and 36 % of locations in cluster 2. Then Table 4 present the positions by proportion of locations in each year analysed.

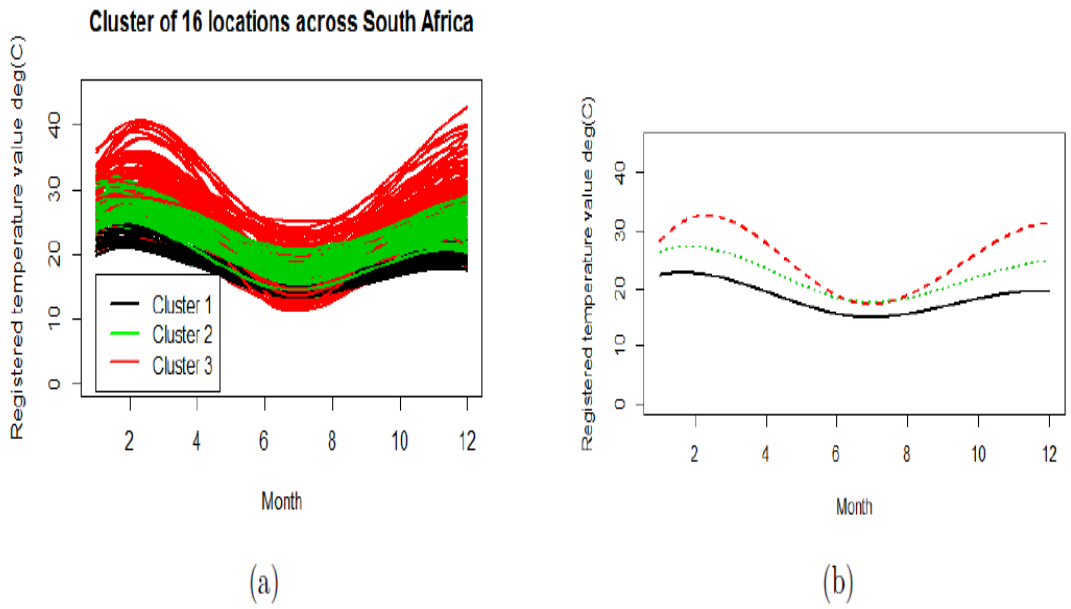

Figure 6

Left Panel Registered Monthly Maximum Temperatures at 16 locations by Cluster

Right Panel is the Mean Curve of Each Cluster

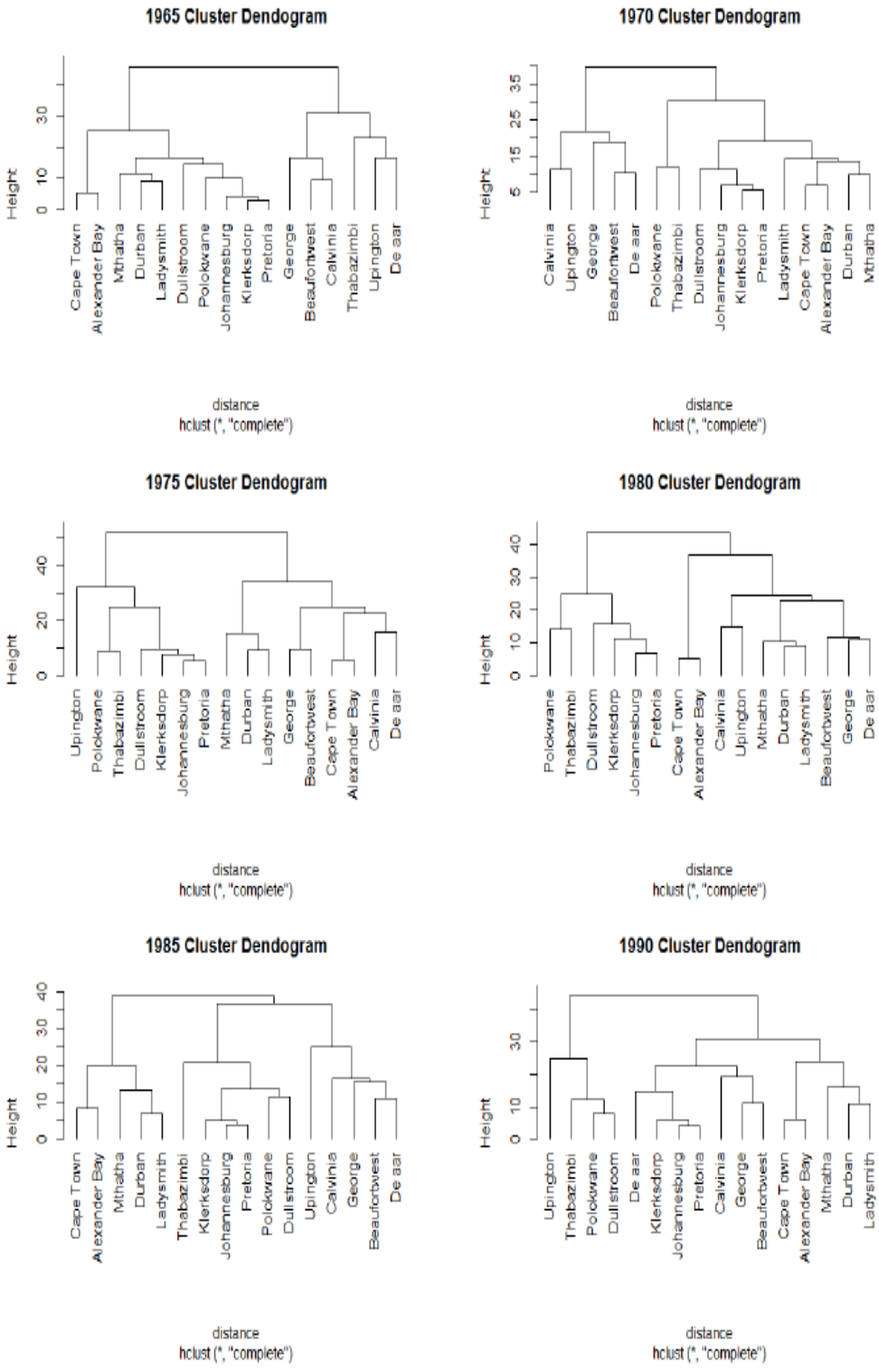

Figure 7

***Hierarchical* Clustering of Monthly Maximum Curves at All 16 Locations in South Africa for 1965 - 1990 with an interval of 5 years**

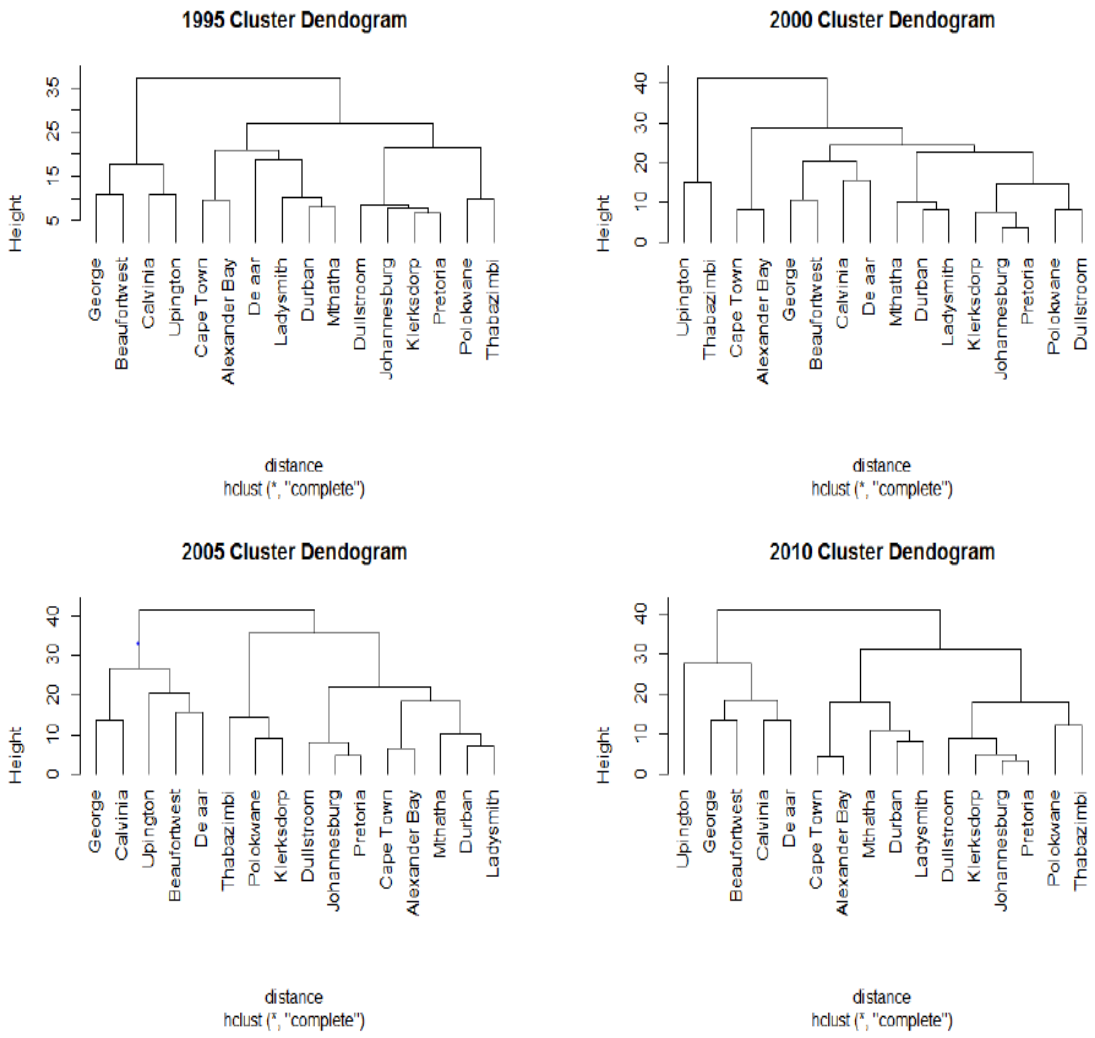

Figure 8

***Hierarchical* Clustering of Maximum Monthly Curves at All 16 Locations in South Africa for 1995 - 2010 with an interval of 5 years**

Table 2

Yearly Proportions 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

Table 3

Proportions by Cluster in All Locations

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

In order to obtain greater insights into the clustering data, we display our cluster results on the map for each year in this study between 1965 to 2010, with a gap of 5 years, in Figure 9, and 10 below. In Figure 9:

(a) we observe that the three locations in Northern Cape are grouped in one cluster;

(b) we observe that the locations close to Indian Ocean are grouped together with the inland locations expect Thabazimbi;

(c) We observe that all locations close to Indian and Atlantic Ocean are grouped in one cluster together with Dullstroom;

(d) all coastal locations and inland locations are grouped in both cluster 1 and 2 except Alexander Bay and Cape Town (coastal location) which are grouped in cluster 3;

(e) locations both inland and coastal are spread across in cluster 1 and 2 except Alexander Bay and Cape Town (coastal location) which are in cluster 3;

(f) 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.

The spatial visualisation reveals that Cape Town and Alexander Bay are always grouped together in one cluster, and also Johannesburg and Pretoria are always in the same cluster. This suggests that in all the years analysed Cape Town and Alexander Bay, and Pretoria and Johannesburg always had similar maximum temperature patterns.

Furthermore, we 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. For the remaining year, such clear consistent cluster memberships were not observed.

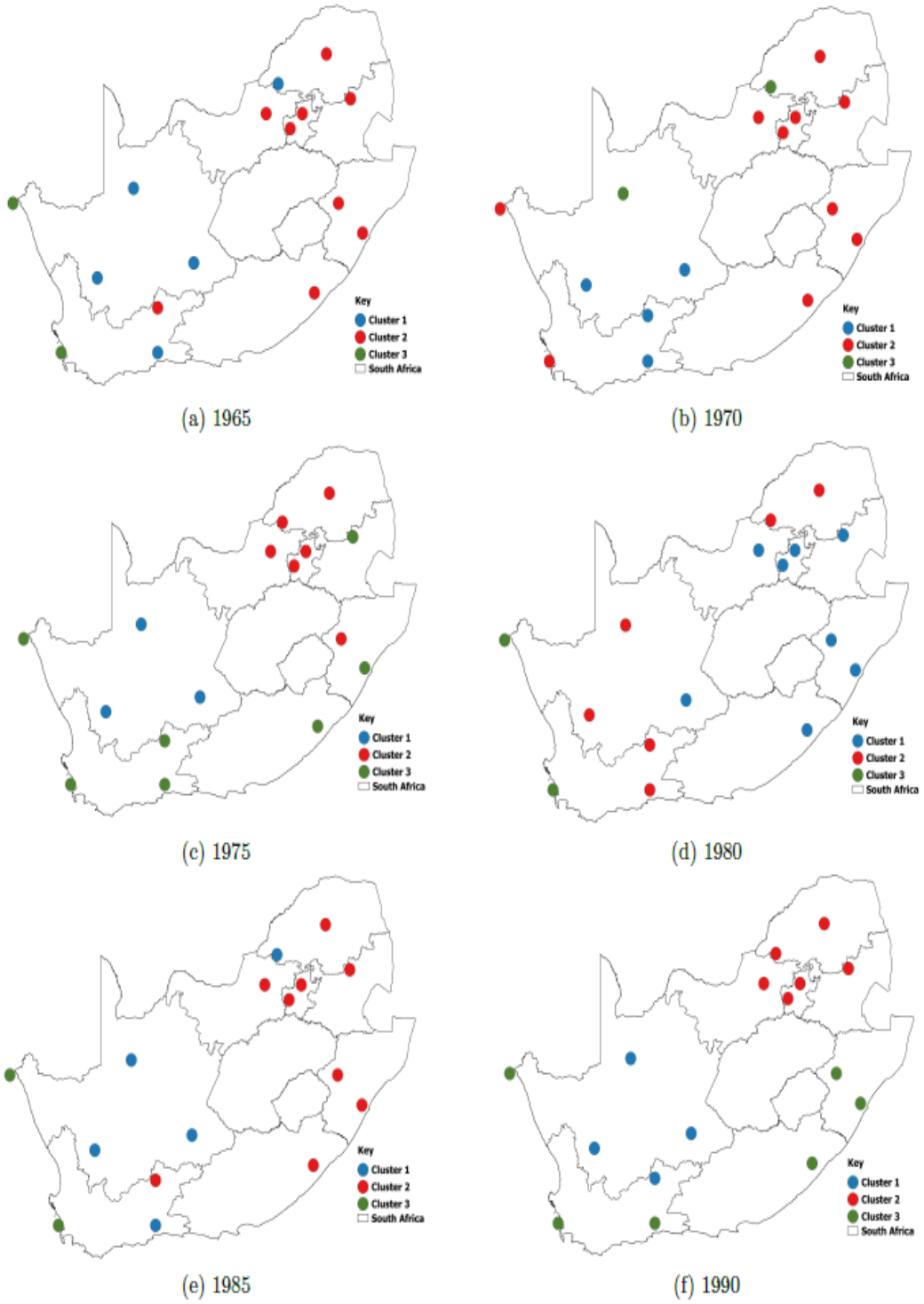

Figure 9

Visualisation of cluster results from 1965 - 1980 with an interval of 5 years

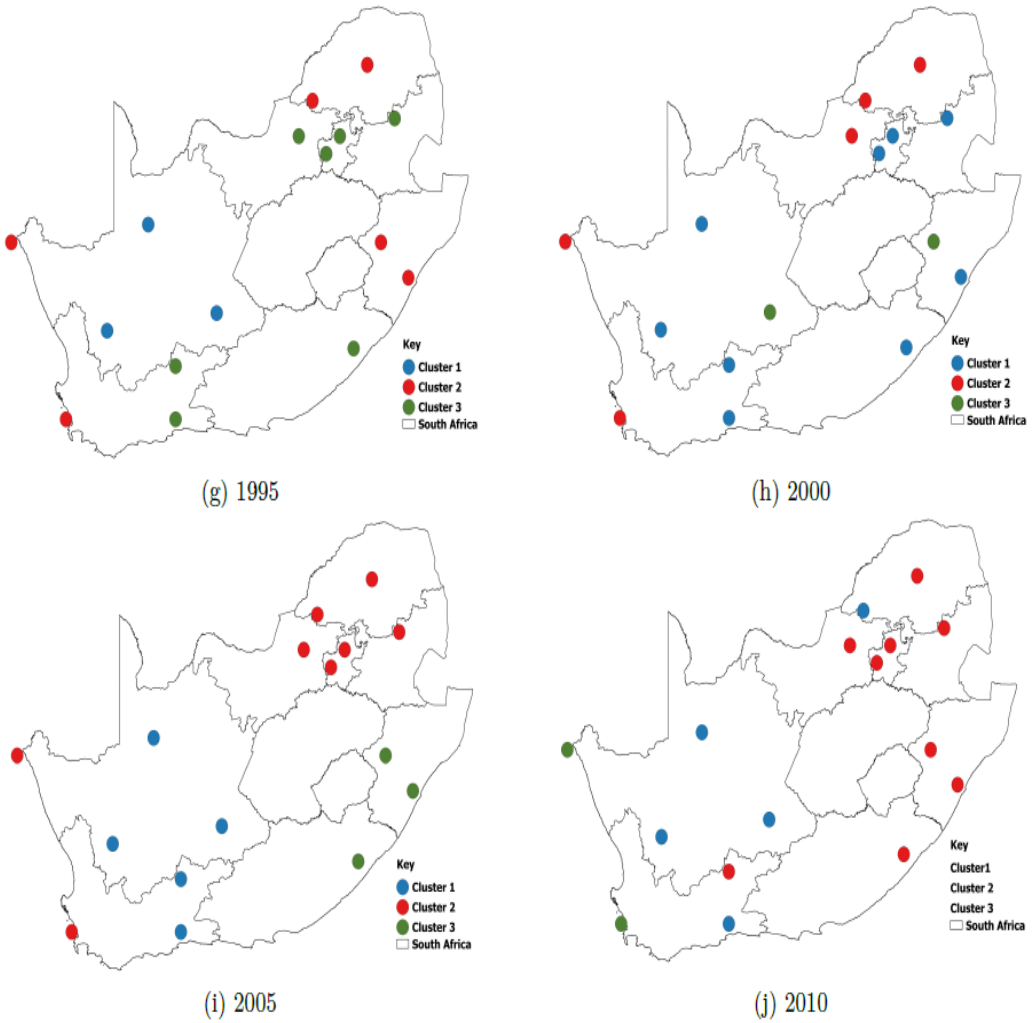

Figure 10

Visualisation of cluster results from 1985 - 2010 with an interval of 5 years

5. Discussion and Concluding Remarks

In this paper we have investigated time and space variation for monthly maximum temperature curves using the methods within the FDA framework, specifically, phase-plane plots, functional principal component and functional clustering methods. The phase-plane plots offered an advantage to the analysis of registered monthly maximum temperature curves, were 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 relevant information. Our analysis using fPCA methods showed that monthly maximum temperature curves of 16 locations spread across South Africa display variation over time, and particularly reveal that temperature pattern were more variable during austral summers (December to March) and less variable during austral winters (June to August).

Functional clustering analysis revealed that there are distinct temperature clusters, with some clusters comprising of consistent locations across the time period analysed, while other locations seem to have less cluster loyalty. The cluster with higher temperature values contain most of inland locations, and ones with average and lower temperature values have coastal locations.

The application of FDA methods on temperature data has shown that more insights can be obtained about temperature data variations both spatially as well as temporally. In this work we focused only on the application of temperature data, however Functional Data Analysis methods can be applied to other weather data to get a more holistic insight into climate induced changes over time and space.

References

Cardot, H. a. (2008). Functional principal components analysis with survey data. In *Functional and Operatorial Statistics* (pp. 95-102). Springe.

Giraldo, R. a. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, 403-421.

Hadjipantelis, P. Z.-G. (2018). Functional Data Analysis for Big Data: A Case Study on California Temperature Trends. In *Handbook of Big Data Analytics* (pp. 457-483). Springer.

Hall, P. a.-G. (2009). Estimation of functional derivatives. *The Annals of Statistics*, 3307-3329.

Levitin, D. J. (2005). Introduction to functional data analysis. *Canadian Psychological Association*, 135.

Levitin, D. J. (2007). Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne*, 135.

Marron, J. S. (2015). Functional data analysis of amplitude and phase variation. *Statistical Science*, 468-484.

Olorunmaiye, J. a. (2016). Models of hourly dry bulb temperature and relative humidity 0f key selected areas in Nigeria for engineering applications. *Nigerian Journal of Technology*, 360-369.

Preston-Thomas, H. (1990). The international temperature scale of 1990 (ITS-90). *metrologia*, 3.

Ramsay, J. O. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 351-363.

Ramsay, J. O. (2006). *Functional data analysis.* Wiley Online Library.

Ramsay, J. O. (2007). *Applied functional data analysis: methods and case studies.* Springer.

Ramsay, J. O. (2009). *Functional data analysis with R and MATLAB.* Springer Science \& Business Media.

Wang, J.-L. a.-M.-G. (2015). Review of Functional Data Analysis. *arXiv preprint arXiv:1507.05135*.

Zhang, Y. a. (2014). Joint clustering and registration of functional data. *arXiv preprint arXiv:1403.7134*.

Ziervogel, G. a.-H. (2014). Climate change impacts and adaptation in South Africa. *Wiley Interdisciplinary Reviews: Climate Change*, 605-620.