A Review Work on Spatial Extension of Weather Forecasts

Course Code: MTH643A

Submitted by

Anirban Ghosh(221271), Khyati Singh(221332)

Rajdeep Adhya(221385), Rohit Dutta(221396)

Supervisor

Dr. Arnab Hazra

Assistant Professor

Department of Mathematics and Statistics,

Indian Institute of Technology, Kanpur

Submission Date: 14 November, 2023

Acknowledgement

We wish to extend my heartfelt appreciation and gratitude to Dr. Arnab Hazra, our supervisor, for her unwavering guidance and support, consistent efforts, and motivating encouragement, all of which were instrumental in enabling me to navigate the challenges encountered during this project. Additionally, we are deeply thankful to him for equipping us with a substantial understanding of the subject matter.

Contents

Abstract

In this study, machine learning techniques were applied to forecast weather conditions at 113 sites across the continental United States from 2014 to 2017. The dataset includes both forecasted and observed weather information, supplemented by data from the National Oceanic and Atmospheric Administration Physical Sciences Laboratory. **The objective is to enhance the accuracy of weather predictions by leveraging predictive modeling and spatially extended forecasts**. Employing spatial extension techniques, visual representations were generated to assess the precision of the extended forecasts. This study offers valuable insights into the effectiveness of advanced weather forecasting techniques, providing a detailed exploration of prediction accuracy across diverse geographical locations in the United States. **In this project we tried to reproduce the paper titled "A spatial extension of weather forecasts", by Benjamin Schweitzer, Robert C. Garrett, Nichole Rook, Thomas J. Fisher (2019) [1].**

1 Introduction

Predicting weather conditions typically involves forecasts targeted at specific locations like airports but is commonly interpreted as a regional forecast.The practice of spatial extrapolation(for eg. Kriging) a weather forecast is commonplace in every day interpretation of weather forecasts. Weather forecasts are made for specific point locations. In this project, we extracted daily high temperatures from the nearest grid to Fort Wayne airport and developed a forecasting model for the high temperatures of that specific grid. The model takes into account the high temperatures of 113 airports provided in the data expo as covariates for prediction. In the 2018 Data Exposition (Snow & Martinez 2023)[2], hereafter referred to as the Data Expo, weather forecasts and records for 113 locations across the United States from 2014 to 2017 were supplied. Our goal is to spatially expand a provided forecast to cover regions without a supplied forecast using observed weather from the Physical Sciences Laboratory (PSL - https://psl.noaa.gov/), within the National Oceanic and Atmospheric Administration (NOAA) as a validation set, we examined the accuracy of expanded forecasts and compared it to the forecasts provided. We explored two modeling techniques to expand the provided forecasts spatially. One approach involved employing LASSO penalized regression(Tibshirani 1996)[3], while the other utilized a more broadly applicable spatial technique (Pebesma and Bivand 2005 [4]; Bivand et al. 2013 [5]) derived from the initial LASSO results. While our framework is theoretically applicable to any location on Earth, its practical accuracy and validity diminish as the distance between a new site and the 113 predictor sites from the Data Expo data increases.

2 Data Description

2.1 Gridded Data

The Global Daily Gridded Climate Data can be accessed through PSL (Physical Sciences Laboratory). We obtained the daily maximum temperatures(in degrees Celsius) for each 0.5 × 0.5 grid within the United States from the CPC Global Unified Temperature dataset(https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html), specifically focusing on the dates that aligned with the provided Data Expo data. We had weather information for 3329 grids in US during the time window. For example, the plot of gridded maximum temperature on 1st September, 2017 of United States is shown in Figure 1.

2.2 Data expo data

The 2018 Data expo data is obtained from the Data expo section of ASA Section on Statistical Computing & ASA Section on Statistical Graphics web page (https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2018) by selecting the year 2018. There are three data files, locations.csv, forecast.dat, and histWeather.csv.

The locations.csv file is a comma separated value file that contains information on the cities for which the forecasts was made. The columns are city, state, longitude, latitude, and AirPtCd.

In the forecast.dat file the first column is the city number, so 1 means Eastport, Maine and 113 means Honolulu, Hawaii. The second column is the date being forecasted, the 3rd column is the

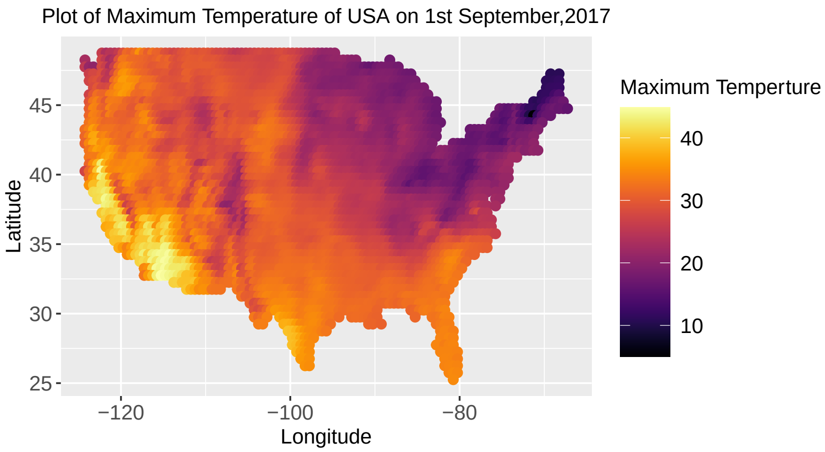


Figure 1: 0.5 × 0.5 gridded maximum temperature on 1st September, 2017 of United States; Here, temperature is given in Celsius scale.

forecasted value. The 4th column indicates what value is being forecast (MinTemp, minimum temperature; MaxTemp, maximum temperature; and ProbPrecip, the probability of precipitation). The 5th column is the date that the forecast was made on.

The histWeather.csv file is a comma separated file with the historic measures of weather from the airports. The main columns of interest are, **AirPtCd** which is the airport code for where the measurement were made and corresponds to the same column in the locations.csv file, **Date** which is the date of the measurement, **MaxTemperatureF** and **MinTemperatureF** which are the maximum and minimum recorded temperature (in Fahrenheit) for the date and **PrecipitationIn** which is the amount of precipitation in inches of water. As, the gridded data obtained have temperature in Celcius scale, we converted the given maximum temperature of histWeather.csv from Fahrenheit to Celcius scale. The 113 airport locations of Data Expo data is shown in Figure 2.

3 Extending Forecasts

We extended the weather forecasts by leveraging the data expo information available for 113 sites and the 0.5 × 0.5 gridded maximum temperature data we acquired. Two modeling techniques were employed: first, a LASSO penalized regression (Tibshirani 1996; James et al. 2015), and second, a

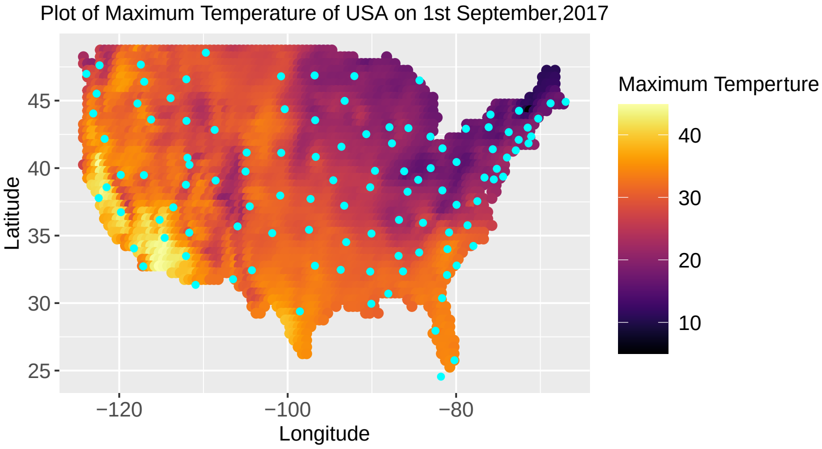


Figure 2: 0.5 × 0.5 gridded maximum temperature on 1st September, 2017 of United States; Here, temperature is given in Celsius scale along with 111 Data Expo sites (shown in (cyan) colored dots) in the contiguous United States (single sites in Anchorage, Alaska and Honolulu, Hawaii excluded in this image) for which we have observed and forecasted daily high and low temperatures

model that utilized the fitted LASSO model to identify pertinent covariates (i.e., airports) within a spatial bubble surrounding each forecasted region.

Note 1. *Two airports situated in single sites in Anchorage, Alaska and Honolulu, Hawaii excluded from all the Figures for better visualization.*

3.1 LASSO penalized regression

The modeling techniques of LASSO is explained through an example based on daily high temperatures (in degrees Celsius) for a specific grid at latitude 40.75 and longitude -85.25, which is nearest to the Fort Wayne airport, Indiana. Since we don't have weather data for this exact location in the Data Expo sites, we use the LASSO model for prediction. We extracted the daily maximum temperature of (latitude: 40.75, longitude: -85.25) from 1st July 2014 to 1st September 2017 as our **response variable** along with the observed daily high temperatures at all the supplied Data Expo sites (as the features set or predictor variables) to train a LASSO Model which performs variable selection (thus picking the important variables of the 113 possible features).

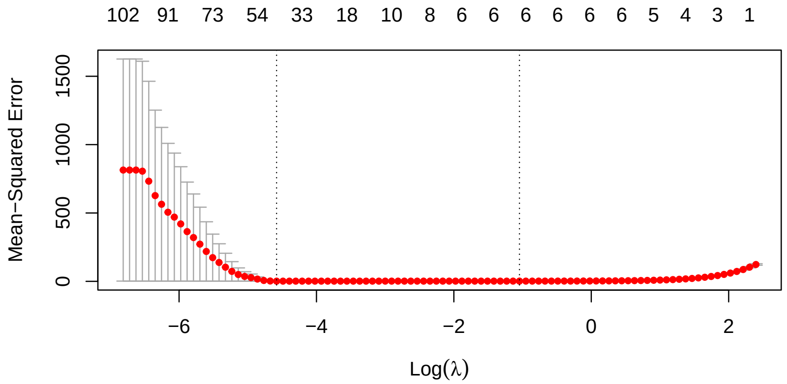


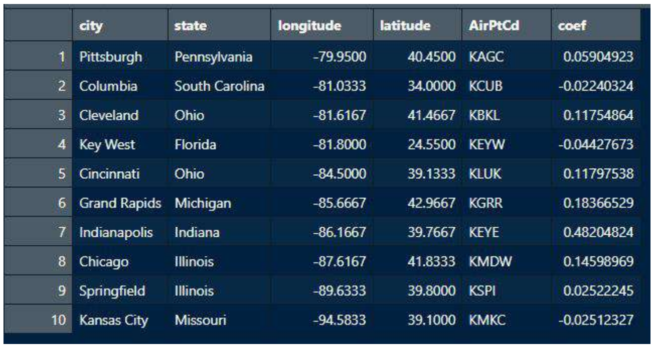
Figure 3: In this plot, the left vertical line shows where the CV-error curve hits its minimum. The right vertical line shows us the most regularized model with CV-error within 1 standard deviation of the minimum.

The LASSO model is given by,

$$argmin_{\beta_0,\ldots,\beta_{113}} \sum_{i=1}^{n}(T_i - \beta_0 - \sum_{j=1}^{113} \beta_j X_{ji})^2 + \lambda \sum_{i=1}^{113} |\beta_j|$$

where, T_i is Observed maximum temperature at (latitude: 40.75, longitude: -85.25) on the ith day, X_{ji} is the observed high temperature at Data Expo site j on the i-th day, β_j is the coefficient for site j and λ is the shrinkage penalty/regularization term.

Now, to tune the λ parameter (James et al. 2015) that minimizes the penalized mean squared error we utilize the **glmnet** package in R (Friedman et al. 2010) to undertake a 5-fold cross-validation study to get the optimal λ value (shown in Figure 3) for which the error sum of squares of the model gets minimized. The value of $\lambda(= 0.01024881)$ for which the CV-error is found to be minimum is considered as the final λ value. The LASSO penalized regression modeling approach performs variable selection and choose covariates with non-zero coefficients. **The selected sites for modelling the temperature of the grid (latitude: 40.75, longitude: -85.25) are KAGC (Pittsburgh), KCUB (Columbia), KBKL (Cleveland), KEYW (Key West), KLUK**



	city	state	longitude	latitude	AirPtCd	coef
1	Pittsburgh	Pennsylvania	-79.9500	40.4500	KAGC	0.05904923
2	Columbia	South Carolina	-81.0333	34.0000	KCUB	-0.02240324
3	Cleveland	Ohio	-81.6167	41.4667	KBKL	0.11754864
4	Key West	Florida	-81.8000	24.5500	KEYW	-0.04427673
5	Cincinnati	Ohio	-84.5000	39.1333	KLUK	0.11797538
6	Grand Rapids	Michigan	-85.6667	42.9667	KGRR	0.18366529
7	Indianapolis	Indiana	-86.1667	39.7667	KEYE	0.48204824
8	Chicago	Illinois	-87.6167	41.8333	KMDW	0.14598969
9	Springfield	Illinois	-89.6333	39.8000	KSPI	0.02522245
10	Kansas City	Missouri	-94.5833	39.1000	KMKC	-0.02512327

Figure 4: Selected covariates from LASSO model along with their corresponding airport name, city name and locations for forecasting maximum temperature of the grid with latitude: 40.75, longitude: -85.25

(Cincinnati), KGRR (Grand Rapids), KEYE (Indianapolis), KMDW (Chicago), KSPI (Springfield) and KMKC (Kansas City) as clearly depicted in Figure 5. Also, the coeficients of the selected covariates in the model along with their corresponding airport name, city name and locations are shown in Figure 4

Since, LASSO is known to over select variables, we can see that model has selected many sites near(latitude: 40.75, longitude: -85.25) but Key West(Florida) having latitude: 24.55, longitude: -81.8 is also selected which is far away from Ford Wayne (the red dot in figure 5). Using these selected sites, we then fit a linear model where the daily high temperature for (latitude: 40.75, longitude: -85.25) is a function of daily high temperatures in the selected sites.

The forecasted high temperature for i^{th} day would be: 2.05149 + 0.05904(x_1) + (-0.02240)(x_2) + 0.11754(x_3) + (-0.04428)(x_4) + 0.11797(x_5) + 0.18366(x_6) + 0.48204(x_7) + 0.14598(x_8) + 0.02522(x_9) + (-0.02512)(x_{10}) where x_1, x_2,.....,x_{10} are the maximum temperature of the covariates(i.e airports) selected for i^{th} day.

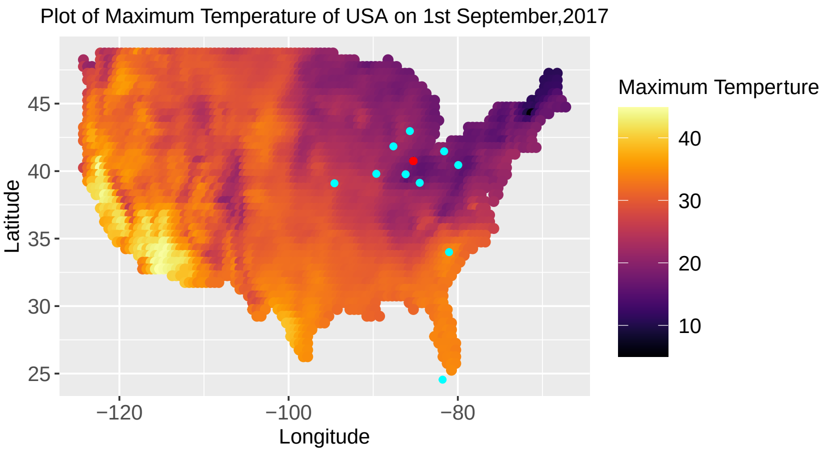


Figure 5: The (cyan) dots are the airports that has been selected by the LASSO model to forecast maximum temperature of the grid (red dot)at latitude: 40.75, longitude: -85.25

3.1.1 Forecasting Based on LASSO Based Model

As we don't have any data of those 113 Data Expo locations beyond 1st September, 2017, we considered the last 30 days of 113 Data Expo data (i.e. from 3rd August till 1st September) as our test data and the rest as training data. Also, we have considered the daily maximum temperature of the grid, having latitude: 40.75, longitude: -85.25 from 1st July, 2014 to 2nd August, 2017 as the response and from 3rd August to 1st September, 2017 as the test data for response. Then we fit the LASSO model described in subsection 3.1. The selected features(covariates) are shown in Figure 6 & 7. Also, we have performed a 30 days(from 3rd August, 2017 to 1st September, 2017) forecast of maximum temperature of latitude: 40.75 & longitude: -85.25, based on the maximum temperature of the selected airports (shown in Figure 6) as covariates and the corresponding lines diagrams are shown in Figure 10.

We have replicated the forecasting procedure on 1st September, 2017 as proposed by LASSO method for every 0.5×0.5 grid over the United States on which we have the response data on daily maximum temperature. The actual and forecasted maximum temperature of 1st September, 2017 has been shown in the Figure 8. Also, from Figure 9 of error in forecasting, we can observed that the prediction is quite good except a particular region(the densed yellow region)

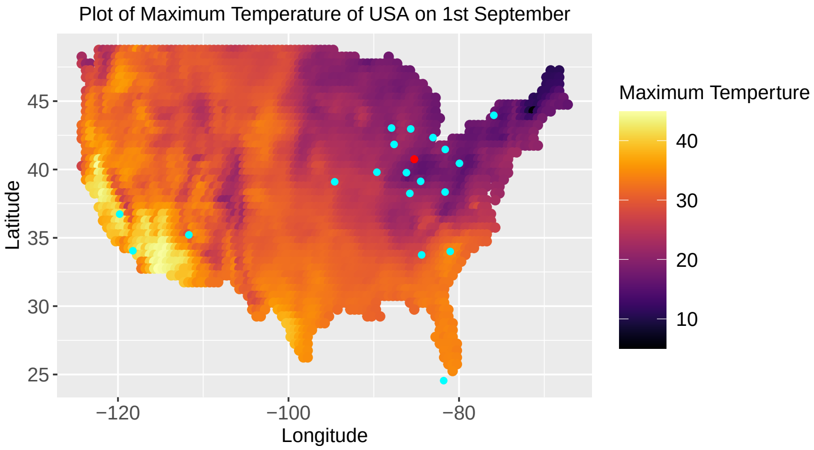
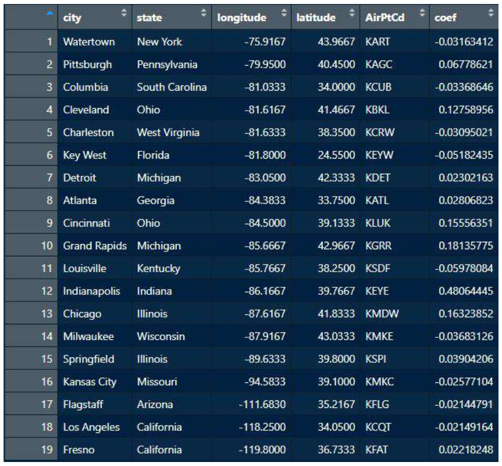


Figure 6: The (cyan) dots are the airports that has been selected by the LASSO model to forecast maximum temperature of the grid (red dot)at latitude: 40.75, longitude: -85.25



	city	state	longitude	latitude	AirPtCd	coef
1	Watertown	New York	-75.9167	43.9667	KART	-0.03163412
2	Pittsburgh	Pennsylvania	-79.9500	40.4500	KAGC	0.06778621
3	Columbia	South Carolina	-81.0333	34.0000	KCUB	-0.03368646
4	Cleveland	Ohio	-81.6167	41.4667	KBKL	0.12758956
5	Charleston	West Virginia	-81.6333	38.3500	KCRW	-0.03095021
6	Key West	Florida	-81.8000	24.5500	KEYW	-0.05182435
7	Detroit	Michigan	-83.0500	42.3333	KDET	0.02302163
8	Atlanta	Georgia	-84.3833	33.7500	KATL	0.02806823
9	Cincinnati	Ohio	-84.5000	39.1333	KLUK	0.15556351
10	Grand Rapids	Michigan	-85.6667	42.9667	KGRR	0.18135775
11	Louisville	Kentucky	-85.7667	38.2500	KSDF	-0.05978084
12	Indianapolis	Indiana	-86.1667	39.7667	KEYE	0.48064445
13	Chicago	Illinois	-87.6167	41.8333	KMDW	0.16323852
14	Milwaukee	Wisconsin	-87.9167	43.0333	KMKE	-0.03683126
15	Springfield	Illinois	-89.6333	39.8000	KSPI	0.03904206
16	Kansas City	Missouri	-94.5833	39.1000	KMKC	-0.02577104
17	Flagstaff	Arizona	-111.6830	35.2167	KFLG	-0.02144791
18	Los Angeles	California	-118.2500	34.0500	KCQT	-0.02149164
19	Fresno	California	-119.8000	36.7333	KFAT	0.02218248

Figure 7: Selected covariates from LASSO model along with their corresponding airport name, city name and locations for forecasting maximum temperature of the grid with latitude: 40.75, longitude: -85.25; Here the model intercept value is 2.143913

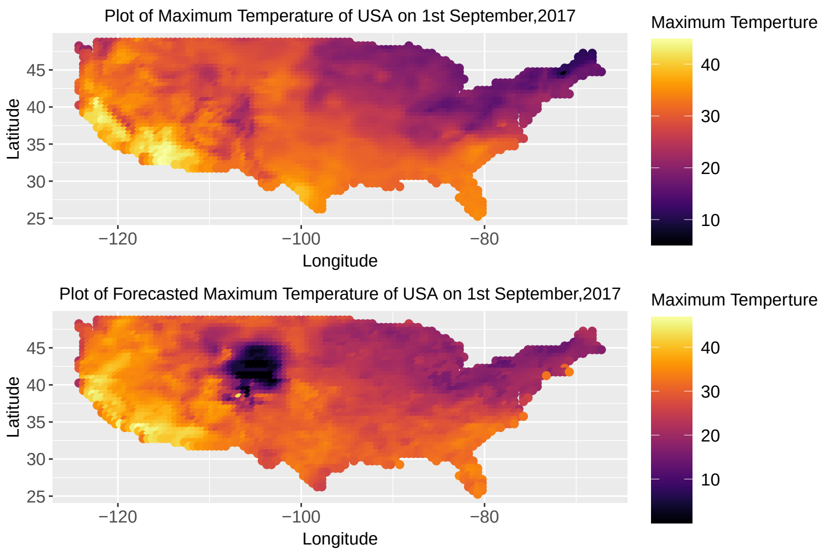


Figure 8: The actual and forecasted maximum temperature of 1st September, 2017

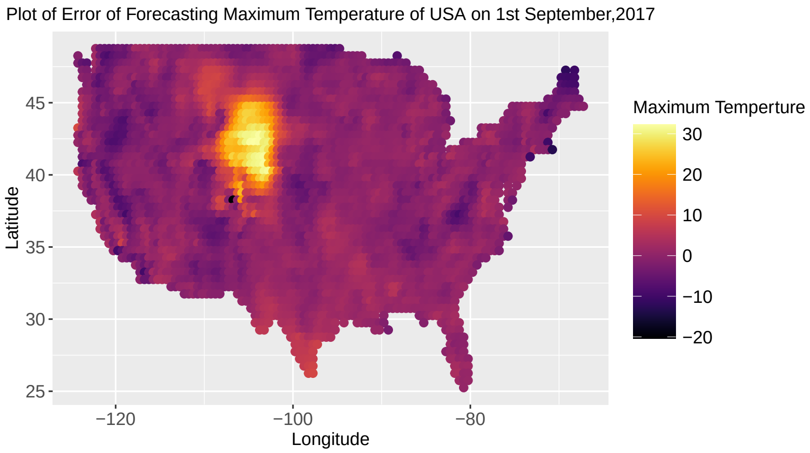

Plot of Error of Forecasting Maximum Temperature of USA on 1st September,2017
Latitude
45
40
35
30
25
−120
−100
−80
Longitude
Maximum Temperture
30
20
10
0
−10
−20

Figure 9: The error in forecasting the maximum temperature of 1st September, 2017

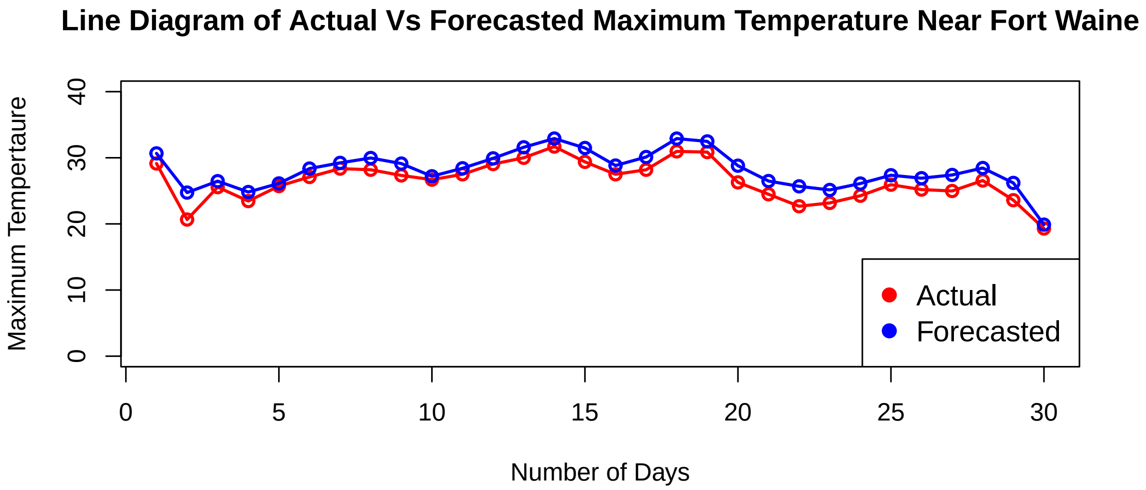


Figure 10: Here red colored line corresponds to the actual observed maximum temperature and blue colored line refers to the forecasted maximum temperature of latitude: 40.75 & longitude: -85.25

3.2 Spatial Bubble Based Model

The example of the location having latitude: 40.75 & longitude: -85.25, (nearest to, Fort Wayne) works well for a multitude of reasons. One, since Fort Wayne is surrounded by peripheral cities with weather forecasts supplied by the Data Expo, it is analogous to how a person may use weather forecasts in nearby cities. Second, although we find that the LASSO technique works reasonably well overall (as seen in the sections that follow), this example also demonstrates the limitations of LASSO. Florida, Arizona, Los Angeles, California are far away from Fort Wayne are in different climate zones i.e., they are at a spatially very distant locations as compared to Fort Waine. It is not very intuitive for Florida, Arizona, Los Angeles or California to be a predictor of weather for Fort Wayne. By looking at the magnitude of the coefficient on that site we see it has relatively little influence compared to the nearby sites in Ohio, Indiana, Illinois and Michigan. As a follow up, in an attempt to address the LASSO over-selection of predictor variables we build a Spatial Bubble Based Model.

The main idea behind this method is, at first we assume a circular region of a given radius around

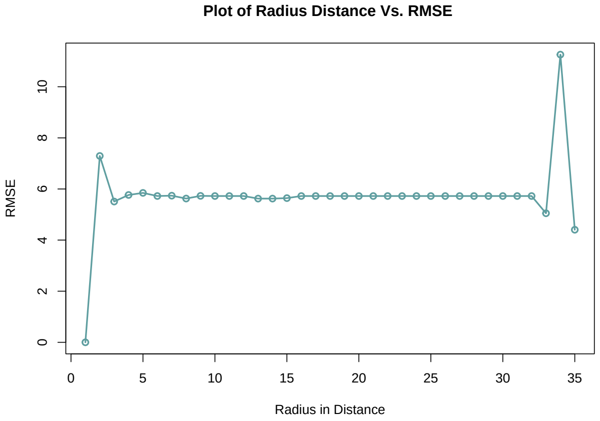


Figure 11: RMSE values corresponding to 30 days forecast for varying choices (different radius) of spatial bubbles

the point we want to forecast and the consider all the airports with equal weight from Data Expo data which fall inside the circle. Finally we fit a LASSO model on those selected covariates to build the final model for forecasting. To judge the optimal length of radius around the point to be forecasted, we perform the same process for a number of possible values of differing radius and select that value of radius for which the root mean square error (RMSE) gives the minimum value.

3.2.1 Forecasting Based on Spatial Bubble Based Model

So, in this model for our example we have assumed a set of circles around latitude: 40.75 & longitude: -85.25 (nearest to Fort Waine) having radius in a sequence of numbers extending from radius 2 units to radius 35 units. Now for each circle having a particular radius we find a set of Airports that are within this circle by comparing the euclidean distance between the airports and the point of interest and then taking them as our new set of covariates. Then, we perform Lasso Model on every such choice of radius on the selected set of covariates and make forecast for 30 days maximum temperature each time for the corresponding selected set of airports and compute the RMSE each time(shown in Figure 11). From the Figure 11 we observe that the RMSE is minimum when the radius is of 2 units and then by fitting the LASSO model on the covariates that are selected inside

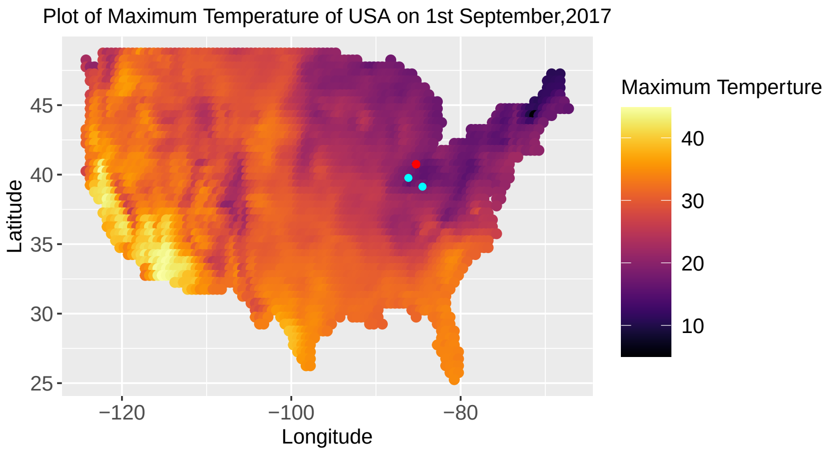


Figure 12: The (cyan) dots are the airports that has been selected by the spatially bubble based LASSO model to forecast maximum temperature of the grid (red dot)at latitude: 40.75, longitude: -85.25

the circular region, finally two airports get selected as final covariates, which are KLUK (cincinnati, Ohio) and KEYE (Indianapolis, Indiana) shown in the Figure 12. Also, we have performed a 30 day(from 3rd August, 2017 to 1st September, 2017) forecast of maximum temperature of latitude: 40.75 & longitude: -85.25, based on the maximum temperature of the selected airports (shown in Figure 12) as covariates and the corresponding lines diagrams are shown in Figure 13. Here, for ith day, the forecasted maximum temperature of latitude: 40.75 & longitude: -85.25 is given by $-1.234501 + 0.1975684(X_{1i}) + 0.8197858(X_{2i})$, where X_{1i} and X_{2i} are the maximum temperature of KLUK and KEYE airports for the ith day respectively.

4 Conclusion

- The LASSO penalized regression model subsets from the 113 airports that are responsible for the weather forecasts.
- Most of the covariates (or, airports) selected by LASSO for modelling the maximum temperature of a particular grid, forms a cluster around the response grid.
- The LASSO method forecasts the maximum temperature quite well but it suffers from over-

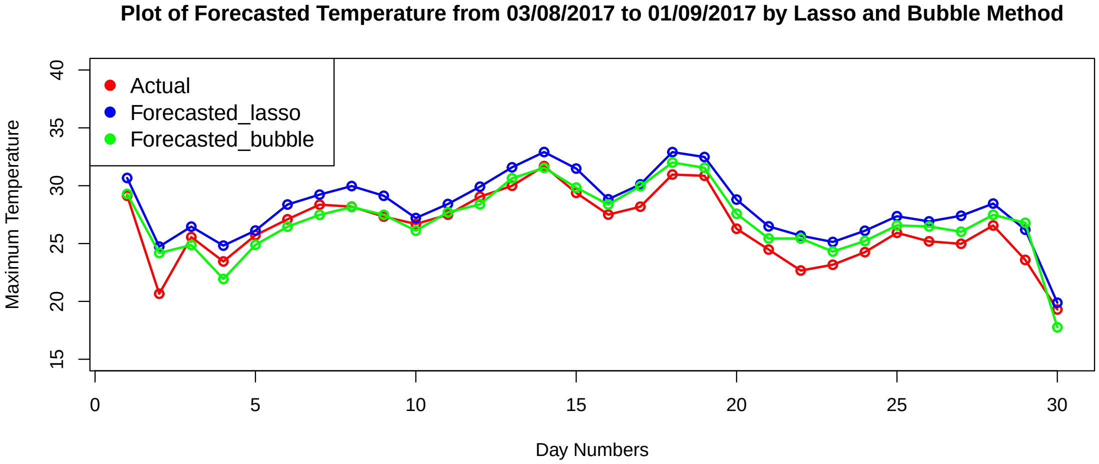


Figure 13: 30 days forecasted maximum temperature of latitude: 40.75 & longitude: -85.25

fitting (or, over variable selection).

- By using, spatial bubble based method and then applying LASSO on the selected airports, we can overcome the over fitting issue and prediction accuracy also increased.

5 Worklog

- Paper finding : Everyone, the current paper reviewed by us is found by Anirban Ghosh.
- Data finding : Khyati Singh
- Data cleaning : Rajdeep Adhya, Khyati Singh, Anirban Ghosh
- LASSO model : Khyati Singh, Anirban Ghosh
- Circular Region based model : Rohit Dutta, Rajdeep Adhya
- Comparison : Rohit Dutta, Rajdeep Adhya
- R codes for Final Plots : Rajdeep Adhya
- Report : Khyati Singh, Rohit Dutta

6 GitHub Link

All the data sets, .Rdata files, R scripts are uploaded in the following link:

https://github.com/rohitdutta22/Spatial$_project.git$

References

[1] Benjamin Schweitzer, Robert C Garrett, Nichole Rook, and Thomas J Fisher. A spatial extension of weather forecasts. *Computational Statistics*, pages 1–15, 2023.

[2] Mine Cetinkaya-Rundel and Wendy Martinez. The 2018 data challenge expo of the american statistical association, 2023.

[3] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[4] Edzer Pebesma and Roger S Bivand. S classes and methods for spatial data: the sp package. *R news*, 5(2):9–13, 2005.

[5] Roger S Bivand, Edzer J Pebesma, Virgilio Gómez-Rubio, and Edzer Jan Pebesma. *Applied spatial data analysis with R*, volume 747248717. Springer, 2008.