Kriging Statistical Methods and Applications 2 Project

Dylan Sain

May 2022

1 Introduction

One of the most common things I hear as a Computer Scientist interested in machine learning is the disparity towards classical statistical models and the more 'advanced' machine learning ones. However the one thing that I learned a lot about in class was that simple linear regression was just the tip of the iceberg for understanding the kinds of things regression can predict and explain. For this project I wanted to explore a more advanced regression, one that we did not have a chance to cover in class. Kriging is one such model that specializes in spacial statistics, or attempting to understand and predict the space between points on a data set. I decided to research this idea, understand it and test it on Lat, Lon data trying to predict the temperatures in Colorado for a month.

2 Kriging Theory

[5] [4] First I had to understand the basic ideas of Kriging. Just to recap, a base linear regression works by trying to find a β such that we can solve the minimization problem:

$$y = x\beta + \epsilon \tag{1}$$

For this example one might say that we could predict temperature by doing something like:

$$y = \beta_1(lon) + \beta_2(lat) + \epsilon \tag{2}$$

However we already know there is a major relationship in space between a x and y value, hence Kriging. Kriging adds a term to the basic linear regression that uses a thing called a variogram to predict values in a spacial sense. This turns our previous problem into a new one:

$$y = x\beta + h + \epsilon \tag{3}$$

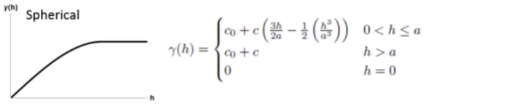
Where h is the new kriging variable that utilizes the variogram and a weight matrix to solve the problem. The basic idea is that it uses a weighted sum of the data using this formula:

$$\hat{Z} = \sum_{i=1}^{N} \lambda_i z(x_i) \tag{4}$$

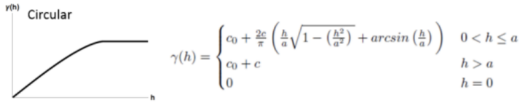
Where the predicted value ($\hat{Z}$) is equal to the sum of the sum of the weights (λ_i) times the values at i ($z(x_i)$). Unfortunately this means that Kriging is very computationally expensive as for each data point that needs to be predicted a new weight matrix must be calculated. The magic happens in a process called Variography and in the creation of a variogram. Essentially using the formula:

$$variogram(distance_h) = \frac{||(value_i – value_j)||}{2} \tag{5}$$

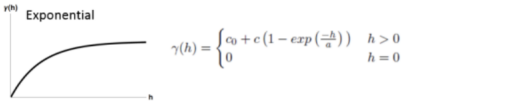
This formula encapsulates the most important idea in basic spacial reasoning, things closer together are more alike then thing farther apart. Using this formula one needs to fit a variogram model to the data. These models include: Circular, Spherical, Exponential, Gaussian, and Linear.



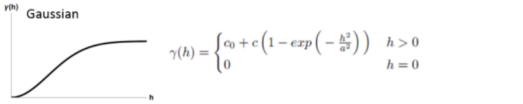
Spherical semivariance model illustration



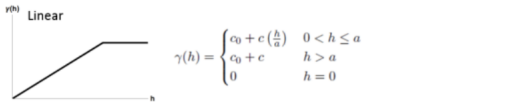
Circular semivariance model illustration



Exponential semivariance model illustration



Gaussian semivariance model illustration



Linear semivariance model illustration

With a variogram model chosen and fit to the data the Kriging process starts. It uses this model to understand the spacial difference between two points and

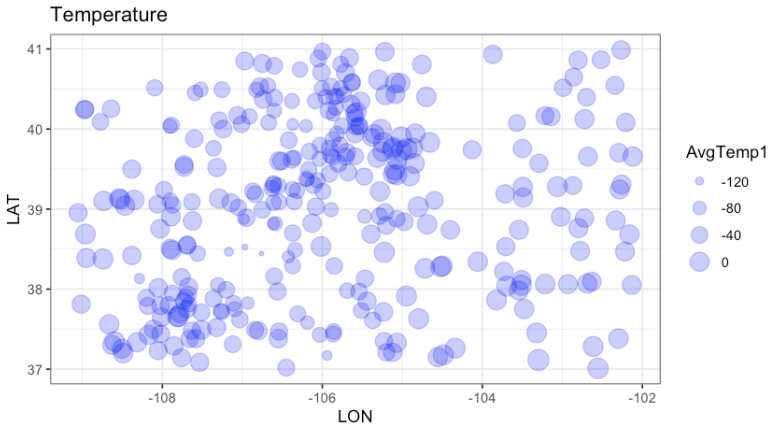
weights are calculated. From there a point can predicted anywhere as long as the weights have been predicted for that x.

3 Data Wrangling

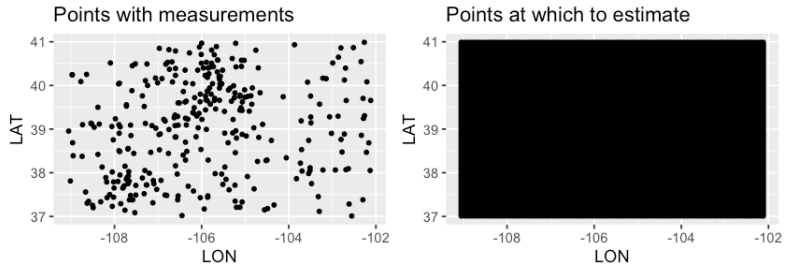
[6] [3] After I understood the process of Kriging I needed to be able to get data on the state of Colorado. The GHCN (Global Historical Climatology Network) has been collecting data on the climate for almost 175 years. It has thousands of stations across the globe and was perfect for getting the Lat, Long and temperature data that I needed for this project. One of the largest problems for this project was getting the data into a format I could use. Unfortunately the main problem was that the temperature data, elevation and latitude and longitude data were all in separate text files. I had to filer in one data set by state and use the IDs of that set to reduce the other. Once then I could merge the two data sets into a reasonable number of files to collect. I was able to reduce it further by filtering by a single year and only taking TMAX and TMIN as observations. This ended with around 331 stations around Colorado. I wanted to predict the temperature and thus averaged the temperatures over the course of a month. One of the first things I noticed was the inclusion of some NA values:

```
      ID                 LON              LAT             ELEV
Length:331         Min.   :-109.1   Min.   :37.01   Min.   :1037
Class :character   1st Qu.:-107.3   1st Qu.:38.05   1st Qu.:1672
Mode  :character   Median :-106.0   Median :39.12   Median :2395
                   Mean   :-105.9   Mean   :39.04   Mean   :2348
                   3rd Qu.:-105.1   3rd Qu.:39.94   3rd Qu.:2957
                   Max.   :-102.1   Max.   :40.99   Max.   :3542
                   NA's   :6        NA's   :6       NA's   :6
   AvgTemp1           AvgPrcp1
Min.   :-127.887   Min.   : 0.000
1st Qu.: -45.000   1st Qu.: 2.613
Median : -25.500   Median : 5.581
Mean   : -26.554   Mean   : 6.787
3rd Qu.:  -4.492   3rd Qu.: 9.774
Max.   :  34.032   Max.   :25.355
NA's   :12         NA's   :8
```

These could have very quickly messed up the variogram calculations later on and as such I elected to remove all stations with an NA value, leaving me with 319 stations instead of the original 331.



Looking at the initial values I saw we had a pretty decent view of the entire state with good data centered around the mountains were things were bound to get interesting with the heavy changes in elevation. Next step was to create a spacial data set that would be used as predictors. I used the elevatr library to find the elevations for a data set inside Colorado.

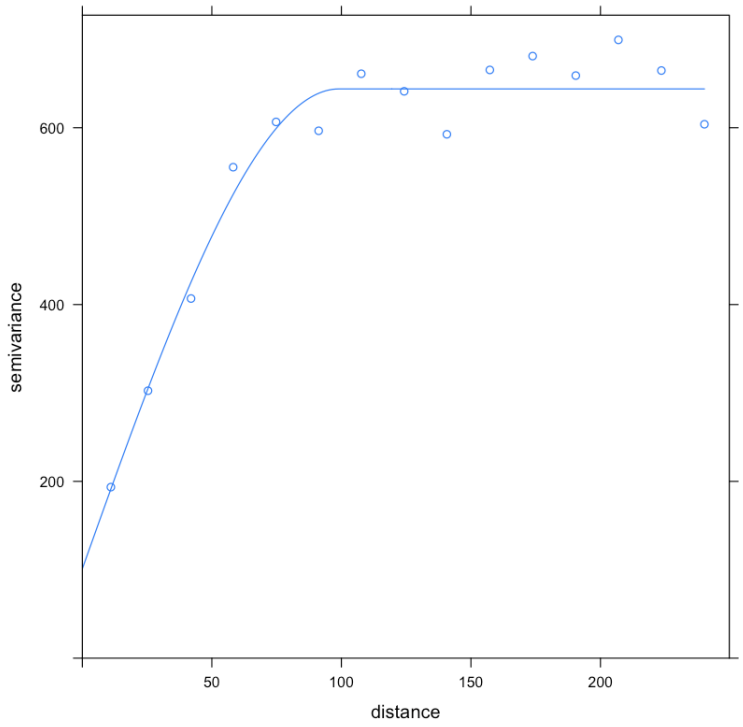


Unfortunately due to the naturally square shape of Colorado, this graph doesn't end up all that interesting to look at, but its good to be able to visualize the points at which the model will estimate. Overall the model was used to estimate 10,000 lat, lon pairs through out the state of Colorado.

4 The Model

[2] [1] Now that the data was in a use able format I had to fit the variogram. I tested multiple different variogram models, however due to the sudden leveling

out at around distance level 100, it became obvious that the spherical variogram model was the best model for this data set.

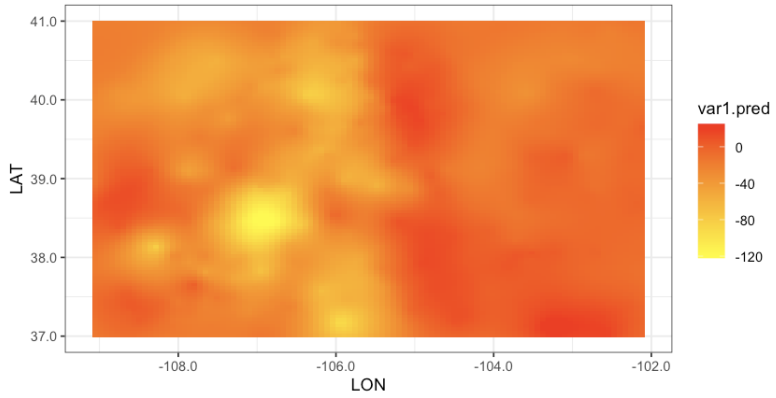


With a variogram model chosen I simply had to choose my equation and run the kriging. I ran it using two different equations. A base model with just the kriging estimator attached to it $AvgTemp \sim 1$ and a full model with both Lat and Elevation $AvgTemp \sim LAT + ELEV$.

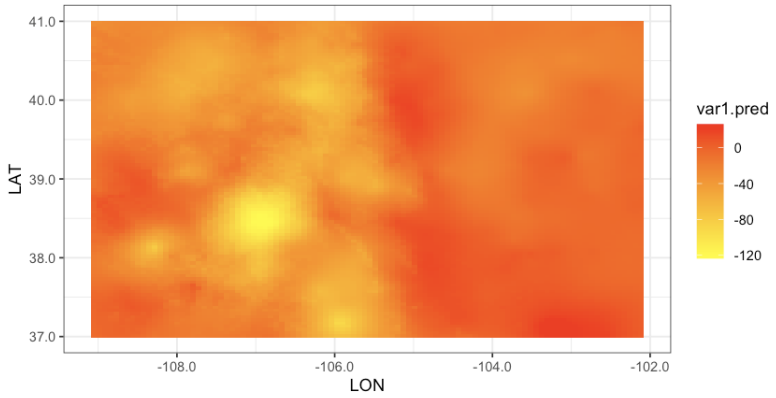
5 Results

There were the models for the month of January in 2015.

The base model



The full model



At first glance the models are extremely similar, however there are some differences in the full model that explain some variability in the data that the base model doesn't have. Both Elevation and Latitude are instrumental in temperature readings. A higher elevation tends to correspond to a lower temperature and same with Latitude. We can see that around the 105 Lon, and 40 latitude there is a sharper increase in temperature. This is most likely due to a sharp decrease in elevation that was not accounted for the in base model.

6 Next Steps

I would like to explore further the kind of analysis that Kriging offers. While there was a change in the graphs between the base and the full model would have liked to see what other data and predictors might have affected the model as a whole. I also would like to explore other time periods and wee if there is a yearly change in the data.

7 Conclusion

Overall this project was successful as I was able to understand Kriging and work with a real life data set collected across the years. I learned a lot about how Kriging worked and the reasoning behind spacial analysis.

8

References

[1] How kriging works.

[2] Kriging interpolation.

[3] Global historical climatology network daily (ghcnd), Dec 2021.

[4] Nabil A. Introduction to kriging in r, Oct 2015.

[5] PhD Aisha Sikder. Building kriging models in r, Oct 2020.

[6] Spatial Reasoning and Spatial Reasoning. R: Reading amp; filtering weather data from the global historical climatology network (ghcn).