Considering Anthropogenic Factors in PM 2.5 Machine Learning Models

Emily Chang (ec5ug@virginia.edu)
University of Virginia

An Unexplored Area

A literature review found that environmental features are over-represented in PM 2.5 predictive models.

While human activity is often the main culprit of PM 2.5 emissions, models fail to account for anthropogenic factors.

Data Examined

Data was split across season as training models separately by season improved performance.

Each seasonal model was trained on two sets of data: (1) environmental and (2) environmental and anthropogenic.

Environmental Data

- temperature
- dewpoint
- relative humidity
- precipitation
- wind speed
- wind direction
- pressure
- prior day's PM 2.5

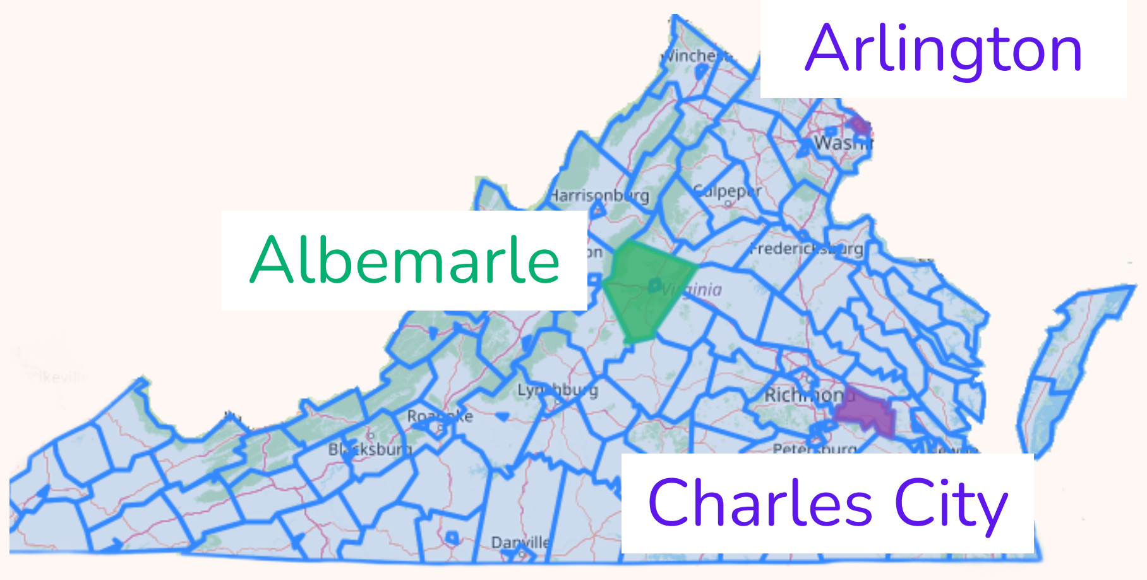
Anthopogenic Data

- unemployment
- population density
- buildings built

Study Area

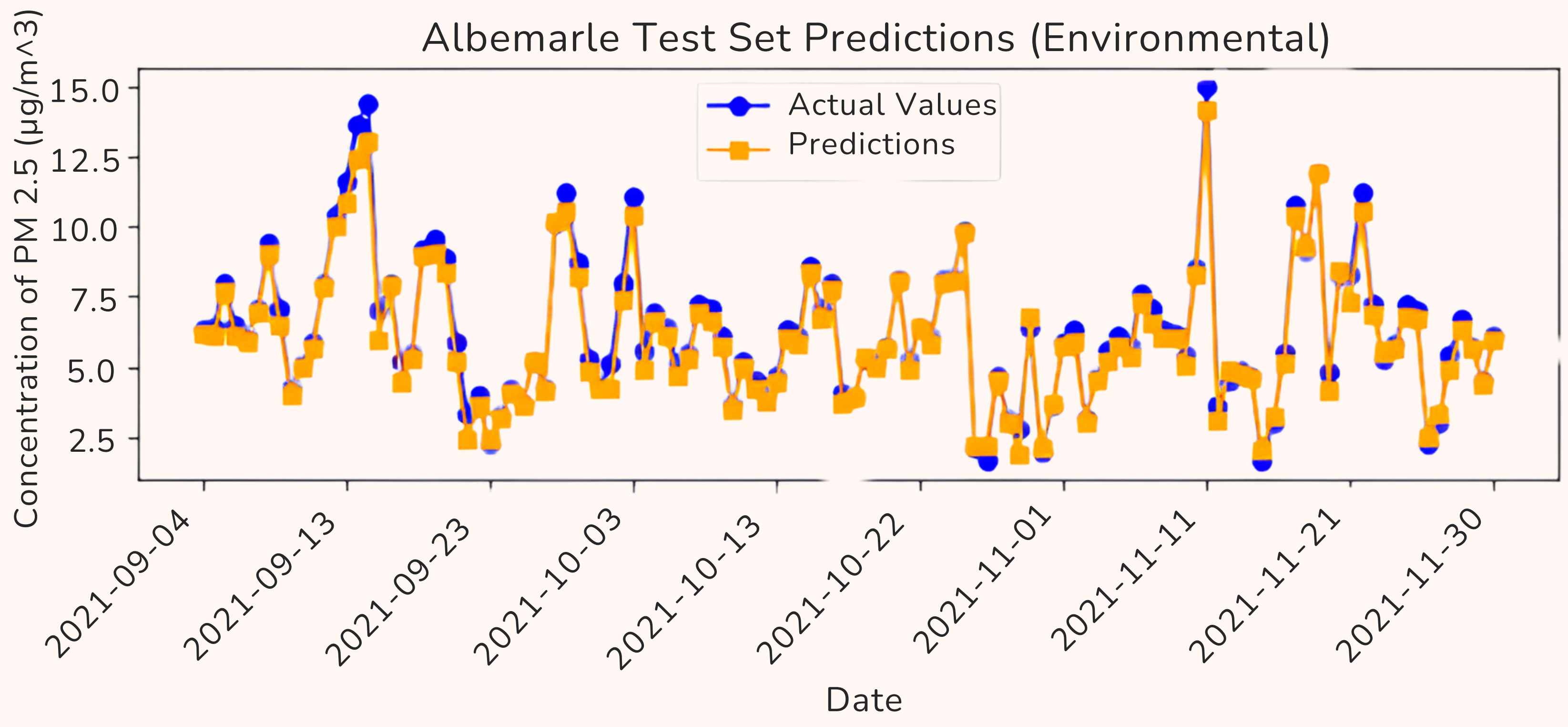
Traditional models (e.g. linear regression) and deep learning models (e.g. LSTMs) were trained on data from Albemarle County.

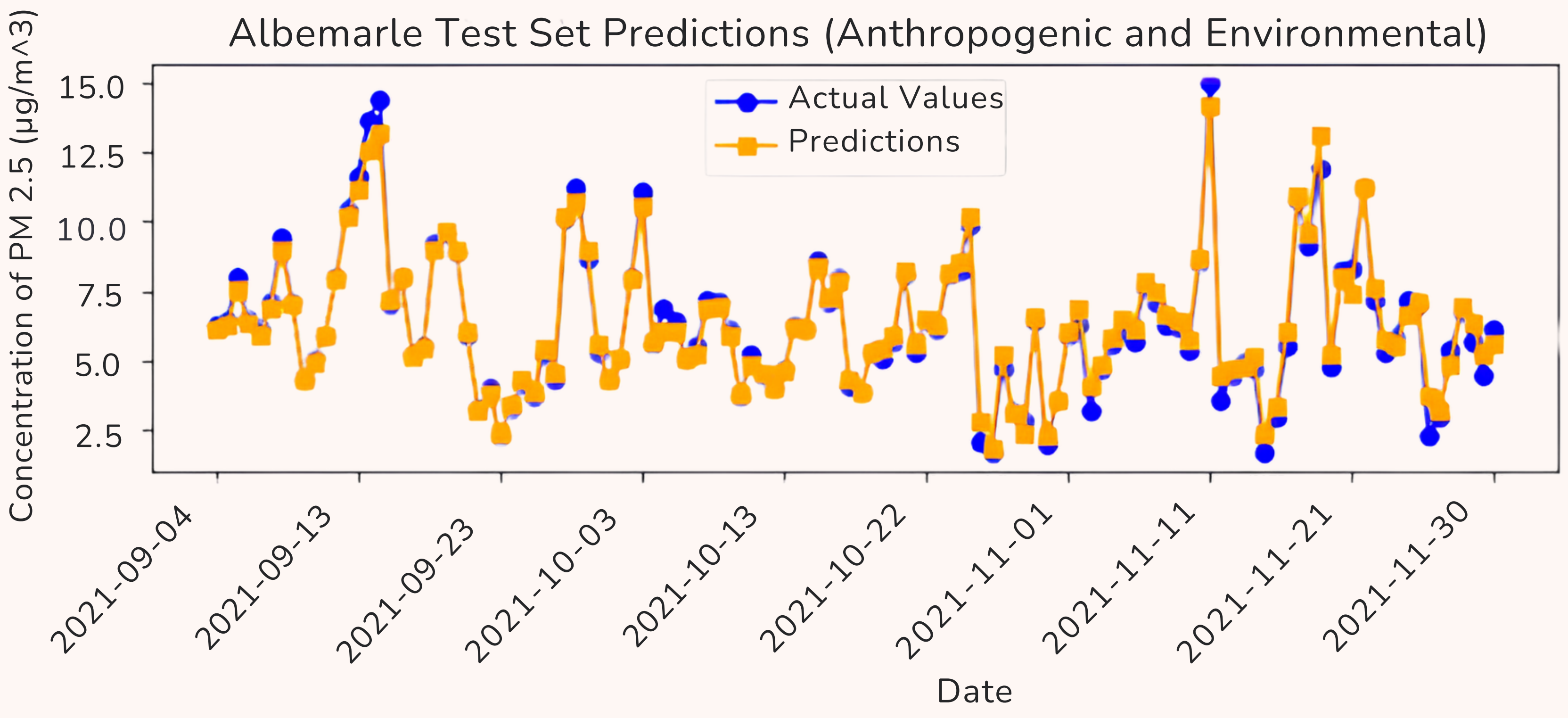
To test the model's ability to generalize across counties, these models were evaluated on a rural sample: Charles City County, and an urban sample: Arlington County.



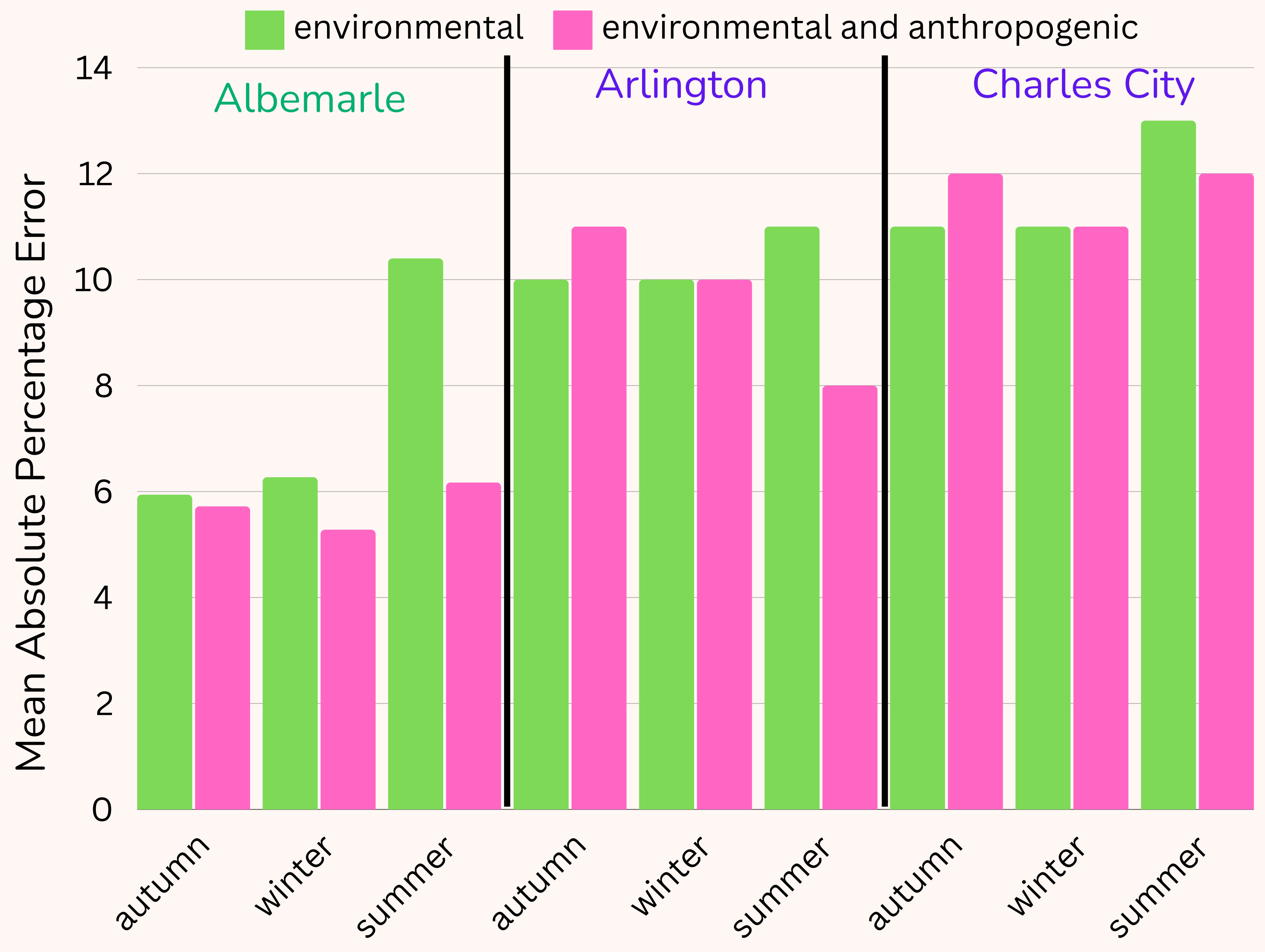
LSTM is king

LSTMs trained on either dataset are able to accurately predict a year's worth of PM 2.5 values. The autumn models are depicted below.



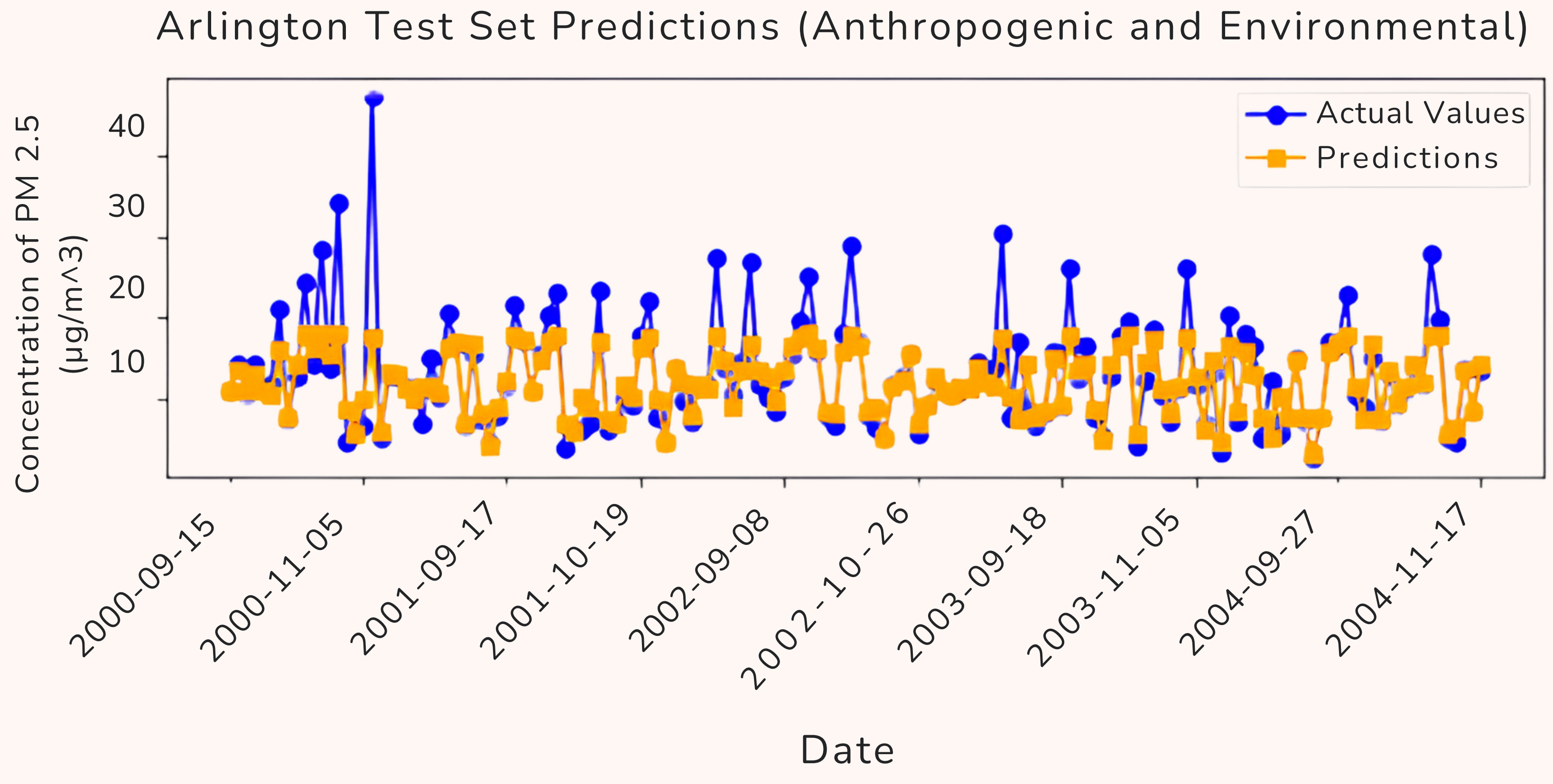


Anthropogenic and environmental data can improve performance



Poor Generalization

- Anthropogenic and environmental data consistently improves performance when the data is in distribution.
- Models trained on anthropogenic and environmental data may overfit to the training data as the models cannot predict spikes in pollution in Arlington and Charles City.



This project has been part of the "Machine Learning for Virginia" project at UVA. Alexander Davis and Kenny Zhang are recognized for their contributions.