Using Machine Learning to Forecast Air Quality in Beijing

KEVIN LIMKRAILASSIRI

MENTOR: JAN ZIKEŠ

Motivation

Poor air quality in Beijing is a well-known concern causing adverse health affects and affecting the quality of life of every citizen.

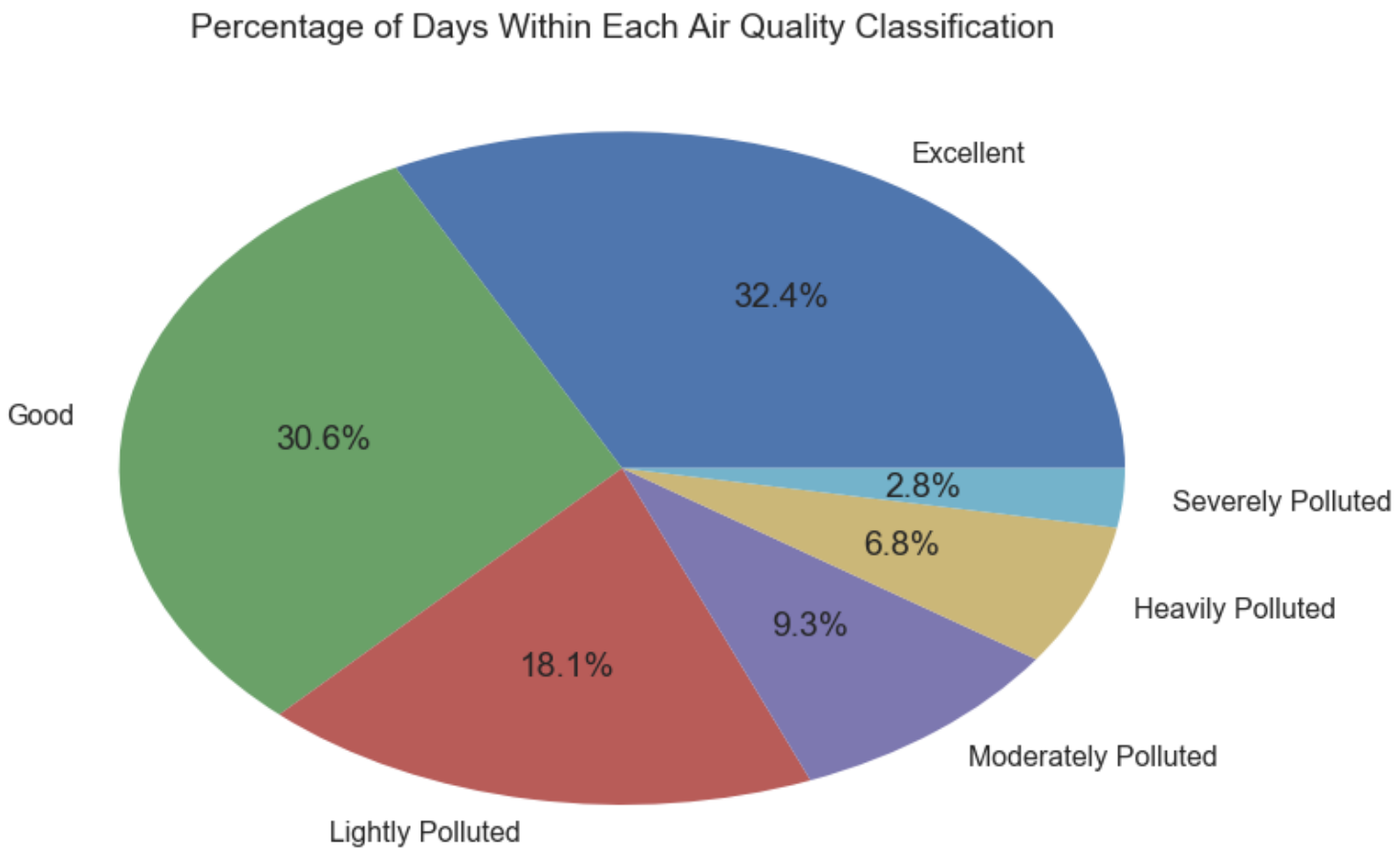
The US Embassy and Chinese government have established measurement centers to provide hourly PM2.5 data to the public.

Beyond merely instantaneous measurements of air quality, forecasting air quality several days ahead would allow citizens to anticipate days when poor air is predicted.

Herein, we employ linear regression and a type of recurrent neural network called long short-term memory to generate predictions of air quality using past air quality and weather data in Beijing.

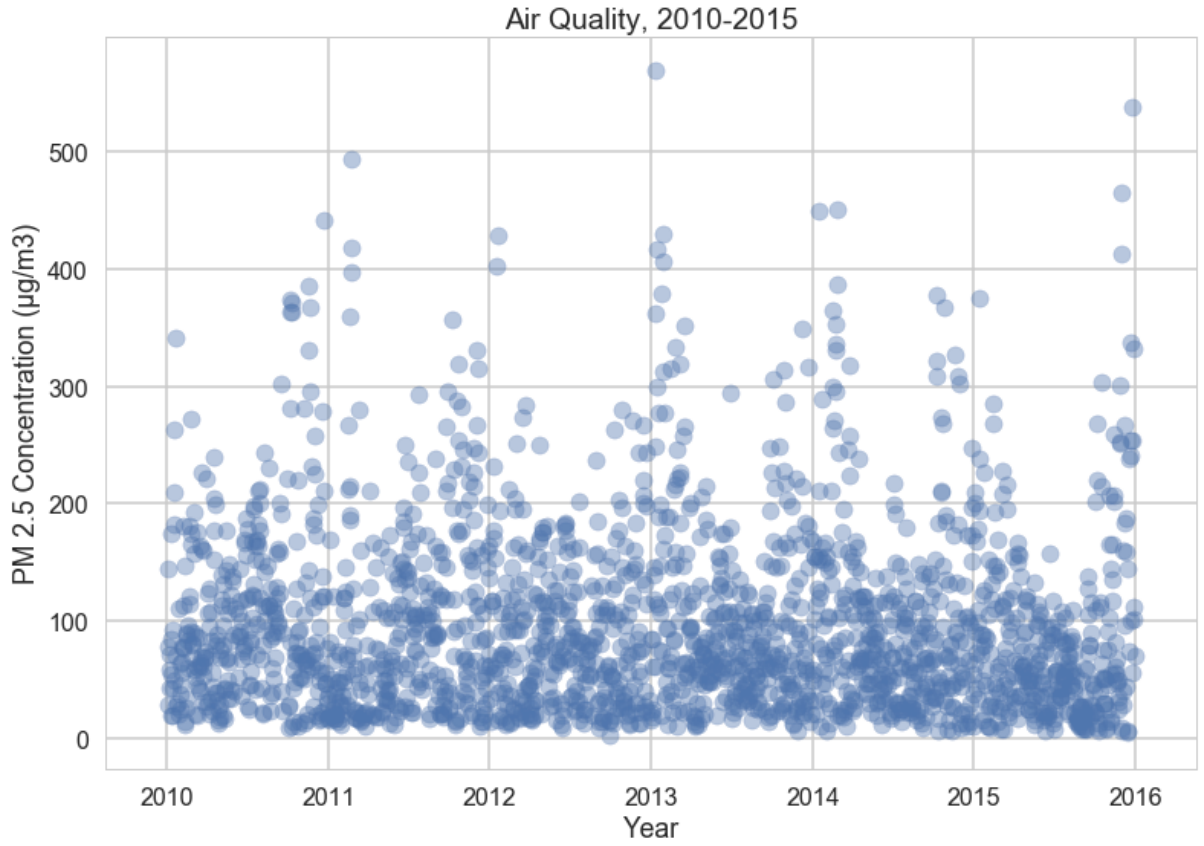
Exploratory Data Analysis

Beijing is Facing a Major Crisis in Air Quality

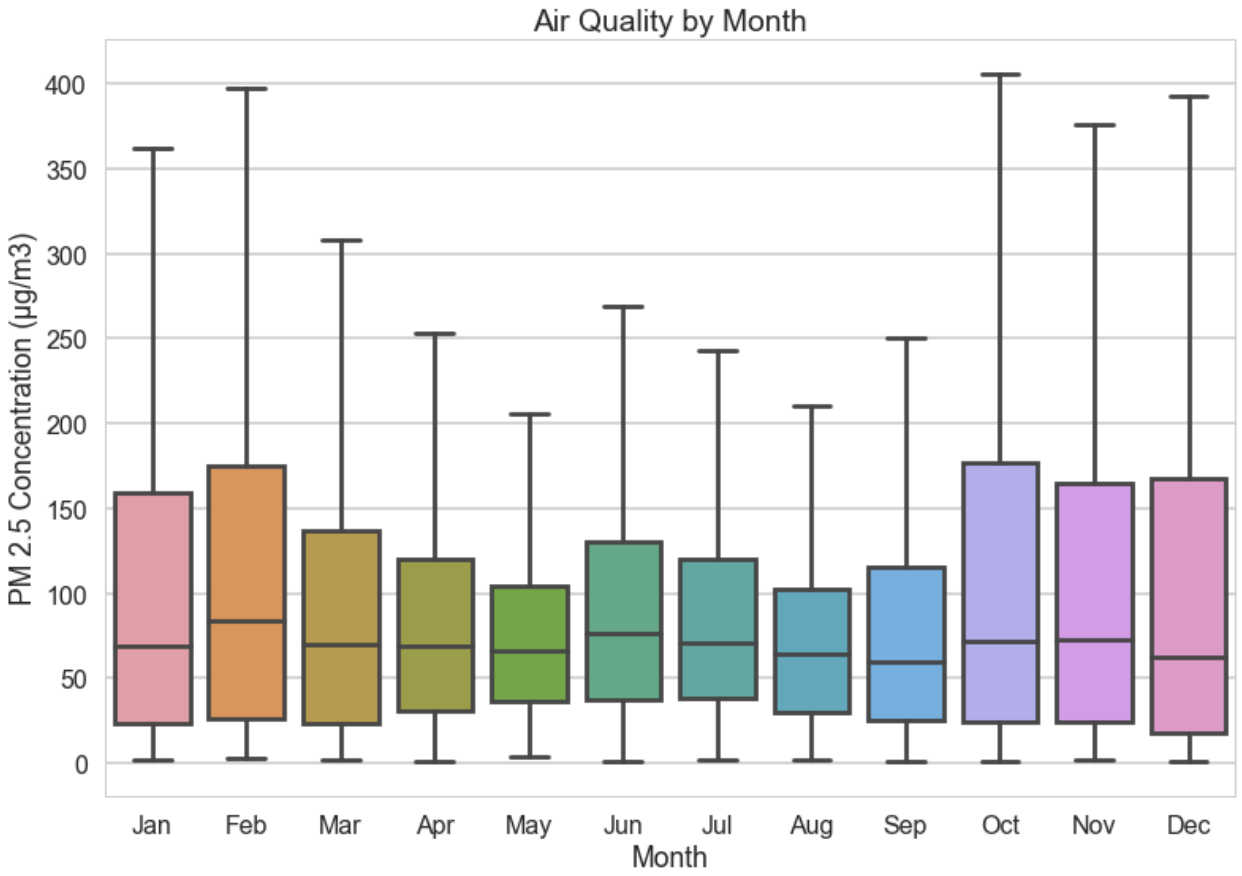


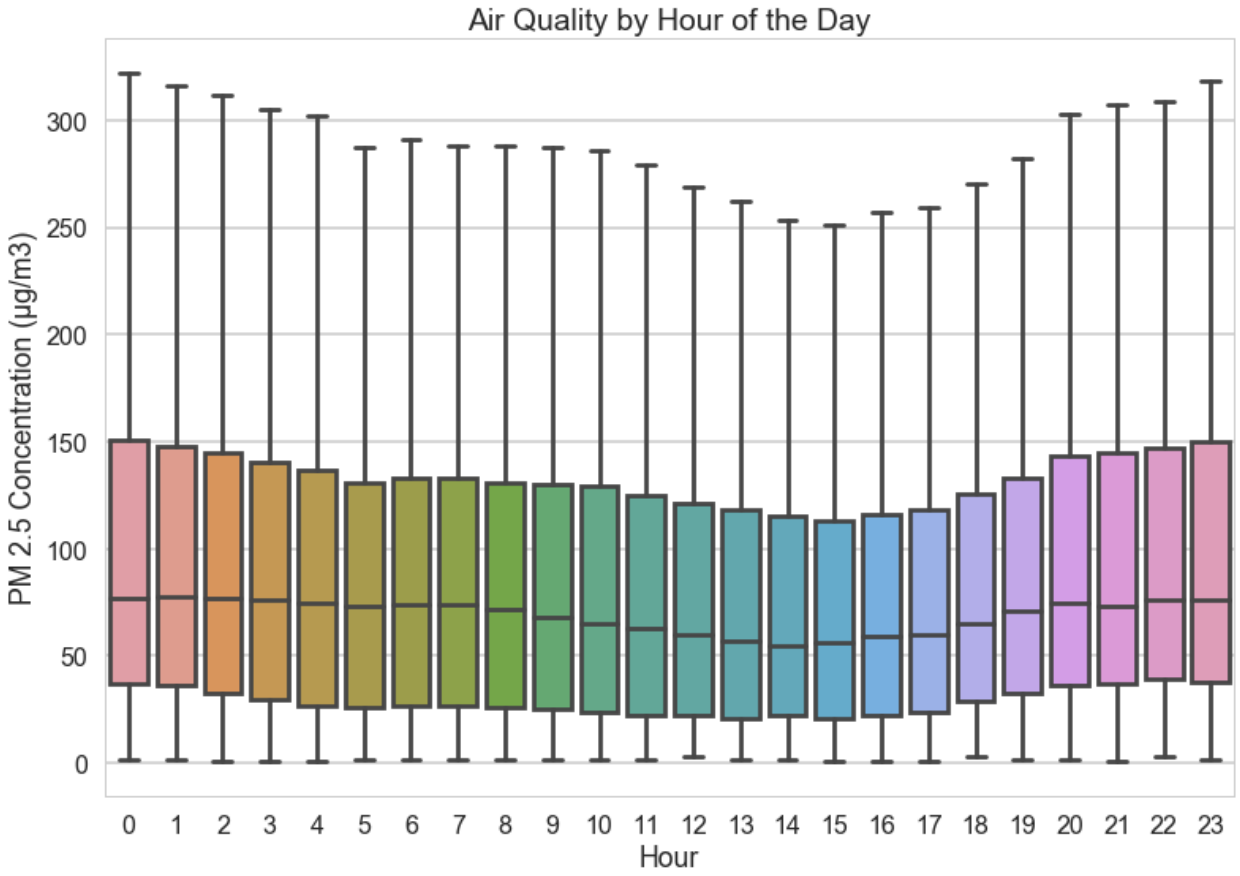
Air Quality Looks Seasonal… and Predictable?

The seasonality of air quality can be learned by LSTM models.

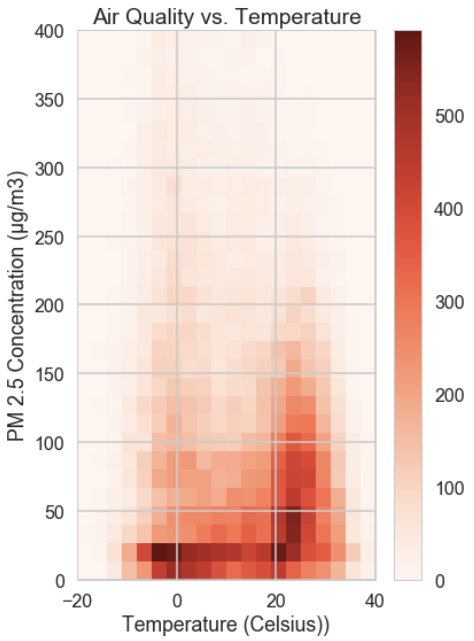


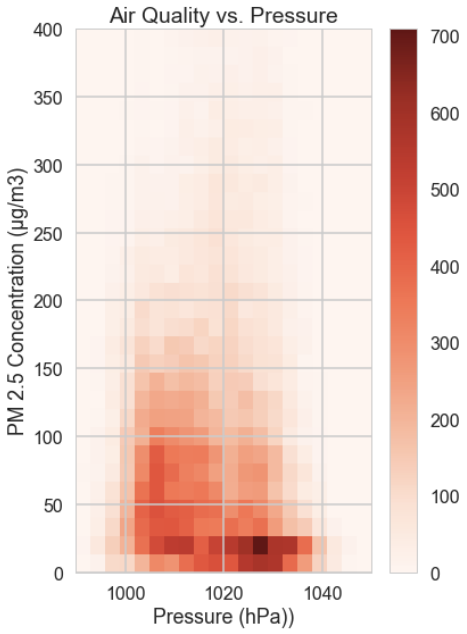
Air Quality by Month and Hour

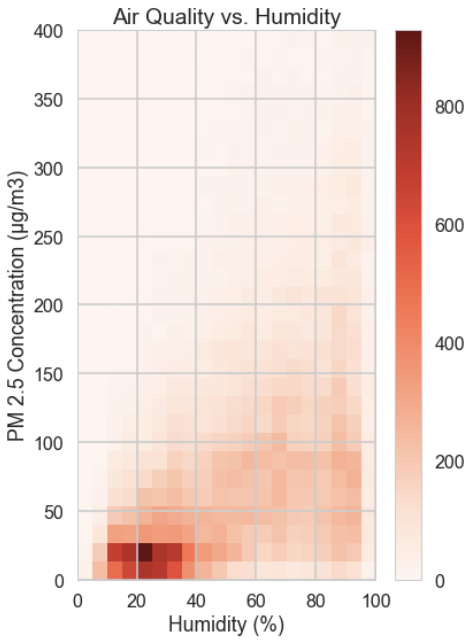


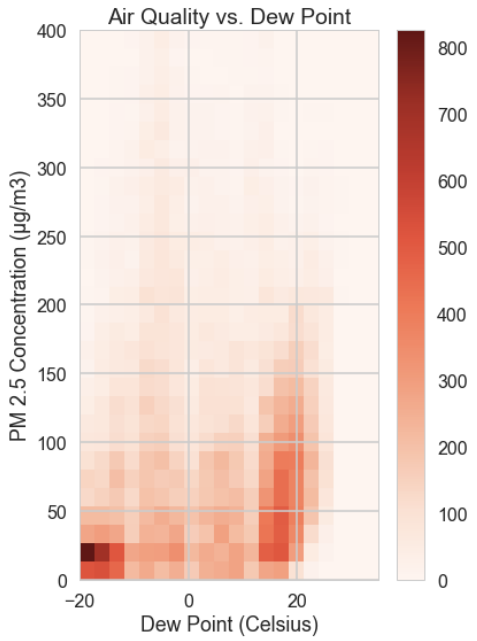


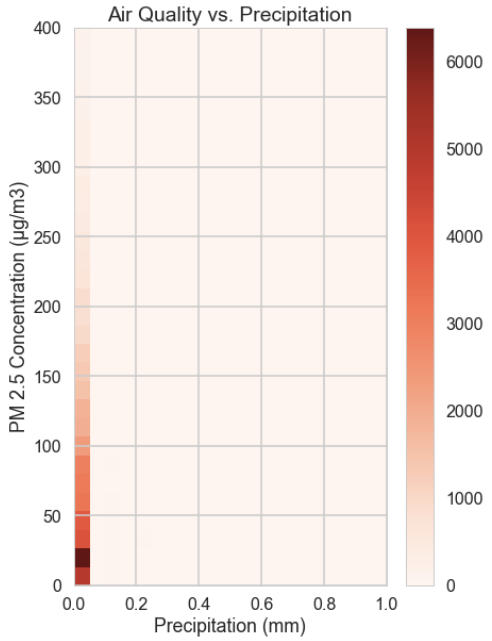
Air Quality Correlated with Weather



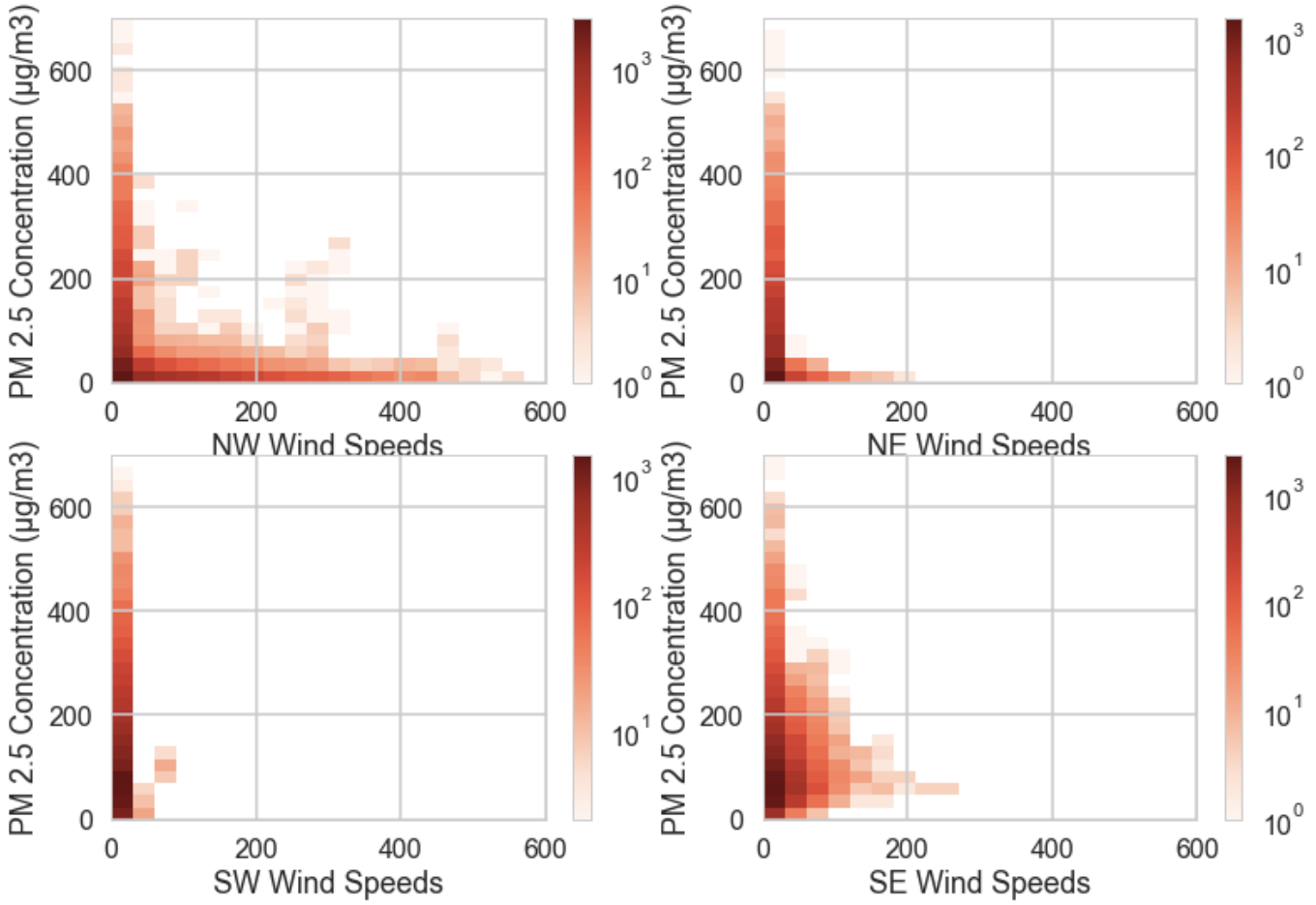








Wind Speed and Direction Have a Big Influence on Air Quality



Modeling

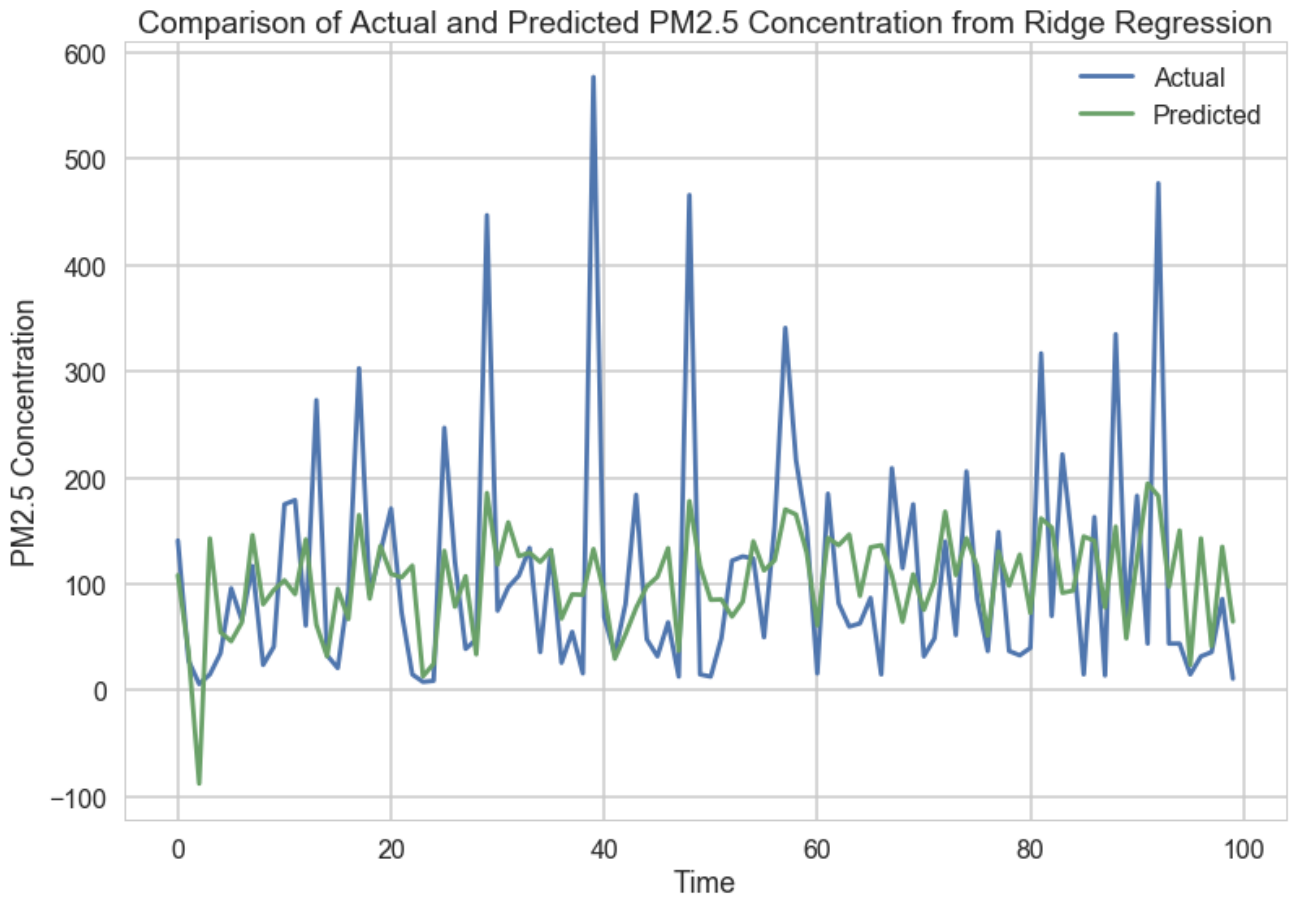
Linear Regression

We employ Ridge and Lasso modules from sci-kit learn.

Accounting for 9 features,

- R^2 = 0.242
- RMSE = 80.69 µg/m^3
- Humidity most strongly correlates with air quality

Predictions are bound in the range of 0 to 200 µg/m^3.



Machine Learning *via* Long Short-Term Memory (LSTM)

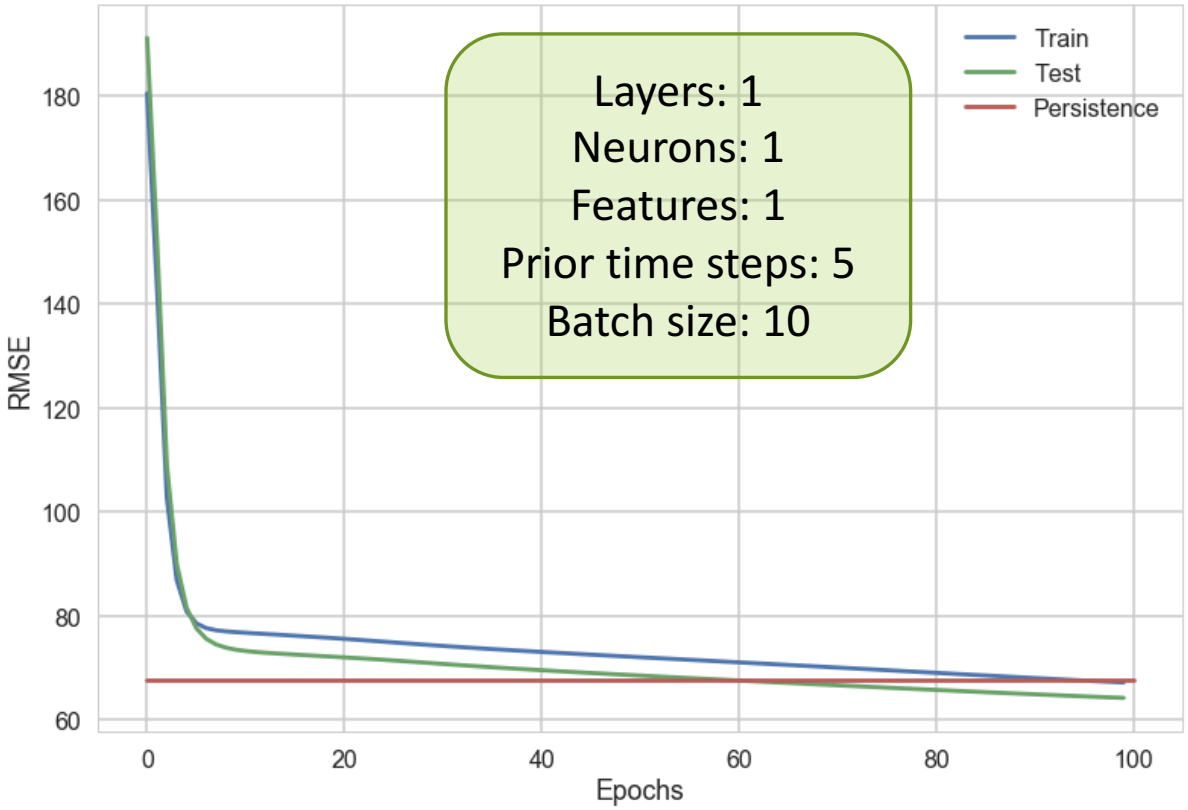
Recurrent neural networks (RNN) perform a sequence of the same operation wherein the output of one operation is retained as the input of the next operation.

- Thus, RNNs can retained memory while training on a set of data!

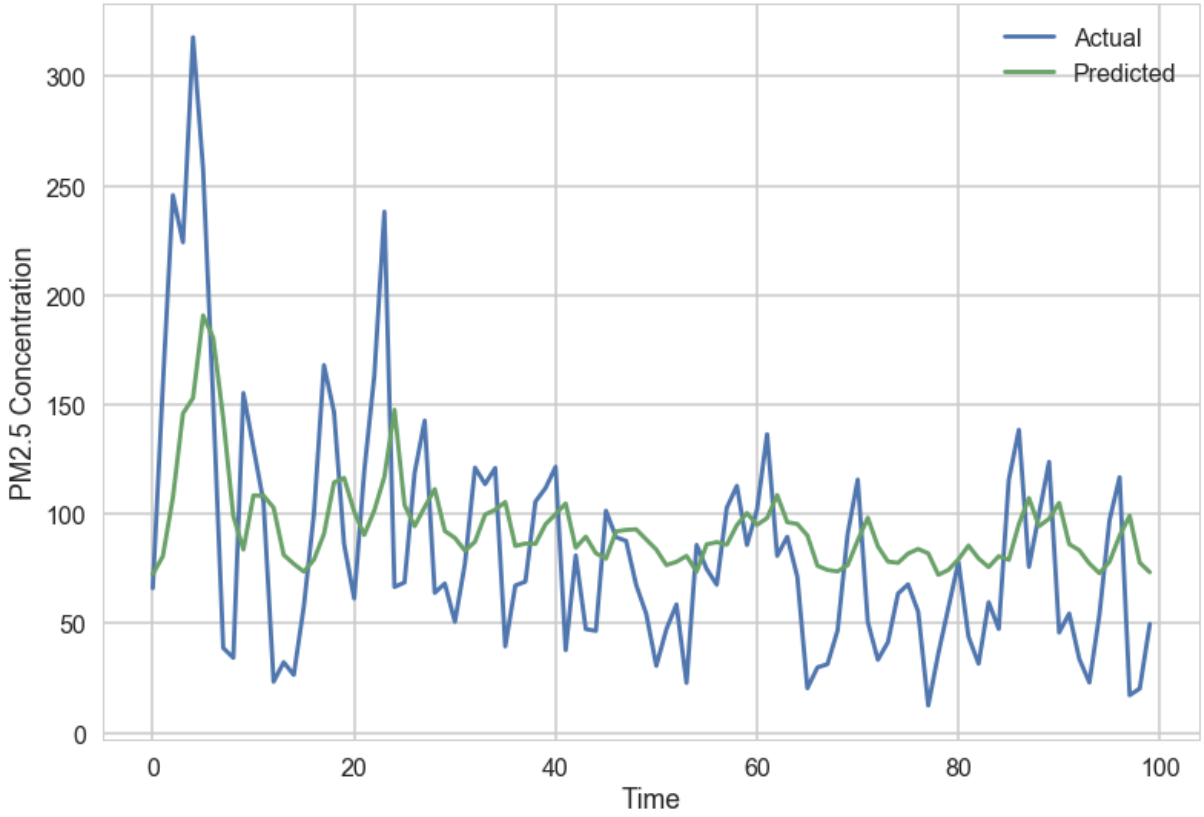
However, RNNs typically suffer from vanishing gradients, in which memory recorded several iterations prior is lost in favor of memory recorded recently.

Long short-term memory (LSTM) avoids this issue as it is capable of retaining long sequences of information!

A first attempt at LSTM…

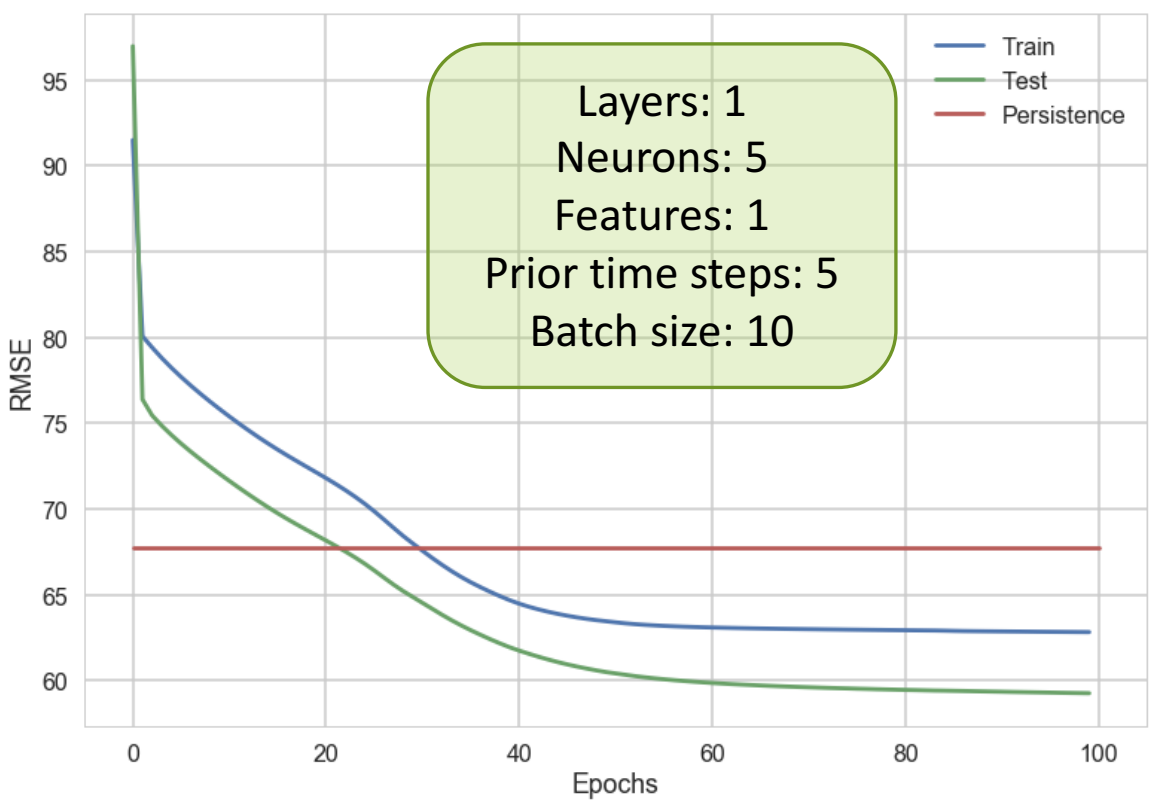


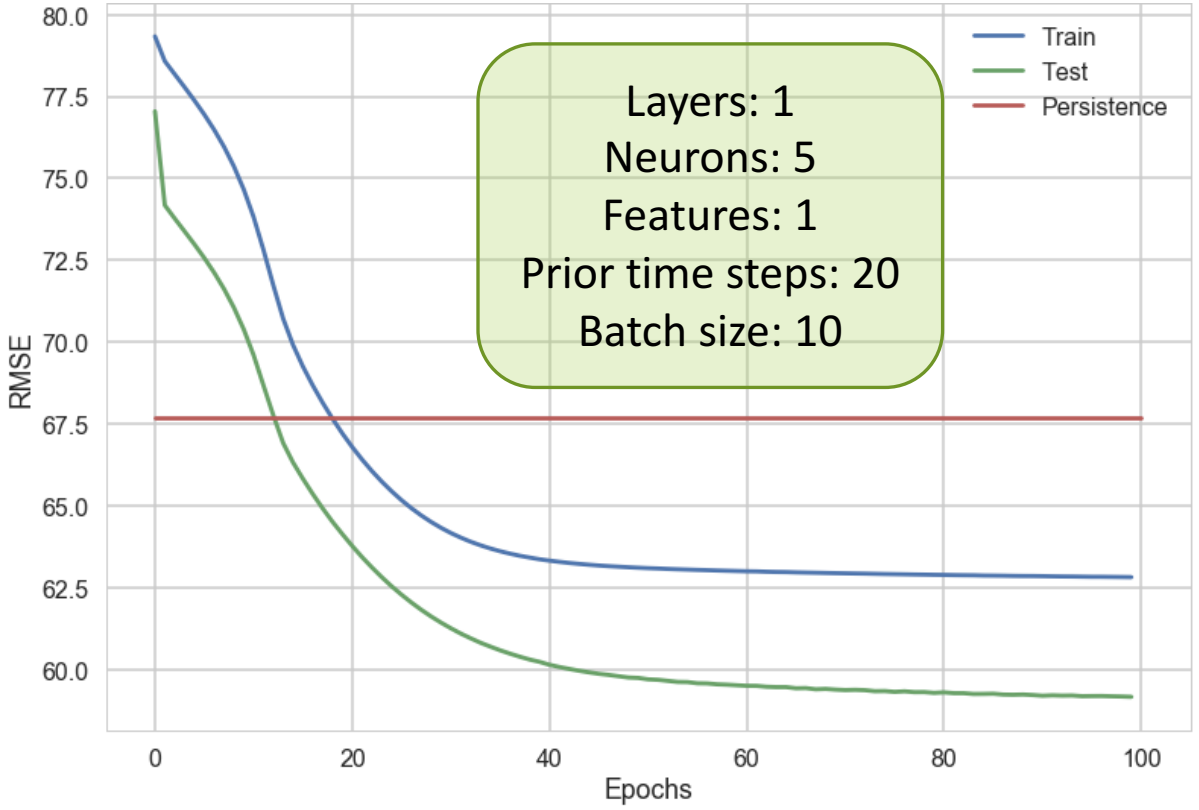
Layers: 1
Neurons: 1
Features: 1
Prior time steps: 5
Batch size: 10



More designing and training…

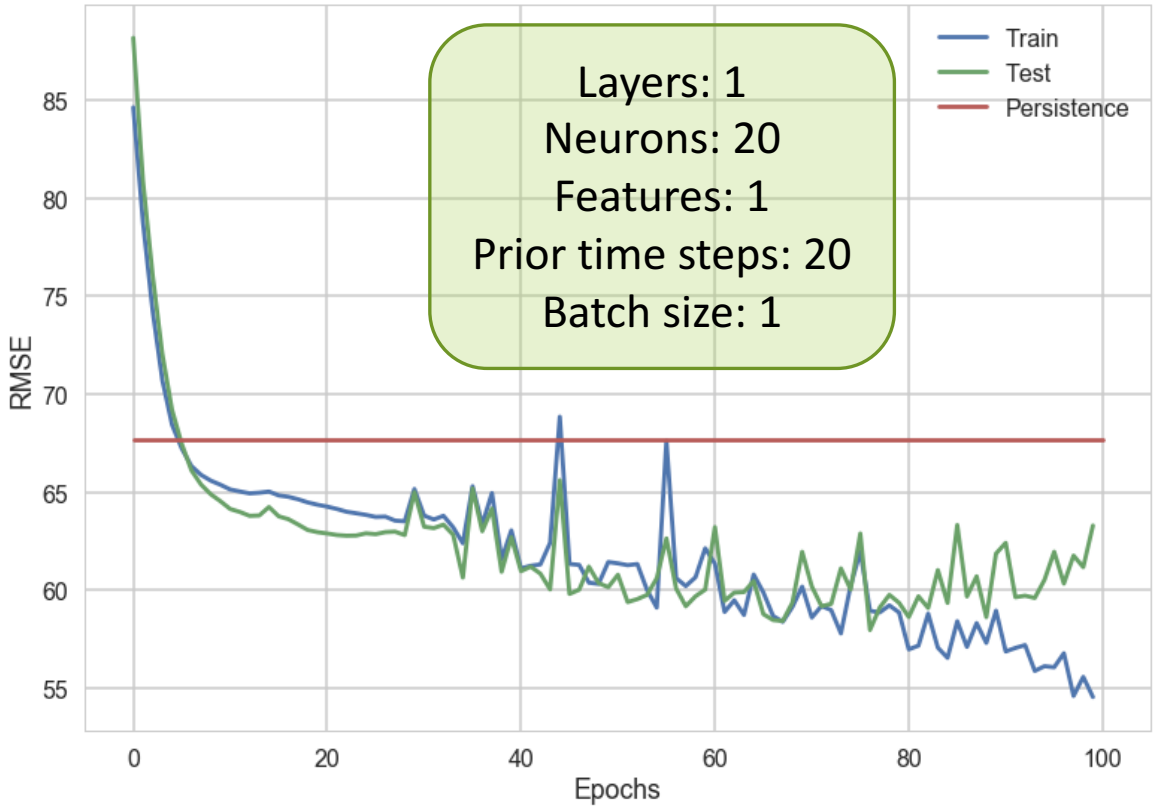
Varying prior time steps between 5 and 20.

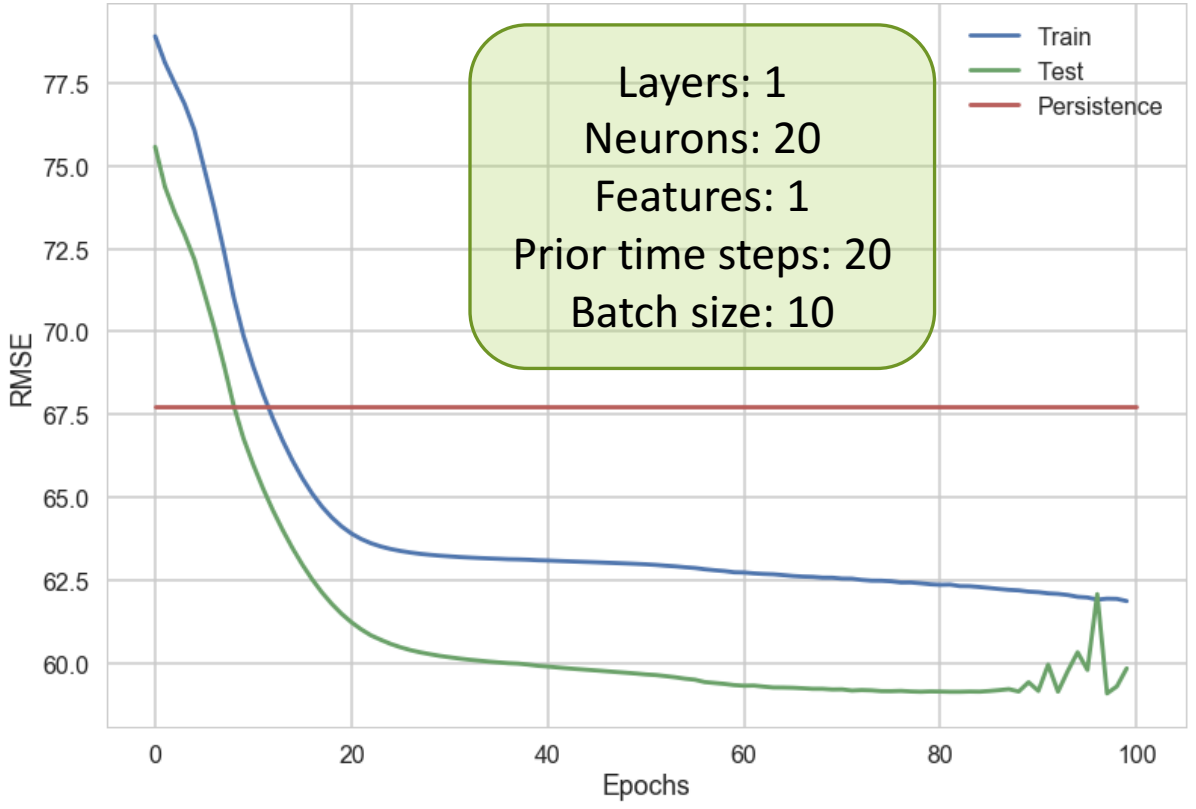




And more designing and training...

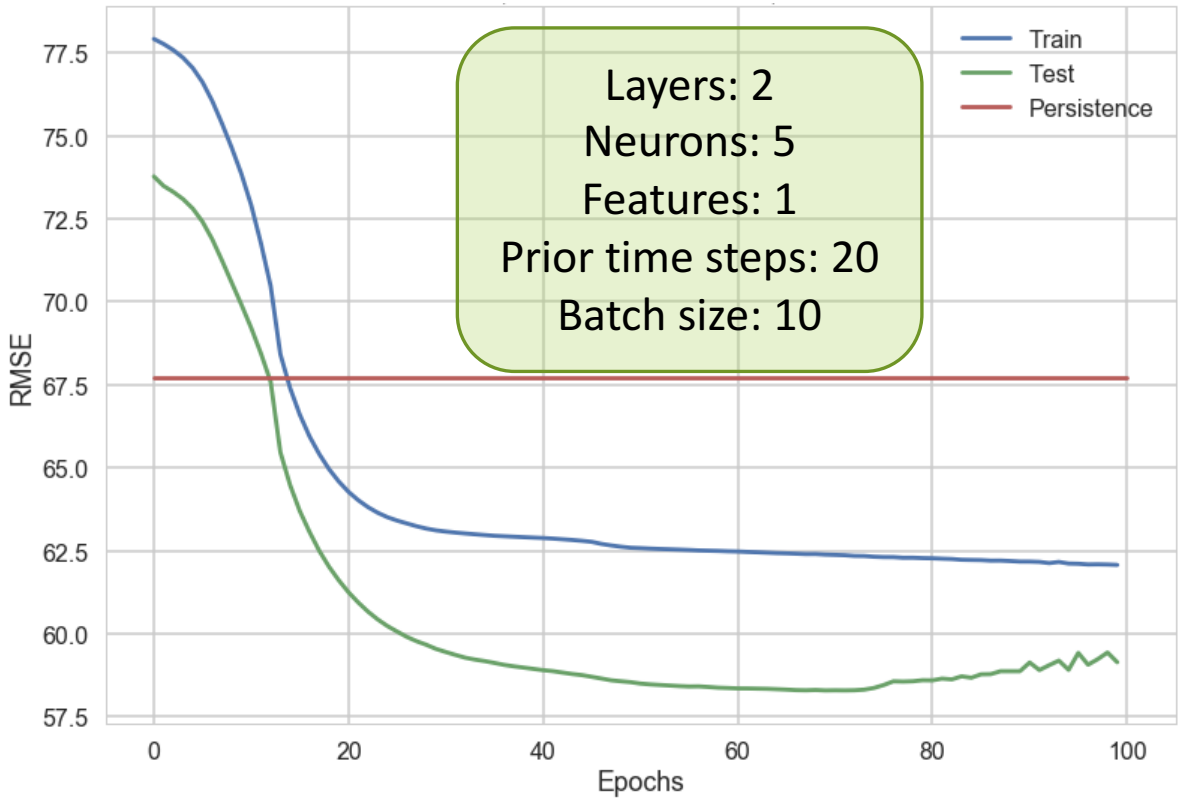
Varying batch size between 1 and 10.

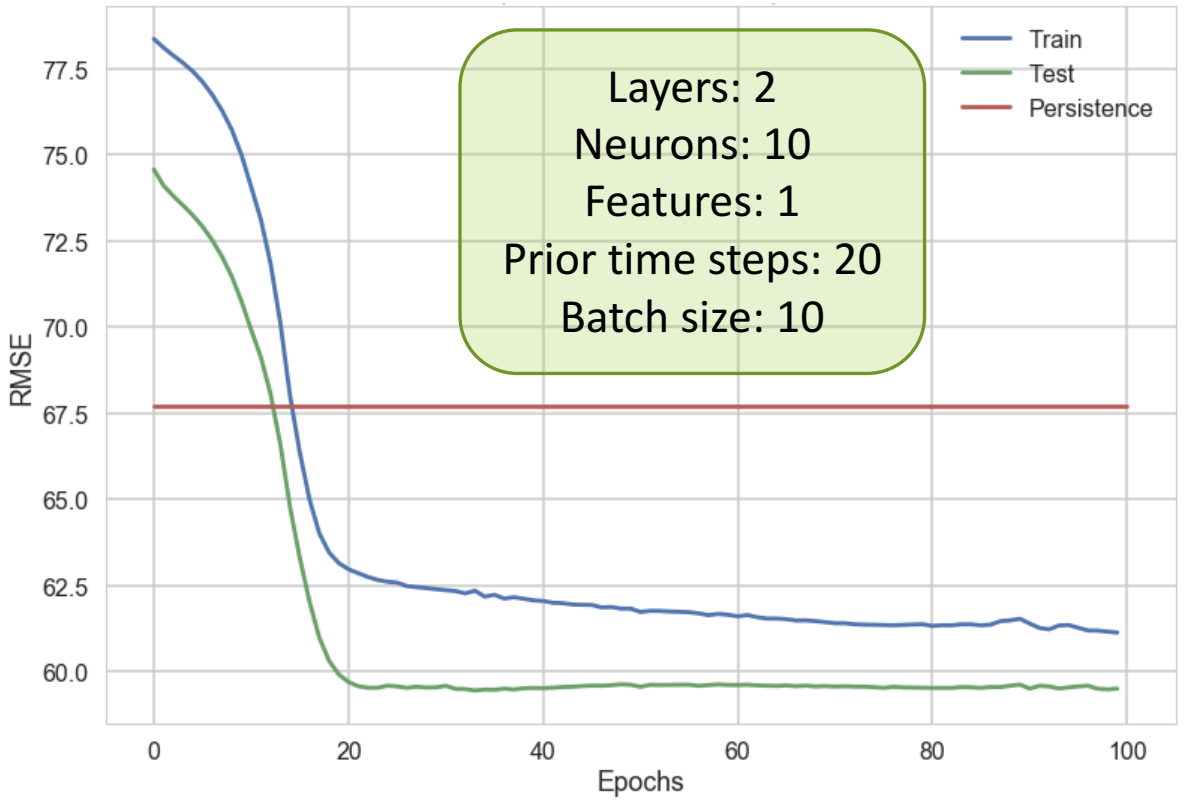




…and more designing and training…

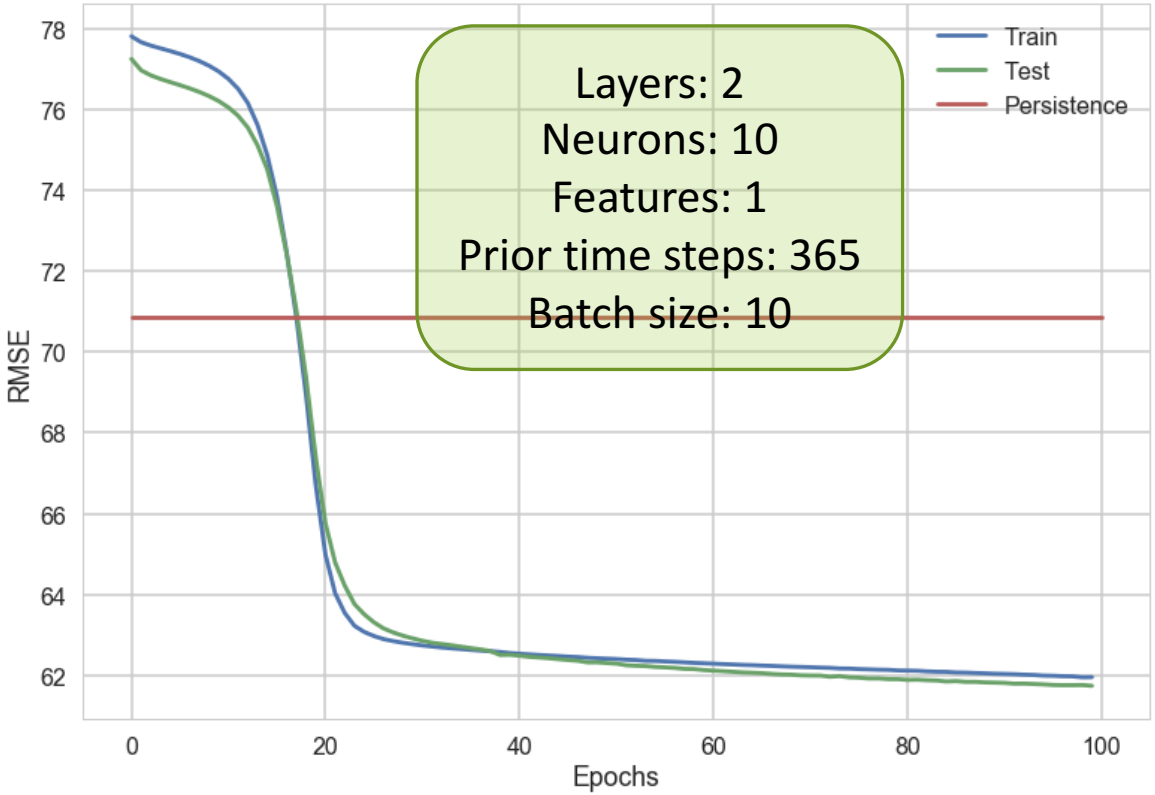
Varying neurons between 5 and 10.

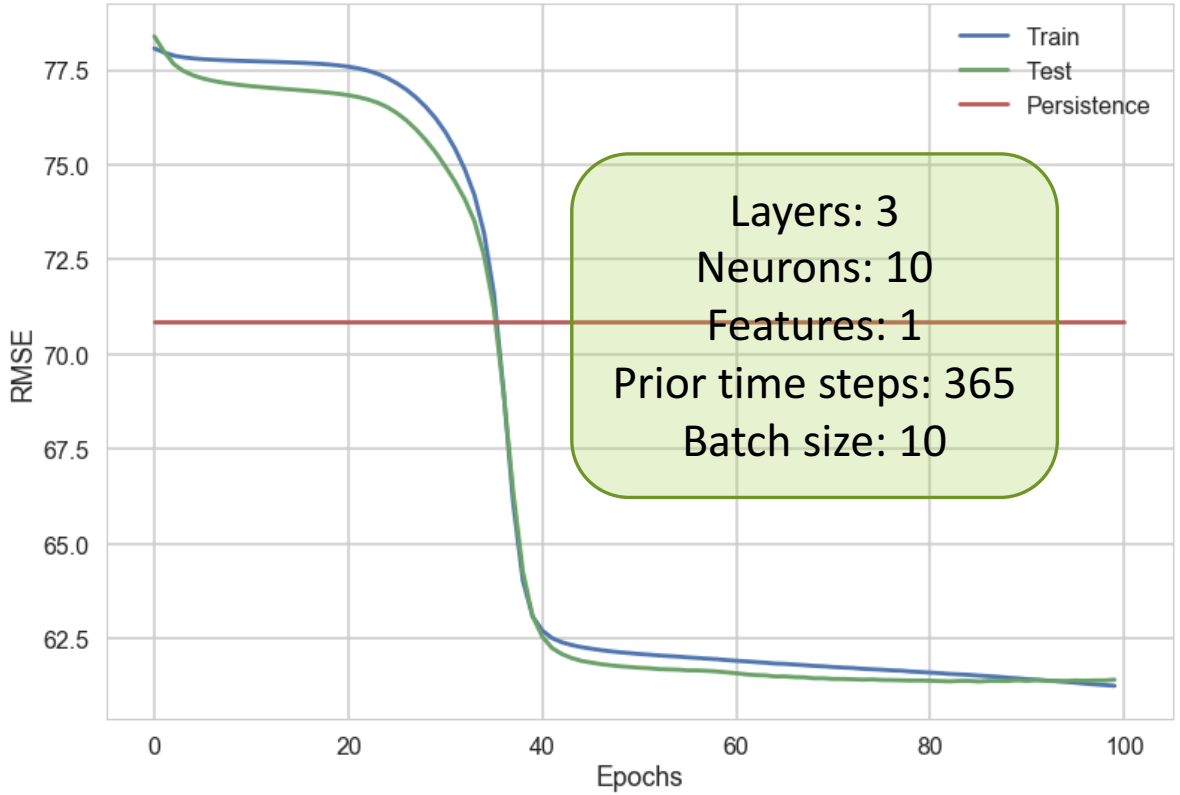




Until the fit is just right!

Varying LSTM layers between 2 and 3 with 365 prior time steps.





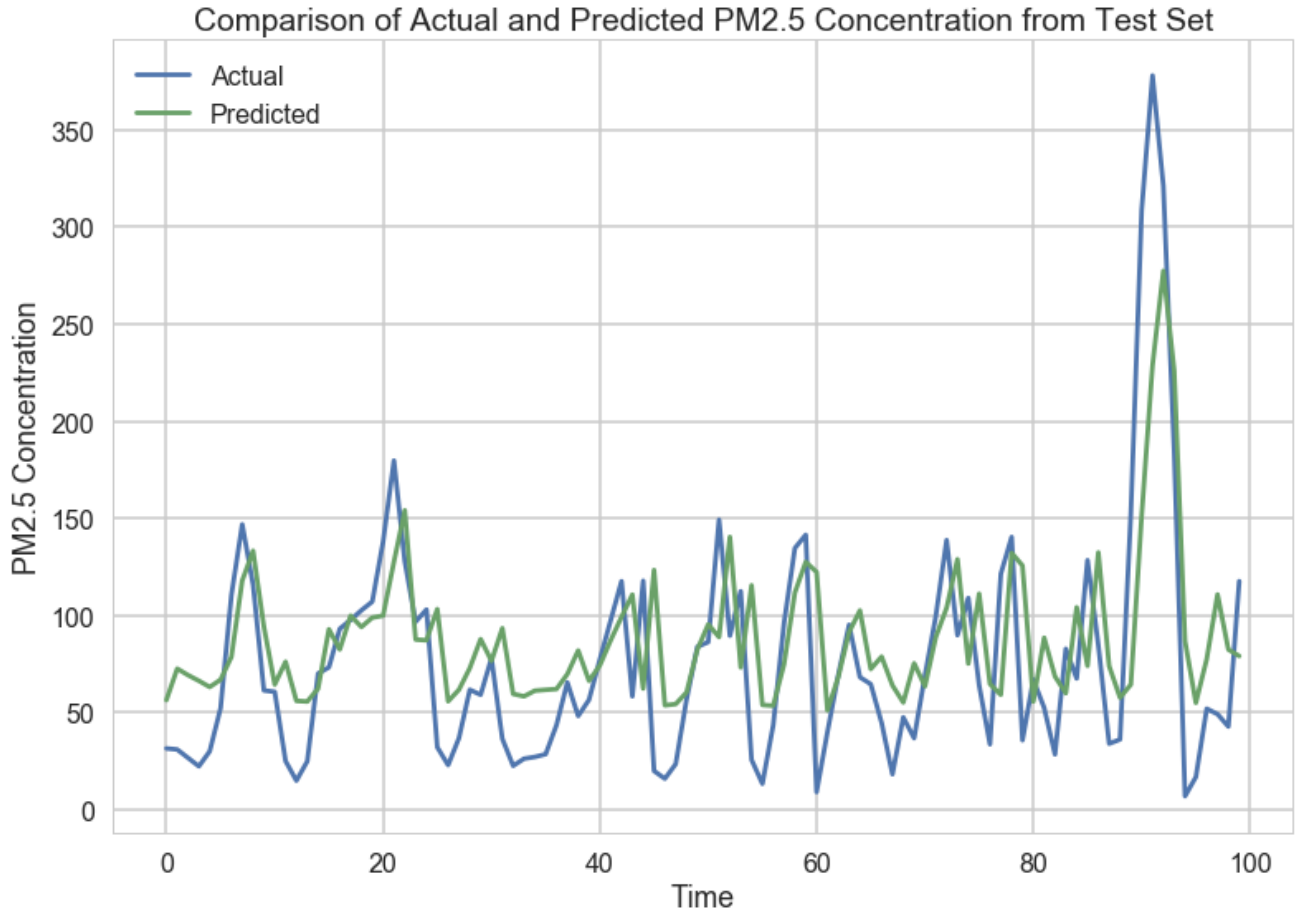
A summary of all LSTM models studied

Test #	LSTM layers	Neurons	Prior Time Steps	Batch size	Features	Profile of fit	Test RMSE
1	1	1	5	10	1	Underfitted	64.26
2	1	5	5	10	1	Underfitted	59.26
3	1	5	20	10	1	Underfitted	59.19
4	1	20	20	1	1	Overfitted	57.97
5	1	20	20	10	1	Underfitted	59.08
6	2	5	20	10	1	Underfitted	58.29
7	2	10	20	10	1	Underfitted	59.46
8	2	10	365	10	1	Good fit	61.74
9	3	10	365	10	1	Good fit	61.37
10	3	10	365	10	3	Overfitted	61.40

How well do the predictions track the actual data?

The model does not track well for PM2.5 concentration below 50 µg/m^3.

However, the model does try to track sharp peaks in the actual data.

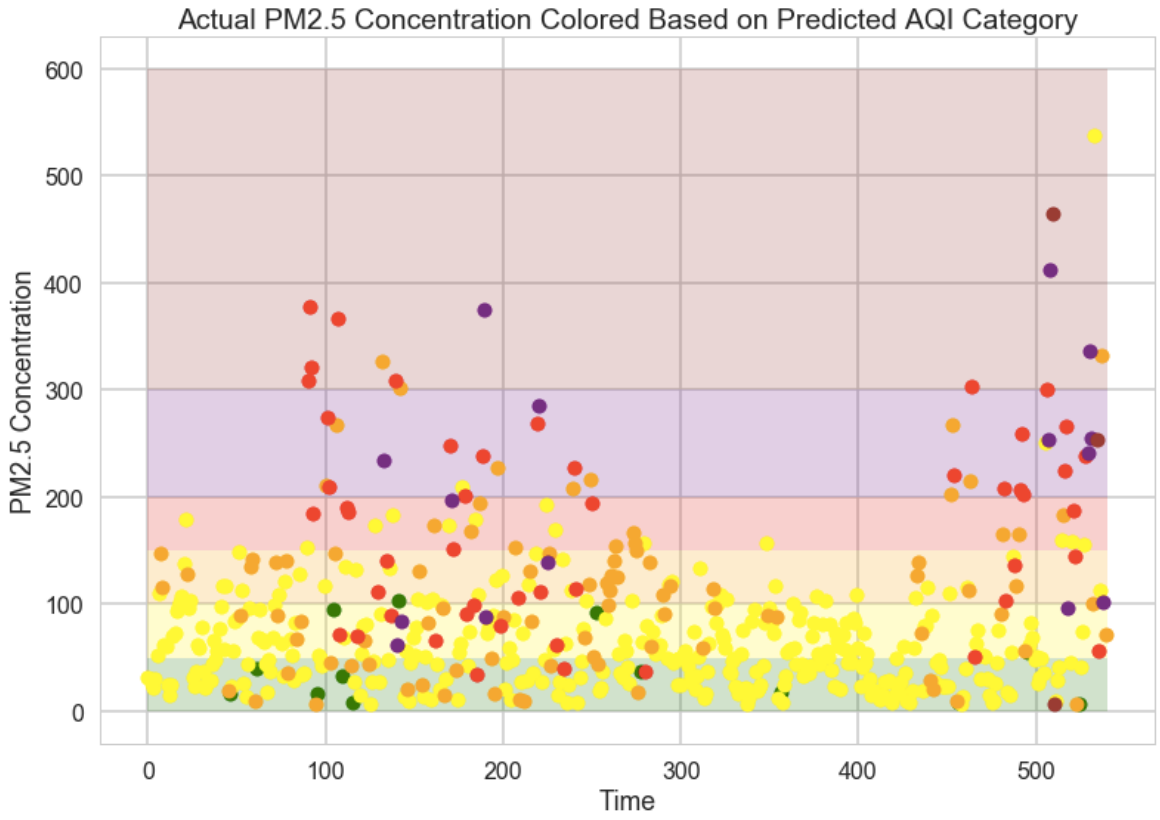


Two points of comparison

The World Air Quality Index Project accurately predicted the AQI classification of **22.9%** of data points for the two week span ending on January 12, 2018.

A team of Microsoft researchers generated a forecast with RMSE of **30 µg/m³** for predictions 1-6 hours into the future and RMSE of **64 µg/m³** for predictions 7-12 hours into the future.

The LSTM model accurately predicted the AQI classification **32.6%** of the time with RMSE **61.37 µg/m³**. Very respectable!



Conclusions

Linear regression yielded an R^2 score of 0.242 with RMSE of 80.69 µg/m^3.

LSTM models with 2 or 3 layers and 365 prior time steps provided the best fit to the data.

- RMSE for the 3 layer model is 61.37 µg/m^3.
- AQI classification was correctly predicted 32.6% of the time.

Recommendations:

- More data such as hourly carbon emissions data from local factories and residential areas.
- More LSTM layers, neurons, and prior time steps, which means more computing power to test the LSTM model within a reasonable time.

References

[1] Jia, H. et al. Peering into China's thick haze of air pollution. American Chemical Society, https://cen.acs.org/articles/95/i4/Peering-Chinas-thick-haze-air.html

[2] Stromberg, J. "What Does the Unbelievably Bad Air Quality in Beijing Do to the Human Body?" Smithsonian.com, www.smithsonianmag.com/science-nature/what-does-the-unbelievably-bad-air-quality-in-beijing-do-to-the-human-body-22655/

[3] Roberts, D. "Opinion: How the US Embassy Tweeted to Clear Beijing's Air." Wired, www.wired.com/2015/03/opinion-us-embassy-beijing-tweeted-clear-air/

[4] Tie, X. et al. Severe Pollution in China Amplified by Atmospheric Moisture. Scientific Reports 7 (2017).

[5] Graves, A. et al. A Novel Connectionist System for Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009).

[6] "Beijing Air Pollution: Real-time PM2.5 Air Quality Index (AQI)." World Air Quality Index, www.aqicn.org/

[7] "Urban Air." Microsoft Corporation, www.microsoft.com/en-us/research/project/urban-air/