Michael Shoemate

Applied Multivariate Analysis Final

Problem 1

The most important GDT values to explain dynamics of national GDT in Mexico are provided on the left. The variables that are not useful to explain national GDT are on the right, ordered by decreasing significance.

Most Important GDT

Colima
Distrito Federal
Estado de México
Jalisco
Nuevo León
Sonora
Tabasco
Veracruz

Not Useful GDT

Oaxaca
Nayarit
Tamaulipas
Yucatán
Sinaloa
Chiapas
Baja California
Morelos
Quintana Roo

The variables considered useful to GDT explain 98% of the variation, where Mexico GDT is thresholded by its mean. I performed the backward variable selection procedure on a logistic regression model with all 17 GDT predictors. At each step, the state with the least effect on national GDT was removed.

Dengue Confusion Matrix	**Actual Low**	**Actual High**
Predicted Low	238	5
Predicted High	2	110

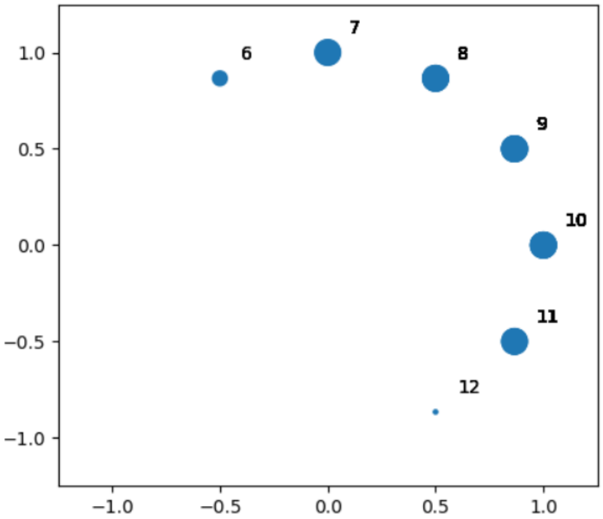
In this scenario I wanted to compare the importance of predictors without the effect of missing values, so I dropped any records with missing values. Attempting to impute missing values before applying the selection procedure would unnecessarily bias the results toward states with complete data. In recent years, all states have complete data, so this is not a significant factor for this form of analysis.

I also used Factor Analysis to look for groups of variables and found (after standardization and Varimax[1]) three latent variables that represent groupings of cities. Morelos, Nuevo León, and Tamaulipas were significant variables in the first loading, Baja California, Sonora and Sinaloa were significant variables in the second loading, and Jalisco, Colima and Nayarit significant in the third loading. Significant is defined as an absolute value greater than .7. These groupings of variables also happen to be near each other. It should be noted that variables considered similar by factor analysis are less likely to be selected together by the previous logistic regression selection procedure.

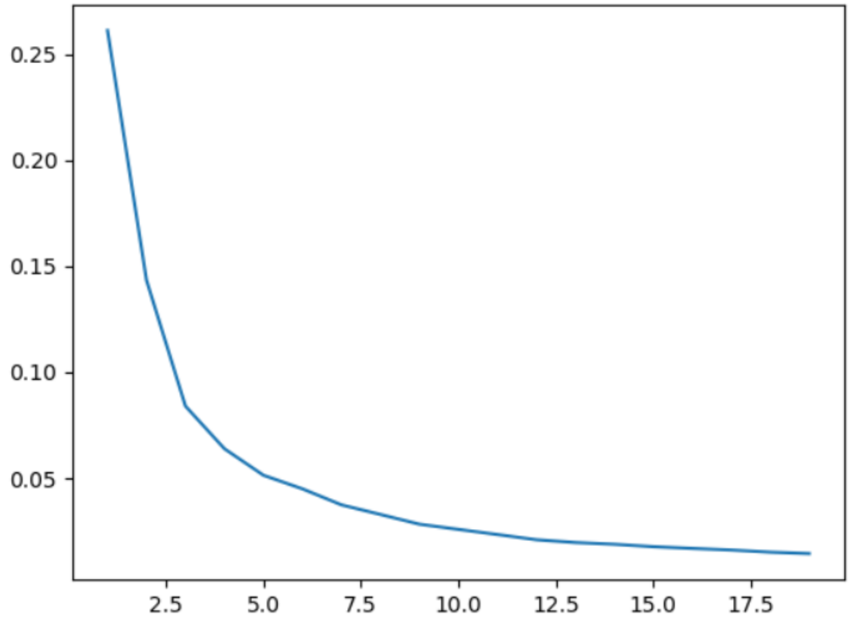
Question 2

Considering this is data on disease frequency, it would be much more useful to identify clusters that specifically represent outbreaks. Therefore if I perform clustering on the original data, my centroids are likely not going to have as useful of an interpretation. Throughout the remaining analysis, the data is weighted by the disease index.

I performed clustering with a variety of perspectives on the data. First off, I wanted to know if there was a "Dengue season" of the year. Since month is a circular variable, considered month an angle and projected into a rectangular space. The mean in the rectangular space is then mapped back to the polar space and transformed back to a numeric month. Among records considered 'disease events,' there was a marked tendency towards the end of the year. The middle of flu season is 9.198, or mid-to-late September. The plot shows flue frequencies on a circle, labeled by month. This analysis considers only events with dengue index greater than the .8 quantile.



I also performed clustering on the dates with KMeans, weighted by the national dengue index. Using the weights, dates is clustered under consideration of the severity of disease events. The plot measures total distance (Y) for n initial cluster centers (X).



Based on the bend, I chose five as the estimated number of national outbreaks. The clusters determined from this method are provided below. Ultimately the centroids indicate the months that national Dengue was worst among 2004, 2007, 2009, 2012 and 2014. Other years represented in the data had weaker dengue seasons. The null values in the data did not affect this analysis, because only the country data was used.

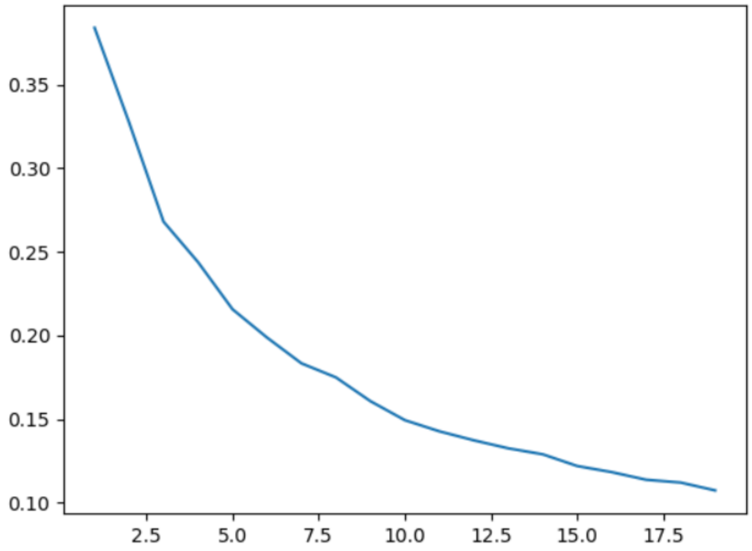
Year	**Month**
2004	5
2007	7
2009	9
2012	9
2014	12

To answer the question, “are there states that tend to be affected similarly each year?” I performed clustering on the GTD state columns. In this perspective on the data, each observation is a time series. Unfortunately, different states may experience similar patterns within each season, but have lagged or offset responses. I compensated for this by performing agglomerative clustering with distances based on dynamic time warping (DTW). The dynamic time warping distance metric is credited to Alex Minaar [2]. I modified the DTW metric to work with SKlearn’s pairwise distance metric and used the complement as an affinity matrix for agglomerative clustering. The sequences compared by DTW don’t necessarily have to be the same length, but I ultimately dropped the columns at the end that were too sparse, and imputed missing values on the rest.

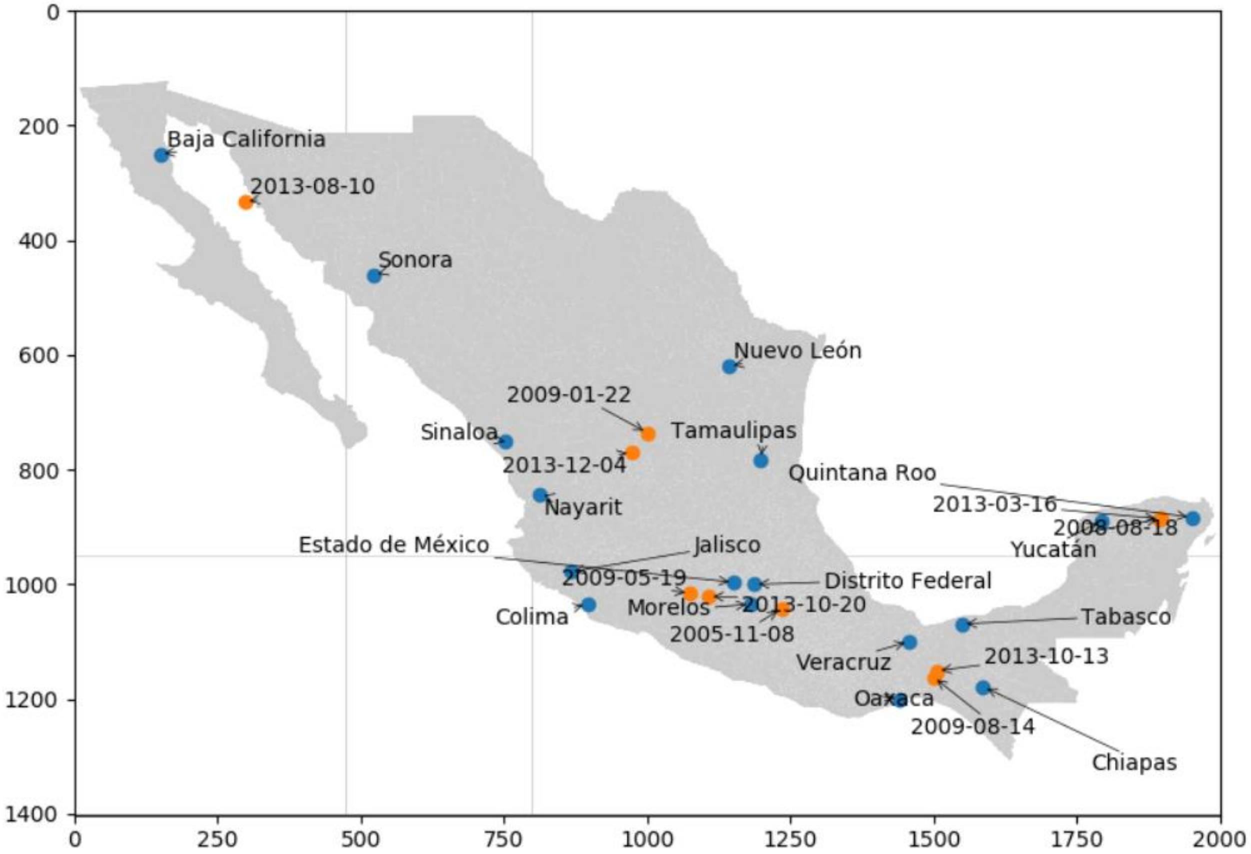
In any given dengue season, the behavior of one city may be an indicator for any other city in the group.

1	Estado de México	Veracruz				
2	Colima	Morelos	Nayarit	Oaxaca	Sonora	Yucatán
3	Baja California	Chiapas				
4	Distrito Federal	Nuevo León				
5	Sinaloa	Tabasco				
6	Jalisco	Quintana Roo	Tamaulipas			

Finally, I wanted to identify regional outbreaks with respect to time. To accomplish this, I first melted the data to three columns: ‘Date’, ‘Location’, and ‘value’. I also wanted to preserve locality, as outbreaks may spread geographically, so I replaced ‘Location’ with ‘Latitude’ and ‘Longitude’ representing each of the states. I performed KMeans clustering again, but this time the weighting parameter is the regional dengue index ‘value’. There isn’t a strong bend in the total distance (Y) from cluster centers as I vary the number of clusters (X), so I chose 10.



The ten centroids that best describe Dengue outbreaks are shown in yellow. The locations I gave to the centers of the states are labeled in blue. The graph shows for instance, the region of Quintana Roo and Yucatàn suffered a Dengue outbreak in 2013 and 2008. Notice that most of the detected outbreaks occurred during the flu season discussed previously.



References

[1] Ben FrantzDale. *Other Algorithms*. https://en.wikipedia.org/wiki/Talk:Varimax_rotation

[2] Alex Minaar. *Time Series Classification and Clustering with Python*. http://alexminnaar.com/time-series-classification-and-clustering-with-python.html

[3] Addicted04. *Municipalities of Mexico.* https://commons.wikimedia.org/wiki/File:Municipalities_of_Mexico_(equirectangular_projection).svg