The School of Mathematics and Statistics
Statistics Project and Dissertation
2024-25

Using Earth Observation to Identify Coherence in Surface Temperature of the Caspian Sea.

Abhinav Singh - 3041866
Msc Data Analytics, University of Glasgow, United Kingdom

ABSTRACT

The Lake Surface Water Temperature is analysed for the largest lake in the world- Caspian Sea, recorded using remote sensing satellite as a part of the ARC-Lake project. The dataset utilised consisted of 1990 time-series of surface temperature recordings, each with 405 bi-monthly recordings over the span of years 1995-2012. Exploration of functional time series reveals no significant trend over the period and seasonality is predominant. Although we acknowledge the potential for unobserved long-term trends given the limited dataset, this study focuses on the dominant seasonal variation in surface temperature. Seasonality was modeled using the Smoothing Spline Theorem with a saturated spline, where the optimal regularization parameter (λ) was selected to refit the spline, yielding a smoother functional curve that gives a better generalization. Due to higher dimensionality of the data Functional Principal Component Analysis was utilized, first two modes of variation were able to capture over 95% of the total variability thus were chosen to be retained and the resultant scores were utilized for clustering. Finally, clustering results reveal that there exists five distinct regions that depict different temporal patterns during a given season. This division highlights the influence of geographic location on lake temperature dynamics and it is safe to conclude that the Caspian sea. The analysis concludes revealing a strong seasonal coherence across the Caspian basin however, clustering revealed that its intensity varies by region due to local climatic or some external factors.

1. Introduction

Lake Surface Water Temperature(LSWT) is classified as an Essential Climate Variable which critically contributes to the characterisation of Earth' s climate (ESA, 2024). Many studies have been conducted to assess the impact of its variability on the biosphere and human life in particular. A study conducted for the period 1880-2007 indicates that the primary catalyst of

variability in global mean land temperature (GMLT) is sea surface temperature (SST) variability (Hoerling, Martin; Kumar, Arun -2008). Another study reveals that with each 0.6°C increase in SST, the maximum wind speed of hurricanes increases with 5 knots (Ahrens, 2009). LSWT significantly impacts aquatic life, affects biodiversity and water quality. A research about the increase in water salinity of Caspian Basins was conducted by Moslem, Yazdani (Nov, 2020) that revealed several factors, including climate change, oil extraction, and water diversion, disrupt the Caspian Sea's freshwater inflow, which worsens its freshwater quality by increasing salinity and creating environmental hazards, especially in its southern regions. Sea Surface warming is more swift in sea basins compared to an open ocean (Belkin, 2009). Assessing the impacts of global warming on local marine ecosystems demands fine-scale characterization of spatial and temporal SST dynamics. A Journal on marine systems highlighted the importance of considering crucial features of SST variation like timing patterns and duration of warm season as well as frequency of extreme hot SST days to analyze the sea surface warming dynamics particularly in the southern region of Caspian Sea (Omid Beyraghdar Kashkooli, Mohammad Ghadami ; 2019). The Caspian Sea level fluctuates on multiple scales, with short-term changes reaching 4 m, seasonal variations of 40 cm, and multi-year shifts around 3 m. Researchers pose concerns that global warming through changes in precipitation, evaporation and wind regime will have significant impact on the Caspian Sea level already at the end of the current century (H. Lahijani; S.A.G Leroy, 2023).

Other factors also affect the Sea’s ecosystem, an interesting article presented by Amin, Sadeqi (2004) shows concerns about the Caspian Sea undergoing Aral Sea Syndrome, which is often invoked when lakes dry up due to severe water overuse and mismanagement. Sadeqi in his article mentioned that this has been caused due to increased water consumption in agriculture, hydropower, industry, and urban needs. Additional geopolitical reasons like Russia recently constructing several dams on the Volga and Ural Rivers up north has restricted the freshwater inflow, and global warming has exacerbated the situation.

This study focuses on identifying regions of the Caspian Sea where lake surface water temperature (LSWT) exhibits similar temporal variations, i.e., coherence. The specific objectives are:

- **Identifying coherence in Lake Surface Water Temperature.**

Coherent regions exhibit similar temperature patterns, highlighting any unusual variations can signal local environmental disturbances or errors in measurement.

- **Group together zones with similar temperature patterns (clustering).**

This helps identify areas that behave similarly and may be influenced by common environmental Factors. It also helps policymakers to make conservation efforts uniquely for distinct regions and not treat the water body uniformly.

- **Map the identified groups onto the spatial grid**

Maps make the clustered results much more intuitive and visualising these clustered regions across the lake can pinpoint zones having any unusual variability.

The remainder of this dissertation is structured as follows:

Section 2 provides a detailed description of the study sites, highlighting the regional diversity and its importance on the ecosystem.

Section 3 describes the materials and methods employed. This includes the data preprocessing/transformation techniques, the application of Functional data analysis for time-series modelling, FPCA for dimension reduction and finally the goes over Clustering methodologies.

Section 4 presents and discusses the principal results of the study. It interprets the findings from the modelling technique incorporated, functional principal component analysis and the spatial clustering in the context of the posed research questions.

Section 5 concludes this dissertation by summarising the key findings, acknowledging the study's limitations, and proposing areas for further research.

2. Study Sites

The site selected for study is the largest lake on Earth, while most argue that the Black Sea has a larger surface area to it, as it is connected to the Mediterranean Sea via the Bosporus Strait, therefore it cannot be categorised as a lake which makes Caspian Sea win this title, as it is completely landlocked. The Caspian Sea is located about 27 meters below sea level. It covers an area of 371000 km^2 and the volume being 78200 km^3,which means it entails 40-44% of the world's total lake water volume. The average depth of the sea basin is 208m, whereas the deepest point measured is 1025m in the Southern Region of the lake. It shares borders with five countries - Turkmenistan to the east, Kazakhstan to the northeast, Russia to the northwest, Azerbaijan to the west, and Iran to the south. For contextual comparison, that lake alone is significantly larger than the entire land area of the United Kingdom.

Approximately 130 rivers flow into the Caspian Sea and contribute to this freshwater body. Europe's longest river, The Volga contributes the most with 80 - 90% of freshwater inflow. The morphology of the Caspian Sea floor is highly heterogeneous across its northern, central, and southern parts (Yazdani, Moslem; 2020). It’s home to diverse flora and fauna, with more than 162 fish species many of which are under the IUCN Red List of threatened species including the infamous Caspian seal (IUCN, 2008) and endangered species like sturgeons which are the fish that make the valuable Caviar. A historical estimate states that 90% of the world's total sturgeon population resides in the Caspian Sea while recent studies show that this number has considerably reduced due to overfishing (Dmitry Pokidaev, 2025).

This rich diversity, physical isolation and the immense scale of the Caspian Sea has created a unique ecological system additionally a hub of immense geopolitical and economic importance. Protecting and understanding it helps balance biodiversity conservation with human needs which makes it of great interest for our study. Figure 1 (Amin, Sadeqi ,2004) displays the location of the Caspian Sea on a world map.

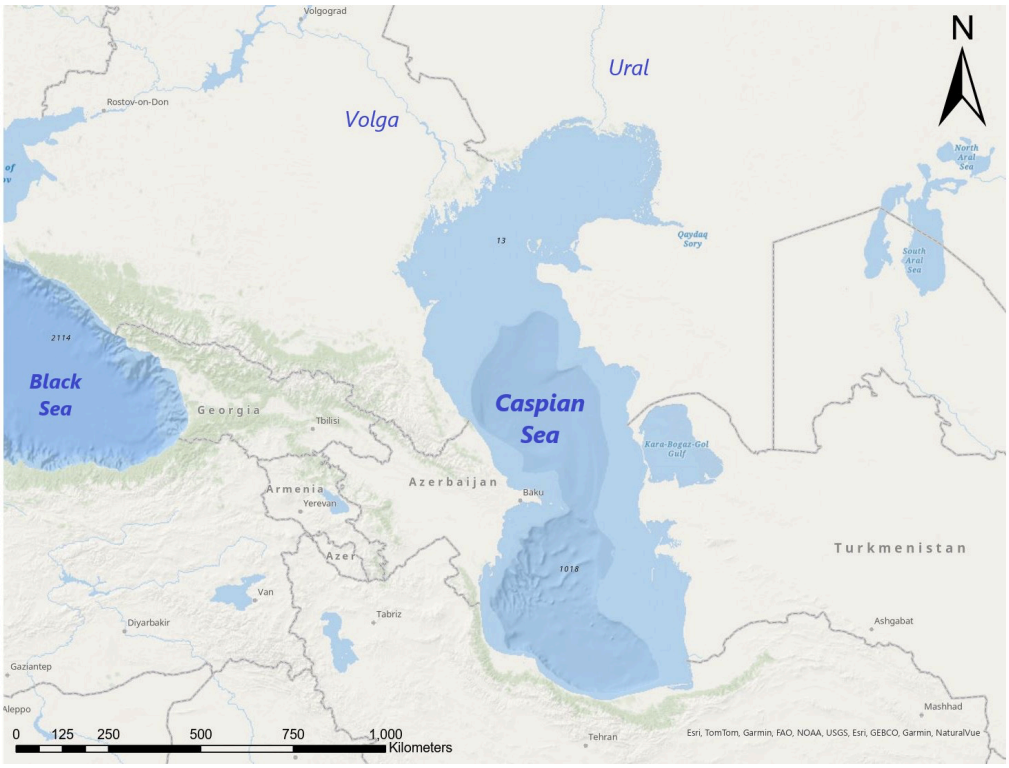

Fig:1 - Geographical location of Caspian sea

3. Materials and Methods

3.1 Data sets

The traditional way of measuring LSWT has been using in situ instruments, which are quite heterogeneous and irregularly distributed. This is where remote-sensing technology bridges the gap and provides continuous temporal series data. The product used in this study is Arc-lake v3 collected by the European Space Agency and has been processed/developed by the Earth Observation and Space Research Division at the University of Reading. The ARC-Lake project was to generate a consistent LSWT time-series from sensors (ATSR-2, AATSR - Along-Track Scanning Radiometer) for more than 250 lakes globally from 1995 to 2012, which is fully and freely available at https://www.laketemp.net/home_ARCLake/.

The dimension of the dataframe available under the "caspian.csv" file is 1990 x 408. Table 1 shows the characteristics of variables under study; it consists of a 1990 row of time series data of LSWT each row having bi-monthly temperature records (8th and 23rd of each month) from June 1995 to April 2012. The dataset also consists of geographic co-ordinates for these grid squares, X-coordinate representing longitude and Y-coordinate representing the latitude. Additionally these pairs of coordinates also map with a unique numeric variable named 'Cellids', this would help uniquely identify a location.

Table 1: Description of the dataset, including variables, units, and description

Variable	Type	Description
Cellids	Numeric	Unique identifier for each spatial location
X & Y	Numeric	Geographical coordinates of the pixel.
Temperatures (datetime columns)	Numeric (floating point)	Bi-monthly temperature values (June 1995-April 2012)

Each of the 1990 time series corresponds to a grid square also known as pixels, each covering approximately three square kilometers of area within the Caspian region. A cross-section of the northern region can be found in Figure 2, where each blue circle corresponds to one grid square.

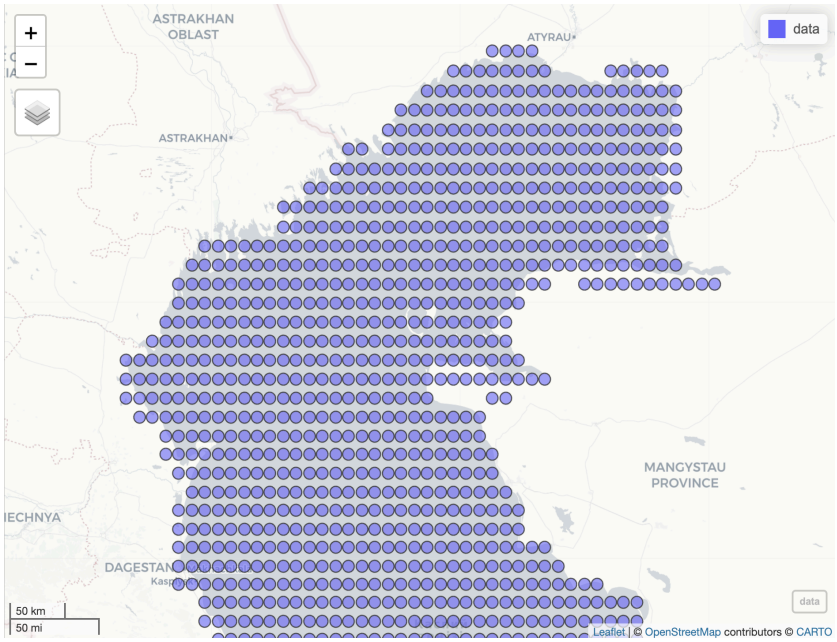

Figure 2: Grid Squares/Pixels spanning the Caspian Region

3.2 Exploratory Analysis

The data is recorded over a continuous variable - time, which is considered to be a smooth process, additionally each pixel having bi-monthly values for approximately 17 years of data makes it a rich and high-dimensional dataset. Thus, it has all the characteristics to be categorised as a functional dataset. Therefore we take a functional data analysis approach to model this

underlying dataset. We utilize the *fda* package in R because it's specifically designed for functional data analysis, turning raw time series into mathematically smooth curves, which would allow us to study the shape, variation, and structure of LSWT patterns.

Initially, the dataset was examined for missing values. No missing entries were identified, and therefore the analysis proceeded without the need for data imputation or cleaning.

The dataset was then transformed using two major steps in preparation for further analysis-

Step-1: A subset of 50 pixels using *simple random sampling* were taken for better graphical interpretability and efficient computing. As shown in Figure 3, since the subset is spatially distributed across most regions of the lake, it is reasonable to conclude that the randomly distributed subset would provide a fair representation of the overall characteristics of the Caspian Sea. This strategy is valuable and commonplace in high dimensional geospatial datasets, where full-resolution analysis can be computationally expensive, while only offering limited additional interpretive benefit.

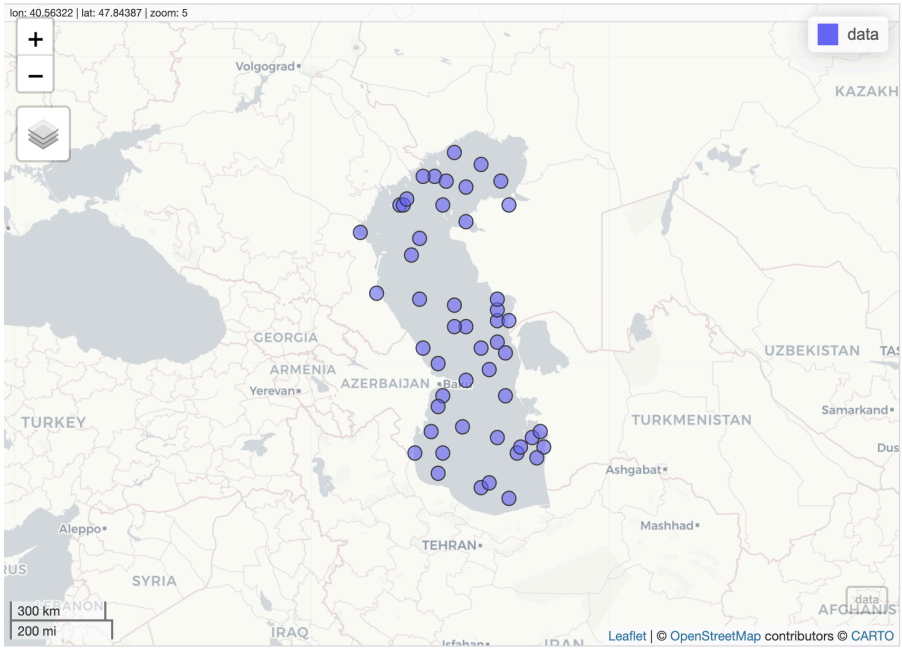

Figure 3: Fifty locations were chosen as a subset, spread fairly evenly across all regions of the Caspian Sea

Step 2: The FDA package in R and functions within the package expects time points in rows and different pixels in columns, however our raw data has it the other way around. Therefore, we had to transpose the existing dataset and only select columns 4 to 408 excluding cellids and X and Y co-ordinate columns to make the data compatible with the package requirements. Therefore, moving forward, this transformed subset will serve as the basis for subsequent analyses.

Figure 4 illustrates a functional time series plot of LSWT for the 50 pixels chosen, providing a visual understanding of temporal variation patterns over time.

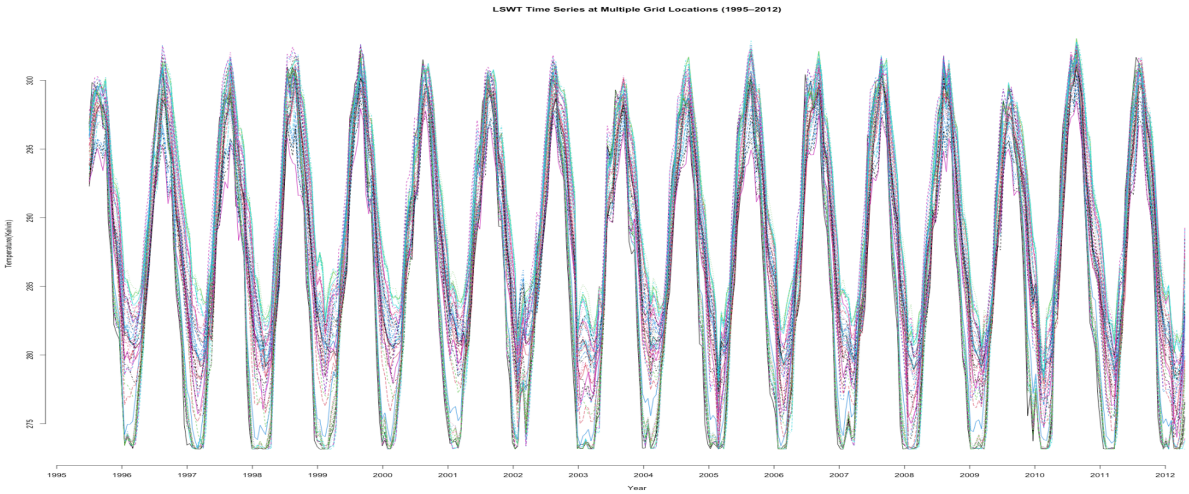

Fig 4: Time Series plot of temperature in Caspian Sea depicting dominant seasonal variation for 1995 to 2012

The plot reveals underlying characteristics, the peak represents the summers and troughs represents the winter period. The cycles repeat very consistently, suggesting a **strong annual periodicity** in the Caspian Sea's surface temperature. Hottest summers were experienced in the summer of 2010 with temperatures soaring up to 301 K that's roughly 28° C, depicting a stark contrast to the preceding year, where summer highs were considerably lower. Figure 5 depicts the overall mean temperature for the given time series shown in red line. Chilliest winters were observed in the winters of 2004 and start of 2005, where overall mean temperatures for selected locations under study dropped to 273 K, equivalent to 0° C.

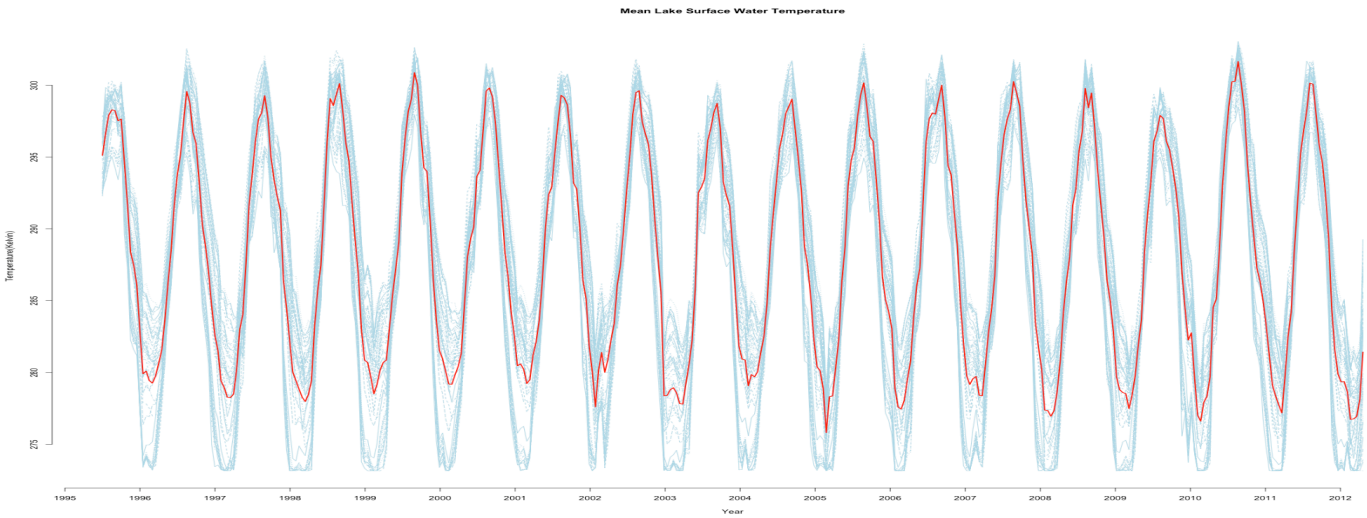

Fig 5: Mean LSWT of Caspian Sea in Red line imposed on the TS plot in blue for 1995 to 2012

Initial examination of the time series plots suggests that the data is primarily driven by a strong seasonal component, while there is no visual evidence of a significant long-term trend. However, the possibility of an unobserved trend cannot be dismissed given the limited dataset. Hence, the scope of this subsequent analysis will focus only on exploring the dominant seasonal variability. The years selected for this exploratory analysis were chosen with seasonality patterns in mind, as they depicted notable variations across seasons that made them suitable candidates for study, whereas year 1996 was taken as the starting point, as it represents the first year in the dataset with a full set of bi-monthly temperature records, unlike 1995 which begins mid-year. Taken together, these years provide a balanced representation of both earlier and more recent seasonal patterns.

Figure 6 shows a TS plot of the seasonal variability year 1996 (topleft), 2005 (topright), 2009 (BottomLeft), 2010 (BottomRight). It can be seen that the summers have broader peaks in recent years compared to that in 1996. The red dotted line at 295 K is plotted just for better comparative study through which it can be observed that all locations/pixels displayed significantly higher temperatures during summer season 2010. For the year 2005 it can be seen that there is a drastic temperature drop during the start of the year. Additionally, some locations displayed rise in temperatures around day fifty for season 1996, whereas for later years those locations tend to have delays in temperature rise, with it getting warmer around the fortnight of March.

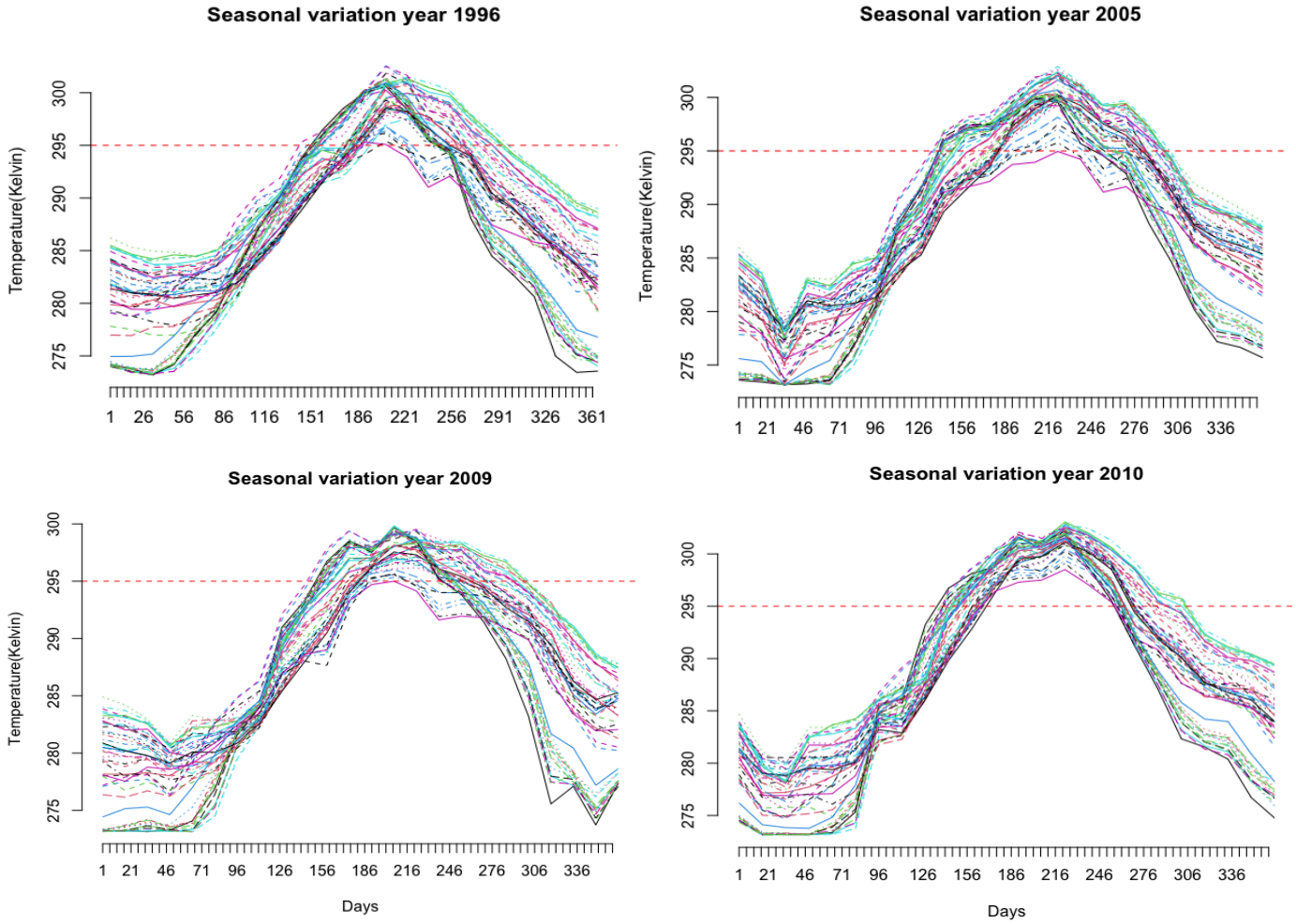

Fig6: TS plot of seasonal variability comparing 1996(topleft),2005(topright), 2009(BottomLeft),2010(BottomRight).

3.3 Methodology

The methodology applied is based on Basis Expansion modelling technique which are utilised when the data represents a non-linear structure. As our data represents seasonal oscillation this method would be best suited for the case.

$$y_{ij} = x_i(t_{ij}) + \varepsilon_{ij}$$

Here, y_{ij} represents observed temperature at pixel i and time j; t_{ij} represents continuum in time and $x_i(t)$ is smooth functional data which can be represented as:

$$x(t) = \sum_{k=1}^{K} \beta_k \cdot \phi_k(t)$$

Where $\phi(t)$ is a basis system for x and β_1, β_2,.....β_k are the coefficents of expansion which is a vector of k x 1. For this research we will utilize the B-splines basis system as it captures any non-repeating anomalies better than other basis systems.

We will briefly go over some essential information about B-Splines. Splines are polynomial segments joined end to end constrained to be smooth at join which are called knots. Order of a spline is equals to the degree of polynomial added 1, i.e. for a quadratic polynomial the order would be 3. Derivatives are continuous only up to Order - 2 in a spline system. This makes cubic splines a popular choice as second order derivative is continuous and we'd utilize the same.

Number of basis functions = order + number of interior knots

An example can be seen in Figure 7 (left) having 3 interior knots and order of polynomial is 2, whilst the Figure 7 (right) represents Cubic spline with 8 number of basis functions, the increase in the number of interior knots represented red vertical lines can be clearly observed.

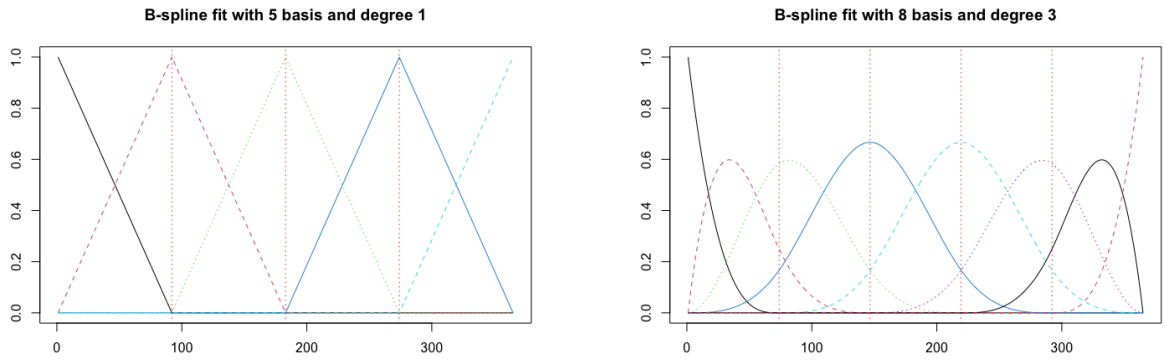

Figure 7: Illustration of different order of spline and number of basis: Spline with 5 basis and order 2 (left), Spline with 8 basis and order 4 (right)

Number of basis function chosen, affects the level of smoothing. Higher the number of basis function, lower the bias i.e it fits the data points more accurately and thus resulting in a

greater variance. Figure 8 depicts the above theory graphically, the scatterplot of temperature data for one location(pixel) with varied number of basis function. A linear polynomial with 5 basis function(TopLeft), cubic polynomial with 8 basis function(TopRight) and finally the plot(Bottom) is a cubic polynomial with saturated basis. A saturated B-spline basis is achieved when you have the maximum number of independent B-spline basis functions possible for a given degree and knot vector. Therefore it becomes crucial to find optimal number of basis function to optimize this Bias-Variance trade off.

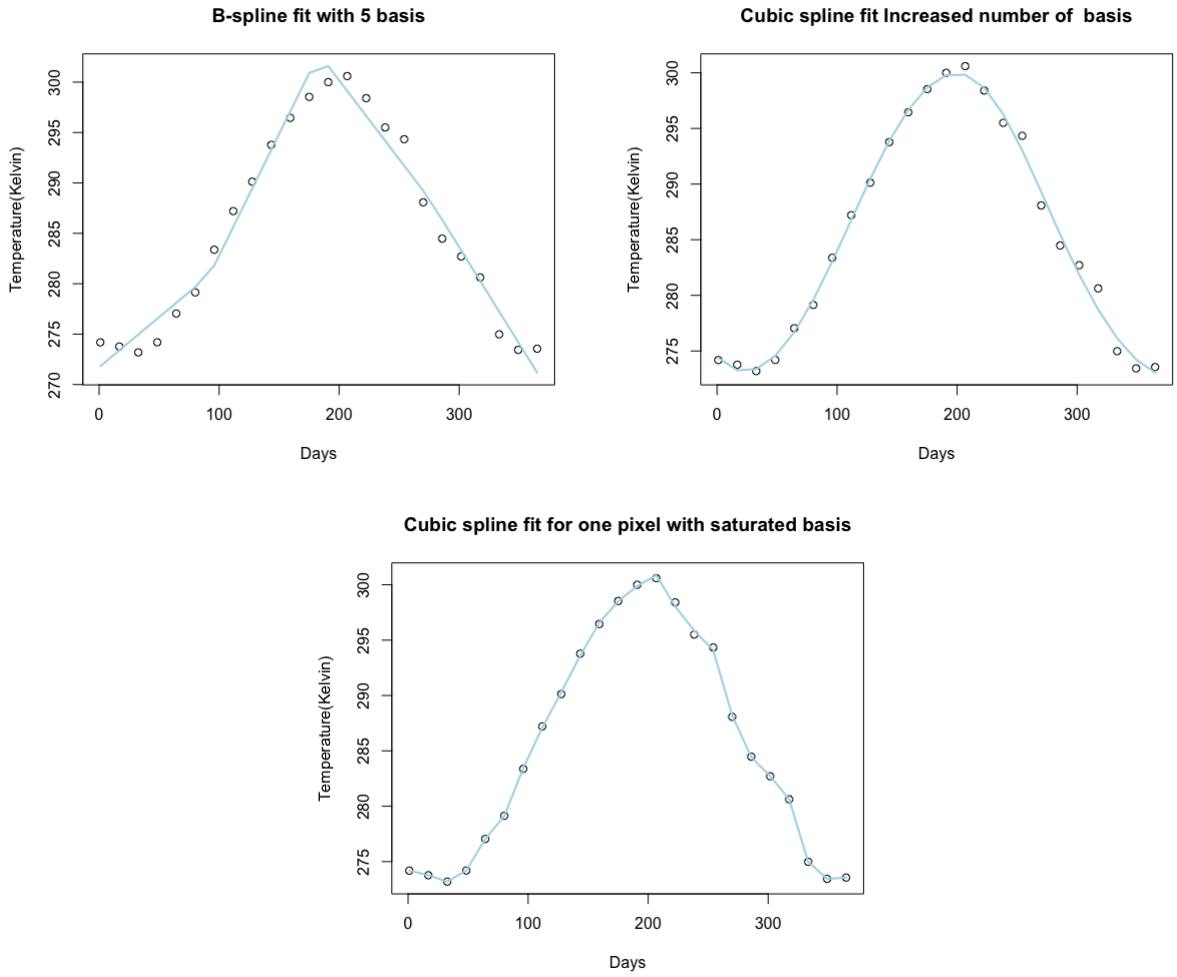

Figure 8: Varied fit on Temperature data for one location for 1996: Linear polynomial with 5 basis function(Topleft), cubic polynomial with 8 basis function(TopRight) and finally the plot(Bottom) is a cubic polynomial with saturated basis

In order to find a balance between this Bias variance tradeoff, to find the optimal number of basis function we'd minimize the Integrated Mean Squared Error.

$$\text{IMSE}\left[\hat{x}(t)\right] = \int \text{MSE}\left[\hat{x}(t)\right]\, dt$$

where $\text{MSE}\left[\hat{x}(t)\right]$ is the sum of variance and squared Bias term-

$$\text{MSE}\left[\hat{x}(t)\right] = \text{Bias}^2\left[\hat{x}(t)\right] + \text{Var}\left[\hat{x}(t)\right]$$

For our research, we utilize a method called Smoothing Spline Theorem (Paul Eilers and Brian Marx, 1996). Here we decide the number of basis beforehand to be as many as the knots making the model complex which increases variability. Now that we have a complex model we prevent this overfitting by introducing a penalty term. It reduces the variance significantly thus the resultant model generalises the data better. We add a penalty to the least square esitmate, which is defined as follows-

$$\text{PENSSE}_\lambda(\mathbf{x}) = [\mathbf{y} - \mathbf{x}(t)]^T[\mathbf{y} - \mathbf{x}(t)] + \lambda J[\mathbf{x}]$$

where:

- $J[\mathbf{x}]$ measures the 'roughness' of $\mathbf{x}$ which is integral of squared term of second derivative of $\mathbf{x}(t)$ defined as -

  $$J[\mathbf{x}] = \int [D^2\mathbf{x}(t)]^2 dt$$

  Second derivative D^2 measures curvature and squaring it penalizes positive and negative curvatures equally and finally integrating it over 't' penalizes total roughness across the entire domain. So, if the second derivative is zero the function is a straight line - perfectly smooth, and if the second derivative is a large value then curve has more wiggles. In FDA package in R we perform this using - int2Lfd(2) where Lfd refers to linear differential operator and (2) refers to the second derivative.

  "A remarkable theorem from variational calculus states that the function that minimizes the Penalised SSE is a natural cubic spline with knots at every data point." (Nychka, D., 1995) Hence, our choice for this paper.

- λ is a non-negative *regularization parameter* that help achieve balance between the SSE part and the penalty part within the PENSSE equation above.

The behavior of the regularization parameter λ can be summarized as follows:

- The function $\mathbf{x}(t)$ is tends to be smoother as λ tends to ∞, i.e. the roughness penalty dominates.
- The function $\mathbf{x}(t)$ tends to fits the data more closely as λ tends to 0, i.e the penalty term has less effect.

Finding the optimal value for smoothing/regularisation parameter λ can be challenging so we utilize generalized cross-validation (Wahba, 1990) method, we compute that using -

$$\text{GCV}(\lambda) = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{\left(m - \sum_{i=1}^{m} h_{ii}\right)^2}.$$

Where, h_{ii} are the diagonal elements from the hat matrix H of order m x m. We consider the value of λ to be the best when it minimizes $\text{GCV}(\lambda)$.

Finally, this optimal value of lambda is used to refit the spline and the resultant smoother function generalizes the data in the best possible way given the trade off.

Having addressed the methodology for the first objective, the following sections outline the methodological approaches employed to solve for the remaining objective i.e. grouping together the regions that depict similar temperature patterns. A sequential procedure was adopted to address the same.

1. Dimension Reduction – FPCA
2. Clustering Algorithm – K-means clustering

Functional Principal Component Analysis

Initially, as we have a higher-dimensional dataset, Functional Principal Component Analysis helps reduce the data into smaller principal scores that capture the main modes of variation. We have functional data $X_i(t)$ i.e our smoothed temperature curves obtained. In PCA we decompose the covariance matrix Σ, similarly for FPCA we have a surface $\sigma(s,t)$ instead. For functions we utilize Kosambi Karhunen–Loève decomposition (Surajit, Ray).

$$\sigma(s,t) = \sum_{i=1}^{\infty} d_i \, \xi_i(s) \, \xi_i(t)$$

Where, d_i represents the amount of variation in direction $\xi_i(t)$. The first few d_i tend to explain majority of the variation and the proportion of variation explained is given by-

$$\frac{d_i}{\sum d_i}$$

FPCA decomposes each function $X_i(t)$ into a mean plus orthogonal functional components which is known by Karhunen–Loève expansion -

$$X_i(t) = \mu(t) + \sum_{i=1}^{\infty} \mathbf{f}_{ij} \xi_i(t),$$

Where, $\mu(t) = E[X(t)]$ and $\mathbf{f}_i$ are the PC scores and $xi_i(t)$ are the principal components.

This technique utilizes finding a probe $\xi_1(t)$ that maximizes

$$\text{Var}\left[\int \xi_1(t) x_i(t) \, dt\right].$$

But we need to constrain-

$$\int \xi_1(t)^2 \, dt = 1.$$

For $\xi_2(t)$ we need to maximize variance subject to the orthogonality condition, orthogonality is required as it allows each components contribution to be independent -

$$\int \xi_1(t)\,\xi_2(t)\,dt = 0.$$

FPC Scores are projections given by:

$$\mathbf{f}_{ij} = \int_{\mathcal{T}} \{X_j(t) - \mu(t)\}\,\xi_i\,dt,$$

These scores summarize the main patterns of variation in the data and can be used in subsequent analyses such as clustering. So instead of clustering full functions, we cluster the score vectors, for which we utilize Kmeans-clustering algorithm.

Partition Clustering Method - K-Means

Kmeans is an unsupervised machine learning algorithm. An unsupervised algorithm refers to a technique where algorithms discover hidden patterns or structures for an unlabelled dataset without any human supervision. Kmeans is a clustering method where the data points are clustered randomly at the beginning into a certain number of disjoint subsets and then the observations are relocated cluster to clusters. The required number of distinct clusters denoted by 'K', needs to pre-determined and the algorithm allocates each observation to one of the distinct sets. We'd utilize the default algorithm within the Kmeans function in R which aims to minimize total within-cluster variation also known as Hartigan-Wong algorithm (UofG, DMML 2024) defined as:

$$\begin{aligned} W(C_k) &= \sum_{\boldsymbol{x}_i \in C_k} d_E(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_{C_k})^2 \\ &= \frac{1}{2|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_k} d_E(\boldsymbol{x}_i, \boldsymbol{x}_j)^2, \end{aligned}$$

where

- $\boldsymbol{x}_i$ is an observation belonging to cluster C_k;
- $\bar{\boldsymbol{x}}_{C_k}$ is the average of all the points belonging to cluster C_k, known as the *cluster centre* or *cluster centroid*;
- $|C_k|$ represents the number of observations in the k-th cluster; and
- $d_E(\boldsymbol{x}_i, \boldsymbol{x}_j)^2$ is the squared Euclidean distance between the p-dimensional continuous variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$d_E(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 = \sum_{l=1}^{p} (x_{il} - x_{jl})^2.$$

The above within-cluster sum of squares $W(C_k)$ quantifies the amount by which observations within a given cluster varies from one another. Ideally, a cluster is classified as good when

$W(Ck)$ is minimized. Now we sum this up over all K clusters which gives the total within-cluster variation given as follows-

$$W(C) = \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_k} d_E(\boldsymbol{x}_i, \boldsymbol{x}_j)^2.$$

As, we need to specify the number of clusters beforehand, we utilize an elbow method which is a line plot with the number of clusters on the horizontal axis and the total within cluster sum of squares value on vertical axis, we essentially choose the value where we observe a bend shaped like an elbow thus its name, the choice made using this method is subjective to different individuals. There's another popular alternative called silhouette plot that helps decide an optimal number of clusters but for this research we stick with the prior method.

The starting point for choosing the cluster centers in Kmeans are sensitive to the clusters obtained so it's important to incorporate an algorithm that takes this into account and which is what Nstart argument in kmeans function does, it runs the algorithm from multiple random starts and then selects the "best" run as the final result. Whilst there is no fixed guideline for the required number of random starts, and hence it is impossible to be certain whether the global minimum of the optimization has been found. However, generally, as the number of variables increases, a greater number of random starts is necessary.

4. Results and Discussion

The bi-monthly temperature observations were represented successfully as smooth functional curves by the Penalised B-Splines approach. The cross validation technique GCV, helped find the optimal choice for regularisation/penalty parameter that led to accurately finding a compromise between overfitting and underfitting. The value of lambda that minimizes GCV is computed to be ninety nine as shown in the generalized cross validation plot of Figure 9.

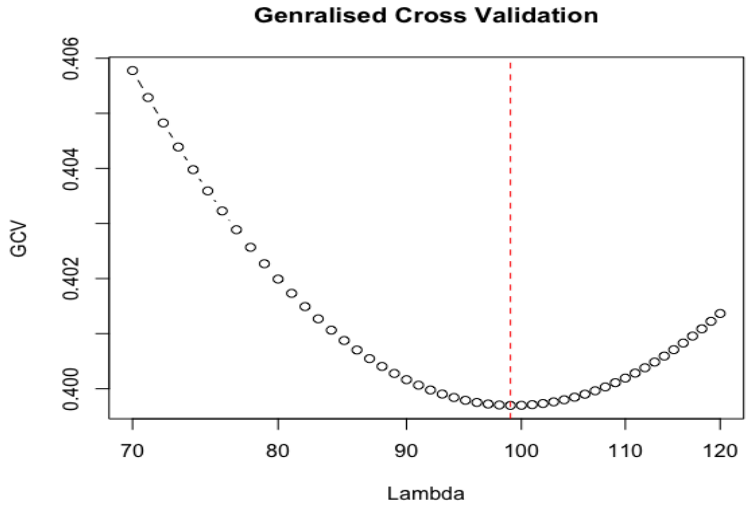

Fig 9: Generalised cross validation plot to find optimal regularisation parameter, lambda equals 99.

In Figure 10(left), it shows the saturated spline fit with knots placed at each bi-monthly data point; the curves seem to have too many wiggles signaling overfitting. The optimal value for lambda managed to obtain a smooth functional curve as shown in Figure 10(right). One can observe better overall smoothness, particularly at the endpoints and peaks, resulting in a robust generalization of the underlying pattern.

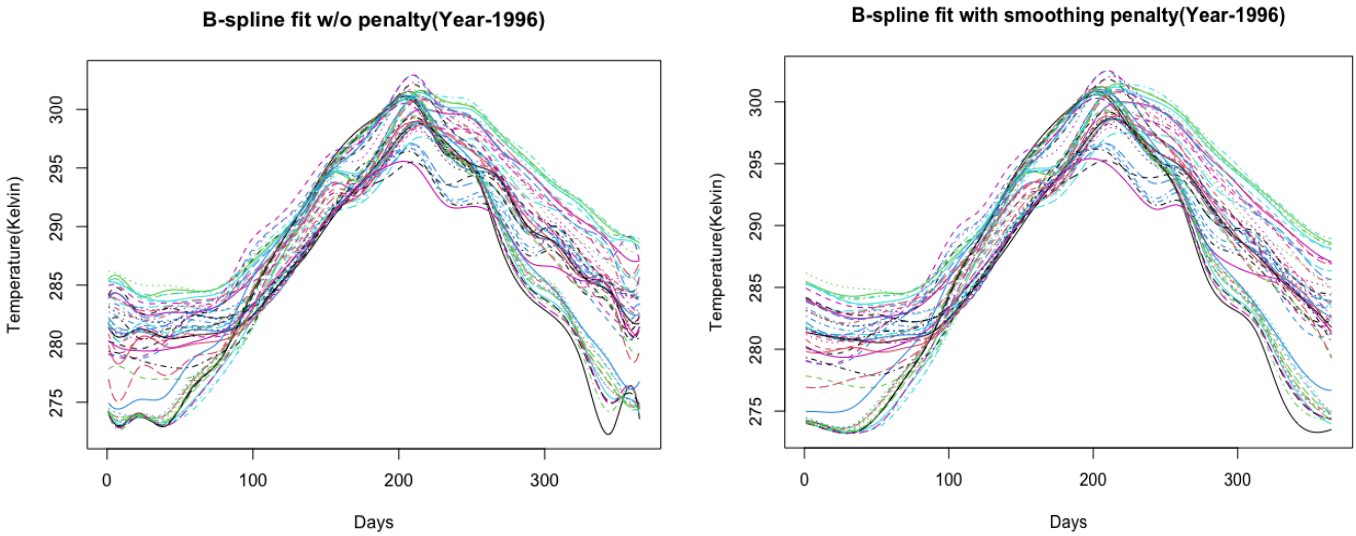

Fig 10: Saturated splines fit without regularisation parameter(left), Splines fit after regularisation(right)

Figure 11, the curve in red depicts the average pattern of smoothed temperature curve over 1996 representing the central tendency in the functional space.

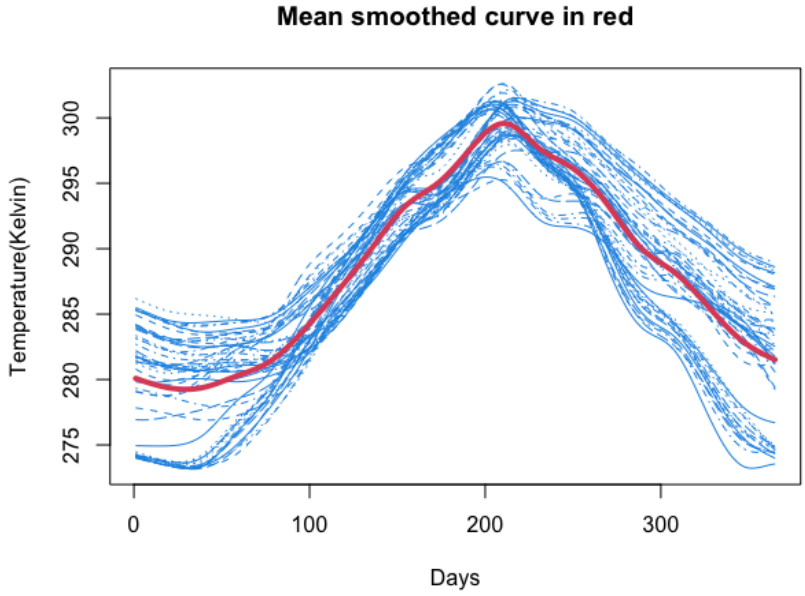

Fig 11: Mean of smoothed functional Time Series of temperature in red

The correlation plot, Figure 12 reveals high correlations along the diagonal represented in dark red which indicates that days that are nearby tend to vary together but as we move further, for

example day 25 when compared with day 180 show weak correlation in light yellow, in context to our problem it means winters are not varying with summers together

.

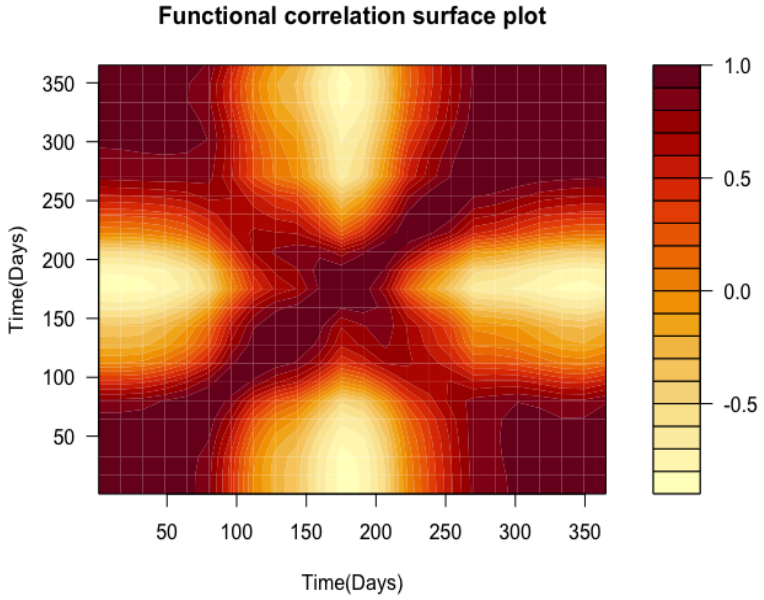

Fig 12: Functional Correlation filled contour plot representing high correlation along diagonal.

Now we move to discuss results for our next research objectives, FPCA results are as follows- The proportion of variance explained by the first two principal components turned out to be approximately 95% represented by the red line in Figure 13 (left) as they provide sufficient summary of how the temperature curves differ from one another.

The Figure 13 (right) shows the two principal components, primary modes of variation. A clear annual cycle can be seen, the black line is the first principal component, capturing about 82% of the total variability. It's positive for roughly the first 120 days and approximately the last 150 days of the year, and negative in the middle whereas the second principal component is represented by a red dashed line which is capturing around 14% of the variability, depicting a stark contrast from PC1.

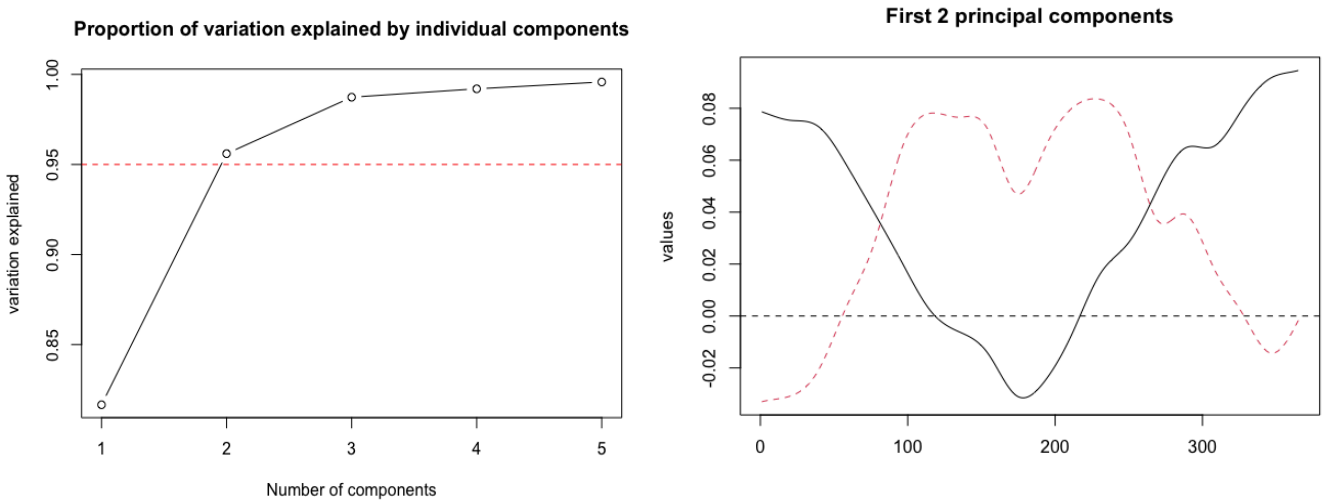

Fig 13: Variation explained by the number of components(left), PC plot for first two components(right)

Additionally, In Fig 14 the solid line represents the mean seasonal temperature curve and (+) and (-) lines depict variations of that PC about mean. For PC1, the width at both the ends i.e start and end of the year is high, the upper + indicating temperatures are warmer than average throughout the entire year and lower - indicating temperatures are lower than average almost the entire year except 100 to 200 day mark.

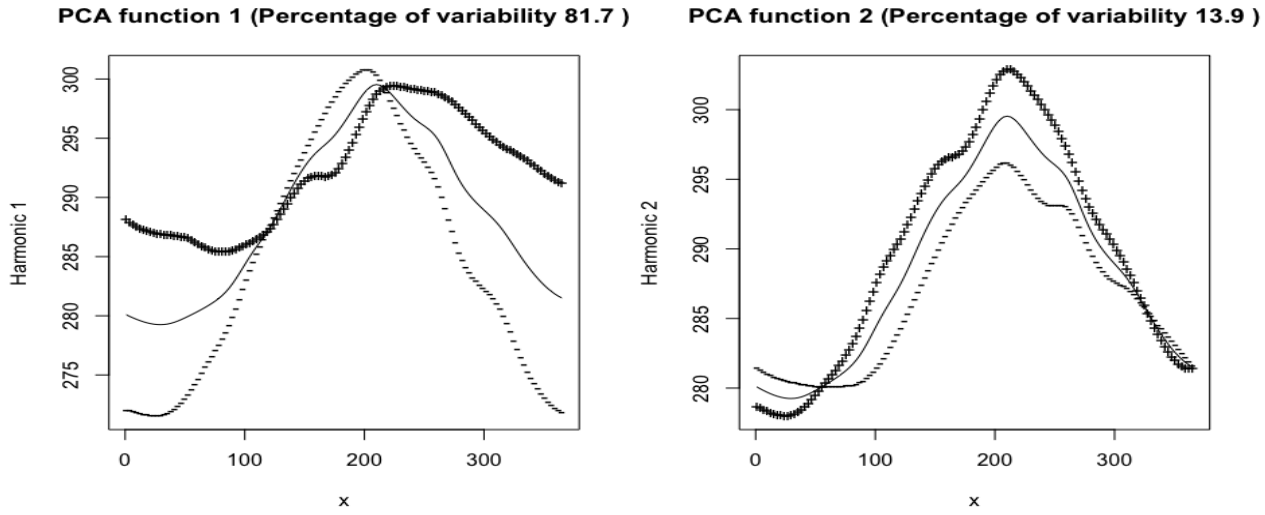

Fig 14: Harmonic Plot for: PC1(left) explaining 81.7% variability and PC2(right) explaining 13.9% variability

Since the choice of initial cluster centers is arbitrary, we set the 'nstart' argument within the Kmeans function to take value 25. Given the moderately large number of variables in our dataset, 25 random initializations should provide a sufficient balance between computational efficiency and reliable optimization.

The result for the optimal value of K is shown using an elbow plot in Figure 15. Based on the elbow method, our initial choice for the number of cluster centers is three as no substantial reduction in the total within-cluster sum of squares is observed beyond this point, and the elbow structure appears sufficiently distinct. Alternative choice could also be five clusters, as beyond that mark, the total within-cluster sum of squares seems to converge, the curve flattens out to a straight line. Although the selection of cluster numbers involves a degree of subjectivity, a more informed understanding can be obtained by exploring the clustering results on both the spatial grid and the functional time series plots.

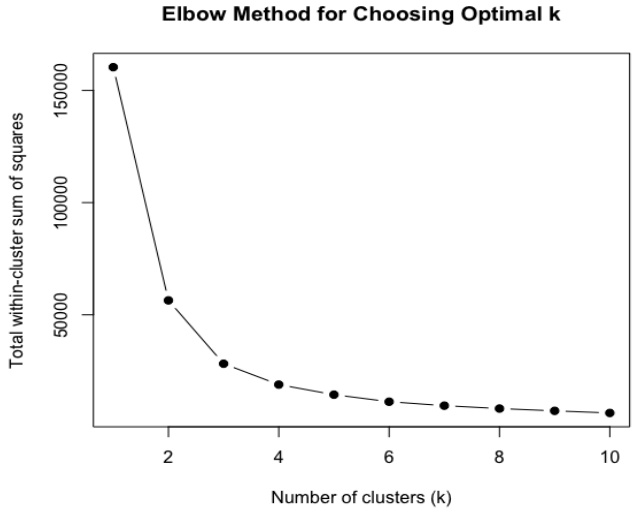

Fig 15: Elbow method for choosing optimal number of clusters with significant bend at K equals 3 and 5

The resultant map on Figure 16 shows spatial zones with 3 cluster centers. Three distinct zones can be observed, Northern part colored in orange, central zone in blue and a southern zone in green. This result backs the findings obtained by Yazdani, Moslem in his paper "Geological Position of The Caspian Sea With Emphasis on The Risk of Increasing Water Salinity" in 2020. Although, some observations in the south eastern zone seems to be categorised alongside the central, colored blue. We utilize a clustered time series plot to further look for any irregularities.

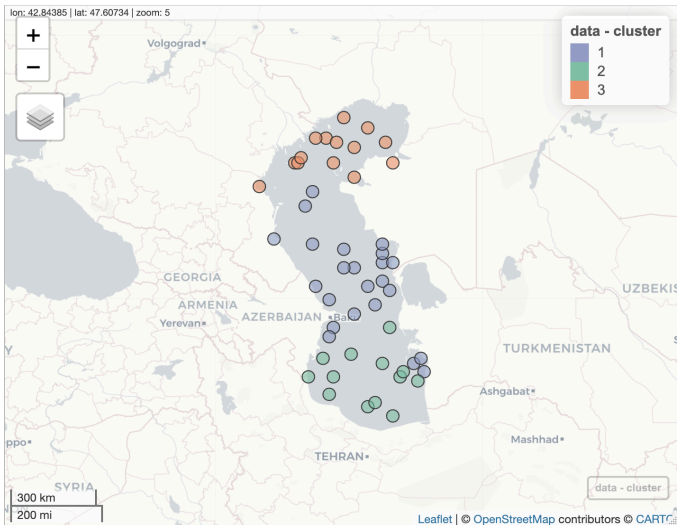

Fig 16: Spatial Map of Caspian with observations having clustered in 3 distinct regions, north, central and south

Figure 17 (left) reveals the smooth functional time series plot categorised in three clusters. It can be seen that the observations in the northern region(orange) show colder temperatures with the curvature ends being lower than others and so for the summer highs. Whereas the Southern region in green shows warmer temperatures throughout which is expected as it's closer to the equator. Although some curves within the central region(blue) tend to stand apart from the majority where few curves hit the highest peaks and few others contradict with the lowest peaks during summer highs. That calls for looking at different numbers of cluster centers. Figure 17 (right) displays the functional time series plot with 5 cluster centers which reveals much clear and distinct curves with the southeast having the highest peaks during summers and the Central East having lowest highs during summer period.

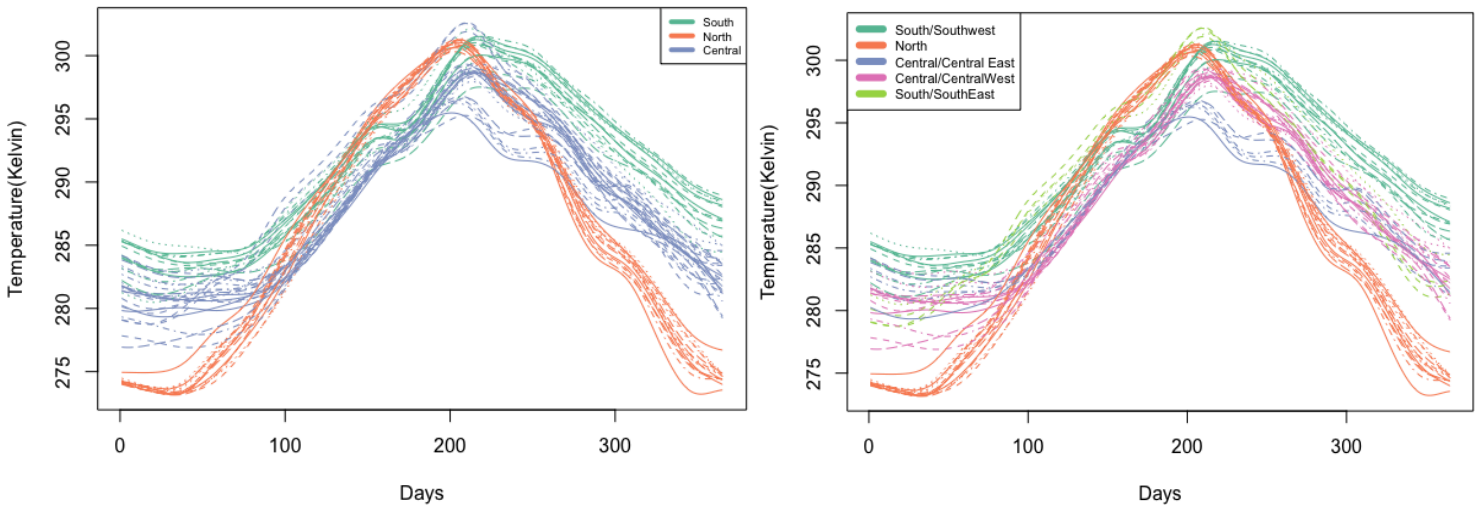

Fig 17: Smoothed functional curve clustered with 3 clusters(left) and 5 clusters(right)

The spatial map on Figure 18 also shows better clustering; one can notice that the few initial clusters in the south eastern regions that were categorised under the central region previously now have their own cluster center, colored with pink depicting different temporal patterns in that region. Additionally the central region now has two distinct regions within it, with east represented in blue and the west represented in orange.

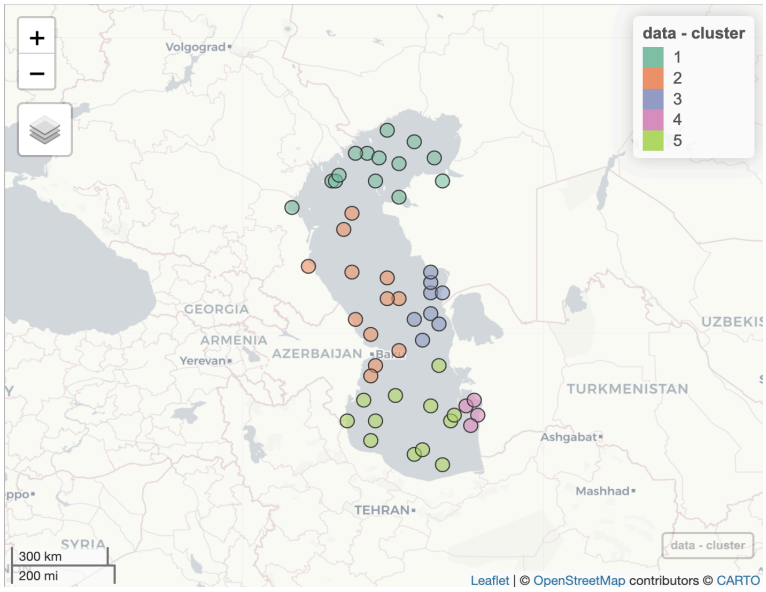

Fig 18: Spatial Map of Caspian with observations having clustered in 5 distinct regions

5. Conclusion, Limitation and Future Work

5.1 Conclusion

Lakes are freshwater sources with no exchanges to open waters that's why even minor variability could cause implications to its rich ecosystem. Consequently, detailed analysis is critical to protect and sustain their ecosystem. The main aim of this study was to characterize dominant features of spatiotemporal variability patterns of LSWT in the Caspian Sea and cluster the regions with similar temporal behavior based on the satellite data available to us for the period June 1995 to April 2012.

Our main findings in this study can be summarized as follows:

- Modelling Seasonality using the Smoothing Spline resulted in a smooth functional curve that generalizes the variability. The strong similarity and coherence across different chosen locations suggest a **consistent annual seasonal cycle**.
- FPCA of the smooth functional curves revealed that the first two modes of variability captured **95% of the total variability**, highlighting that most of the seasonal variation can be explained with just a few dominant patterns.
- Clustering with five centers provided clearer and more **distinct temporal patterns**. The south region exhibited the highest peaks during summer, southeastern to be particular,

and the central region split into two distinct clusters: central-east with lower summer highs and central-west with intermediate patterns, whilst the northern regions showed colder temperatures with lowest curvature near the ends.
- **Intra-cluster coherence is more profound** than inter-cluster coherence.

Identifying such coherent zones is valuable for understanding regional climate impacts and help guide policies reducing the severity of climate-related consequences in regions likely to face unavoidable impacts.

5.2 Limitations and Future work

- Sampling Technique - This study utilized simple random sampling technique to obtain a subset of 50 square grid locations. While this technique provides an unbiased sample, it is susceptible to sampling variability. Future work could employ other alternative sampling techniques like **stratified random sampling** or **systematic sampling**.

- Exclusion of long term Trend Analysis - The scope of this project was focused on modelling the dominant seasonal variability rather than long-term temperature trends. The **Seasonal Mann-Kendall Trend Test** detects overall trends by analyzing each season separately to account for seasonal variability. Code for which can be found in the appendix section of R file, performed solely for personal sanity, as it was beyond the scope of this report. Results obtained align with findings in existing literature about the Trend patterns of Caspian Sea. That code could be utilized as a base for building further.

- **Smoothing Penalty Term** - This report utilized int2lfd(2) as a penalisation term for smoothing the saturated spline, i.e penalising the second derivative, another popular alternative is harmonic penalty that could potentially give better generalisation, avenue for future work.

- **Modelling seasonality for different years** - This report only modelled seasonality for 1996, to serve as a detailed case study within the project's time constraints. Other years could be modelled using the same approach as future work.

References

1. Hoerling, Martin, Kumar, Arun, Eischeid, Jon, Jha, Bhaskar, 2008, What is causing the variability in global mean land temperature? , Geophysical Research Letters, vol, 35,pp:1-5.

2. Belkin, I.M., 2009. Rapid warming of large marine ecosystems. Prog. Oceanogr. 81, 207–213.

3. Aherns, Donald, C., 2009, Meteorology today: An introduction to weather, climate, and Environment, Brooks/cloe.

4. Yazdani, Moslem. (2020) "Geological Position of The Caspian Sea With Emphasis on The Risk of Increasing Water Salinity," n.d.

5. Amin Sadeqi (2024). *Is the Caspian Sea Going to Experience Aral Sea Syndrome?* Hydrological Sciences.

6. Omid Beyraghdar Kashkooli; Mohammad Ghadami (2019) "Spatiotemporal variation of the southern Caspian Sea surface temperature during 1982–2016" , Journal of Marine Systems 193(2019) pg:126-136.

7. H. Lahijani; S.A.G Leroy;(2023) "Caspian Sea level changes during instrumental period, its impact and forecast: A review", Earth Science Reviews, Vol 241

8. Merchant, Christopher and MacCallum, Stuart (2018): Lake Surface Water Temperature ARC-Lake v3 (1995-2012). University of Reading. Dataset. https://doi.org/10.17864/1947.186

9. Paul Eilers and Brian Marx, 1996: Flexible Smoothing with BSplines and penalties.

10. Nychka, D., 1995. *Smoothing spline theory*. University of Wisconsin–Madison, Department of Statistics.

11. Surajit, Ray., Functional PCA. University of Glasgow, Department of Mathematics and Statistics.

12. UofG, Data Mining and Machine Learning. Partitioning Cluster Analysis: Week 7

13. R-project.org. (2025). *R: Seasonal Mann-Kendall Trend Test*.

14. European Space Agency(ESA) Climate Office. (2024). *What is an Essential Climate Variable?*

15. Dmitry Pokidaev (April 2025) Caspian Sturgeon Population Declines 90% Amid Ecological Crisis, The times of central Asia

16. International Union for Conservation and Nature (IUCN , 2008)