Predicting Particulate Matter Concentration Levels in Fresno County

Helena Card
Advisor: Dr. Steve Chung
California State University, Fresno
Department of Mathematics

Introduction

The goal of this study is to model the relationship between particulate matter and meteorological variables using multiple linear regression, support vector machines, and neural networks. We hope to discover which model preforms the best at predicting particulate concentration levels.

Pollution can have many adverse health effects, including respiratory disease, cardiovascular disease, and cancer. Pollutants can occur naturally, such as volcanic ash, and can be created, like the smoke from a car exhaust pipe. Here, we focus our research on two types, PM2.5 and PM10 particulates. PM2.5 particles are mostly derived from combustion of gasoline, oil, or fuel and are 2.5 micrometers or smaller in diameter. As PM2.5 particles are very small, about 28 times smaller in diameter than a strand of hair, these particles can cause health issues and potentially enter the bloodstream. PM10 particles have a diameter of 10 micrometers or smaller. They have some of the same emission sources as PM2.5; as well as industrial emissions, dust, ash, and pollen. Nazif et al. (2018) completed a study using regression and multivariate models to predict PM10 and PM2.5 levels in Malaysia. We apply this experiment to data from Fresno County from 2015-2021, focusing on climate factor and additionally, seasonal effects. Several previous studies such as: Latha and Badarinath (2005), Outapa (2019), Zhang et al. (2006), and Chandu and Dasari (2020) found that PM levels were greatly influenced by seasonal changes.

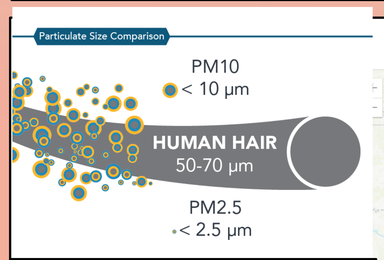


Figure 1: Source: California Air Resource

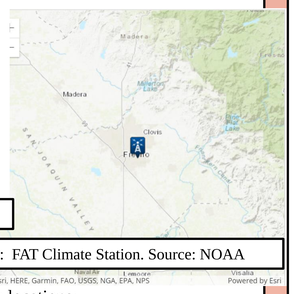


Figure 2: FAT Climate Station. Source: NOAA

Study Area

PM data was primarily collected from two locations:

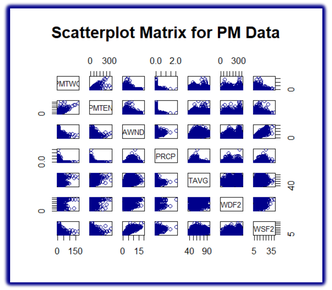
- 908 N Villa Ave, Clovis, CA 93611(36.84, -119.71)
- 3727 N First St., Fresno, CA 93726 (36.78, -119.77)

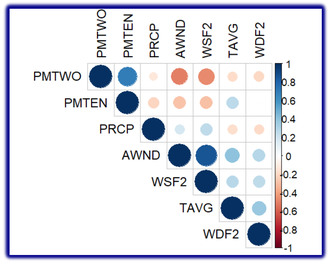
Climate data was collected from:

- Fresno Yosemite International Climate Station, CA US 36.77°, -119.72° (Source NOAA)

Research Methods

Day et al. (2015) stated that changes in meteorological characteristics such as rainfall, wind speed, or temperature can affect the severity of air pollution problems. Thus, we have focused on these three main climate factors in our study. The National Centers for Environmental Information (NOAA) has 50 climate stations in Fresno County. The data provided by the Fresno Yosemite International Climate Station used in this study includes the following daily summaries: average temperature (TAVG), average wind speed (AWND), precipitation (PRCP), direction of fastest 2-min wind (WDF2), and fastest 2-min wind speed (WSF2), along with the PM2.5 and PM10 data from 2015-2020. This dataset was used for model development. Data from 2021 was used to test the predictive models. The dataset was also split into two seasons, the hot season from June to August and the cold season from September to May, these results will be included in our journal at a later date. Data was also split by year for further analysis by time series plot. Missing data was omitted from analysis.



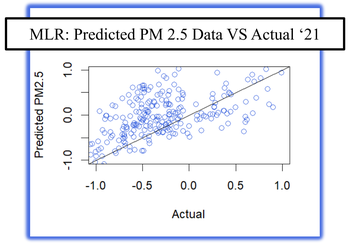


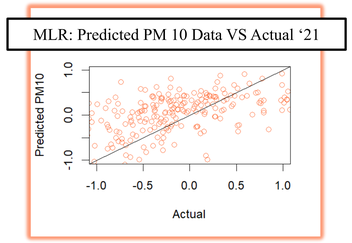
Results

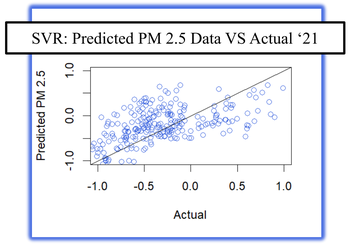
We compared three models to find which best predicts 2021 PM data: Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Neural Network (NN). We used data collected from 2015-2020 to train each tool.

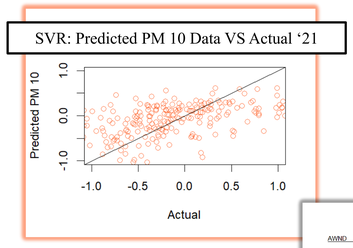
Multiple Linear Regression

Our MLR showed a p-value: < 2.2e-16 for both PM2.5 and PM10. Thus we can conclude that our results are statistically significant.

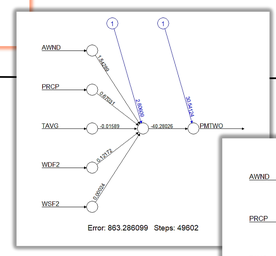


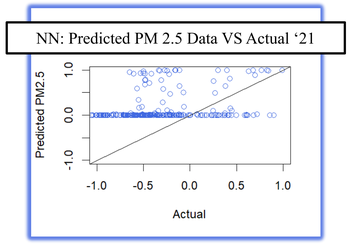


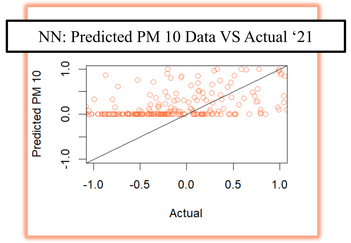




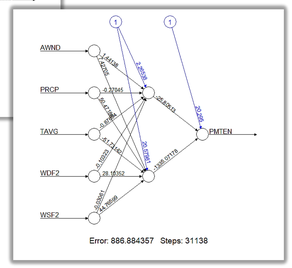
Support Vector Regression







Neural Network



Conclusions

Table Results from PM 2.5

Full Data Model	R2	MSE	MAE	PA	IA
MLR	**0.47412**	**0.77238**	0.57369	**0.26696**	0.86162
SVR	0.46429	0.80747	**0.54763**	0.15315	0.90733
NN	0.32301	0.92816	0.69906	0.12825	**0.91844**

Table Results from PM 10

Full Data Model	R2	MSE	MAE	PA	IA
MLR	0.56092	**0.68096**	0.59371	**0.29420**	0.85658
SVR	**0.57282**	0.69898	**0.56723**	0.20632	0.88853
NN	0.53671	0.79519	0.68373	0.10392	**0.93376**

When comparing the results of MLR, NN, and SVR, the models with higher R^2, PA, and IA and lower MSE and MAE preform the best.

The bolded variables in the tables above indicate the best results. We can see the results for PM2.5 show MLR as the best model. For PM10 it is less obvious whether MLR or SVR preform the best, however MLR preforms better by a small margin. Thus, we conclude that MLR is the best model for predicting particulate matter concentration levels in Fresno County, based on the full set of data. We continue our research on seasonal effects in our journal to follow.

Literature Cited

Primary Resource:

Nazif, A., Mohammed, N. I., Malakahmad, A., & Abualqumboz, M. S. (2018). Regression and multivariate models for predicting particulate matter concentration level. Environmental Science and Pollution Research, 25, 283-289.

For a full list of references, please follow QR code:

Acknowledgments

This work was supported by a U.S. Department of Education grant Title V Promoting Postbaccalaureate Opportunities for Hispanic Americans (P031M210039) for the project Finish in Five. Special thanks to Dr. Steve Chung and all the mentors of Finish in Five at Fresno State.