Applying functional data analysis and clustering methods on weather forecast data in the U.S.

Chuyuan Lin[*] Ying Yu[*] Yifan Wu[*] Jiguo Cao[*]

Abstract

The objective of this study is to investigate the potential covariates correlated to the weather prediction performance in the U.S, especially to explore the spatial and time effects in the prediction accuracy. We performed the functional principal component analysis (FPCA) and time series clustering techniques to divide 50 U.S. states into clusters. Cluster-specific characteristics of weather prediction performance were visually detected and cluster-to-cluster differences were quantified in order to identify this most and least predictable U.S. states. Then we conducted a functional analysis to capture the main pattern of variance in the prediction error over time and further investigate how other weather-related variables correlate with the prediction accuracy.

Key Words: Weather Prediction, Functional Data Analysis, Functional Principal Component Analysis (FPCA), Functional Linear Regression, Time-series Clustering

1. Introduction

1.1 Background

Various human being activities, such as agricultural, fishery, industrial production or daily traveling, are affected greatly by the climate events and weather variations (Adams et al. [1990]). Teisberg et al. commented that the consumption of the U.S. electricity is significantly correlated with the local temperature (Teisberg et al. [2005]). Accurate weather forecasts usually provide a tremendous help and instruction to the preparations of the weather-sensitive industries and activities. Modern climatology and weather forecast techniques focus on predicting upcoming weather conditions based on current climate measures, such as temperature, humidity and air pressures, etc. Despite advances in meteorology and satellite technologies, there are still significant uncertainties in weather forecasts. Bauer and Thorpe (Bauer et al. [2015]) suggested that the understanding of the climatic process and the input of statistical expertise are equally important to reduce these uncertainties. Then a big question arises to statisticians is how to improve the accuracy of weather forecasts based on a comprehensive statistical analysis and modeling.

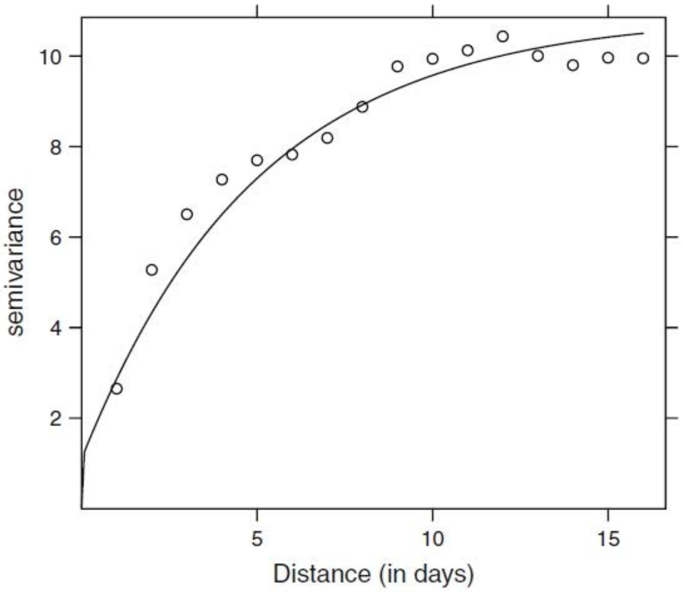
During the past decades, different statistical models were developed to achieve higher resolution of spatio-temporal predictions with higher accuracy. In the 20th century, two popular temperature forecast models were parametric SARIMA model (Box and Jenkins [1976]) and non-parametric kernel predictor (Collomb [1983] Györfi et al. [1989] Bosq [1996]). In 2000, a functional autoregressive (FAR) model with functional predictors was implemented by Besse, Cardot and Stephenson (Besse et al. [2000]), which could produce an entire annual temperature trend one year ahead with a substantial reduction in mean square error (MSE) compared to the traditional SARIMA model. In addition, with the growth of the popularity of

[*] Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6

machine learning, some nonparametric techniques such as support vector machine (SVM) and neural network were also applied to weather forecasting and have demonstrated moderate performance (Radhika and Shashi [2009]). Another widely used technique on weather prediction is the spatio-temporal model. Spatio-temporal model is a type of statistical methodology that can estimate or predict the value of response variable at an unobserved location and a future time (Hengl et al. [2012]), which helps us to incorporate both time and spatial factors at the same time.

However, there are challenges still remain unsolved. A big challenge facing scientists today is how to improve the prediction accuracy of the regions with extreme weather and wide annual temperature gap. Hengl et al. [2012] discovered that compared to the Mediterranean region with mild climate, the weather prediction of the mountainous part in Croatia is less accurate with more variations. This challenge motivated us to investigate the relationship between prediction performance and weather stability. In other words, we aimed to determine a set of weather-related variables (i.e. time, geographical location and other weather measures) that can represent the weather stability, and explore their effects on the prediction performance.

Figure 1: Marginal experimental variograms for residuals



Another big challenge exists in most of the statistical analysis is that the forecast is less accurate when we try to predict more days ahead, and the improvement of long-term weather prediction is in a slow process as it takes ten years to increase weather forecast skills by only one day (i.e. 4-day forecast in today is as accurate as 3-day forecast a decade ago) (Bauer et al. [2015]). Figure 1 (Hengl et al. [2012]) demonstrates the problem of worse prediction due to increasing time span. Scientists have expressed their lack of confidence on weather forecasts when time span gets larger, saying that 47% of the respondents to the weather forecast survey do not believe the forecast result for 7 to 14 days later (Lazo et al. [2009]). Since people are most interested in short-term weather forecasts as it provides direct guidance on planning day-to-day activities (Lazo et al. [2009]), we only evaluated the overall accuracy of 1-day forecast in this study.

1.2 Datasets

Our data contain 3-year weather forecast and historical measurements records across 113 U.S. cities from September 2014 to August 2017. Historical weather records comprise different weather measures in each city, such as temperature, humidity and sea level pressure, etc. The forecast weather records consist different measures of weather that were forecast over the 3-year period, including minimum temperature, maximum temperature, and the probability of precipitation, and specify the date that was forecast and the date that the forecast was made on. The geographical information of the cities for which the forecast was made is also available. Each city is documented with its corresponding state, geographical coordinates (i.e. longitude and latitude) and airport code (AirPtCd). AirPtCd provides information regarding the airport closets to the origin of the city, where is the place that the historical data was measured. We also accessed the external data source to get the geographical coordinates of the airport in order to calculate the empirical distance from the airport to city. Details of variables were summarized in Table 1.

To evaluate the prediction performance, we defined our response variable as the absolute value of the prediction error for the minimum temperature:

$$\varepsilon_t = |T_t^{real} - T_t^{fore}|,$$

where T_t^{real} and T_t^{fore} are the real and forecast temperatures at time point t, respectively.

1.3 Objectives

Based on our motivations, the goals of our analysis are to focus on:

1. Exploring variations in weather forecast accuracy across different geographical locations in the U.S., and identifying the most and least predictable regions.

2. Explaining how prediction performance changes over time.

3. Investigating how the prediction performance is affected by or correlated to different weather measures, such as sea level pressure, precipitation or humidity, etc.

Table 1: Description of variables in the dataset

Variable Category	Variable Name	Description
Historical Weather Information	Date	Date that has been forecast
	Min_TemperatureF	Real minimum temperature measured in Fahrenheit.
	Max_TemperatureF	Real maximum temperature measured in Fahrenheit.
	Mean_TemperatureF	Average of minimum and maximum temperature.
	Mean_DewpointF	
	Mean_Sea_Level_PressureIn	
	Mean_VisibilityMiles	
	Mean_Wind_SpeedMPH	
	PrecipitationIn	
Forecast Weather Information	City_Index	The city where the forecasts were made on.
	Forecast_Date	The date that was forecast.
	Forecast_Made_On	The date that the forecast was made on.
	Forecast_Value	Indicate what value is being forecast.
Geographical Information	City	
	State	
	Longitude	
	Latitude	
	AirPtCd	Airport code of the airport closest to the origin of the city.
	Dist_To_AirPt	Distance from airport to city.

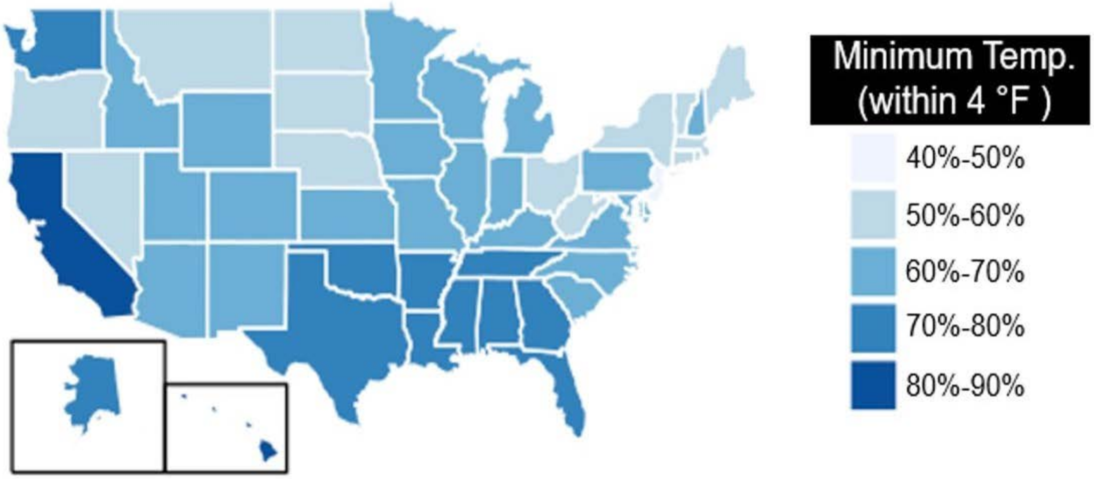
2. Exploratory Data Analysis

In order to get the sense of data before moving forward, we first performed exploratory data analysis (EDA), which helps to empirically detect trends in data and plays as a foundation of our further studies. The following sections explore the variation of prediction error from four different aspects based on our intuition and basic knowledge on weather forecasts, supported by data summary statistics and plots. These explorations motivated us to find potential methods to explain and model the discovered phenomena in data.

2.1 Geographical Pattern

To investigate our first objective, we generated following graph to compare the forecast accuracy of different geographical locations across the U.S. We consider the prediction is accurate if $\varepsilon_t < 4$ (i.e. the prediction error is within 4 Fahrenheit), and the accuracy was evaluated as the percentage of prediction satisfying this condition within each state. More blue represents regions with higher prediction accuracy and less blue represents regions with lower prediction accuracy.

Figure 2: Prediction accuracy of minimum temperature (F) for each state



In general, neighboring states tend to have similar prediction performance. However, for those states which are not close to each other, such as Washington, California, Hawaii and Florida, they also have similar prediction accuracy because they share the similar weather conditions. Therefore, states are not only clustered by their geographical locations, but also by the similarity of climate conditions. For example, coastal states with mild climate are more likely to be clustered together and have better forecast performance than the inland states with more extreme weather. We referred this as the ”spatio-climate effect”, a term we created to address the joint effect of geographical location and climate in weather forecasts. To illustrate this spatio-climate effect, we utilized clustering methods to divide the U.S. into different regions based on weather prediction performance, and then identified the most and least predictable U.S. states.

2.2 Seasonal Pattern

Season plays an important role in determining future climate expectations. Intuitively speaking, the cold season will cause significant uncertainty in forecast guidance, and is expected to be less predictable than the warm season. This is well illustrated in Figure 3, which shows that the performance of weather prediction varies over time from September 2014 to August 2017. The red and blue regions represent winter (December to February) and summer (June to August) period, respectively. The variation of ε_t shows periodicity within each year; specifically, the prediction is more variable in winter compared to summer.

Figure 3: Absolute prediction error (F) vs Forecast date

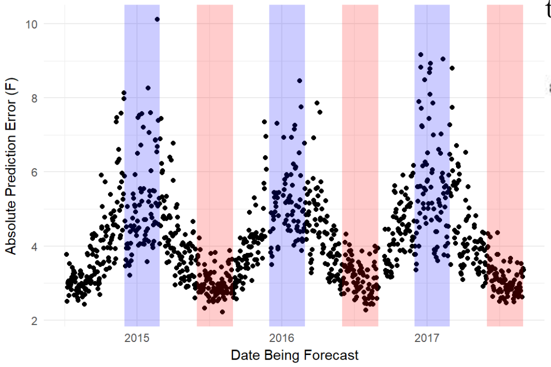


Figure 4: Real mean temperature (F) vs Mean Absolute prediction error, where the dot represents the mean absolute prediction error and the line corresponds to its 95% Confidence Interval

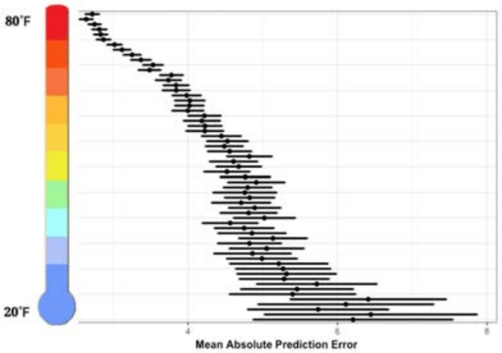


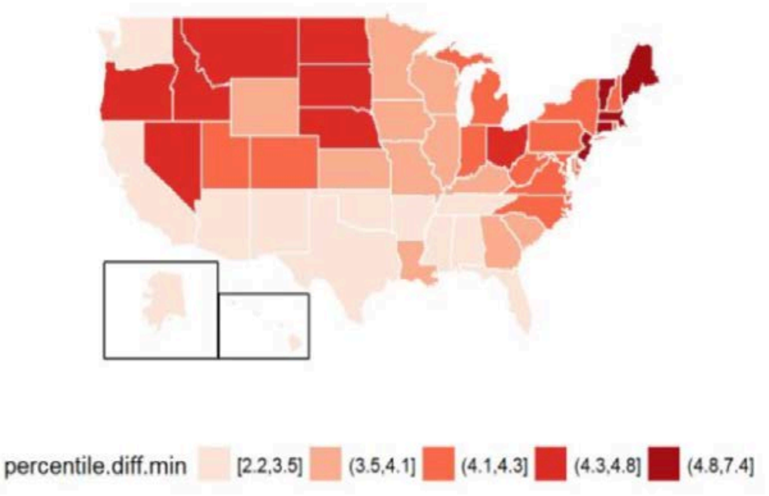
Figure 4 shows the relationship between the real minimum temperature and prediction error ε_t, demonstrating an increasing trend in ε_t with larger variability when the real minimum temperature decreases from summer to winter.

For the analysis of data with time-dependent variable, we applied a functional analysis to incorporate time factor into models by treating our response as a function of time, which helps us to explore how the prediction performance changes over time.

2.3 Other Weather-related Variables

In addition, the absolute prediction errors are suspected to have a significant correlation with some weather-related variables, such as local mean temperature, humidity level, sea level pressure and visible miles, etc. As an example, the following two graphs show that the minimum temperature is one of the potential covariates that will affect the prediction performance.

Figure 5: Average minimum temperature across 50 U.S. states



Moreover, in Figure 5, the average real minimum temperature across the U.S. shows an approximately opposite pattern as Figure 2, indicating that the real minimum temperature is negatively correlated with the prediction error. Regression model, such as random forest, was used to explore the significant covariates correlated with weather prediction accuracy, and candidate covariates were selected from the historical weather information dataset (see Table 1).

3. Methods & Results

3.1 Clustering

To explore the pattern of variations within minimum temperature prediction accuracy, the time series curves of ε_t for each state were first parameterized and smoothed by B-spline. Then, two different unsupervised clustering methods were applied on those 50 smoothed curves.

3.1.1 Parameterization and Smoothing on Time Series Data

In the original data, the absolute prediction error was observed over time in each state. Thus, to study the patterns of absolute prediction error for each state over time, the time series data can be considered as a function of time $\epsilon(t)$.

To reduce the noise and capture the main pattern of the time series data, we fitted the data in each state to a smoothed curve as a linear combination of several spline functions. The idea of data-to-curve transformation is almost the same as the linear regression. Consider the time t $\in$ [a,b] with M distinct interior points $\xi_1, \xi_2, ..., \xi_M$ that partition the [a,b] as a(ξ_0) $<$ ξ_1 $<$ ξ_2... $<$ ξ_M $<$ b(ξ_{M+1}); then the spline function with a degree d will be fitted on each interval $[\xi_i, \xi_{i+1}]$ with $d-1$ continuous derivatives on the open interval (a, b), where $i = 0, 1, ..., M$. In this case, we used B-spline, a type of spline function, to approximate the functional data $s(t)$. With a degree d and M interior points $\xi_1, \xi_2, ..., \xi_M$, $M + d + 1$ Schoenbergs B-spline basis functions $(B_1, B_2, ..., B_{M+d+1})$ forms the linear space, and the $s(t)$ will be approximated as a linear combination of the basis functions as

$$s(t) = s(t, \beta) = \sum_{l=1}^{M+d+1} \beta_l B_l(t), \qquad (1)$$

where the $\beta = (\beta_1, \beta_2, ...\beta_{M+d+1})'$ is the coefficients of the corresponding basis functions (Curry and Schoenberg, 1966). Similar to the coefficient estimation in linear regression, to estimate the β, we first transferred the observed time point $t_1, t_2, ..., t_n$ to a $n \times (M+d+1)$ matrix B with row vector $B_i = (B_1(t_i), ..., B_{M+d+1}(t_i))$. Under the assumption that the $B'B$ is non-singular, the β is estimated using least squared error as

$$\hat{\beta} = \text{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - s(t_i, \beta))^2 = [B'B]^{-1} By, \qquad (2)$$

where y $= (y_1, y_2, ..., y_n)$ is the observation of the response variable on time $(t_1, t_2, ..., t_n)$.

For the distinct interior points setting, we used 17 distinct interior points to divide the 3-year period into 18 time intervals with the same data amount, so that each time interval contains 2 months of data. After obtaining the smoothed curves of $\epsilon_i(t), i = 1, ..., 50$ for 50 U.S. states, we utilized two unsupervised clustering methods to group states which show similar performance on $\epsilon_i(t)$.

3.1.2 Time Series Clustering

Clustering is a method to group a set of objects that are similar in the same group (cluster). Provided the time-dependence feature of time series data, the conventional clustering techniques are not suitable to identify meaningful clusters. Clustering time-series data has been widely used in different applications, such as stock market data and medical data (Aghabozorgi et al. [2015], Aggarwal and Reddy [2013]). In time series clustering, dynamic time warping (DTW) is one of the metrics for measuring the similarity between two time series. DTW is calculated using a dynamic programming algorithm that tries to find the optimum warping path between two series under certain restrictions (Aghabozorgi et al. [2015]).

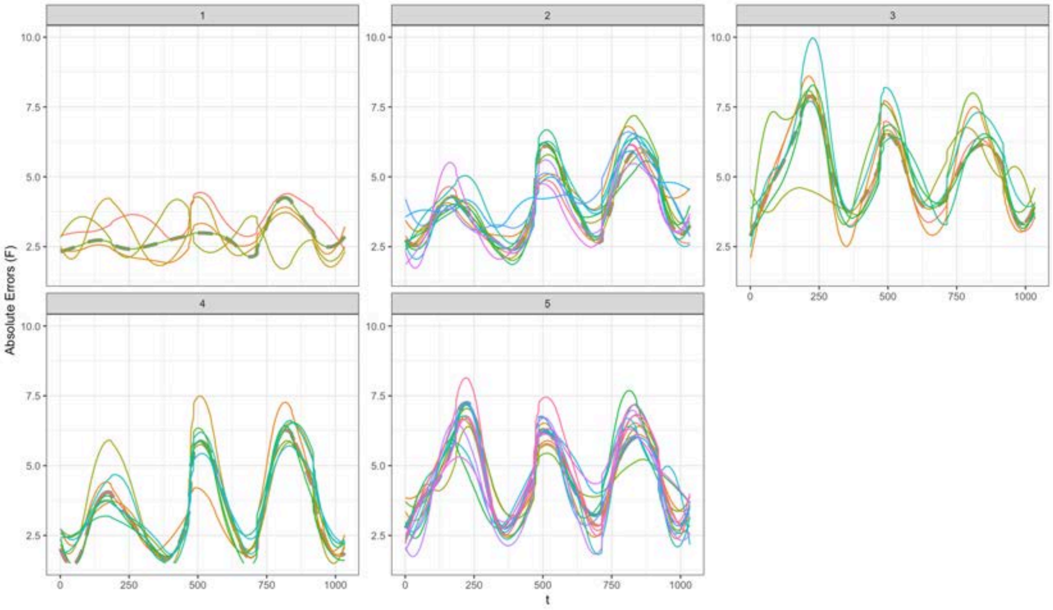
Our first step was to smooth the raw data as described in section 3.1.1. After curve parameterization and smoothing, we obtained 50 smoothed time series curves and our goal was to group similar curves into a few clusters. The dtwclust package, developed for the R statistical software, provides implementations for a range of time-series clustering algorithms (Sardá-Espinosa [2017]). We fitted the time series clustering using dtwclust package with DTW as the distance function to cluster 50 time series into k clusters. Note that the number of clusters k must be specified in advance. The algorithm intends to find the medoids that are centrally located in clusters, where each centroid is a time series. A range of k values, from 3 to 10, were tested and evaluated using Dunn index, Davies-Bouldin index and Silhouette index in the table below where a higher value of Dunn index and Silhouette index are preferred and a lower value of Davies-Bouldin index is preferred (Sardá-Espinosa [2017]). From Table 2, these three indexes suggest to pick $k = 6$. However, we noticed that there is a cluster with only one time series (Arizona) in it. We combined it with another cluster, resulting in a $k = 5$ clusters in the final model.

Table 2: Table of different clustering evaluation metrics for each number of cluster, k

Index	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Dunn index	0.10	0.10	0.06	0.15	0.08	0.11	0.10	0.14
Davies-Bouldin index	2.07	1.67	1.95	1.04	1.55	1.38	1.51	0.87
Silhouette index	0.33	0.27	0.19	0.22	0.15	0.16	0.15	0.24

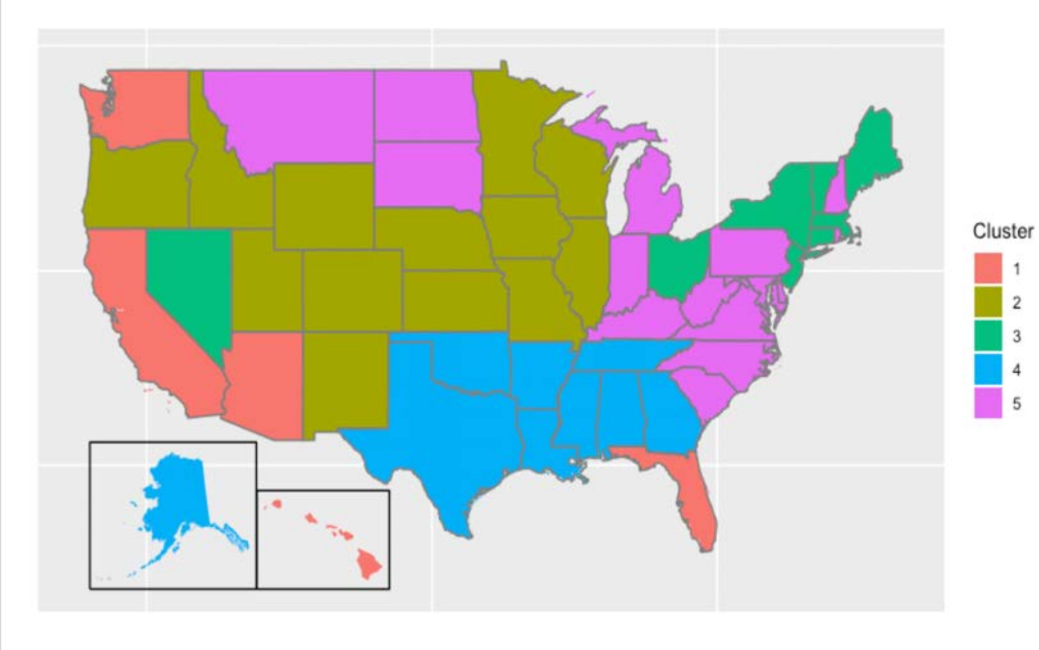
As seen in Figure 6 below, five clusters with its centroid and individual series were plotted where the dotted lines indicate the centroid and the solid lines indicate the individual series. From the figure, we can also see 5 different patterns for each cluster. A common pattern across all clusters is that the absolute prediction errors, ε_t, tend to be higher in the winter and lower in the summer. Cluster 1, on average, has lower and more stable absolute prediction errors over the span of 3 years. The prediction errors for Cluster 2 and Cluster 4 tend to increase over the years, whereas the prediction errors for Cluster 3 tend to decrease over time. Last but not the least, the pattern of prediction errors for Cluster 5 is similar from year to year.

Figure 6: Time-series clustering with k=5 clusters based on absolute prediction errors



Another way to assess the goodness-of-clustering is to map the states in each cluster on the actual U.S map. As discussed in the earlier section, states that are close to each other are expected to have a similar weather. Moreover, some states are not neighbours but have similar climate effects, such as Hawaii and Florida. We hope that our clustering method is able to capture these two characteristics. From Figure 7, we can observe that most of the states within the same clusters are geographically close to each other. For cluster 1, California, Florida and Hawaii, which are the hotter states compared to the others in the US, are clustered into the same group. This is consistent with our hypothesis where states with similar climate should be in the same group.

Figure 7: Time-series clustering with 5 clusters plotted on the U.S map



3.1.3 K-means Clustering on Functional Principal Component (FPC) Scores

Instead of clustering the states on time series data, we proposed another clustering method on functional data. The second clustering method focuses on applying the K-means clustering method on FPC scores. The key steps are:

1. Conduct the principal component analysis (PCA) into the curves
2. Obtain the principal component (PC) scores of the curves from the first few PCs who can explain more than 90% of the variation of the curves
3. Use the K-means method to cluster the obtained PC scores

The motivation of this clustering method comes from a clustering method created by Abraham et al. [2003]. This method clusters the curves by involving the K-means clustering method to the coefficients of B-splines which approximate the smoothing curves. The key steps are

1. Given n group of time series data where n_i observations in the th group, i = 1 ... n. In th group for any i in 1 to n , we approximate the observations $\{(y_j, t_j)|j = 1..n_i\}$ to a smooth curve $y_i(t)$ as a linear combination of B-spline.
2. Cluster the data into k groups by using K-means clustering the estimated B-spline coefficients $\{\hat{\beta}_i|i = 1...n\}$ into K groups.

To classify $\{\hat{\beta}_i|i = 1...n\}$, the estimated coefficients of n $s(t)$, into k groups, the procedure of k-means clustering is to search for k partitions, $\{C_1, C_2, ..., C_k\}$ with center vectors $\{c_1, c_2, ..., c_k\}$ which minimize

$$\frac{1}{n}\sum_{j=1}^{k}\sum_{\hat{\beta}_i\in C_j} \|\hat{\beta}_i - c_j\|^2 \tag{3}$$

where the $\|\cdot\|$ is defined as the Euclidean norm (Hartigan and Wong [1979]). A strong consistency property has been proved in this method, which is that the calculated center $\{c_1, c_2, ..., c_k\}$ has strong consistency to a unique center $\{c_1^*, c_2^*, ..., c_k^*\}$

, which indicate that by finding an appropriate function basis space to approximate the data to curves, the procedure and the result of algorithm will be stable when getting more time series data; then the calculated center of the clusters, $\{c_1, c_2, ..., c_k\}$, will converge to the unique $\{c_1^*, c_2^*, ..., c_k^*\}$ (Abraham et al. [2003]).

Instead of conducting the K-means methods on B-spline coefficients, we clustered the curves refer to the largest variation direction, which was applying the K-means method to the FPC scores. Functional principal component analysis (FPCA) is an extension of PCA to the functional data $x(t)$, where t is a continuous variable (Ramsay et al. [2009]). Given a set of functional data with N smoothing curves, $\{x_i(t)|i = 1...N, t \in [a, b]\}$, the first step of FPCA is to estimate the covariance function as

$$v(s,t) = \frac{1}{N-1}\sum_i [x_i(s) - \bar{x}(t)][x_i(t) - \bar{x}(t)], \tag{4}$$

where the s and t share the same domain $[a, b]$. Using the Karhunen-Loeve decomposition (Fukunaga and Koontz [1970]), the $v(s, t)$ can be decomposed as

$$v(s,t) = \sum_{j=1}^{\infty} d_j \xi_j(s)\xi_j(t), \tag{5}$$

where the $\xi_j(t)$ is the eigenfunction and usually with the restricted condition as $\int \xi^2(t)dt = 1$, and the d_j is the eigenvalue of $\xi_j(t)$. Similar to PCA in multivariate, the d_j is proportional to the percentage of variation that the $\xi_j(t)$ explains. Finally, the j^{th} PC score of the functional data $x_i(t)$ can be calculated as

$$\rho_{ij} = \int \xi_j(t)[x_i(t) - \bar{x}(t)]. \tag{6}$$

We can reorder the eigenfunctions following the size of their eigenvalues from largest to the smallest, and gain the first p PCs with largest eigenvalues that can explain most of the variation (i.e. >90%) in the curves. Then, each $x_i(t)$ can be rewritten and approximated as

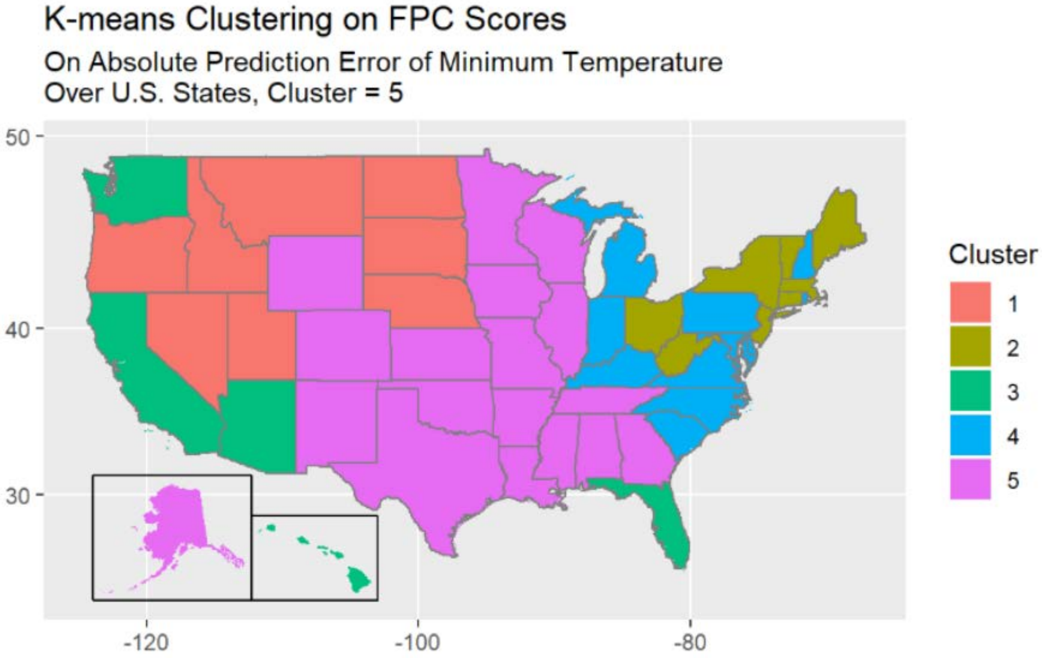
$$\begin{align} x_i(t) &= \bar{x}(t) + \sum_{j=1}^{\infty} \rho_{ij}\xi_j(t) \tag{7} \\ &\approx \bar{x}(t) + \sum_{j=1}^{p} \rho_{ij}\xi_j(t) \tag{8} \end{align}$$

In other words, FPCA provides a new group of basis functions $\{\bar{x}(t), \xi_1(t), ..., \xi_P(t)\}$, and reform the functional data into a linear combination of the basis functions, where the coefficient of $\bar{x}(t)$ is always 1 and the coefficient of the $\xi_p(t)$ is the score of the p^{th} PC of the corresponding curve.

Similar to the clustering B-spline coefficients using K-means, this clustering method is applying the K-means clustering on the coefficients of the eigenfunctions. Therefore, this method may also have the consistency property, as the strong consistency property stands for all kinds of basis functions (Abraham et al. [2003]), and the $\bar{x}(t)$ should not affect the clustering because it exists in all the curves with the same coefficients 1.

We plotted the result of clustering on FPC scores in Figure 8 by labeling the clusters in different colours.

Figure 8: Functional Data Clustering with k=5 Clusters

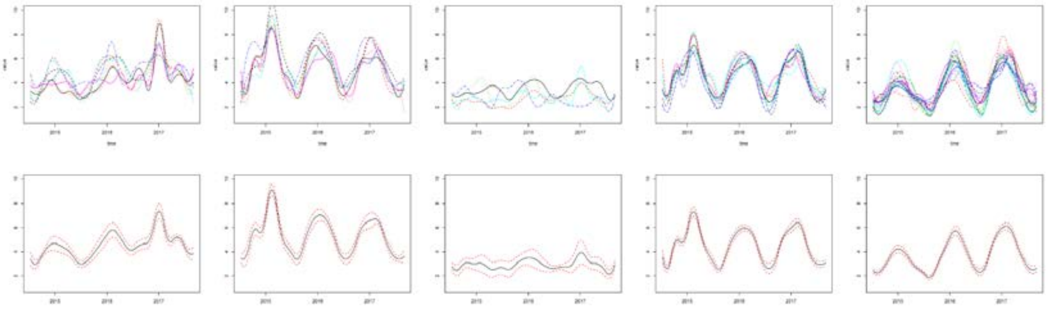

K-means Clustering on FPC Scores
On Absolute Prediction Error of Minimum Temperature
Over U.S. States, Cluster = 5
50
40
30
-120
-100
-80
Cluster
1
2
3
4
5

Key features we found in Figure 8 are:

- cluster 1 groups the northwest inland states in the U.S.;
- cluster 3 groups west coast states, including Washington, California and Arizona, and states with mild climate, such as Hawaii and Florida;
- cluster 5 groups most of the middle inland states in the U.S;
- Northeast states are mostly grouped together.

To investigate the patterns in each cluster, we plotted both the original smoothed curves and their mean curves with 95% confidence interval in Figure 9. Starting from the left of the figure, the clusters were plotted in order from 1 to 5. In Figure 9, the five plots in the first panel are the smoothed curves of each state in the corresponding cluster, and the five plots in the second panel are the mean curves of respective clusters with 95% confidence interval indicated by red dotted lines.

Figure 9: Functional Data Clustering: K-means on FPC Scores with k=5 Clusters



In Figure 9, we observed that all clusters have a higher prediction error and wider confidence interval during the winter, which implies the extrapolation problem of minimum temperature prediction in a cold environment. Moreover, Some special characteristics exist in each cluster,

- Cluster 1 and 5 shows an increasing trend of prediction error over time, whereas cluster 1 has larger fluctuation.
- Cluster 3 has visually lower and more stable absolute prediction error than the other clusters;
- Cluster 2 and 4 are similar to each other, and both of them have larger fluctuation than the other 3 clusters with a little stepwise pattern from the end of 2014 to the start of 2015.

To quantify cluster-to-cluster difference in prediction accuracy, we integrated the mean curves of each cluster and ordered the 5 clusters by the magnitude of integration (area under curves), allowing us to identify the most and least predictable states in U.S. The ranking results show that cluster 3 has the best prediction performance with the smallest integral value, while cluster 2, which contains Connecticut, Maine, Massachusetts, New Jersey, New York and Vermont has the worst prediction performance with the largest integral value.

Table 3: Cluster integration ranking results

Cluster	Rank	Overall Integral	Representative States
CL3	1	3397.3	California, Florida
CL5	2	4284.2	Alaska, Texas
CL4	3	5247.7	Michigan, Pennsylvania
CL1	4	5248.2	Nevada, North Dakota
CL2	5	6155.2	New York, Massachusetts

3.2 Concurrent Functional Linear Model

To investigate the correlations between different weather measures and ε_t, the concurrent functional linear model was conducted. The concurrent functional linear model is a simple extension of linear regression to the functional data.

3.2.1 Model Description

Unlike simple linear regression that both the response variables and predictors are single values, in the concurrent functional linear model, we aimed to regress multivariate covariates on the functional responses on the basis that all the responses and covariates will be considered as objects with a function of time(Ramsay et al. [2009]). This means that given N observed objects with q covariates, we observe that the values of all response variables and the covariates are continuous functions of variable t. Let $\{y_i(t)|i = 1...N\}$ be the functional value of response variable for the i^{th} observation, and let $\{x_{ij}(t)|i = 1...N, j = 1...p\}$ be the functional value of the j^{th} covariate for the i^{th} observation. Let $Z(t)$ be the Nq matrix that contains all the $x_{ij}(t)$ functions, then the concurrent functional is defined as

$$y(t) = Z(t)\beta(t) + \epsilon(t), \tag{9}$$

where the $\beta(t)$ is the coefficient function vector of the covariates. To deal with the multicollinearity problem between the corvariates, the $\beta(t)$ is estimated by minimizing sum squared error with the weighted regularized criterion as

$$\text{LMSSE}(\beta) = \int [y(t) - Z(t)\beta(t))]'[y(t) - Z(t)\beta(t))] + \sum_{j}^{p} \lambda_j \int [L_j\beta_j(t)]^2 dt \tag{10}$$

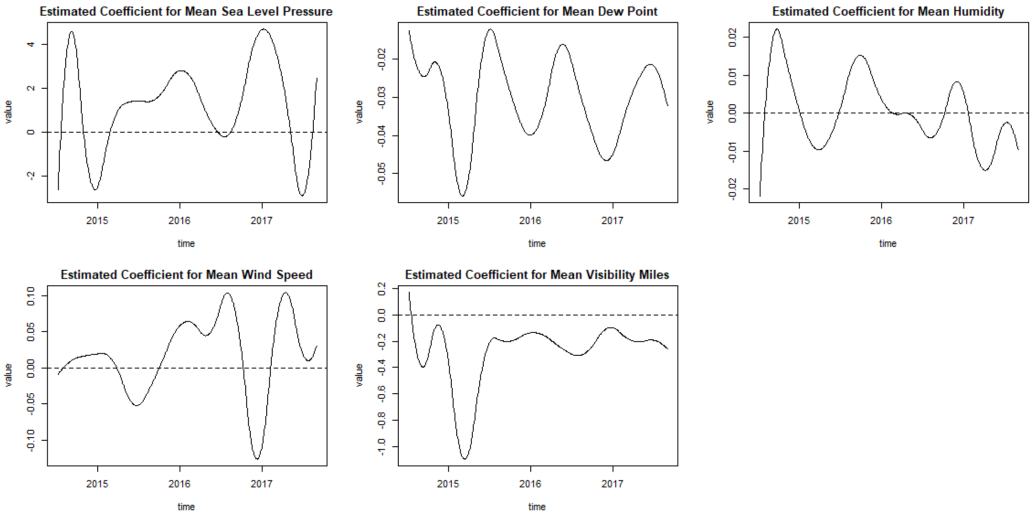
3.2.2 Results

According to the LMSSE form, the concurrent functional linear model can be understood as an extension of the ridge regression model (Ramsay et al. [2009]); therefore, we implemented a point-wise ridge regression model on every observed date with a common penalty λ in all covariates over time. In the model, the response variable was the absolute prediction error ε_t in each state, and our interest covariates were the mean sea level pressure, dew point, humidity, wind speed and visibility miles.

The first step in our model construction was to choose the penalty term λ. We first conducted the cross-validation to choose the optimal λ with the smallest mean squared error on each observed date, and then used the median of the all the λs as the final λ. Then we estimated the coefficient of each covariate over time, as shown in Figure 10.

According to Figure 10, the dew point and visibility miles were negatively correlated to the absolute prediction error in most of the past three years, and the effect of sea level pressure and wind speed were positive in most of the time. Moreover, some periodic pattern may exist in the effect of dew point and humidity. The negative effect of the dew point may be relatively stronger in winter than in summer. Besides, the effect of the humidity has a relatively large positive effect in fall, but has a relatively large negative effect in spring and summer.

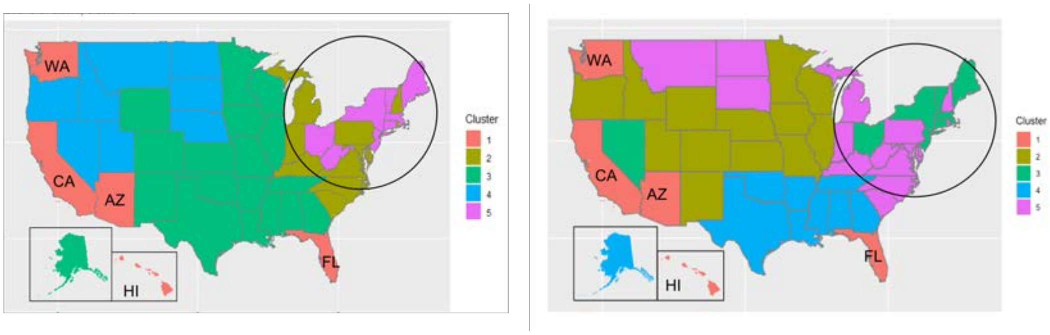
Figure 10: Estimated coefficients of covariates in concurrent functional linear model



4. Conclusion

By comparing the results of the clustering methods based on time series curves and FPC scores, we are able to identify some interesting similarities. Figure 11 shows the U.S. map of time-series clustering (the left panel) and FPCA clustering (the right panel). We observed that all of the two clustering methods group Hawaii, Washington, California, Arizona and Florida into the same cluster (Cluster 1 in both methods coloured in red in Figure 11), showing that these 5 states share some distinct characteristics from others, where the absolute prediction errors are more stable throughout the years compared to other clusters. This can be further verified by the time series plots in Figure 6 and 9 as well as Table 3.

Figure 11: Clusters in Time Series Data V.S. Cluster in Functional Data



Moreover, both methods show that the states highlighted in circle in Figure 10 are grouped into the same cluster. This provides us with more confidence that these states are different from the rest of the states. One plausible reason may be that these states have extreme weather during winter and summer, which makes it more difficult to predict.

The concurrent functional linear model provides us with more information about how the prediction errors are associated with other covariates. The result of the estimated coefficients suggests that the effect of the covariates change over time. In Figure 10, some interesting jumps of the estimated coefficients were observed, which might be related to the missing values or some special weather events existed in the corresponding time. For instance, the most obvious jumps in the estimated coefficients of sea level pressure, dew point and visibility miles were observed at the beginning of 2015 in Figure 10.

5. Discussion

There are some limitations in our study. As we defined in the earlier section, prediction error is the absolute value of the difference between the real minimum temperature and the forecast temperature, so by taking the absolute value on the prediction error, we lose some information regarding the problem of directional bias; for example, the overestimation and underestimation of the prediction errors are not well addressed in this study. In addition, the concurrent functional linear model only considers the same-day effect of covariates on the response variable but does not include the information from the past up to the point. This model can be further improved by using the general functional regression model (Ramsay et al. [2009]).

To involve the spatial correlation, future studies can focus on implementing the spatio-temporal model described in (Hengl et al. [2012]) to incorporate both temporal and spatial components simultaneously. In this case, we can further investigate how other weather-related variables, such as humidity, wind speed and sea level pressure, will affect the prediction performance over time as well as space.

6. Acknowledgement

The authors are most appreciative of the organizers of 2018 JSM Data Expo who made this happen. We also thank Dr. Peijun Sang, PhD candidates Yuping Yang and Zhiyang Zhou, faculty members and graduate students in the Department of Statistics and Actuarial Science at Simon Fraser University who provided helpful suggestions relating to this project.

References

Christophe Abraham, Pierre-André Cornillon, ERIC Matzner-Løber, and Nicolas Molinari. Unsupervised curve clustering using b-splines. Scandinavian journal of statistics, 30(3):581–595, 2003.

Richard M Adams, Cynthia Rosenzweig, Robert M Peart, Joe T Ritchie, Bruce A McCarl, J David Glyer, R Bruce Curry, James W Jones, Kenneth J Boote, and L Hartwell Allen Jr. Global climate change and us agriculture. Nature, 345(6272): 219, 1990.

Charu C Aggarwal and Chandan K Reddy. Data clustering: algorithms and applications. CRC press, 2013.

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. Information Systems, 53:16–38, 2015.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. Nature, 525(7567):47, 2015.

Philippe C Besse, Hervé Cardot, and David B Stephenson. Autoregressive forecasting of some functional climatic variations. Scandinavian Journal of Statistics, 27 (4):673–687, 2000.

Denis Bosq. Nonparametric statistics for stochastic processes: estimation and prediction, volume 110. Springer-Verlag, 1996.

George EP Box and Gwilym M Jenkins. Time series analysis: forecasting and control, revised ed. Holden-Day, 1976.

Gérard Collomb. From non parametric regression to non parametric prediction: Survey of the mean square error and original results on the predictogram. In Specifying Statistical Models, pages 182–204. Springer, 1983.

Keinosuke Fukunaga and Warren LG Koontz. Application of the karhunen-loeve expansion to feature selection and ordering. IEEE Transactions on computers, 100(4):311–318, 1970.

Lázló Györfi, Wolfgang Härdle, Pascal Sarda, and Philippe Vieu. Nonparametric curve estimation from time series, volume 60. Springer-Verlag, 1989.

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.

Tomislav Hengl, Gerard BM Heuvelink, Melita Perčec Tadić, and Edzer J Pebesma. Spatio-temporal prediction of daily temperatures using time-series of modis lst images. Theoretical and applied climatology, 107(1-2):265–277, 2012.

Jeffrey K Lazo, Rebecca E Morss, and Julie L Demuth. 300 billion served: Sources, perceptions, uses, and values of weather forecasts. Bulletin of the American Meteorological Society, 90(6):785–798, 2009.

Y Radhika and M Shashi. Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering, 1(1):55, 2009.

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

Alexis Sardá-Espinosa. Comparing time-series clustering algorithms in r using the dtwclust package. R package vignette, 12, 2017.

Thomas J Teisberg, Rodney F Weiher, and Alireza Khotanzad. The economic value of temperature forecasts in electricity generation. Bulletin of the American Meteorological Society, 86(12):1765–1772, 2005.