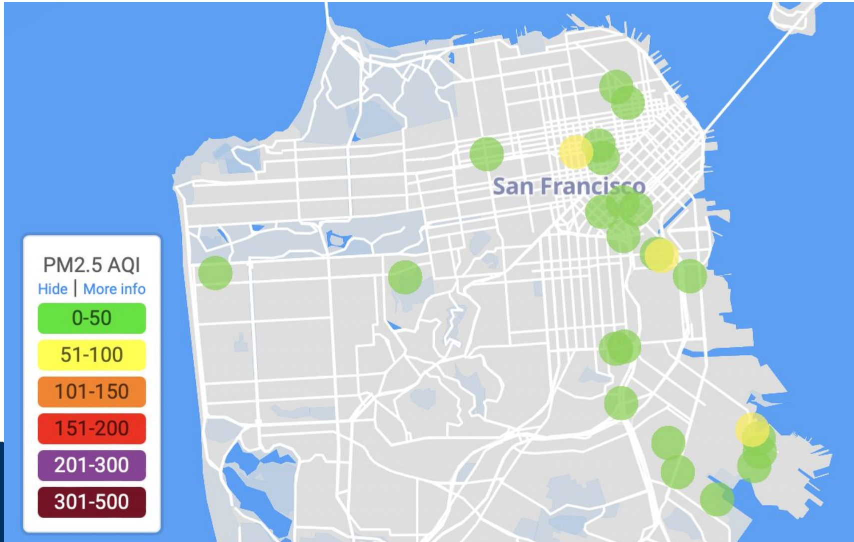
Statistical Linear Modeling for Hourly PM2.5 Multi-Step Time Series Forecasting in Eastern San Francisco

Eleanor Kim



Berkeley
UNIVERSITY OF CALIFORNIA

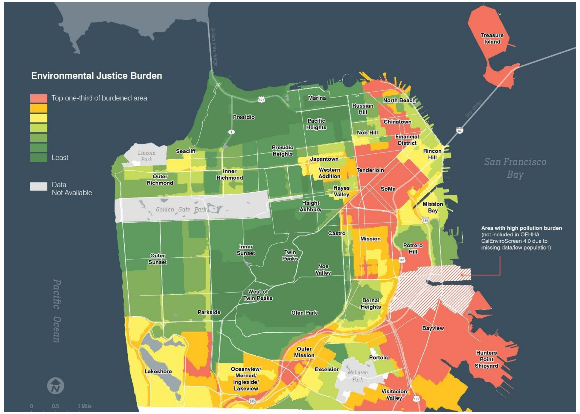
Background

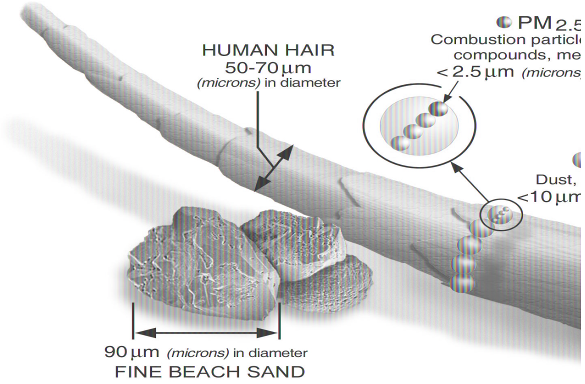
brightline DEFENSE

clarity

CALIFORNIA AIR RESOURCES BOARD

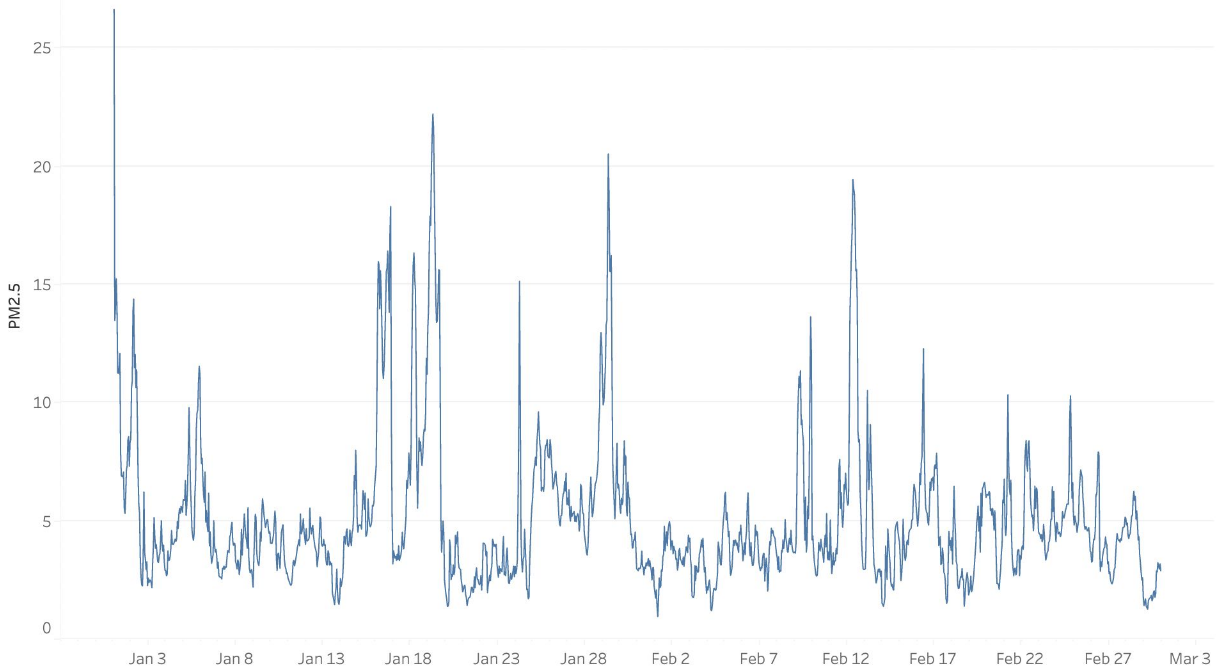
BAY AREA AIR QUALITY MANAGEMENT DISTRICT





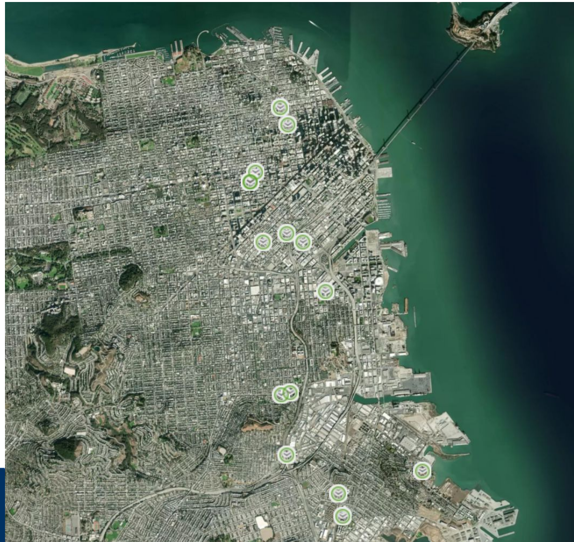
Berkeley
UNIVERSITY OF CALIFORNIA

Monitoring PM2.5 in San Francisco



- August 2020 – February 2024
- 15 sensors
- 5 neighborhoods in SF
- Hourly average PM2.5

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	3.96	6.27	8.44	10.02	241.42



Berkeley
UNIVERSITY OF CALIFORNIA

Guiding Questions

- **Can we produce a predictive model for PM2.5 with high accuracy?**
- How can we optimize the model's accuracy with respect to the number of lag steps included in the model?
- What are the individual contributions of factors like humidity, temperature, seasonality, neighborhood?
- How do anomalously high PM2.5 values affect the model?

Berkeley
UNIVERSITY OF CALIFORNIA

About the Data

```
'data.frame':	546743 obs. of  10 variables:
 $ ID                            : chr  "AYHT7CF7" "ATWJ3V74" "ALZ4PSJB" "A2BPKVSX" ...
 $ Datetime                      : POSIXct, format: "2024-02-15 23:00:00" "2024-02-15 23:00:00" "2024-02-15 23:00:00"
 $ PM2.5_Hour_MassConc_Calibrated: num  7.39 5.86 5.93 9.42 5.86 4.04 4.03 3.61 6.35 6.32 ...
 $ Temperature                   : num  12.1 13.1 11.7 11.1 11.6 ...
 $ Humidity                      : num  83.4 81.2 86.2 89 85.7 ...
 $ Latitude                      : num  37.8 37.8 37.8 37.7 37.7 ...
 $ Longitude                     : num  -122 -122 -122 -122 -122 ...
 $ Device_Name                   : chr  "Fitness SF" "Howard & 9th" "BAAQMD Co-Location" "The Box Shop" ...
 $ Neighborhood                  : chr  "SoMa" "SoMa" "Potrero Hill" "BVHP" ...
 $ Season                        : chr  "Winter" "Winter" "Winter" "Winter" ...
```

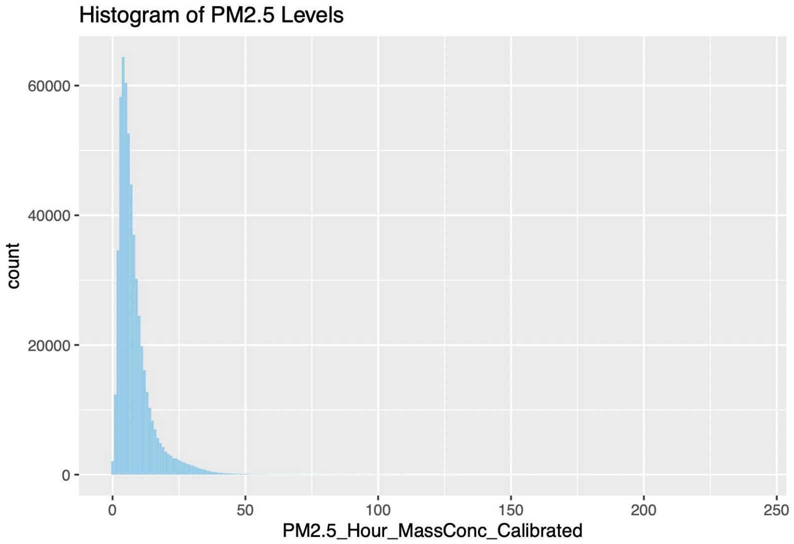
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    3.96    6.27    8.44   10.02  241.42 
```

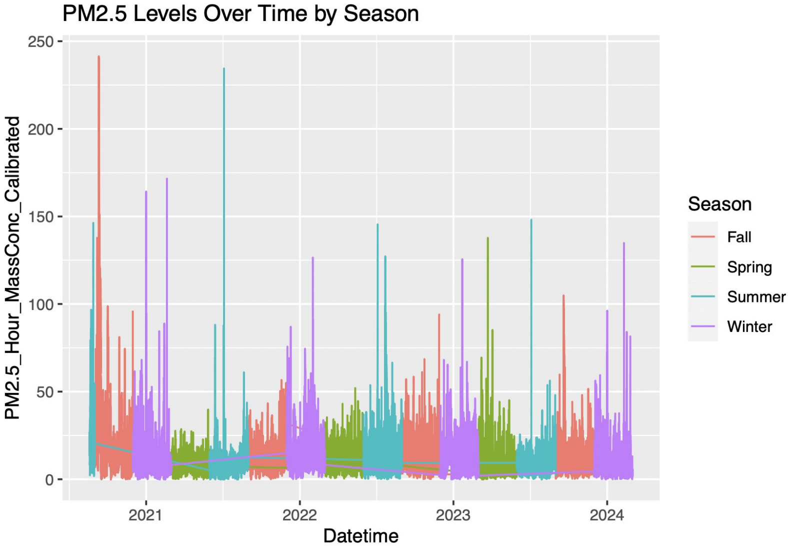
Berkeley
UNIVERSITY OF CALIFORNIA

Methodology

- Partial regression & Frisch–Waugh–Lovell Theorem
- Model selection criteria (Mean Absolute Percentage Error, Adjusted R^2, Akaike Information Criterion)
- Ordinary Least Squares
- Weighted Least Squares
- Ridge (Weighted)
- Lasso (Weighted)
- Leave-one-out Formula
- Outlier Detection (Cook's Distance & Leave-one-out)

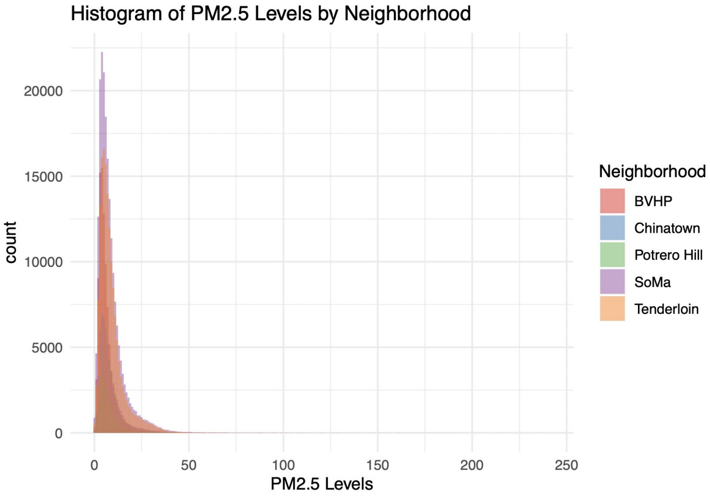
Exploratory Data Analysis

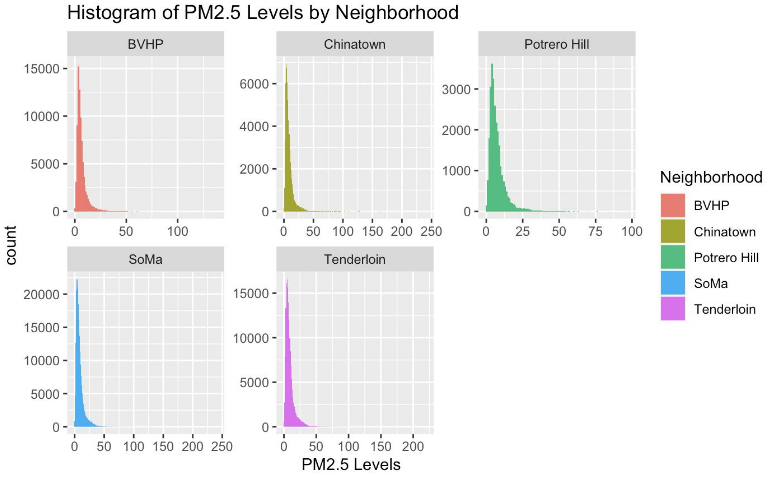




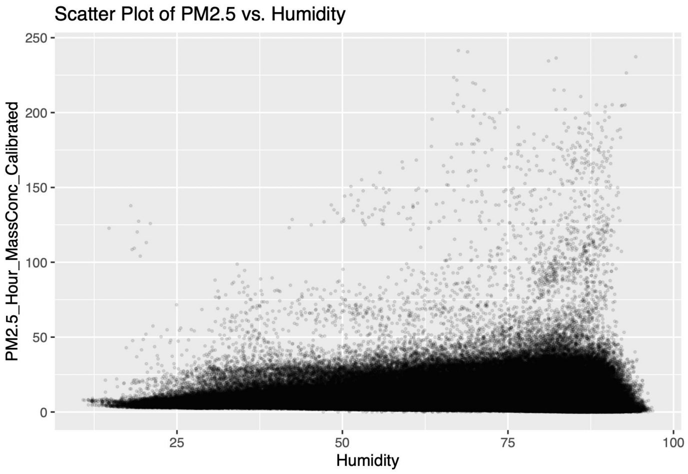
Berkeley
UNIVERSITY OF CALIFORNIA

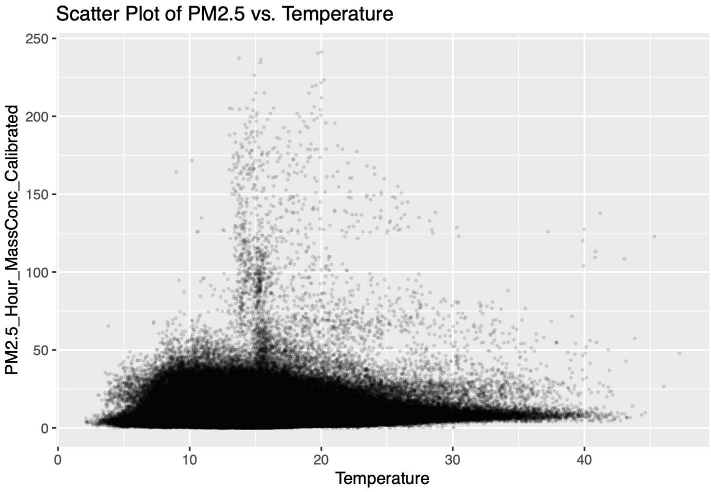
Exploratory Data Analysis





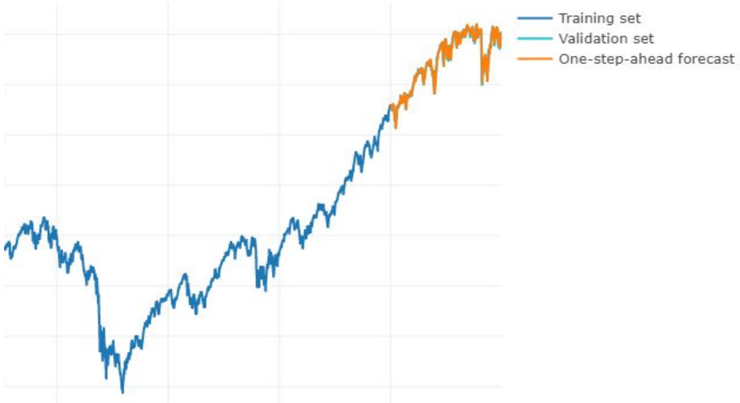
Exploratory Data Analysis





Objective

- Minimize MAPE
- Maximize Adjusted R^2
- Minimize AIC



$$\begin{aligned} PM2.5_{t,i} = \quad & \beta_0 + \beta_1\, PM2.5_{t-1,i} + \beta_2\, PM2.5_{t-2,i} + \ldots + \beta_l\, PM2.5_{t-l,i} + \\ & Weather_{t,i}\, \boldsymbol{\beta}_{l+1} + Season_{t,i}\boldsymbol{\beta}_{l+2} - Neighborhood_{t,i}\, \boldsymbol{\beta}_{l+3} + \epsilon_{t,i} \end{aligned}$$

Feature Engineering

- How can we optimize the model’s accuracy with respect to the number of lag steps in the model?
- What are the individual contributions of Humidity, Temperature, Seasonality, and Neighborhood?
    - Should we include dummies for Season & Neighborhood?
    - Should we include Humidity and/or Temperature in our model?

Berkeley
UNIVERSITY OF CALIFORNIA

Iterative approach to determine the optimal number of lag steps for time series forecasting.

What's the optimal ℓ?

$$PM2.5_{t,i} = \beta_0 + \boxed{\beta_1\, PM2.5_{t-1,i} + \beta_2\, PM2.5_{t-2,i} + \ldots + \beta_l\, PM2.5_{t-l,i}} + Weather_{t,i}\, \boldsymbol{\beta}_{l+1} + Season_{t,i} \boldsymbol{\beta}_{l+2} - Neighborhood_{t,i}\, \boldsymbol{\beta}_{l+3} + \epsilon_{t,i}$$

- Loop through different lag steps to train and evaluate models.
- Evaluation metrics: Mean Absolute Percentage Error (MAPE) and Adjusted R-squared
- Normalize performance metrics for comparison
- Calculate weighted average to identify the lag step with the highest performance
- Confirm optimal lag step frequency across multiple random seeds

$\longrightarrow \ell = 6$

Berkeley
UNIVERSITY OF CALIFORNIA

Comparison of two methods to understand **Neighborhood** contributions to the model.

$$
\begin{aligned}
PM2.5_{t,i} = \quad & \beta_0 + \beta_1\, PM2.5_{t-1,i} + \beta_2\, PM2.5_{t-2,i} + \ldots + \beta_l\, PM2.5_{t-l,i} + \\
& Weather_{t,i}\, \boldsymbol{\beta}_{l+1} + Season_{t,i} \boldsymbol{\beta}_{l+2} - \boxed{Neighborhood_{t,i}\, \boldsymbol{\beta}_{l+3}} + \epsilon_{t,i}
\end{aligned}
$$

- Full Regression Model (OLS): Includes neighborhood variable along with other predictors.
  - Coefficients reflect associations with outcome while accounting for other predictors.
  - Evaluation metrics include MAPE, Adjusted R-squared, and AIC.
- Partial Regression Analysis (FWL): Uses Frisch-Waugh-Lovell Theorem to isolate neighborhood effects.
  - Separate regression models for each neighborhood dummy variable after obtaining residuals from full model.
  - Coefficients represent unique contributions of neighborhoods while controlling for other predictors.

Berkeley
UNIVERSITY OF CALIFORNIA

Comparison of two methods to understand **Seasonality** contributions to the model.

$$PM2.5_{t,i} = \beta_0 + \beta_1 \, PM2.5_{t-1,i} + \beta_2 \, PM2.5_{t-2,i} + \ldots + \beta_l \, PM2.5_{t-l,i} + Weather_{t,i} \, \boldsymbol{\beta}_{l+1} + \boxed{Season_{t,i} \boldsymbol{\beta}_{l+2}} - Neighborhood_{t,i} \, \boldsymbol{\beta}_{l+3} + \epsilon_{t,i}$$

- Full Regression Model (OLS): Includes neighborhood variable along with other predictors.
  - Season dummy coefficients are statistically significant indicating that PM2.5 predictions have varied effects by season
- Partial Regression Analysis (FWL): Uses Frisch-Waugh-Lovell Theorem to isolate neighborhood effects.
  - The variation in PM2.5 levels that is not explained by these predictors is not significantly associated with the different seasons

Berkeley
UNIVERSITY OF CALIFORNIA

Should we include Humidity and/or Temperature in our model?

Model_Includes	MAPE	Adj_R2	AIC
Neither	16.25	0.85774	392308
Humidity	16.43	0.85783	392253
Temperature	16.36	0.85781	392271
Humidity and Temperature	16.44	0.85785	392246

For every 1 %-point increase in humidity, the predicted PM2.5 value increases by 0.003 ug/m^3,

For every 1 C° increase in temperature, the predicted PM2.5 value decreases by 0.006 ug/m^3,

(holding all else constant)

Model Selection

Can we produce a predictive model for PM2.5 with high accuracy?

Potential Models

- Ordinary Least Squares
- Weighted Least Squares
- Ridge Regression
  Lasso Regression
- Weighted Ridge
- Weighted Lasso

Selection Criteria

- Minimize MAPE
- Maximize Adjusted R^2
- Minimize AIC

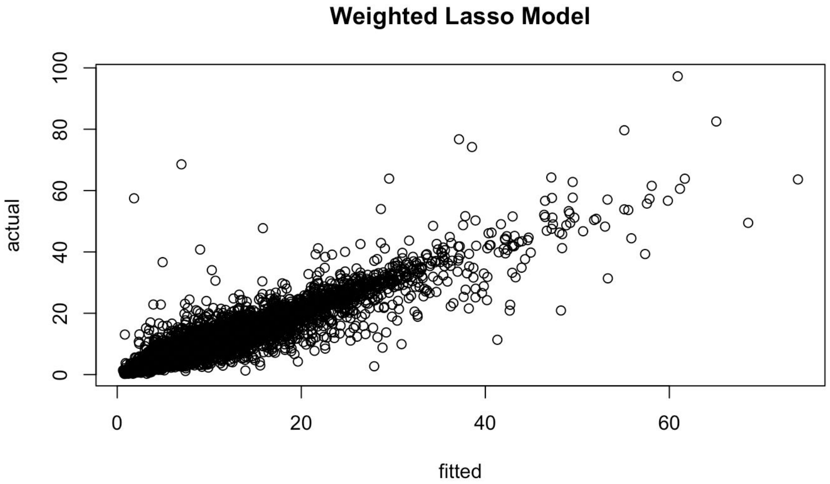
Check Conditions

- Linearity
- Homoskedasticity
- Independence of Errors
- Normality of Residuals
- Variance Inflation Factors
- Multicollinearity
- Outliers
- Model Fit

$$
\begin{aligned}
PM2.5_{t,i} = {} & \beta_0 + \beta_1\, PM2.5_{t-1,i} + \beta_2\, PM2.5_{t-2,i} + \beta_3\, PM2.5_{t-3,i} + \beta_4\, PM2.5_{t-4,i} \\
& + \beta_5\, PM2.5_{t-5,i} + \beta_6\, PM2.5_{t-6,i} + \beta_7\, Humidity_{t,i} + \beta_8\, Temperature_{t,i} \\
& + \beta_9\, Spring_{t,i} + \beta_{10}\, Summer_{t,i} + \beta_{11}\, Winter_{t,i} + \beta_{12}\, Chinatown_{t,i} \\
& + \beta_{13}\, Potrero\ Hill_{t,i} + \beta_{14}\, SoMa_{t,i} + \beta_{15}\, Tenderloin_{t,i} + \epsilon_{t,i}
\end{aligned}
$$

Model	MAPE	Adj_R_2	AIC
OLS	16.37	0.8580	31145
WLS	16.37	1.0000	31145
Ridge	16.54	0.8769	31350
Lasso	16.39	0.8779	31164
Weighted Ridge	16.48	0.8774	31249
Weighted Lasso	16.35	0.8779	31172

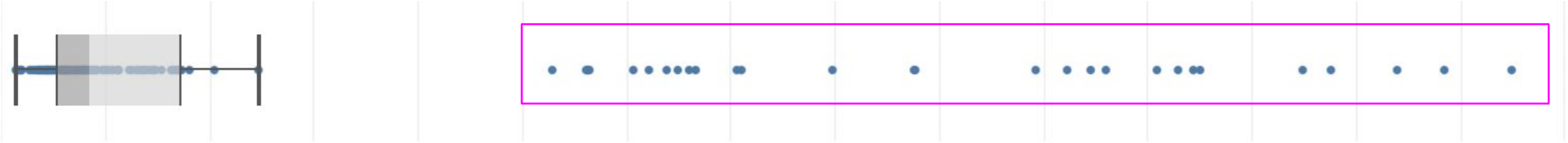
$$
\begin{aligned}
PM2.5_{t,i} = &\beta_0 + \beta_1\, PM2.5_{t-1,i} + \beta_2\, PM2.5_{t-2,i} + \beta_3\, PM2.5_{t-3,i} + \beta_4\, PM2.5_{t-4,i} \\
&+ \beta_5\, PM2.5_{t-5,i} + \beta_6\, PM2.5_{t-6,i} + \beta_7\, Humidity_{t,i} + \beta_8\, Temperature_{t,i} \\
&+ \beta_9\, Spring_{t,i} + \beta_{10}\, Summer_{t,i} + \beta_{11}\, Winter_{t,i} + \beta_{12}\, Chinatown_{t,i} \\
&+ \beta_{13}\, Potrero\ Hill_{t,i} + \beta_{14}\, SoMa_{t,i} + \beta_{15}\, Tenderloin_{t,i} + \epsilon_{t,i}
\end{aligned}
$$



(Intercept)	0.2191
PM2.5_Hour_MassConc_Calibrated_lag_1	0.8085
PM2.5_Hour_MassConc_Calibrated_lag_2	0.0654
PM2.5_Hour_MassConc_Calibrated_lag_3	0.0190
PM2.5_Hour_MassConc_Calibrated_lag_4	0.0344
PM2.5_Hour_MassConc_Calibrated_lag_5	-0.0004
PM2.5_Hour_MassConc_Calibrated_lag_6	0.0179
Humidity	0.0022
Temperature	-0.0091
factor(Season)Fall	0.2101
factor(Season)Summer	0.1060
factor(Season)Winter	0.1167
factor(Neighborhood)BVHP	0.0857
factor(Neighborhood)Potrero Hill	0.0055
factor(Neighborhood)SoMa	-0.0110
factor(Neighborhood)Tenderloin	0.0272

Sensitivity Analysis

How do anomalously high PM2.5 values affect the model?



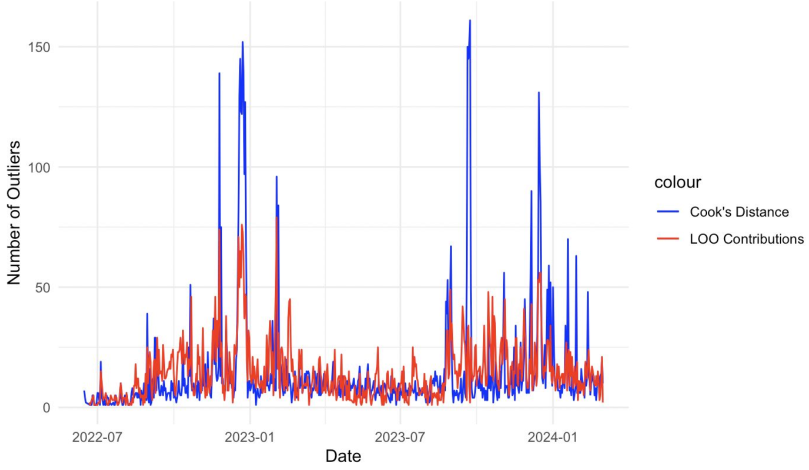
- Outlier detection
    - Cook's Distance

      $$\text{cook}_i = \text{standr}_i^2 \times \frac{h_{ii}}{p(1-h_{ii})}$$
    - Leave-one-out formula

      $$\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + \gamma_{(n+1)}\hat{\varepsilon}_{[n+1]}$$
- Test which set of outliers to leave out to will produce better performing OLS model

Outlier Results

Comparison of Top 10% of Outliers



Comparison of Performance Criteria

Model	MAPE	Adj_R2	AIC
OLS	16.371	0.8580	31145
WLS	16.370	1.0000	31145
Ridge	16.540	0.8769	31350
Lasso	16.392	0.8779	31164
Weighted Ridge	16.480	0.8774	31249
Weighted Lasso	16.346	0.8779	31172
OLS w/o Cook's Outliers	14.745	0.7796	11247
OLS w/o LOO Outliers	14.883	0.9074	12280

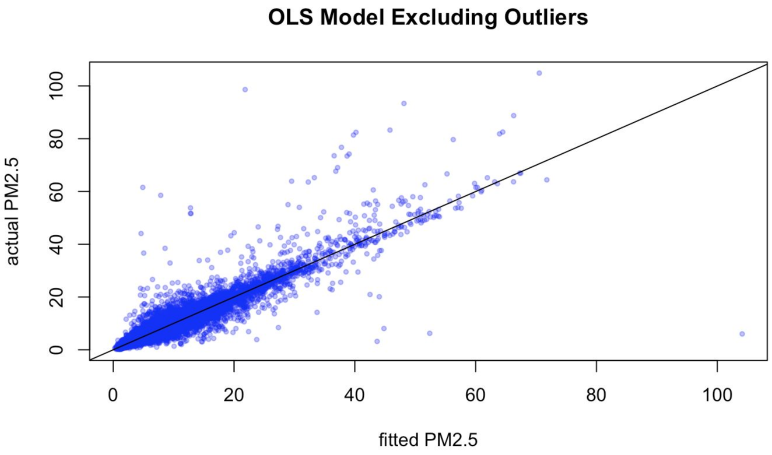
Berkeley
UNIVERSITY OF CALIFORNIA

“Best” Predictive Model

$$
\begin{aligned}
PM2.5_{t,i} = &-2.43 + 0.75\,PM2.5_{t-1,i} + 0.15\,PM2.5_{t-2,i} + 0.026\,PM2.5_{t-3,i} + 0.04\,PM2.5_{t-4,i} \\
&+ 0.007\,PM2.5_{t-5,i} + 0.0048\,PM2.5_{t-6,i} + 0.029\,Humidity_{t,i} + 0.075\,Temperature_{t,i} \\
&- 0.29\,Spring_{t,i} - 0.72\,Summer_{t,i} - 0.22\,Winter_{t,i} - .44\,Chinatown_{t,i} \\
&- 0.28\,Potrero\ Hill_{t,i} - 0.29\,SoMa_{t,i} - 0.34\,Tenderloin_{t,i} + \epsilon_{t,i}
\end{aligned}
$$

Features include

- 6 steps back
- Temperature
- Humidity
- Seasonal effects
- Neighborhood effects



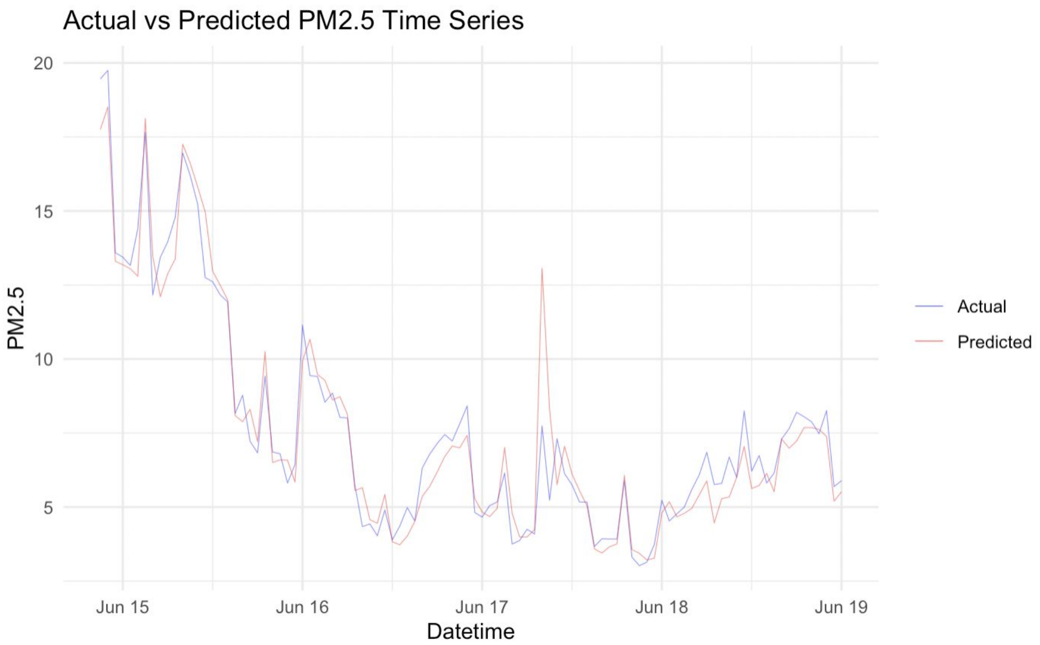

```
Residuals:
    Min      1Q  Median      3Q     Max
-98.111  -0.559  -0.056   0.465  76.778

Coefficients:
                                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      -2.4305866  0.0794405 -30.596  < 2e-16 ***
PM2.5_Hour_MassConc_Calibrated_lag_1              0.7465049  0.0040363 184.947  < 2e-16 ***
PM2.5_Hour_MassConc_Calibrated_lag_2              0.1459185  0.0051736  28.204  < 2e-16 ***
PM2.5_Hour_MassConc_Calibrated_lag_3              0.0264004  0.0049120   5.375 7.70e-08 ***
PM2.5_Hour_MassConc_Calibrated_lag_4              0.0401059  0.0051453   7.795 6.57e-15 ***
PM2.5_Hour_MassConc_Calibrated_lag_5              0.0070125  0.0046542   1.507    0.132
PM2.5_Hour_MassConc_Calibrated_lag_6              0.0047943  0.0036468   1.315    0.189
Humidity                                          0.0291802  0.0006721  43.416  < 2e-16 ***
Temperature                                       0.0751579  0.0023408  32.107  < 2e-16 ***
`factor(lag_data_subset$Season)Spring`           -0.2907651  0.0229080 -12.693  < 2e-16 ***
`factor(lag_data_subset$Season)Summer`           -0.7217566  0.0200783 -35.947  < 2e-16 ***
`factor(lag_data_subset$Season)Winter`           -0.2195454  0.0209274 -10.491  < 2e-16 ***
`factor(lag_data_subset$Neighborhood)Chinatown`    -0.4346814  0.0252152 -17.239  < 2e-16 ***
`factor(lag_data_subset$Neighborhood)Potrero Hill` -0.2788204  0.0236962 -11.766  < 2e-16 ***
`factor(lag_data_subset$Neighborhood)SoMa`        -0.2907601  0.0200524 -14.500  < 2e-16 ***
`factor(lag_data_subset$Neighborhood)Tenderloin`  -0.3439306  0.0198398 -17.335  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.656 on 55046 degrees of freedom
Multiple R-squared:  0.8905,	Adjusted R-squared:  0.8905
F-statistic: 2.984e+04 on 15 and 55046 DF,  p-value: < 2.2e-16
```

Berkeley
UNIVERSITY OF CALIFORNIA

We are able to forecast PM2.5 in San Francisco with **85% Accuracy**



Berkeley
UNIVERSITY OF CALIFORNIA