COVID-19 Forecasting with California Mobility Data

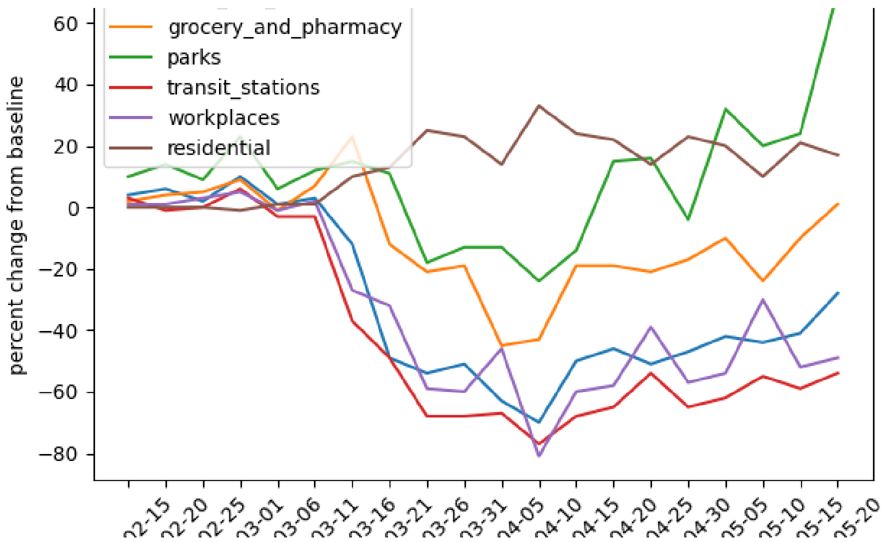
Yanshen Sun

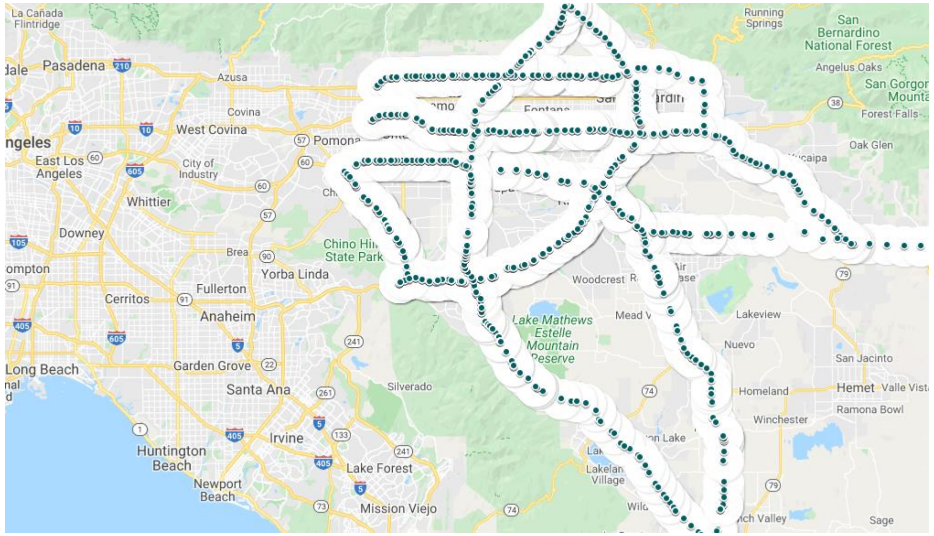
Problem statement

- California county-daily-level
- Human mobility data
  - => number of COVID cases forecasting
  - For day t, predict cases on day t+1 with day t-9, t-8, …, t-1, t

Data

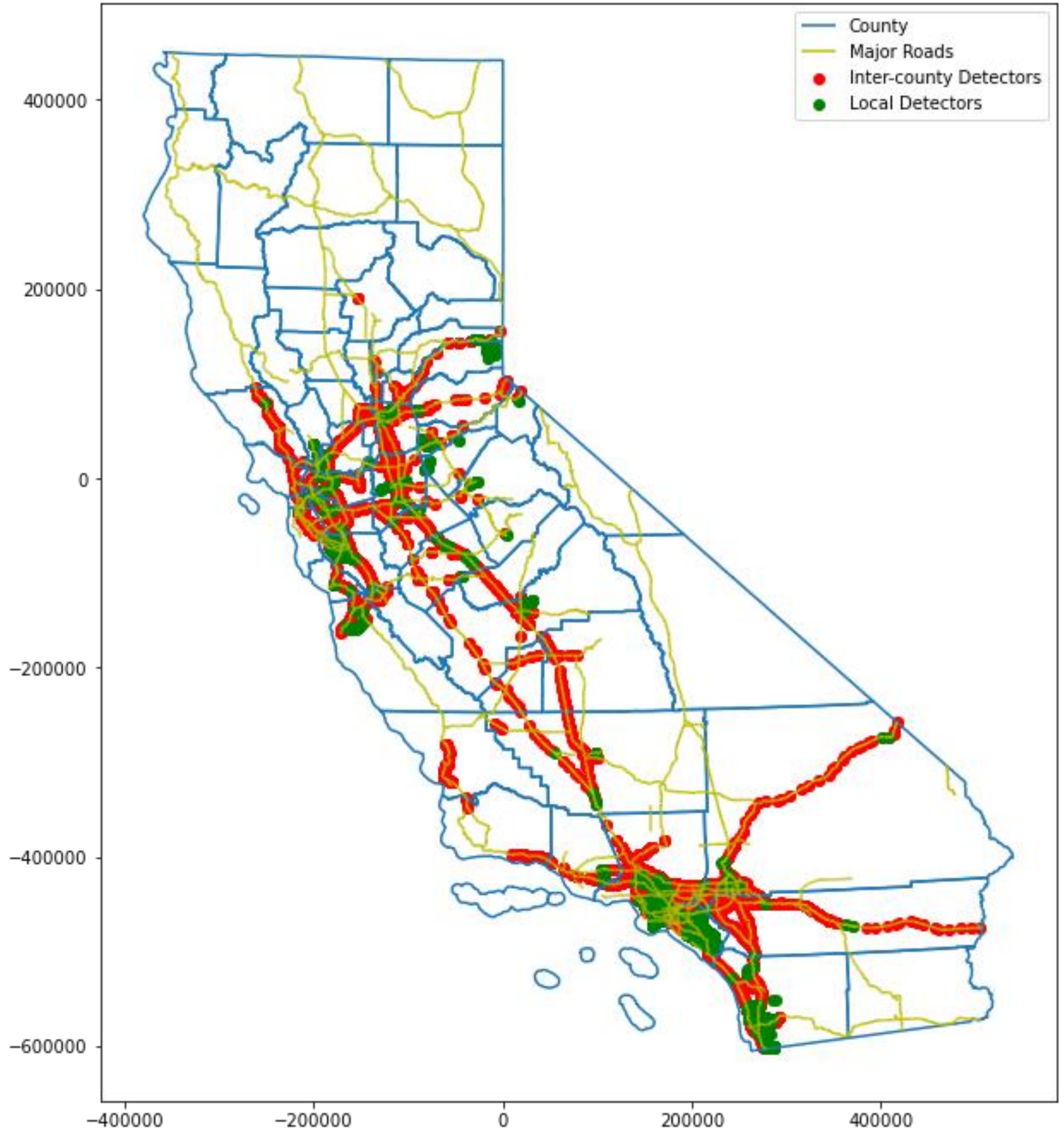
- April 1st ~ May 31st
- New York Times (NYT) COVID-19 dataset: daily case increment
- Google Community Mobility Report: human activities
- PeMS Caltrans traffic data: detector locations and # of cars (traffic volume)
- California county map and arterial road network
- 61 days * 58 counties





Input features

- 86 attributes per day per county
  - Daily traffic volume (2*5*8=80)
    - Local road, inter-county (major) road
    - 4 directions + no direction
    - Count, mean, std, min, 25%, 50%, 75%, max
  - Mobility attribute (6)
- Historical Covid-19 case variation



Dimension reduction

- PCA (60+ features to keep 90% information)
- Random Forest Regressor
- Backward Feature Elimination
- Forward Feature Selection
- 86->42 features

Moran’s-I[1]

- Measure of spatial autocorrelation
- Consider adjacent counties as neighbors
- All attributes are of P-values 0.0
- Strong spatial correlation

	Moran_I	Z_score	P_value
grocery_n_pharmacy	0.131272	108.946091	0.0
ic_n_max	0.170293	141.260832	0.0
ic_s_max	0.122447	101.637609	0.0
ic_w_75	0.132757	110.175405	0.0
ic_e_25	0.074269	61.739311	0.0
ic_s_50	0.102158	84.834992	0.0
ic_n_25	0.149696	124.203135	0.0
Ic_s_max	0.139174	115.489891	0.0
Ic_w_25	0.117173	97.270081	0.0
ic_w_count	0.090619	75.279474	0.0
ic_w_mean	0.034750	29.011929	0.0

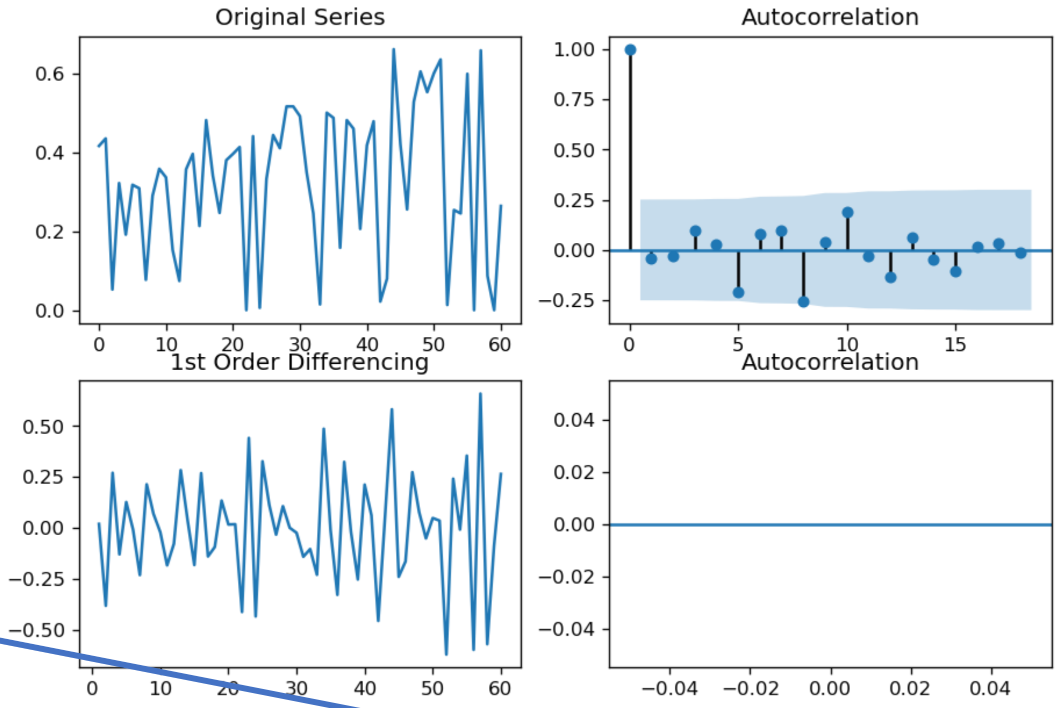
Geographically Weighted Regression (GWR)[2]

- Linear combination of its attributes and its neighbors' attributes

$$y(s) = \beta_1(s)x_1(s) + ... + \beta_p(s)x_p(s) + \epsilon(s)$$

- Current attributes as input, next day's case as output
- 80% train – 20% test 10-fold cross validation
- Test MSE: 0.67

SARIMAX (case study)

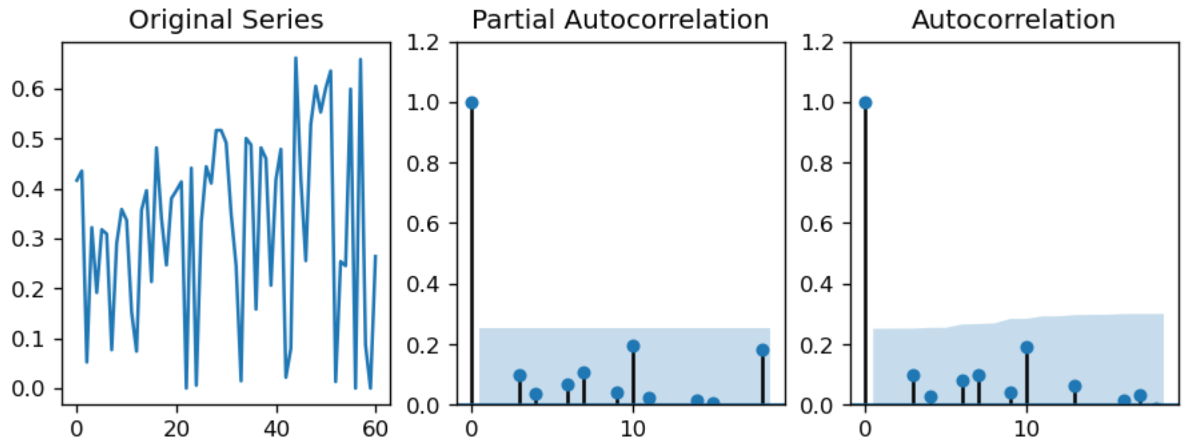


- Los Angeles County

$$(\Delta y_t - \beta_0) = \phi_1(\Delta y_{t-1} - \beta_0) + \theta_1 \epsilon_{t-1} + \epsilon_t$$

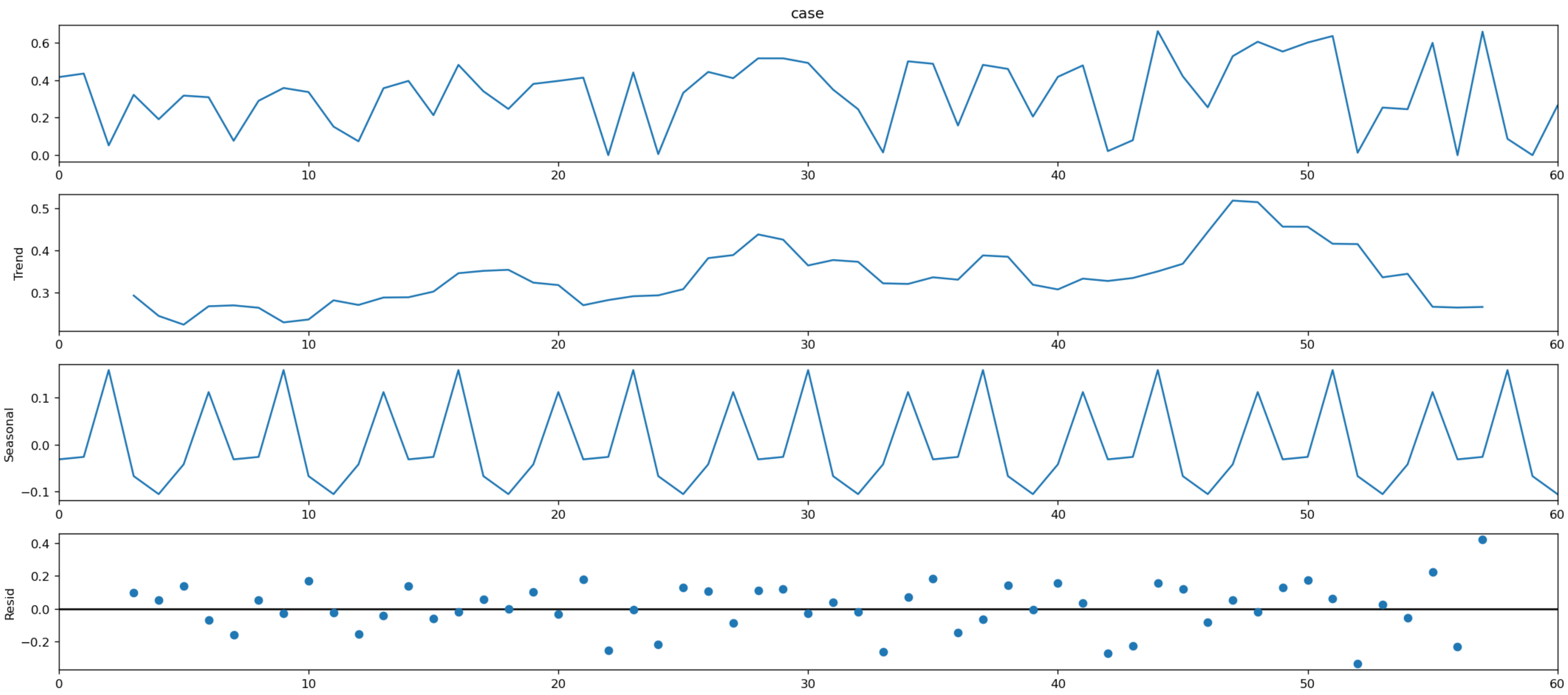
- Order:
  - Differential 0
  - MA: 1~2
  - AR: 1~2
- Multiple inputs

$$y_t = \beta_t x_t + u_t$$
$$(1 - \phi_1 L - \phi_2 L^2) u_t = A(t) + \epsilon_t$$



SARIMAX (case study)

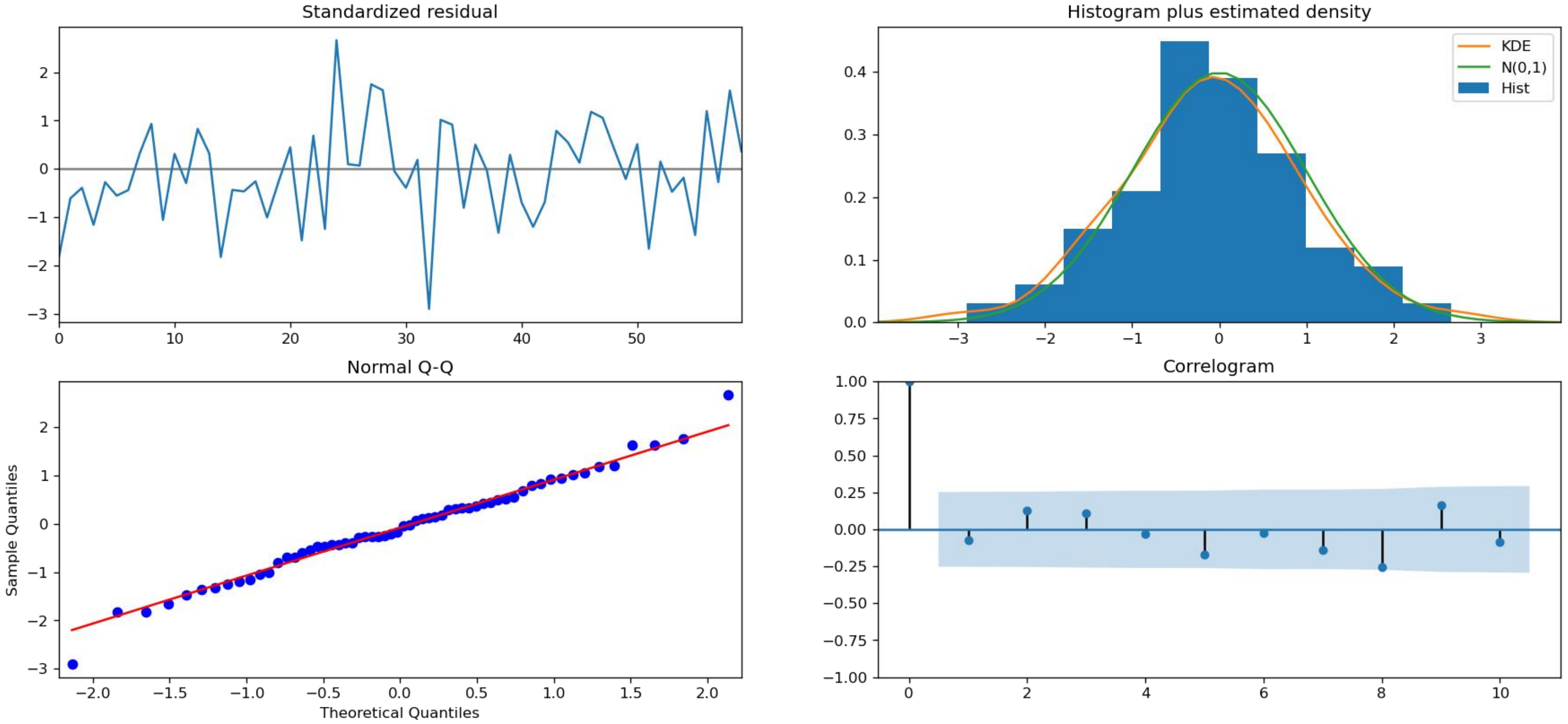
- No Seasonal pattern



SARIMAX (case study)

- Result seems good
- Significant variables:
    - ic_n_std
    - lc_n_min
    - workplaces

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                 D.case   No. Observations:                   60
Model:               SARIMAX(0, 0, 1)   Log Likelihood                  33.827
Date:                Mon, 03 May 2021   AIC                             22.347
Time:                        02:09:12   BIC                            116.592
Sample:                             0   HQIC                            59.211
                                 - 60
Covariance Type:                  opg
```



$$AIC = 2k - 2\ln(\hat{L})$$

AIC = Akaike information criterion

k = number of estimated parameters in the model

$\hat{L}$ = maximum value of the likelihood function for the model

SARIMAX[3]

- One-step-ahead prediction
- Current attributes as input, next day’s case as output
- 80% train – 20% test
- Different models for each county
- Average MSE loss: 0.12

Results

- Merging temporal and spatial model with a linear model
  - MSE: 0.015
- Temporal model + linear model
  - MSE: 0.056
- Spatial model
  - MSE: 0.029

Conclusion

- Goals achieved
  - Spatial-temporal model performs better than spatial/temporal models
  - Spatial relations of counties does improve the results
  - Human mobility relates to Covid-19 cases, especially inter-county travels
- Drawbacks
  - Should find a better way to merge different models

Reference

- [1] Moran, P. A. P. (1950). "Notes on Continuous Stochastic Phenomena". Biometrika. 37 (1): 17–23. doi:10.2307/2332142. JSTOR 2332142
- [2]Brunsdon, C., S. Fotheringham, and M. Charlton (1998). Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician) 47 (3), 431-443.
- [3] Friedman, B. and Roley, V. (1980), "Models of Long-Term Interest Rate Determination", Journal of Portfolio Management, 6 (Spring), 35-45.

Thank!