Drug Overdoses in the United States

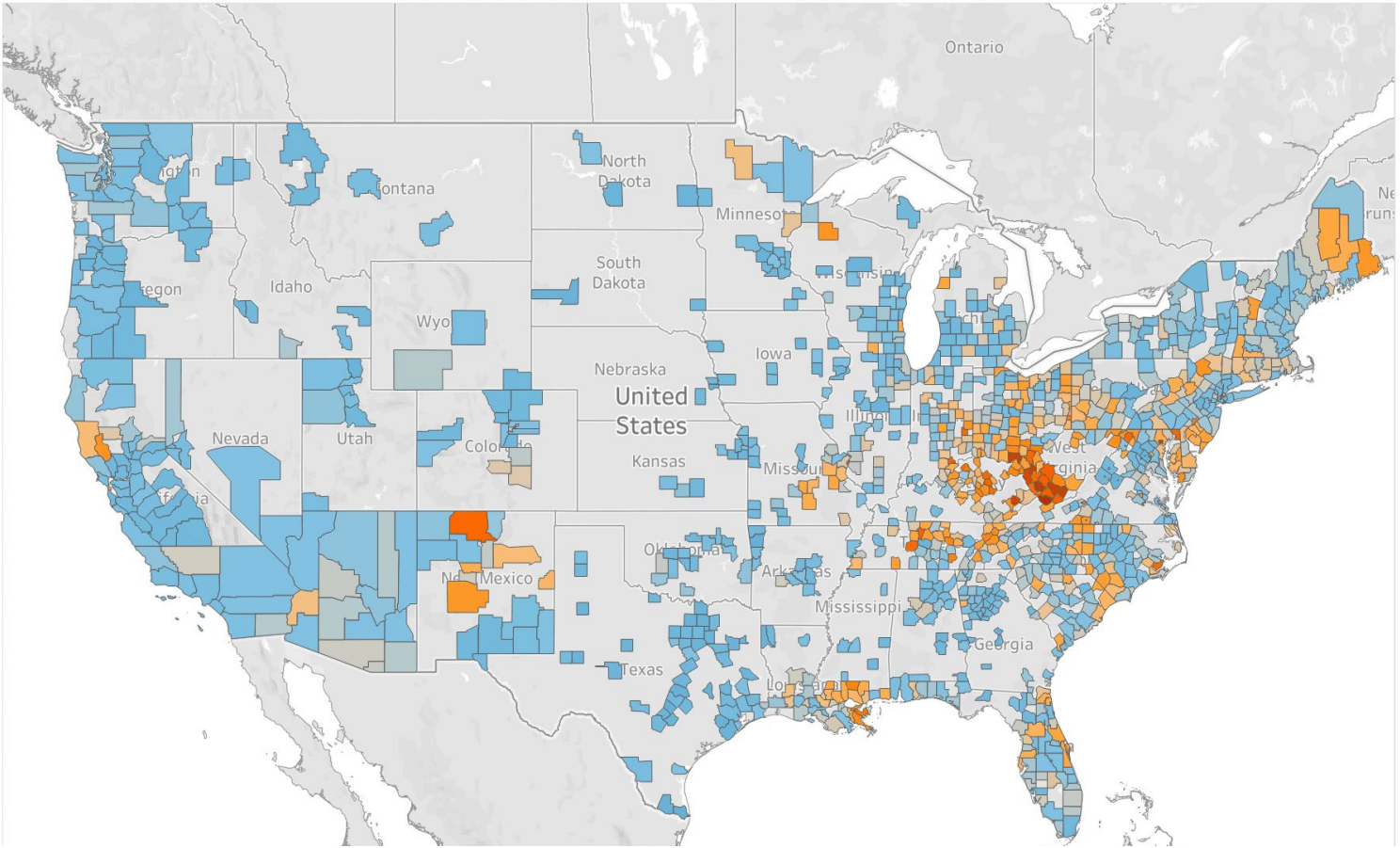
Wei Deng, Ashlyn Jew, Rhiann Zhang, Meera Duggal

Project Goal

Which characteristics of U.S. counties can best **explain** drug overdose rates?

Previously...

Overdose Rate - 2020



→ Main Dataset

- raw counts of deaths related to drug overdose
- (from CDC)

→ **Missing Data**

Previously...

→ Procured over 40 other county characteristics on

- demographics
- health
  - ex) % Uninsured, % Smokers, % Diabetes, ...
- economics
  - ex) % College Educated, % Unemployed, Median Household Income, ...

Year	County	Overdose_Rate_per_100k	Population	Pct_Age_lt_18	Pct_Age_gte_65	Pct_Black	Pct_Native	Pct_Asian	Pct_Pacific_Islander	...	Pct_Unemployed	Pct_Poverty	Median_HHI	Jail_Pop	Urbanicity	Spatial_Mean
2011	Baldwin County, AL	12.853677	179878.0	23.100000	17.000000	10.000000	0.500000	0.600000	0.000000	...	9.0	13.4	50144.0	743.36	2.0	21.667119
2011	Calhoun County, AL	11.884853	114081.0	23.400000	15.000000	20.000000	0.400000	0.800000	0.100000	...	10.3	20.9	39060.0	504.25	2.0	17.854349
2011	Cullman County, AL	12.416807	81778.0	23.300000	15.600000	1.700000	0.500000	0.200000	0.000000	...	9.5	18.2	39395.0	312.43	1.0	34.421638
2011	De Kalb County, AL	15.411559	69380.0	25.500000	14.300000	2.000000	0.800000	0.400000	0.100000	...	11.7	20.3	35487.0	204.54	1.0	15.905439
2011	Escambia County, AL	42.000263	37434.0	23.000000	15.200000	32.100000	3.100000	0.300000	0.000000	...	10.7	25.5	32838.0	192.75	1.0	13.108417
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Project Goal

Which characteristics of U.S. counties can best **explain** drug overdose rates?

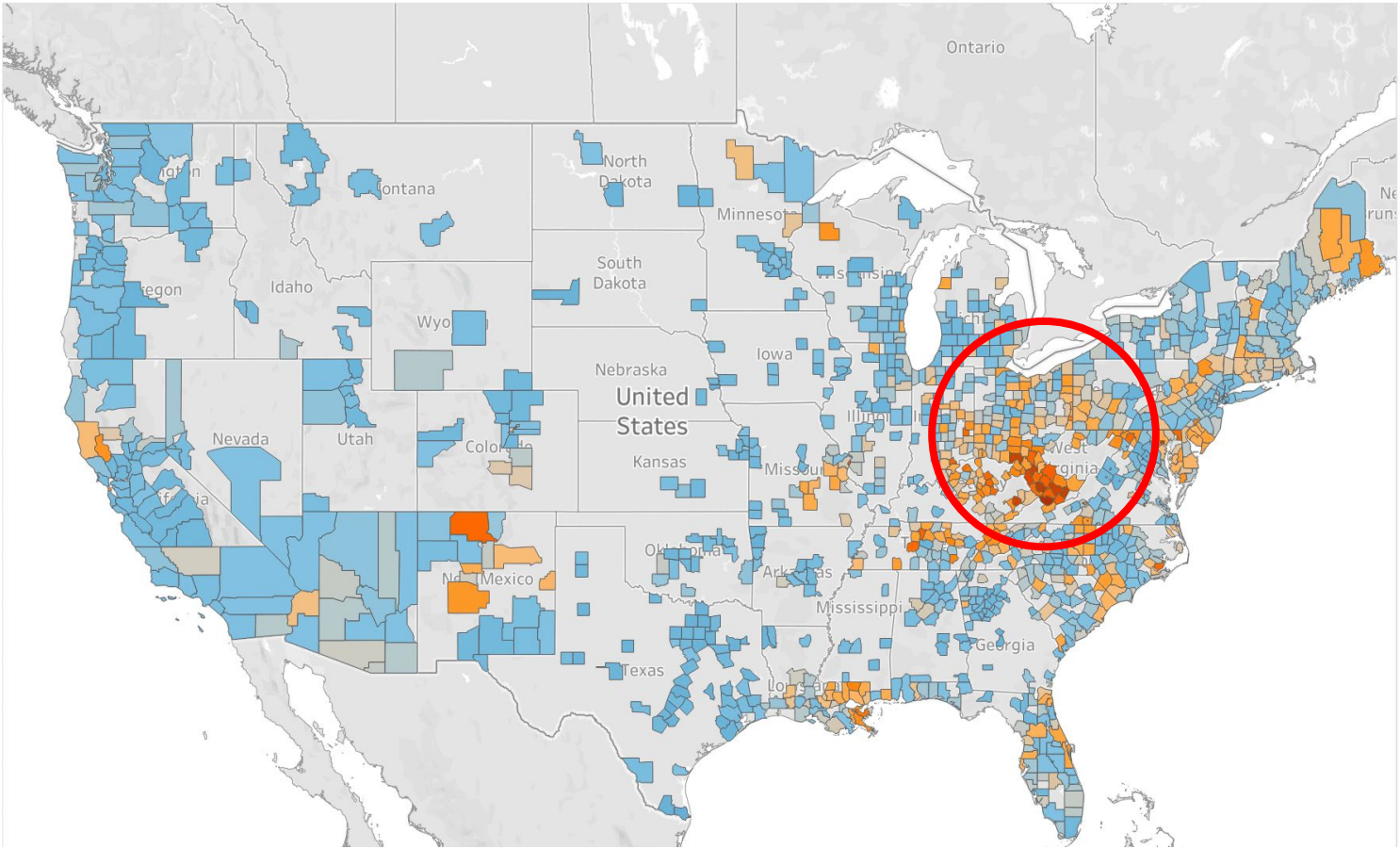
Determining these characteristics will...

→ allow us to fill in our U.S. map with estimated overdose rates

→ grant policy makers insight as to what factors to focus on as well as which communities should be addressed more urgently when trying to counteract the drug overdose epidemic

Previously...

Overdose Rate - 2020



→ Possible clustering of overdose rates

→ Therefore, we explored the geospatial aspect extensively

Geospatial Component: Global Moran's I

→ We introduce Moran's I to determine if there is a spatial relationship in our data
- measures how similar one county is to all other counties

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \sum_{i=1}^{n} z_i^2}$$

→ $z_i = (x_i - \bar{x})$ is the deviation of an attribute (i.e. our overdose rate) from the mean for county i

→ $w_{i,j}$ is the spatial weight between county i and j

→ n is the total number of counties

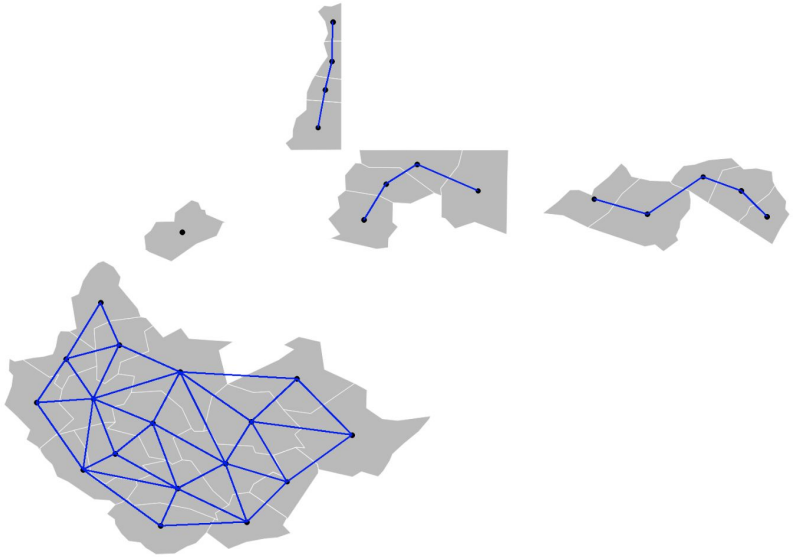
Moran's I

→ In the previous equation, we used the queen weights (depicted on the right)

→ Our Moran's I for overdose rates is **0.461**
- Moran's I values range from -1 to 1
- As a general rule of thumb, a Moran's I value of above 0.3 or below -0.3 is deemed significant

→ With the Moran's I of **0.461** there is a spatial autocorrelation between the counties, so we concluded that it is important to have this component in our model

Ex. Queen Weights for West Virginia



$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are contiguous} \\ 0 & \text{if } i \text{ and } j \text{ are not contiguous} \end{cases}$$

Ordinary Least Squares Modeling

$$y_{it} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{itk} + \epsilon_{it}$$

drug overdose death rate at county i, for year t

global intercept coefficient

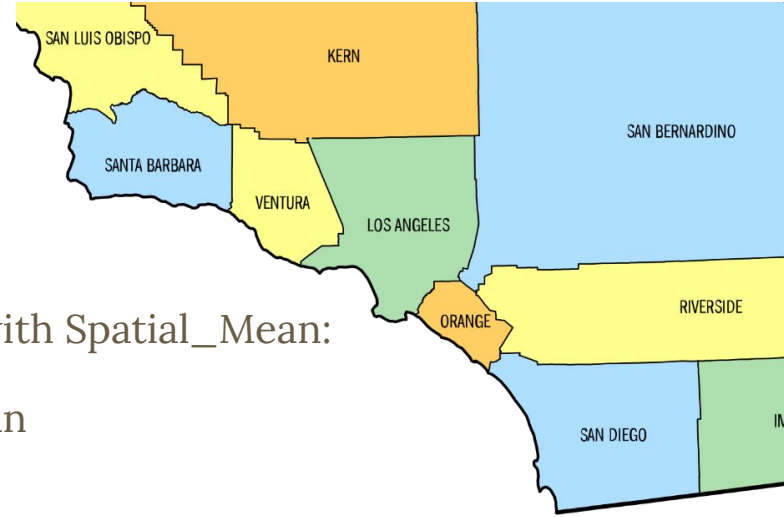
global k-th regression coefficients for the k-th explanatory variable

k-th explanatory variable at county i, for year t

random error term associated with county i, for year t

* Model coefficients were computed using all data throughout all years.
However, when estimating our overdose rates, we input data for each county subsetted by year

Baseline Model: Geospatial Component



→ We introduce a spatial component to our model:

- **Spatial_Mean**: *Average* overdose rate of the counties that are adjacent to the focal county

→ For our baseline model, we ran an OLS model only with Spatial_Mean:

- log_Overdose_Rate_per_100k ~ Spatial_Mean
- Adjusted R^2: 0.451
  - Accounts for a considerable amount of the variation in our data!
- Normalized RMSE: 0.407 / (5.144 - 0.803) = 0.0938
  - Already a pretty good fitting model!

Backwards Stepwise Feature Selection

→ Utilizing 5-fold cross-validation with RMSE as our metric

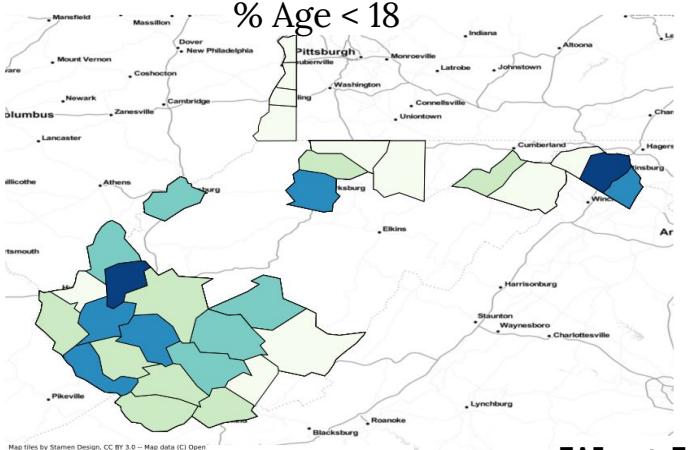
- best model has 8 features

→ Chosen Model:

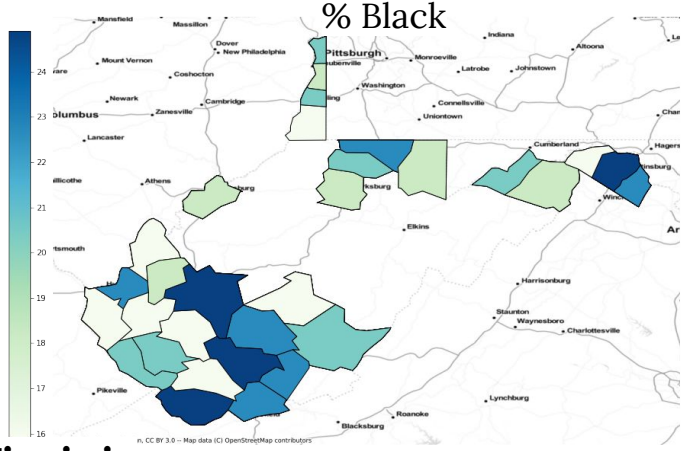
- Spatial_Mean (Average overdose rate of adjacent counties)
- PrimCarePhys_per_100k (# of primary care physicians per 100k residents)
- Pct_Uninsured (% of uninsured residents)
- Pct_Child_in_1ParentHH (% of children in 1 parent households)
- Pct_Poverty (% of residents in poverty)
- Pct_Black (% of Black residents)
- Pct_Age_lt_18 (% of residents that are less than 18 years old)
- Potential_Years_Lost (Years of potential life lost before age 75 per 100k residents)

Spatial Components of Predictors

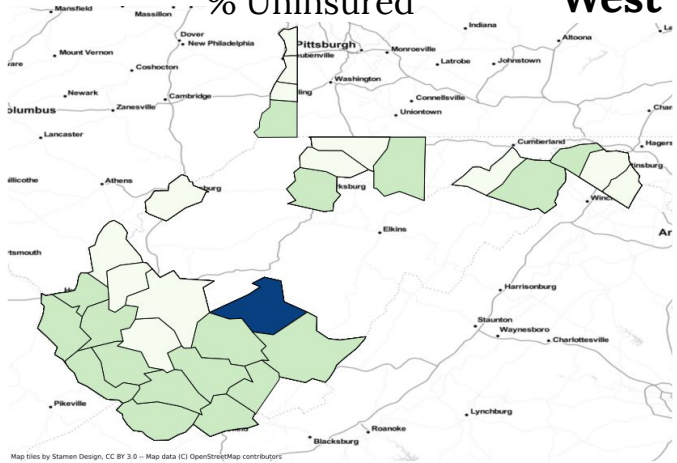
% Age < 18



% Black

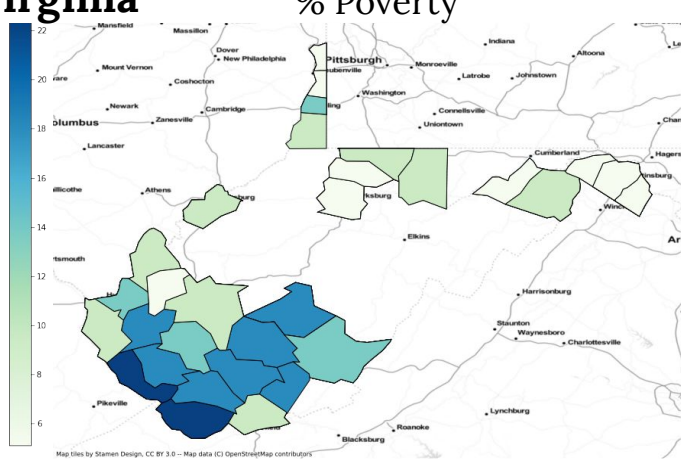


% Uninsured



West Virginia

% Poverty

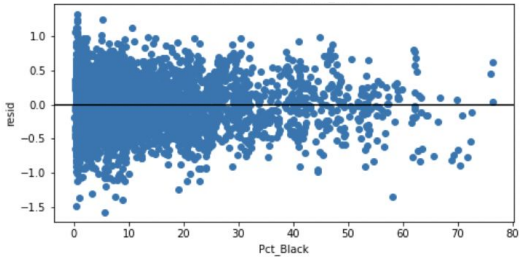


Predictors	Global Moran's I
% Age < 18	0.598
% Black	0.649
% Uninsured	0.526
% Poverty	0.483
Potential Years Lost	0.617
% Children in 1 Parent Household	0.342
Primary Care Physicians per 100k	0.204

Data Transformations

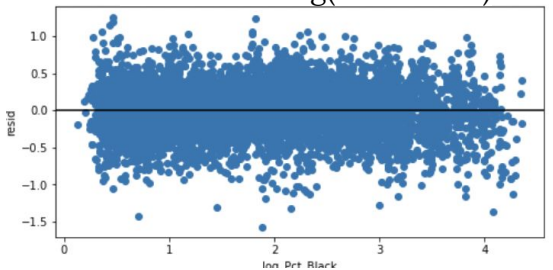
→ These are the residuals before and after we transformed them in our model

Residual vs. % Black

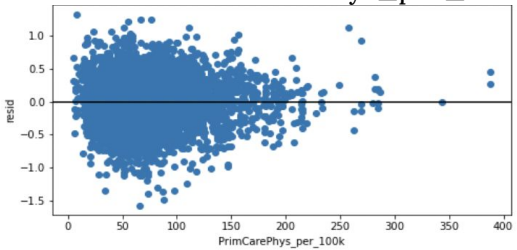


log(x + 1)

Residual vs. log(% Black + 1)

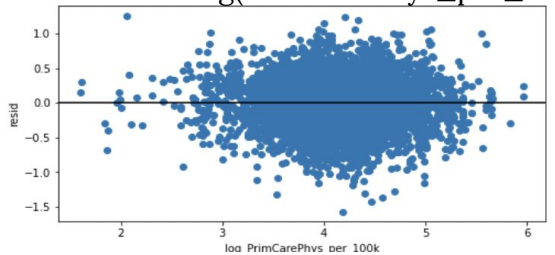


Residual vs. PrimCarePhys_per_100k

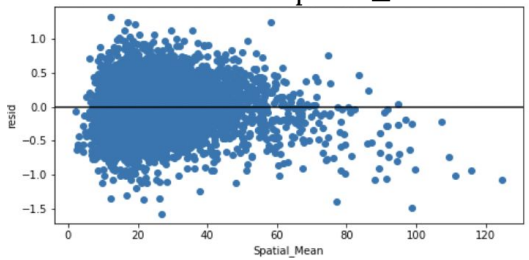


log(x)

Residual vs. log(PrimCarePhys_per_100k)

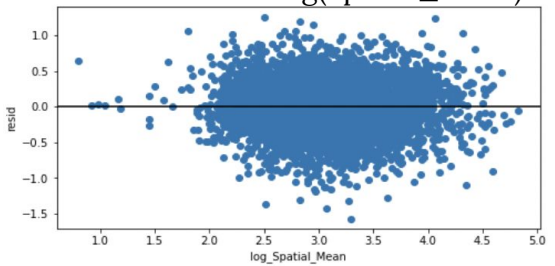


Residual vs. Spatial_Mean



log(x)

Residual vs. log(Spatial_Mean)



Model Performance

OLS Regression Results

Baseline Model:	Pre-transformation:	Post-transformation:	
		R-squared:	0.626
Adj. R^2: 0.460	Adj. R^2: 0.613	Adj. R-squared:	0.625
		F-statistic:	1195.
		Prob (F-statistic):	0.00
		Log-Likelihood:	-2011.4
AIC: 6125	AIC: 4221	AIC:	4041.
		BIC:	4101.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.6932	0.083	20.379	0.000	1.530	1.856
Pct_Age_lt_18	-0.0233	0.002	-13.950	0.000	-0.027	-0.020
log_Pct_Black	-0.0566	0.006	-9.750	0.000	-0.068	-0.045
Potential_Years_Lost	9.327e-05	3.41e-06	27.313	0.000	8.66e-05	0.0001
Pct_Uninsured	-0.0106	0.001	-10.409	0.000	-0.013	-0.009
log_PrimCarePhys_per_100k	-0.0923	0.010	-9.291	0.000	-0.112	-0.073
Pct_Child_in_1ParentHH	0.0084	0.001	9.019	0.000	0.007	0.010
Pct_Poverty	-0.0119	0.001	-7.987	0.000	-0.015	-0.009
log_Spatial_Mean	0.5550	0.012	46.119	0.000	0.531	0.579

Normalized CV-Average RMSE

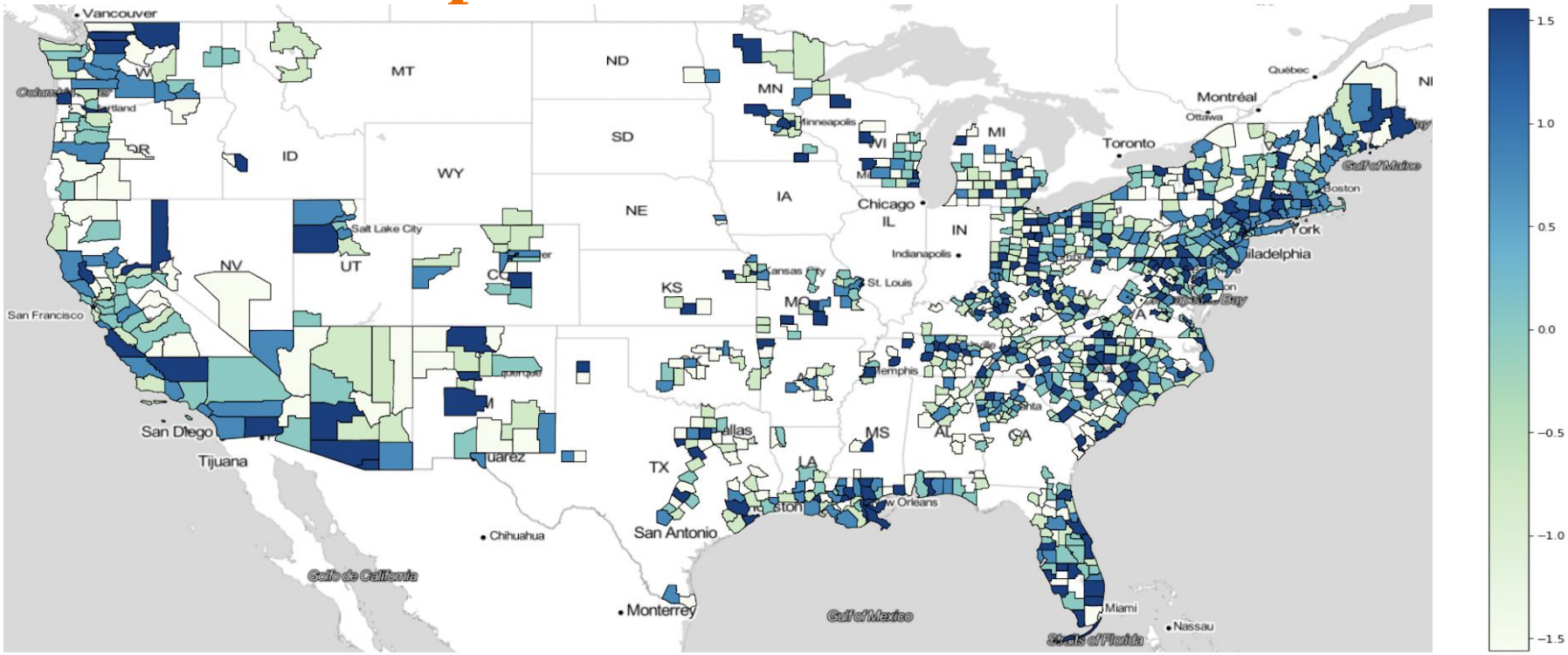
Training: 0.07876 Testing: 0.07887

→ Seeing improvements performance in each of our subsequent models

→ All features are significant

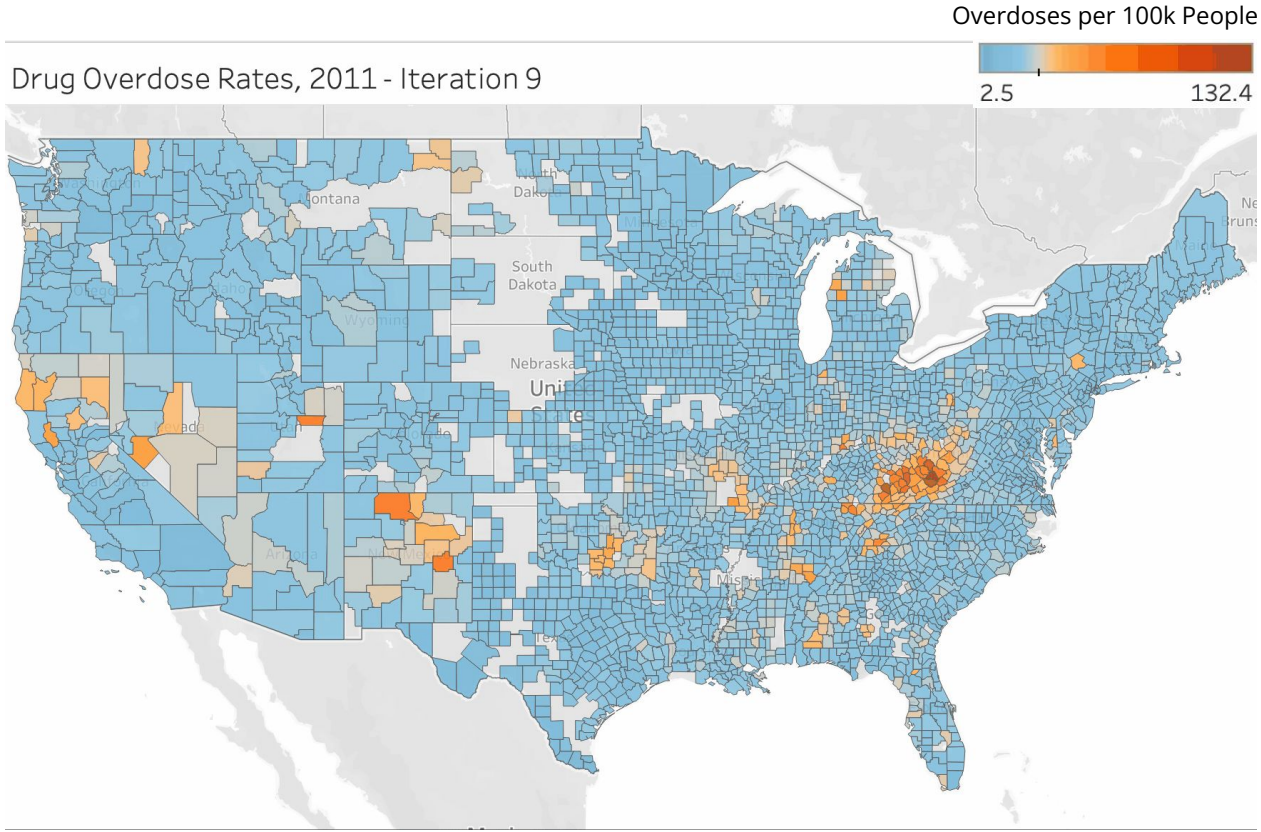
→ Normalized training and testing RMSE both close to zero!

Residual Map for 2020



→ The Moran's I for the residuals are -0.098. Thus, our model has successfully accounted for the spatial component in our data.

Filling in our Map Sequentially

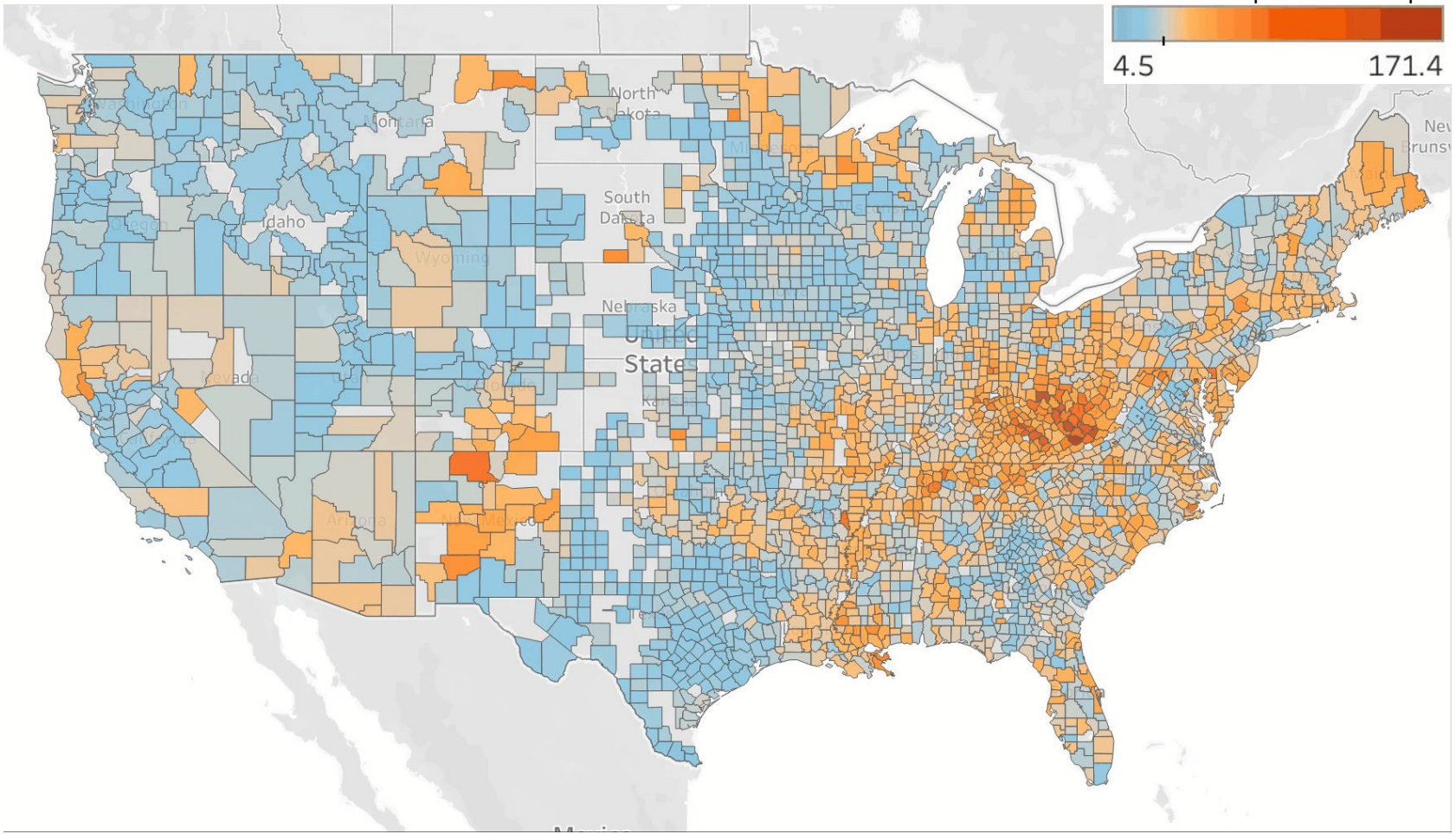


Estimating using estimates:

- Use model to estimate missing adjacent county overdose rates
- Calculate spatial component for filled-in missing counties
- Repeat until we reach a point when we can no longer estimate new missing counties

* Still have some missing counties since they do not have data for other covariates

Drug Overdose Rates, 2020 - Iteration 9



Conclusion

→ Now that we have filled in our entire U.S. map with estimates, policy makers can get more insight for those areas that had missing overdose rates

→ Policy makers should focus on

- counties with severe overdose rates and their surrounding counties
- improving health care availability:
  - % Uninsured, # Primary Care Physicians
- these demographics:
  - Youth < 18 yrs, Black population, Poverty, Single Parent Households

→ Given the performance of our model, we see that there is still room for improvement

- Possibly leaving out predictors that are important
- Can explore other models:
  - Weighted Least Squares or Time Series
- This is a complex societal issue that is still being researched

finale.

-warm