Stat 222 Project: Geospatial Modeling of Drug Overdose in the United States

Individual Progress Report
Written By Wei Deng

Group Members:
Meera Duggal
Ashlyn Jew
Rhiann Zhang

March 20, 2022

I Exposition

The primary goal of our project is to assess the relationship between aspects of counties in the United States (including demographics, location, incarceration rates, and opioid dispense rates) and their corresponding drug overdose rates using publicly available county-level data. We also hope to add a geospatial component into our modeling, estimating a county's drug overdose rate based on surrounding counties.

With this project, we hope to inform public policy decisions related to drug overdoses at a national level, appealing to law makers and government agencies about the factors that most contribute to drug overdoses.

II Data

i Data Sources

The data we are using are mainly collected and maintained by divisions of the United States government. These include the Center for Disease Control and Prevention [1][2], U.S. Census Bureau [3], and the U.S. Bureau of Labor Statistics [4]. Other data (such as incarceration rates) were collected by independent organizations such as the Vera Institute of Justice [5].

ii Data Description

The outcome variable of interest in our data is the drug overdose rate. We labeled this as ”Cruder Rate” in our data set, which is a name derived from the ”Crude Rate” provided by the CDC. Since for many of the counties, this rate was left blank or labeled ”Unreliable”, but still retained the raw numbers, we estimated the rate as the number of overdoses per 100,000 people.

The independent variables we collected and merged to our data set include population by gender and age, poverty rate, urbanicity, opioid dispense rate, and incarceration rate. All of our data are at the county level, over recent years (mainly 1999–2020). All together, there are almost 16000 observations in our data set.

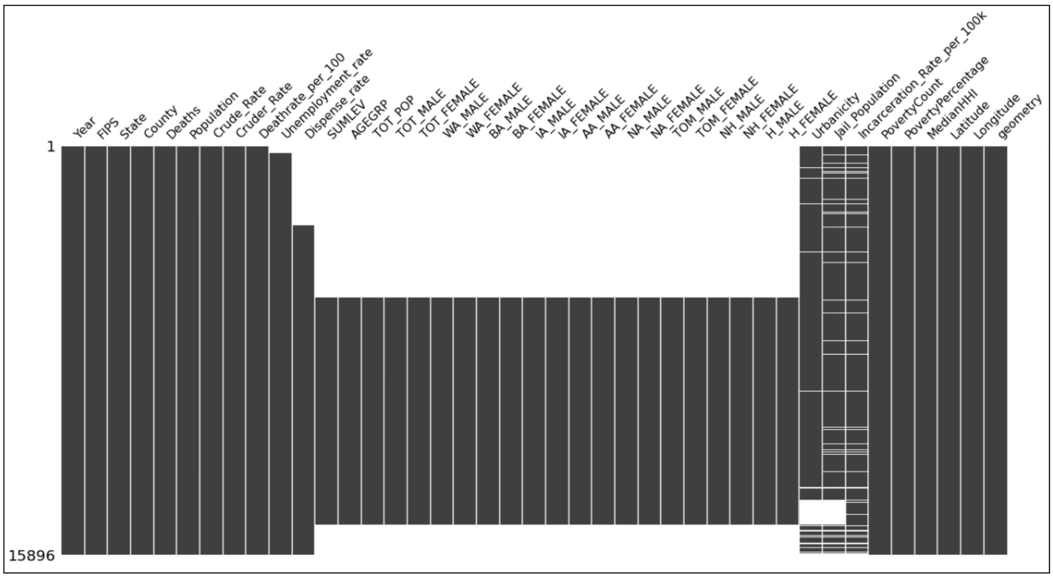


Figure 1: For each variable, whether we have data or not, sorted by year (top is 1999)

From Figure 1, we see that we have a lot of missing data for the earlier years, as well as some for the most recent years. As such, we decided to focus our analysis on the years in which we have all of the data present, 2010–2019.

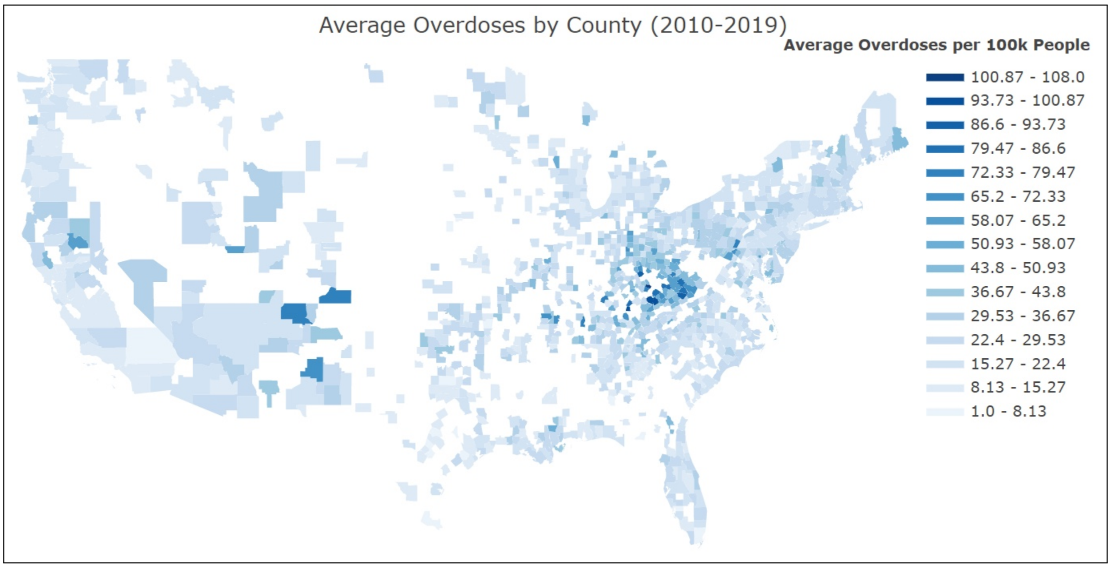


Figure 2: Our dependent variable, drug overdose rate, is missing in many counties

Although we are missing drug overdose rates for many counties for the years of interest, as shown by Figure 2, the counties for which we have data account for more than 60% of the total U.S. population. As such, we can still make some valid inferences about drug overdose rates. We also see that there is a concentration of high drug overdose rates in the midwest/south area of the country, which suggests that there might be a geospatial relationship for drug overdose rates (i.e. a county surrounded by others with high drug overdose rates has a high drug overdose rate itself).

Based on intuition and popularly purported reasons for drug overdose, we decided to focus on county-level variables such as unemployment rate, opioid dispense rate, and incarceration rate. In Appendix 1, we compare the the average drug overdose rates in the counties of West Virginia (the state with the highest drug overdose rate by far) with the average values for unemployment, opioid dispensary, and incarceration. Additionally, we see that there appears to be clusters of higher and lower drug overdose rates among the counties, suggesting that there may be a geospatial component to a county's drug overdose rate. We also see that it appears incarceration rate is more correlative of a factor to drug overdose rate compared to unemployment rate and opioid dispense rate. Lastly, these plots illustrate the amount of missing values in some counties in our data set.

Before doing any numeric analysis, we standardized all of our independent variables to put the variables on a similar scale and lessen the impact of outliers.

III Methods

i Aspatial Ordinary Least Squares Regression

Without taking into account the spatial relationship of our data, we fit ordinary least squares regression models of drug overdose rate on subsets of our independent variables. We wished to start with a simple, interpretable model that we could use a baseline with which to compare more complex models later. In Appendix 2, we see the output of the OLS regression of drug overdose rate (”Cruder Rate”) on year, number of white males, urbanicity (ordinal variable from 1-4, 4 being urban), unemployment rate, dispense rate, incarceration rate, poverty percentage, and median household income. These are variables that we picked manually, based on our intuition of what factors could have a strong relationship with drug overdose rates based on initial exploratory data analysis. The general OLS regression formula is given by:

$$C_i = \alpha + \sum_{k=1}^{p} \beta_k X_{ik} + \epsilon_i$$

where C is the estimated drug overdose rate, α is the intercept, p is the number of features in our model, β_k is the coefficient associated with the k^{th} feature, X is the set of features, and ϵ is the error term.

ii Best Subset Selection

Because we have so many different combinations of features to consider in our OLS regression model, we implemented an exhaustive search for the best subset of features. The metric we used to compare between the different models was AIC. The model with the best performing AIC was the one with the following subset of features: year, unemployment rate, opioid dispense rate, African American male, total male, non-Hispanic male, jail population, incarceration rate, poverty county, and median household income. The results of the OLS regression on this subset of features can be seen in Appendix 3.

Looking at the chosen subset of features, there seems to be obvious multicollinearity. This is confirmed by looking at the respective VIF (variance inflation factor) scores for each of the features. In the future, we will research methods to reduce our model's multicollinearity. Such methods could include ridge/LASSO regression and principal component analysis.

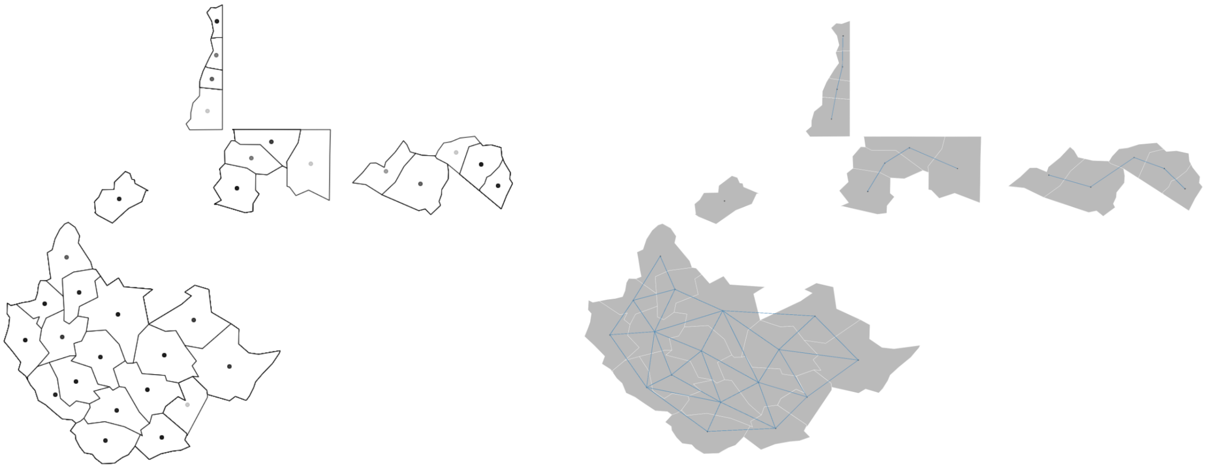
iii Moran's I

In order to determine whether or not there is a spatial relationship between the drug overdose rates of counties, we calculated Moran's I, a measure of spatial autocorrelation, similar to Pearson's correlation coefficient. Moran's I ranges from -1 to 1, with 0 indicating no spatial relationship, 1 indicating positive spatial relationship (counties with similar drug overdose rates are close to each other) and -1 indicating negative spatial relationship (counties with similar drug overdose rates are far from each other). The general formula for Moran's I is given by:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

where n is the number of observations, z_i is the standardized value of the variable of interest $(x_i - \bar{X})$ at location i, and w_{ij} is the spatial weight between locations i and j (the spatial relationship between points i and j) [6]. With our implementation of Moran's I, we chose to use Queen's weights, meaning that a spatial relationship can be established by one county with all the other counties adjacent to it. We did this by using the GeoPandas library, which allowed us to construct polygons to represent the county boundaries. Then, we could compute a centroid for each polygon (county) that represented the middle. Next, we could use the Queen's weight method to calculate spatial weights for our Moran's I calculation. A visual representation of this process can be seen with West Virginia in Figure 3. Calculating the global Moran's I for all of our available counties, we obtained a value of 0.46. This indicates that there is a moderate positive autocorrelation between county proximity and drug overdose rate (i.e. a county being close to another correlates with drug overdose rate).

In Appendix 4, we can see, on a local level, what Moran's I means for West Virginia and how the magnitude of drug overdose rates tend to cluster, leading to higher autocorrelation.



(a) Centroids

(b) Queen's weights

Figure 3: West Virginia Counties

IV Results

i OLS Regression

For the particular regression shown in Appendix 2 (Aspatial OLS with manually selected features), we see a low p-value for the F-statistic, meaning that it is statistically significant (under 0.05 level of significance) that our model explains more of the variation in drug overdose rate than a model with no features would. Taking a look at the significance of individual features, we see that all of our features seem to be statistically significant except for urbanicity, incarceration rate, and poverty percentage.

After running the regression again with the best selected feature with the output in Appendix 3, we see that under a significance level of 0.05, all of our features are statistically significant. Interestingly, we see some of the same features from our first OLS regression that weren't significant become significant in this new model.

ii Moran's I

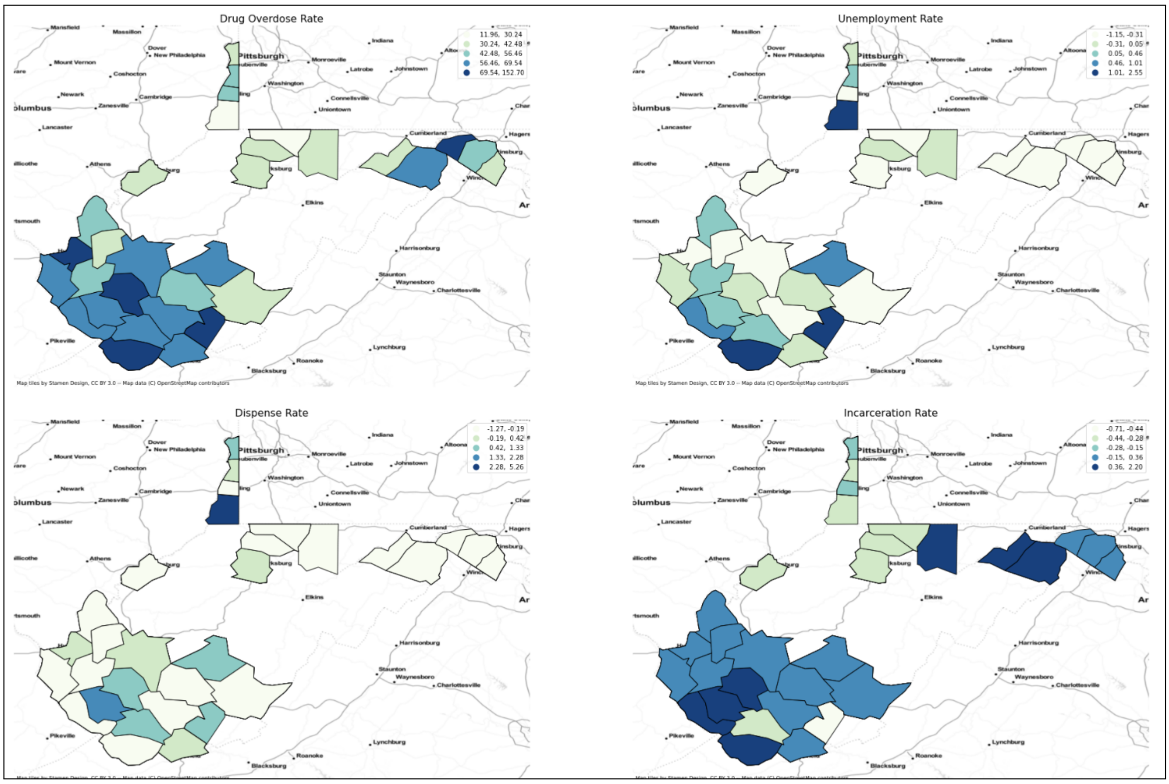
In Appendix 4, we can take a closer look at Moran's I at a local level. Looking at the scatterplot specifically, we recognize that the plot is sectioned into four quadrants: high-high (HH), high-low (HL), low-high (LH), and low-low (LL). The number of points in each of these quadrants represents the spatial relationship pairings of each of the counties of West Virginia, and the slope drawn as the best fit line for these points is Moran's I. We see that the slope is positive, meaning that county proximity is positively correlative with drug overdose rates. We can see this illustrated on the two subsequent plots, where the highest concentration of drug overdose rates correspond with the most spatial correlation.

V Conclusion

In this initial analysis/model building stage of our project, we were able to set a baseline for our future, more complex models with the simple linear regression. We identified the features that are going to explain most of the variation in drug overdose rates and will be looking for ways to reduce the dimensionality/multicollinearity. We were also able to confirm our suspicion that there is a spatial correlation with drug overdose rates among counties. This will allow us to implement a spatial component to any models we construct in the future to estimate drug overdose rates.

Future challenges will include implementing a spatial component to our modeling process, both at the global level and local level. Additionally, we will also build models for different years to see if certain features become more or less important as time goes on and compare each model's performance to decide whether we should pursue different methods. We will also be looking into ways to transform our data in a way so that they are useful for our models. Lastly, we hope to be able to make inferences about which factors are most related to drug overdose rates in our data.

VI Appendix



Appendix 1: Average values for West Virginia (2010–2019)

Dep. Variable:	Cruder_Rate	**R-squared:**	0.240
Model:	OLS	**Adj. R-squared:**	0.240
Method:	Least Squares	**F-statistic:**	299.8
Date:	Sun, 20 Mar 2022	**Prob (F-statistic):**	0.00
Time:	05:11:24	**Log-Likelihood:**	-9572.3
No. Observations:	7591	**AIC:**	1.916e+04
Df Residuals:	7582	**BIC:**	1.923e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-360.5609	11.472	-31.430	0.000	-383.049	-338.073
Year	0.1790	0.006	31.439	0.000	0.168	0.190
WA_MALE	-0.0450	0.011	-4.009	0.000	-0.067	-0.023
Urbanicity	-0.0230	0.013	-1.815	0.069	-0.048	0.002
Unemployment_rate	0.1494	0.015	10.241	0.000	0.121	0.178
Dispense_rate	0.3134	0.014	23.191	0.000	0.287	0.340
Incarceration_Rate_per_100k	0.0088	0.011	0.830	0.406	-0.012	0.029
PovertyPercentage	-0.0275	0.018	-1.572	0.116	-0.062	0.007
MedianHHI	-0.1330	0.020	-6.772	0.000	-0.171	-0.094

Omnibus:	3914.240	**Durbin-Watson:**	1.362
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	41611.534
Skew:	2.240	**Prob(JB):**	0.00
Kurtosis:	13.559	**Cond. No.**	2.36e+06

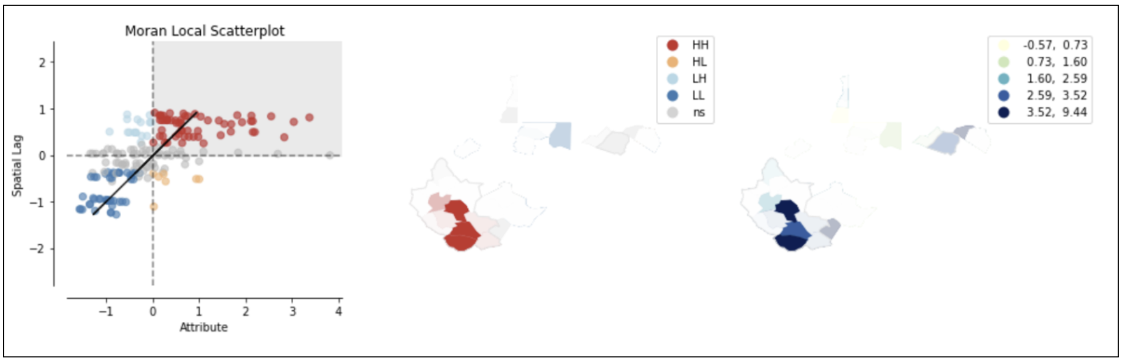
Appendix 2: OLS output of Drug Overdose Rate on manually selected features

Dep. Variable:	Cruder_Rate	**R-squared:**	0.247
Model:	OLS	**Adj. R-squared:**	0.246
Method:	Least Squares	**F-statistic:**	249.0
Date:	Sun, 20 Mar 2022	**Prob (F-statistic):**	0.00
Time:	06:11:30	**Log-Likelihood:**	-9537.4
No. Observations:	7591	**AIC:**	1.910e+04
Df Residuals:	7580	**BIC:**	1.917e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-367.9896	11.136	-33.045	0.000	-389.819	-346.160
Year	0.1827	0.006	33.057	0.000	0.172	0.194
Unemployment_rate	0.1521	0.014	10.784	0.000	0.124	0.180
Dispense_rate	0.2985	0.014	21.930	0.000	0.272	0.325
AA_MALE	0.0982	0.024	4.073	0.000	0.051	0.145
TOM_MALE	-0.1459	0.037	-3.920	0.000	-0.219	-0.073
NH_MALE	0.2223	0.032	7.024	0.000	0.160	0.284
Jail_Population	-0.0997	0.031	-3.231	0.001	-0.160	-0.039
Incarceration_Rate_per_100k	0.0231	0.011	2.017	0.044	0.001	0.046
PovertyCount	-0.1429	0.033	-4.313	0.000	-0.208	-0.078
MedianHHI	-0.1742	0.015	-11.574	0.000	-0.204	-0.145

Omnibus:	3866.583	**Durbin-Watson:**	1.362
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	40325.914
Skew:	2.211	**Prob(JB):**	0.00
Kurtosis:	13.390	**Cond. No.**	2.30e+06

Appendix 3: OLS output of Drug Overdose Rate on best subset selected features



Appendix 4: Local Moran's I, West Virginia

Spatial lag scatterplot (middle), spatial correlation (middle), drug overdose rate (right)

References

[1] Centers for Disease Control and Prevention, National Center for Health Statistics (2021) *Multiple Cause of Death*. Data are from the Multiple Cause of Death Files, 1999-2020, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/mcd-icd10.html.

[2] Centers for Disease Control and Prevention (2021) *U.S. opioid dispensing rate maps.* https://www.cdc.gov/drugoverdose/rxrate-maps/index.html.

[3] U.S. Census Bureau (2021) *County population by characteristics: 2010-2019.* https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html.

[4] U.S. Bureau of Labor Statistics (2021) *Local Area Unemployment Statistics. U.S. Bureau of Labor Statistics*. Retrieved at https://www.bls.gov/lau/#tables.

[5] Institute, Vera (2021) *Vera-Institute/Incarceration-Trends: Incarceration trends dataset and Documentation*. Github. Retrieved at https://github.com/vera-institute/incarceration-trends.

[6] Rey, Arribas-Bel, Wolf (2020) *Geographic Data Science with Python*. Retrieved at https://geographicdata.science/book/intro.html.