NC STATE UNIVERSITY
Department of Computer Science

Air pollution prediction: Time-series clustering & ARIMA

Aditya Upadhyay, Daxkumar Amin, Priyance Mandlewala, Nilay Kapadia, Vaibhava Lakshmi, Vishrut Patel
North Carolina State University

1. Introduction

While dealing with time series forecasting, regression methods such as ARIMA, LSTM and SVR are used. With our data, pre-processing is essential to be carefully done before any of the above methods are employed. K-means with Dynamic Time Warping is explored for this reason.

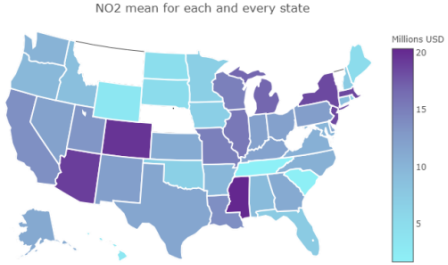

Figure 1. Air pollutant level in USA

2. Data description

- 28 attributes for 2000-2016 of 47 states from the U.S
- 4 pollutants – NO2, SO2, CO, O3
- Data trimmed for model training
- Generalized based on :
  Time – Hourly to Daily
  Location – County to State

References

[1] Niennatrakul V, Rantanamahatama, On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping. 2007 IEEE Seoul.
[2] Jhun I, Coull BA, Schwartz J, Hubbell B and Koutrakis P 2015 The impact of weather changes on air quality and health in the United States in 1994–2012 Environ. Res. Lett. 10 084009

3. Technical section

- With vanilla Autoregressive Integrated Moving Average and no clustering, identified that prediction is skewed - data is state-wide and dense (w.r.t time).
- K-means on state-wise time series, forming clusters of related states.
- Used Dynamic Time Warping (O(n2)) as the distance measure in K-means. Euclidean distance is incompatible with time series.
- Regression method - ARIMA employed. Now, on each of the cluster centroids.

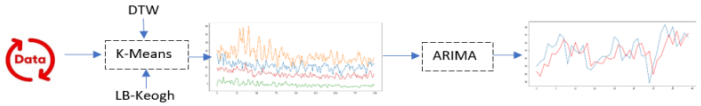

Figure 2. Implementation

- Additionally implemented K-means with DTW on monthly time series and achieved season based clusters. ARIMA can be used on these clusters for forecasting season based pollution values.

4. Results

Cluster	States	RMSE
1	IN, LA, MS, TX, NC, CT, WI, GA, UT, NM	2.227
2	DC, IL, MI, NJ, NY, PA, VA, MA	2.349
3	NV, TN, SC, IA, ME, WY, ND, SD, OH, HI, RI	0.158
4	FL, KS, KY, NH, MD, AR, ID, OR, DE, AL, WA, MN	1.209

Table 1. Cluster formed by K-Means

- **Intra-cluster distance:** 7125.49827
- **Inter-cluster distance:** 856.600
- **Total Sum Square (Variance):** 7982.098

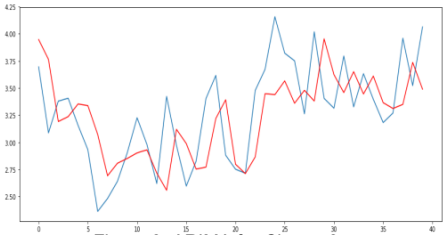

Figure 3. ARIMA for Cluster 3

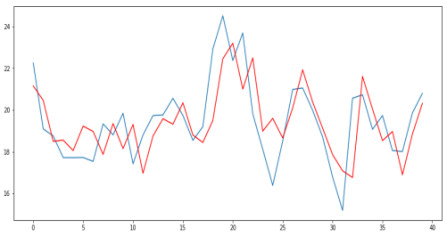

Figure 4. ARIMA for Cluster 3

5. Parameter choices

K-MEANS	Clusters: 4 Iterations: 10
DTW	Alignment: 5 LB_Keogh reach: 5
ARIMA	Lag order: 5 Degree of differencing: 1

Table 2. Parameter Choices

6. Alternative Methods

Exploratory studies on our data were performed with time series regression methods.

- ARIMA without clustering - Trained and tested on the entire dataset. Provides an RMSE of **279**.
- Learned that more pre-processing is required before predicting.
- To test, manually extracted state specific data.
- LSTM - Trained and test on data from the state of Arizona for 2000-2016. ***RMSE of 3.378 using 30 hidden layers and an epoch of 5.***
- LSTM approach, a RNN method remembers past predictions and provides good results. However, this requires processing power. But proves that data handling provides better results.

7. Conclusions

- K-Means clustering followed by ARIMA can be used for time series prediction.
- We observe that pollutant levels follow seasonal behavior that is clustered using K-Means clustering. States that follow similar pollutant level behavior were clustered together.
- Finally, ARIMA can be used to create time-series regression over the clusters for prediction