# Random Walks and Rehab: Analyzing the Spread of the Opioid Crisis

Ellen Considine
Suyog Soti
Emily Webb
University of Colorado Boulder
Boulder, Colorado
USA
ellen.considine@colorado.edu

Advisor: Anne Dougherty

## Summary

We classify 69 types of opioid substances into four categories based on synthesis and availability. Plotting use rates of each category over time reveals that use of mild painkillers and natural alkaloids has stayed relatively constant over time, semi-synthetic drugs have declined slightly, and synthetic drugs such as fentanyl and heroin have increased dramatically. These findings align with reports from the CDC. We select 54 of 149 socioeconomic variables based on their variance inflation factor score (a common measure of multicollinearity) as well as on their relevance based on the public health literature.

To model the spread of the opioid crisis across Kentucky, Ohio, Pennsylvania, West Virginia, and Virginia, we develop two completely different models and then compare them.

Our first model is founded on common modeling approaches in epidemiology: SIR/SIS models and stochastic simulation. We design an algorithm that simulates a random walk between six discrete classes, each of which represents a different stage of the opioid crisis, using thresholds for opioid abuse prevalence and rate of change. We penalize transitions between certain classes differentially based on realistic expectations. Optimization of parameters and coefficients for the model is guided by an error function inspired by the global spatial autocorrelation statistic Moran's I. Testing our model via both error calculation and visual mapping illustrates high accuracy over many hundreds of trials. However, this model does not provide much insight into the influence of socioeconomic factors on opioid abuse

*The UMAP Journal* 40 (4) (2018) 353–380.

rates, because incorporating those factors does not significantly change the model results.

Our second model makes up for this deficiency. Running spatial regression models on socioeconomic predictors (including total drug use rate), we explore the spatial patterning of the opioid crisis as the result of a *spillover effect* and of spatially-correlated risk factors, using spatial lag, spatial error, and spatial Durbin models. While all models confirm significant spatial signals, the spatial Durbin model performs the best. We also calculate the direct, indirect, and total impacts of each predictor variable on opioid abuse rate. *Far and away, the most important variable in all models is the total drug use rate in each county.* The average result (across all seven years) is that a unit increase in total illicit drug use rate would raise the opioid abuse rate by 52%. This is quite realistic, given a CDC statistic that in 2014, 61% of drug overdose deaths involved some type of opioid. By contrast, an ordinary linear regression reports only a 37% increase in opioid abuse rate per unit increase in total drug use rate. Statistical measures such as the Akaike Information Criterion and Likelihood Ratio Test verify the superiority of our spatial models.

To predict possible origins of the opioid epidemic in each of the five states, we run a Monte Carlo simulation of our random walk model from 2000 to 2010. We map these counties and discuss their arrangement in the context of our other findings. The random walk finds that the opioid crisis most likely started in Montgomery, Kentucky, which conclusion aligns with research that opioid abuse is more prevalent in rural communities than in urban ones.

To forecast spread of the opioid crisis from 2017–2020, we use both the random walk and spatial regression models. The two models display surprisingly minimal deviance from each other, especially in 2019 and 2020. The random walk predicts that the number of counties above the illicit opioid use threshold will go down within the next seven years, which aligns with the idea that the opioid epidemic follows the spillover effect seen in infectious epidemiology.

Due to our assumption that the socioeconomic indicators change linearly, the second model's error significantly increases after about 4–5 years. The random walk, on the other hand, operates on a healthy tension between wanting to cluster together and randomly assigning classes. Near the initial date, it clusters more; but the randomness starts to compound rather quickly. For this reason, the random walk has lowest errors near the 4–7 year mark. This means that the best strategy to predict the future would be the spatial regression 1–3 years out, and the random walk for the 4–7 year range. Predicting anything beyond this point will have high error.

# Introduction and Problem Statement

The deadly consequences of abusing prescription narcotic pain-relief medications, heroin, and synthetic opioids are affecting people in all 50 states and across all socioeconomic classes. The opioid epidemic claims the lives of 115 people in the United States every day [Health Resources

& Services Administration 2019]. Through healthcare costs, rehabilitation treatment, lost productivity, and criminal justice involvement, the opioid crisis is costing the U.S. federal government an estimated $78.5 billion each year [National Institute on Drug Abuse 2019]. Our team is presented with the following modeling tasks:

- characterize the spread of the opioid epidemic throughout Kentucky, Ohio, Pennsylvania, Virginia, and West Virginia and analyze resulting patterns;
- incorporate socioeconomic factors in our model and analyze the associations, if any, between them and opioid abuse rates; and
- use results from these models to recommend public policy strategies to combat the opioid epidemic.

To perform these analyses, we are limited to data for 2010–2016 from the American Community Survey (ACS), which provides socioeconomic indicators, and from the National Forensic Laboratory Information System (NFLIS) on illicit drug use. All data are provided at the county level.

To characterize the spread of the opioid crisis throughout these five states, we develop two models:

- The first simulates a random walk through stable, endemic, and epidemic stages.
- The second is a standard collection of spatial regression models.

After describing these two modeling approaches and their results, we report forecasts from both models on the future spread of the opioid epidemic, and compare the results. This dual-pronged approach provides diverse insights into the nature of the opioid crisis, and helps us to identify strategies for government intervention.

# Etiology of the Opioid Crisis

## A Brief Timeline of the Epidemic

Opioids emerged into the non-cancer pain market following studies in the 1990s indicating that pain was inadequately treated. Pharmaceutical companies and medical societies were reassured (by somewhat erroneous studies) that opioids were not addictive [Rummans et al. 2018]. Thus, the **first wave** of opioid prescriptions began.

The **second wave** occurred around 2010 as the addiction began to surface. Government organizations placed limits on opioid prescriptions. Many of those already addicted turned to heroin instead, which was often impure or mixed with other drugs, leading to increased deaths [National Institute on Drug Abuse 2018].

The **third wave** came in 2013 with the rise of synthetic opioids, such as fentanyl.

### Risk Factors

Currently, age appears to have the largest impact on susceptibility to addiction. The younger a person, the more vulnerable to addiction: 74% of individuals admitted to treatment programs between 18 and 30 years of age had started abusing drugs before the age of 17. However, the majority of people admitted for heroin and prescription painkiller addictions started using drugs after the age of 25 [Substance Abuse and Mental Health Services Administration 2014].

## Common Epidemiological Models

Several types of models and overarching principles in public health research guide our approach to the given tasks.

### Compartmental Models

A *compartmental model* is commonly used to simplify mathematical modeling of infectious diseases. Populations are divided into compartments with assumptions about the nature of each compartment and the time rate of transfer between them. In the SIR model, the population is partitioned into three groups: susceptible (S), infected (I), and removed (R). S's are individuals who have not been infected, but who are susceptible to infection; I's are those who are infected and are capable of transmitting the disease; R's are people who can no longer contract the disease because they have recovered with immunity, been quarantined, or died.

### Stochastic Simulation

A.A. Brownlea incorporated a time element into modeling the spread of infectious hepatitis in Wollongong, Australia. He simulated random diffusion advancing as a ring from the origin of infection [Emch et al. 2017].

### Our Model

We combine the approaches of dynamic compartmental and stochastic simulation to model the spread of the opioid crisis.

Our needs diverge from the Greenwood model in that both our population size and the probability of "becoming infected" are dynamic, especially once we include socioeconomic factors as predictors—these predictors change over time. Thus, the foundation of our first model is a time-inhomogeneous random walk, where each county acts as an "agent."

# Foundations of the Models

## Terminology

- **Prevalence** is the fraction of people in a population who are sick with a disease at a particular point in time.
- An **epidemic** burdens a disproportionately large number of individuals within a population, region, or community at the same time.
- A disease that is constantly present in a given area is called **endemic**.
- A disease that occurs with intense transmission, exhibiting a high and continued incidence, is called **hyperendemic**.
- A phenomenon exhibits **spatial autocorrelation** if the presence of some factor in a sampling unit makes that factor's presence in neighboring sampling units more or less likely [Klinkenberg 2019].

## Assumptions

1. All counties for which we have no data have either a low abuse rate or no opioid abuse; all counties with illicit opioid cases are reported in our data set.
2. Three months is the minimum period of time in which a county can transition between distinct stages in the opioid epidemic.
3. Education level is a proxy for income and healthcare status, ancestry and language together indicate race, and veteran status can stand in for disability.
4. Linear extrapolation of socioeconomic indicators is acceptable within five years beyond the time span for which we have data (2010–2016).

## Overarching Concepts

A peril in using aggregated data to characterize social or ecological phenomena is the temptation of the *ecological fallacy*: asserting that associations identified at one scale of analysis are valid at either larger or smaller scales [Emch et al. 2017]. We incorporate some logic about individuals' opioid abuse into our model for county opioid rates, but do not claim that patterns at the county level represent the experiences of individuals.

A major question is whether a spatial autocorrelation of a health outcome is due to *spillover* (diffusion) or is simply explained by regions near each other having similar social, economic, and environmental characteristics that result in similar health outcomes. We explore spillover, spatially-correlated risk factors, and a combination of the two.

### Exploring the Drug Report Data

Tracking all 69 drugs tagged in the National Forensic Laboratory Information System (NFLIS) data would be burdensome and probably not produce useful results. So, to diagnose meaningful trends in the NFLIS data, we divide opioids into four categories based on chemical synthesis and availability:

- **Methadone, buprenorphine, etc.**: mild painkillers sometimes used to treat opioid addiction, easily available in clinics and are not as intensely regulated as other opioids.
- **Hydrocodone, oxycodone, etc.**: semi-synthetic opioids, among the most addictive and deadly drugs, contributing to the most overdose deaths from prescription opioids in 2017 [Centers for Disease Control and Prevention 2019]. Their semi-synthetic nature makes them more difficult to produce outside of a laboratory; they are likely abused as prescriptions or through overlooked distribution leaks.
- **Fentanyl, heroin, U-48800, etc.**: the largest category and the one with the highest increases in abuse and overdose deaths in recent years. It includes mainly synthetic opioids. Heroin is included in this category despite its semi-synthetic nature because it is illegal, and thus like other synthetic drugs, it can't be made entirely without the assistance of a professional laboratory.
- **Morphine, codeine, etc.**: natural alkaloids of the opium poppy. They are less potent than their semi-synthetic and synthetic cousins; codeine can be found in varieties of Tylenol painkillers.

Figure 1 shows that the prevalence of the opioids in each category (with the exception of synthetics) remained relatively constant over the six years for which we have data. Synthetic opioids became increasingly prevalent following 2011, in accord with reports from the Centers for Disease Control [Katz and Sanger-Katz 2018].

## Building a Model, Part I: Characteristics of the Epidemic

The discrete nature of both annual and county-aggregated data leads us to consider discrete time and discrete space (DTDS) Markov models. A stochastic process has the *Markov property* if the conditional probabilities of future states in the process depend only on the current state, not on past states. Our intuition is that, like an individual recovering from addiction, a county may regress if it does not make consistent efforts to address drug abuse. Thus, the current situation of a county directly influences its future situation.

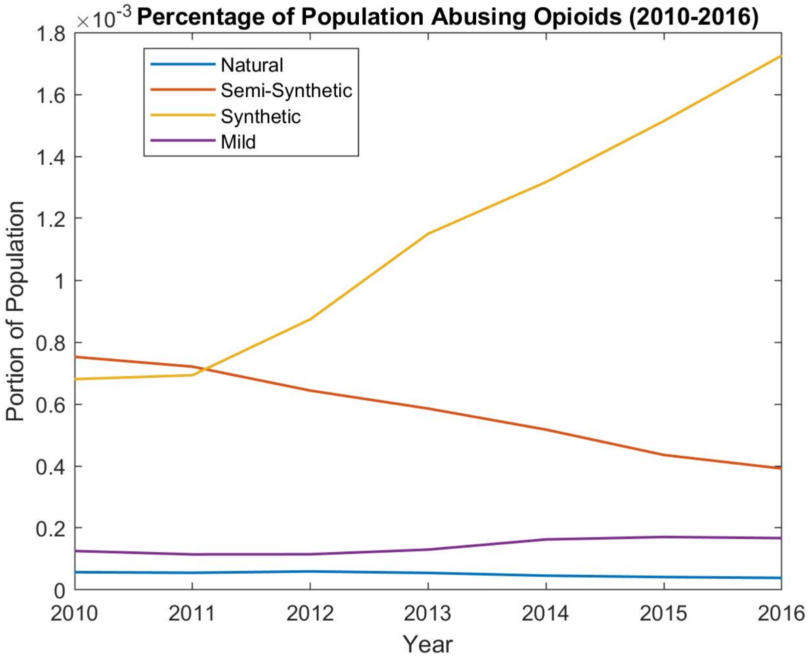

**Figure 1.** Drug trends for each category, 2010–2016.

## Discrete Classes for Prevalence and Rate of Change

Instead of working with the exact prevalence of opioid abuse in each county, we classify each county into "low" or "high" and "increasing" or "decreasing" opioid abuse prevalence.

We choose the median as the cutoff between "low" and "high" prevalence.

We classify all values greater than one standard deviation above the mean to be "increasing," and all values lower than one-half standard deviation below the mean to be "decreasing." Our reasoning for the difference is that the data are from years when the opioid crisis was in full swing, so "normal" in the data set is not the normal in general; and it is harder for a county to rehabilitate than to develop an opioid abuse problem, so smaller changes in the negative direction represent larger changes in the county. Values between these two cutoffs are regarded as "stable."

Counties for which we have no data we classify as "low stable."

## Categories for Our Data

In **Figure 2a**, the vertical line (red) represents the median for our data. In **Figure 2b**, the left (orange) and right (red) vertical lines represent the decreasing and increasing cutoffs. The horizontal axes have a scale of $10^{-3}$.

We have six categories: low and stable (LS), high and stable (HS), low and increasing (LI), high and increasing (HI), low and decreasing (LD), and high and decreasing (HD).

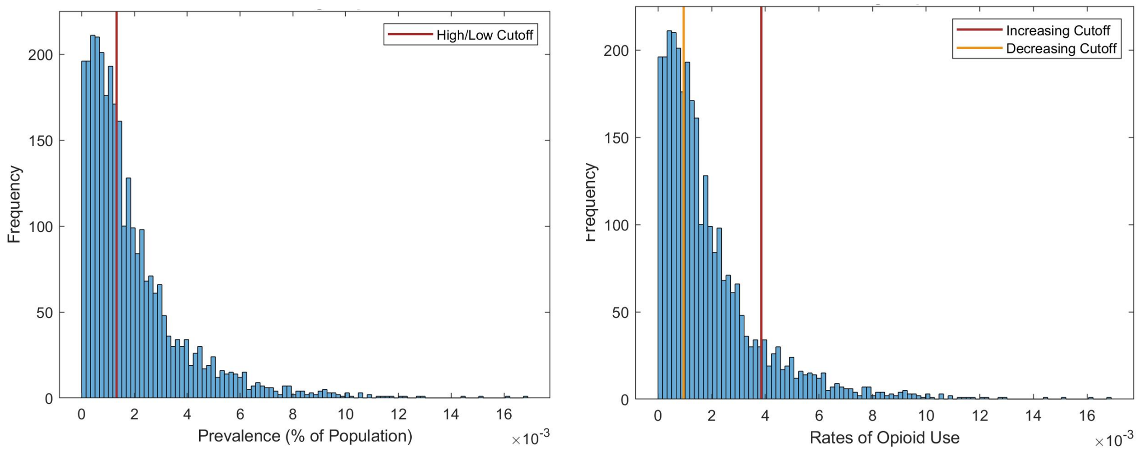

**Figure 2a.** Prevalence.

**Figure 2b.** Rates of change.

**Figure 2.** Cutoffs for prevalence and for decreasing/increasing opioid abuse.

The proportions of counties in each category stayed constant between 2010 and 2016, but the spatial distribution of the categories did not. In fact, the pattern reminded us of the description of Brownlea's hepatitis model as a ring-shaped clinical front expanding radially from an origin [Emch et al. 2017], in this case apparently southwestern West Virginia. The maps in **Figure 3** depict the prevalence of opioid abuse in each county in 2010 and in 2016. Note the "ring-like" expansion of the darker-red sections (which indicate regions of high and stable opioid abuse). The numbers in the legend correspond to the six classes as follows: 1 – Low Stable, 2 – High Stable, 3 – Low Decreasing, 4 – High Decreasing, 5 – Low Increasing, 6 – High Increasing.

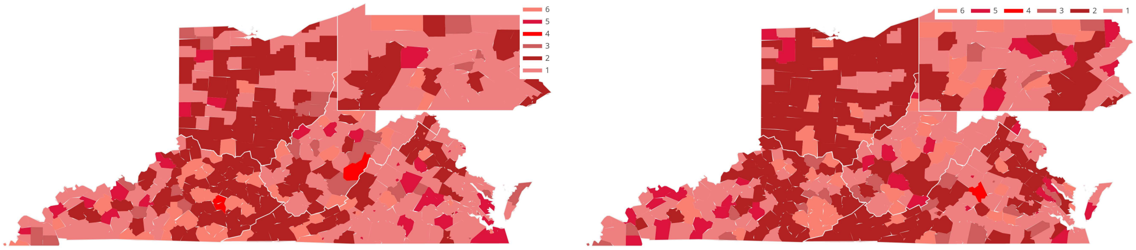

**Figure 3.** Spread of the opioid epidemic from 2010 (left) to 2016 (right).

## How the Model Works

The state (category) of a county at the next time step in our model is influenced by its current state, the states of its neighboring counties, a noise parameter, and random selection from a probability vector. Our Python code performs the following tasks:

1. Initialize the county probability vector, a 1 × 6 vector containing the probabilities that the county will be in the predetermined classes in the

next timestep. It is initialized by adding noise uniformly to each class such that the probability of a country transitioning from one class to any other is greater than zero. The level of noise added is optimized to minimize error (discussed below).

2. Each neighboring county adds a tally to the entry of the county's probability vector corresponding to the neighboring county's class. This tally incorporates inverse distance weighting. In other words, a county directly adjacent adds a tally of 1, a county with one county in between it and the origin county adds a tally of 1/4, etc.

3. Divide each entry in the probability vector by the relevant transition score in the penalty matrix below, which reflects our expectations about how difficult it should be to transition between classes. If our algorithm classifies a county as high stable, but the data say that it should be high increasing, we want the error to reflect the fact that our model was close and not so wrong as to have classified that county as having low prevalence. This need explains the penalty matrix.

$$\left[\begin{array}{c|cccccc} & LS & HS & LI & HI & LD & HD \\ \hline LS & 1 & 4 & 2 & 3 & 2 & 1 \\ HS & 4 & 1 & 3 & 2 & 3 & 2 \\ LI & 2 & 3 & 1 & 2 & 2 & 2 \\ HI & 3 & 2 & 2 & 1 & 4 & 2 \\ LD & 2 & 3 & 2 & 4 & 1 & 2 \\ HD & 3 & 2 & 2 & 2 & 2 & 1 \end{array}\right]$$

4. Scale the probability vector by a coefficient vector that was previously optimized to account for the overall importance of each class, in the same manner as the noise was optimized.

5. Normalize the probability vector so that it becomes a discrete probability distribution.

6. The final probability vector defines the distribution from which we sample to determine the county's class for the next time period.

7. Repeat the steps above for some number of timesteps and then for some number of trials so that each trial walks through the specified number of timesteps.

After running the Monte Carlo simulation detailing possible outcomes, we pick the most likely class of the county at the time that we are most interested in analyzing.

## Evaluation of the Model

Error analysis is tricky for this model because a good result does not necessarily mean a perfect match to the data.

The model results cluster a lot more than the real data do, as seen in **Figure 4**.

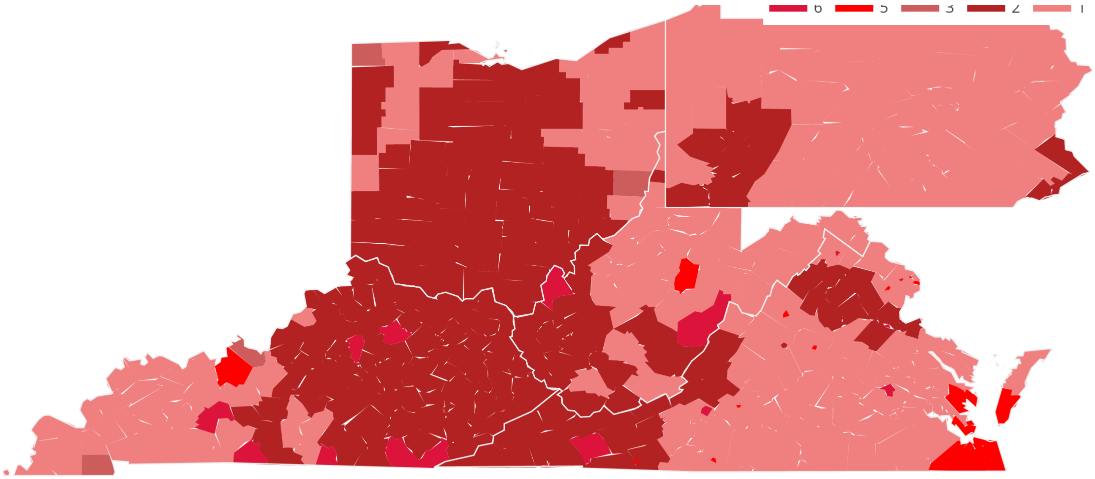

**Figure 4.** Model clustering in 2016.

Instead of trying to characterize clustering at each location, we devise an error function based on overall spatial autocorrelation. We look to Moran's I statistic [Emch et al. 2017] (a standard measure of spatial autocorrelation) for inspiration and present the following formula. For each class, for all counties in that class, our spatial autocorrelation measure is the mean proportion of neighbors with the same class. Then we take the sum over the classes of the difference between the simulated and real spatial autocorrelation measures.

$$A_{\text{class}} = \frac{1}{\text{number of counties of that class}} \times \sum^{\text{counties of that class}} \left( \frac{\#\text{neighbors with same class}}{\#\text{neighbors}} \right),$$

$$\text{Err}_{\text{total}} = \sum^{\text{classes}} \left| A_{\text{predicted}} - A_{\text{real}} \right|.$$

Using this error function to optimize levels of noise and other parameters in our model eradicates the clustering, leaving us with results akin to the those depicted in **Figure 5**. *Note: each time we run the model, we get slightly different results due to the random nature of added noise.*

## Model Testing

The histograms of the six categories stay more or less constant over time, depending on the levels of noise added to the initial probability distribution. Models with more noise tend to diverge at first, but by 2016 the noise compounds so much that simulations with noise have distributions that

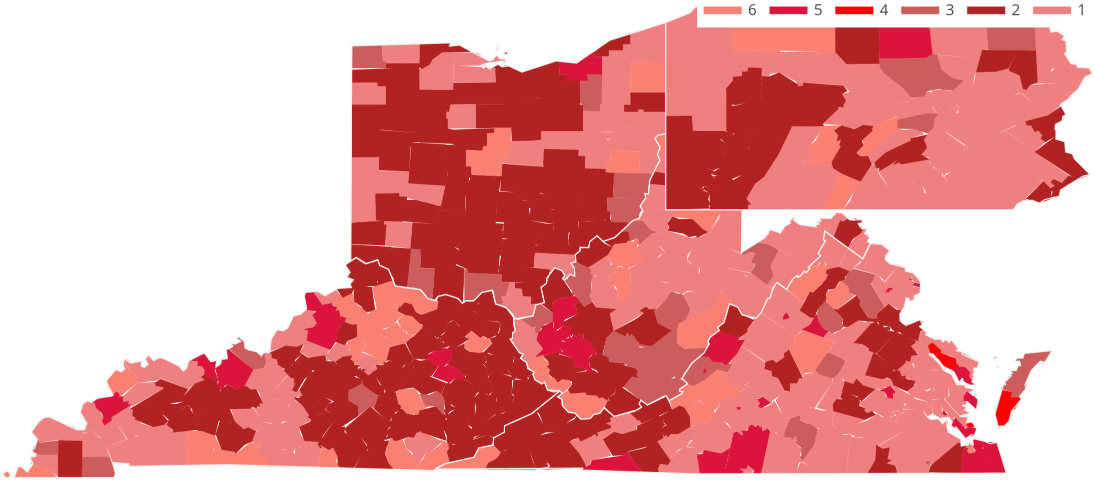

**Figure 5.** Best prediction from the model in 2016.

match the data more closely than simulations without noise. We find that the optimal level of noise is 0.3 (**Figure 6**).

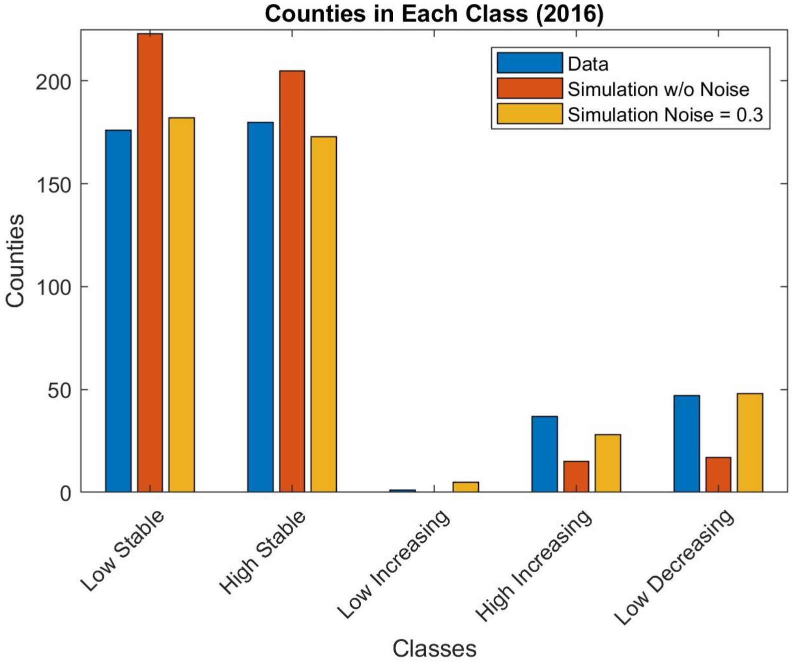

**Figure 6.** Category prevalences for data vs. simulations in 2016.

Because the random walk is on its own after being initialized by the 2010 data, there is no obvious reference point to deduce an optimal timestep size. Optimization of coefficients would fit the model to the data regardless of step size, so we choose a step size of three months (four steps per year) for ease of comparison to the annual data, and with the thought that three months is about the minimum time in which a county could switch category. The final optimization determines the coefficients on each category.

# Building a Model, Part II: Socioeconomic Factors

## Literature

The opioid epidemic is notable for its range across socioeconomic classes. **The three main risk factors appear to be race** [Berezow 2018; Substance Abuse and Mental Health Services Administration 2018; Barbieri 2018], **education** [Substance Abuse and Mental Health Services Administration 2018; Scommegna 2018], and **veteran status** [Recovery First 2019].

Other risk factors include gender (women, despite recent increases, are less likely to abuse opioids than men) [National Institute on Drug Abuse 2018], age (younger people are more at risk) [National Institute on Drug Abuse 2018], and population size (opioid abuse is more prevalent in rural communities than in urban settings) [Keyes et al. 2014]. Disability could be a possible risk factor because of the higher rates of opioid prescriptions associated with chronic pain, but few data on this topic are available.

## American Community Survey Indicators

The subset of the American Community Survey (ACS) data from 2010 to 2016 provided for this problem includes variables about household size and family structure, age and gender distribution, educational enrollment and attainment, veteran status, disability status, residential mobility, place of birth, language spoken at home, and ancestry. It does not include certain statistics that our research indicates would be useful, such as income, unemployment rate, healthcare coverage, and race.

For the sake of consistency across years, we remove variables for which one or more of the years have missing data. This process removes fertility statistics, disability status, citizenship status, world region of birth for foreign born, and several miscellaneous household/family structure variables—in total, 27 of the 149 socioeconomic factors. Because of the discrepancies between the data that we have and the data that we would want, we assume that

- education level is a proxy for income and healthcare status,
- ancestry and language indicate race, and
- veteran status stands in for disability.

Because there was so much information in the survey data regarding household size and family structure, we focus subsequent research on any ties between household/family structure and opioid addiction. While the epidemic affects families of all kinds, we notice an increase in grandparents raising their grandchildren in areas where the opioid epidemic has hit hardest [BAART Programs 2018]. This occurs as parents are separated from their children, both voluntarily and not (e.g., death), due to addiction.

## Model Optimization with Socioeconomic Status

To incorporate socioeconomic factors into our random walk model, we first run a random forest algorithm in `sklearn` [Pedregosa et al. 2011] that classifies each county into one of our six categories based on 23 socioeconomic predictors. Using the feature importance attribute in `sklearn`, we find that the 10 most important socioeconomic factors are total illicit drug use rate, total population, people born in the U.S., American ancestry, Irish ancestry, only English spoken at home, people with some college but no degree, high school graduation rate, Polish ancestry, and people with a graduate or professional degree. Unfortunately, these feature rankings are based on absolute magnitude, so they do not provide insight into the direction of influence on opioid abuse rates.

Our adapted model uses probabilities generated by this random forest classifier to initialize the probability vector for our random walk. The algorithm then proceeds as before.

After making this adaptation, we use our error function to compare the performance of the new model to the old. Once we optimize the coefficients, both models become more accurate, and their performances become comparable.

## Error and Sensitivity Analysis

The random walk model takes on high error immediately after the start of the simulation. With noise, the error peaks around 2013 and then begins a steep descent; without noise, the error continues on an upward path.

Our model supports a healthy tension between the clustering noted earlier and the randomness added later with the inclusion of noise. However, this balance takes about 24 timesteps (6 "years") to reach, which is why the error is higher toward the beginning of the simulation than at the end.

## Possible Origin Locations

To avoid adding bias to our model through reliance on extrapolated socioeconomic factors, we do origin identification analysis with the old model, which did not rely on socioeconomic factors. We run a Monte Carlo simulation to find the possible origin locations, starting the epidemic in each county in 2000, during the height of the first wave of the opioid crisis. The simulation then propagates forward in time until 2010, at which time we compare the results to the given data. The simulations with the lowest amount of dissimilarity to the 2010 data indicate the counties in which the epidemic likely began. **Figure 7** highlights the counties in which the epidemic could have started (purple) and the counties first to contract the epidemic in their respective states (blue: topmost shaded county in each of Pennsylvania, West Virginia, and Virginia, and lowest shaded county in Ohio; and gold: leftmost of three adjacent shaded counties in Kentucky). The most likely origin of the crisis according to the random walk model is Montgomery, Kentucky (in gold).

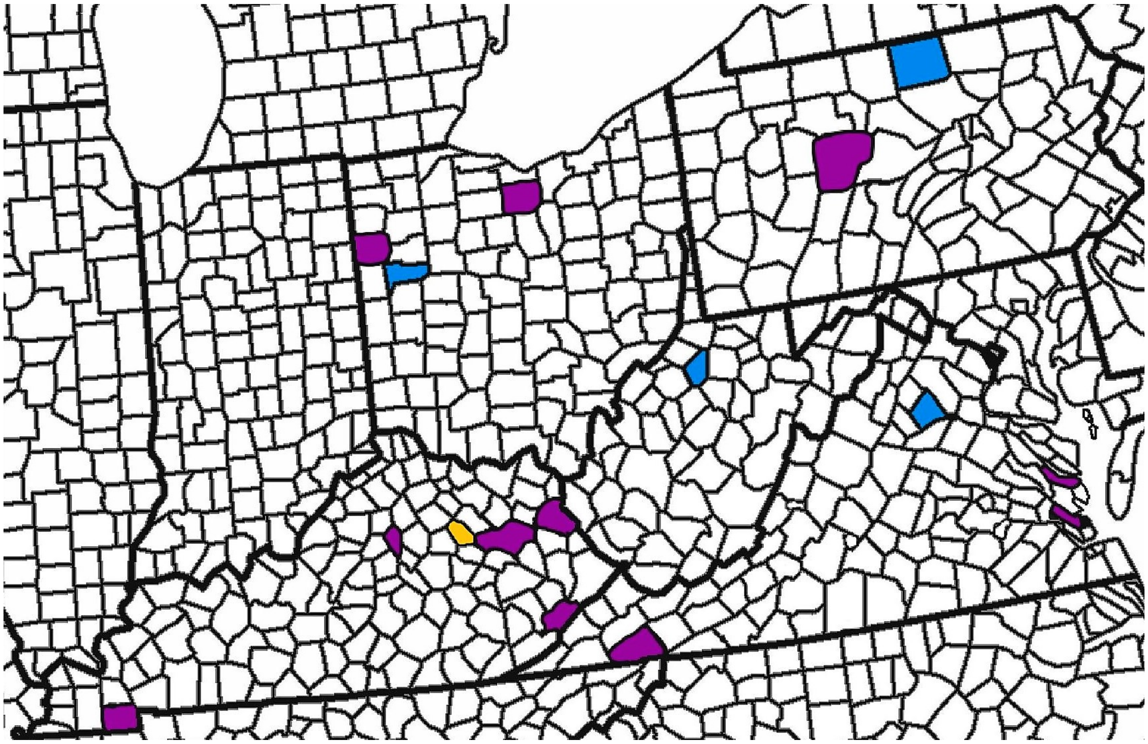

**Figure 7.** Possible origin locations of the epidemic.

The epidemic origin according to this result model is not far from southwestern West Virginia, which we hypothesized earlier as the origin.

# Second Approach: Spatial Regression

We explore the application of spatial regression models to the NFLIS and ACS data [Spielman 2015; Ver Hoef et al. 2017]. Taking into account spatial relationships is important because spatial correlation dramatically reduces the information contained in a sample of independent data, often by a factor of 2 [Waller and Gotway 2004].

Within spatial regression, there are three common types of models:

- **Spatial Autoregressive models (SARs)**, also known as spatial lag models, quantify the spatial dependence of the dependent variable $y$ among neighboring regions, the diffusion or "spillover" effect of $y$ [Sparks 2015]:

$$y = \rho W y + X\beta + u,$$

where $W$ represents the spatial weights (adjacency or distance between regions), $\rho$ is the spatial autoregressive coefficient, and the error $u$ is assumed to be classical (independent of $y$) [Viton 2010].

- **Spatial Error Models (SEMs)** quantify spatial dependence of the residuals. Instead of a spillover effect, they conceptualize spatial error as spatial correlation in one or more unidentified predictor variables:

  $$y = X\beta + u,$$

  $$u = \lambda W u + \nu,$$

  where, in addition to the variables defined above, $\lambda$ is the spatial autoregressive coefficient, and $\nu \sim N(0, \sigma^2)$ [Viton 2010].

- **Spatial Durbin Models (SDMs)** are a combination of spatial lag and spatial error models [Anselin 2003]. By lagging the predictor variables in the model using $W$, we can get a collection of spatial predictor variables in addition to the regular predictors:

  $$y = \rho W y + X\beta + W X\theta + \epsilon,$$

  where in addition to the variables defined above, $\theta$ is a vector of the regression coefficients for the lagged predictor variables and $\epsilon$ is the error [Sparks 2015]. It is also possible to include the lagged predictors in a spatial error model, resulting in a Durbin Error Model (DEM), but that model appears to be less common and we do not explore.

The question of model specification depends on whether we think that opioid abuse results from diffusion or is simply influenced by spatially-varying risk factors. Fortunately, we can rely on statistical tests as well as the combination of our research and intuition to select a model.

## Model Fitting

To perform statistical regression, we use the R package `spdep`. To quantify the spatial weights ($W$ in the model), we use a shapefile of all U.S. counties [United States Census Bureau 2017], subset it to include only the five states of interest, and convert it to a spatial weights object, illustrated in **Figure 8**.

We identify and remove highly-correlated variables, using the method of variance inflation factors (VIF). The formula for VIF is as follows:

$$VIF_k = \frac{1}{1 - R_k^2},$$

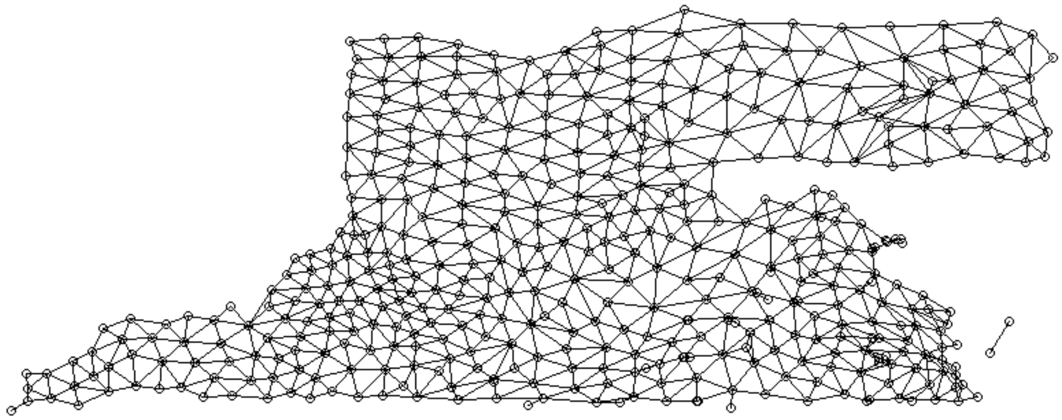

**Figure 8.** Visualization of spatial weights between counties.

where $R_k^2$ is the $R^2$-value obtained by regressing the $k$th predictor on the remaining predictors [Simon et al. 2018]. Thus, if a predictor is highly-correlated with the other predictors, its $R_k^2$ value will be large, the denominator of the VIF expression will be small, and the VIF score will be large. A common heuristic for addressing multicollinearity is to remove predictors with VIF scores over 10. In our data, the group of variables with VIF scores over 10 included many overlapping measures of household/family structure, including educational enrollment and attainment, residential mobility, place of birth, and language. Removing that group of variables left us with 54 predictors, mostly from the ACS but also including total illicit drug use rate (`TotalDrugReportsCounty` from NFLIS divided by total population from ACS), plus our dependent variable: illicit opioid use rate (NFLIS `DrugReports` divided by total population).

We run regressions for spatial lag, spatial error, and spatial Durbin models, using the functions `lagsarlm`, `errorsarlm`, and `lagsarlm` with `type = ‘‘mixed’’`, respectively.

## Statistical Results

Every model (across the seven years) confirms a highly-significant spatial signal in the data. The Likelihood Ratio Test (LRT) is a ratio of the likelihood functions from two nested models, one with more parameters than the other [White 2017]. In this case, the LRT compares spatial linear regression with regular linear regression. All $p$-values on Likelihood Ratio tests and asymptotic $t$-tests for the spatial autoregressive coefficients $\rho$ and $\lambda$ (for SAR and SEM respectively) were $\leq 2^{-3}$. This means that spatial regression gives us significantly more information than regular regression would have.

Regarding spatial model specification, the most direct comparison can be based on maximized log-likelihood [Anselin 2003]. Across all years, the spatial Durbin model had the highest log-likelihood. This makes sense, because we expect the spread of the opioid epidemic to display both spatial lag and spatial error patterning, based on both human interactions and socioeconomic and/or regulatory patterning.

Each SAR and Durbin model also performs a Lagrange multiplier (LM) test for residual autocorrelation. About half of the SAR and Durbin models showed significant residual autocorrelation ($p < 0.01$). This indicates that there may be other spatially autocorrelated variables that could improve our model, and/or our error term is heteroskedastic (has nonuniform variance) [Spielman 2015]. To test the latter, we perform a Breusch-Pagan (BP) test on each spatial Durbin model. All BP tests show highly-significant heteroskedasticity in the residuals. In spatial models, this is often due to spatial units having different population sizes [Spielman 2015].

Unlike the coefficients produced by ordinary linear regression, the coefficients from a spatial lag model do not facilitate interpretation, because a change in one predictor in one region influences response in other regions, which in turn influence the response in the region where the initial change occurred [Sparks 2015]. To account for both direct (local) and indirect (spillover) effects, we used the `impacts` function in `spdep` to calculate the global average impact of a unit increase in each predictor variable. Our results indicate that total illicit drug use rate in a county was by far the strongest predictor of opioid abuse levels in that county.

The Centers for Disease Control reported that in 2014, 61% of drug overdose deaths involved some type of opioid, including heroin [Rudd et al. 2016]. By 2017, this was 67% [Centers for Disease Control and Prevention 2019]. So our finding that a one-unit increase of total illicit drug use rate would raise the opioid abuse rate by 52% seems realistic. For comparison, an ordinary linear regression on the same predictor set gives $R^2 = .74$ but with the coefficient on total illicit drug use rate only .374. It is notable that our average direct impact rate was .372. This result confirms that our spatial model is far superior to a regular regression model in predicting opioid abuse, because it takes into account the indirect impacts of spatial diffusion. The Akaike Information Criterion (AIC), another output from `sarlm` models, formally quantifies this comparison for all the models.

Apart from the illicit drugs rate, none of the other variables has a comparable coefficient size, nor do they seem particularly helpful in informing policy to help address the opioid crisis.

## Sensitivity Analysis

To test sensitivity of our analysis to the socioeconomic variables selected, we use a smaller subset of 23 variables that emphasize educational attainment and ancestry, but also include percentages of grandparents responsible for grandchildren, civilian veterans, those who moved in the last year, birthplace, and speaking only English at home. The results are not significantly different than those from the larger dataset. The spatial Durbin models all performed the best. The average direct, indirect, and total impacts of a unit increase in total illicit drug use rate on opioid use rate were .377, .154, and .531 respectively, each of which are less than .01 different from the corresponding average impacts from the larger model. For the reduced dataset, the consensus of the spatial models is:

- *Education:* percentage with a bachelor's degree is all positively associated with illegal opioid use; percentage of high school graduates and percentage with some college but no degree was mostly negatively associated;
- *Family:* percentage of grandparents responsible for grandchildren is positively associated;
- *Language:* percentage with only English spoken at home is mostly positively associated; and
- *Ancestry:* percentage Dutch and Irish are all positively associated; percentage Norwegian, Ukrainian, and Hungarian are mostly negatively associated.

Note: "mostly" positive/negative indicates that there was at least a 5:2 majority in that category among the seven years' models.

## Future Trends and Comparison to First Model

The `predict.sarlm` function in R allows us to apply trends from our initial results to new data. We used it to predict illicit opioid use up to five years in the future for the spatial lag model. Because we are trying to forecast on the same predictor variables but different observations of the variables in the model (the extrapolated socioeconomic factors), we use the `trend` prediction type [Bivand n.d.].

After making predictions for opioid abuse prevalence, we run these numbers through our six-stage classifier, to make the results comparable to those from our random-walk model. **Figure 9** illustrates the deviation between the forecasts of our first model (random walk) and second model (spatial lag).

The spatial-regression model is heavily dependent on the accuracy of the socioeconomic factors. Because we assume that the SES indicators change linearly over time, the extrapolated indicators are most accurate nearest to the initial data, which means that the spatial regression model is most accurate in years not long after 2016. As noted earlier, the random-walk model has significant error for a few years right after the initial point. After a few years, our previous assumptions regarding the socioeconomic indicators make the spatial regression more and more inaccurate; the accuracy of the random-walk model increases as tension between randomness and clustering reaches a balance.

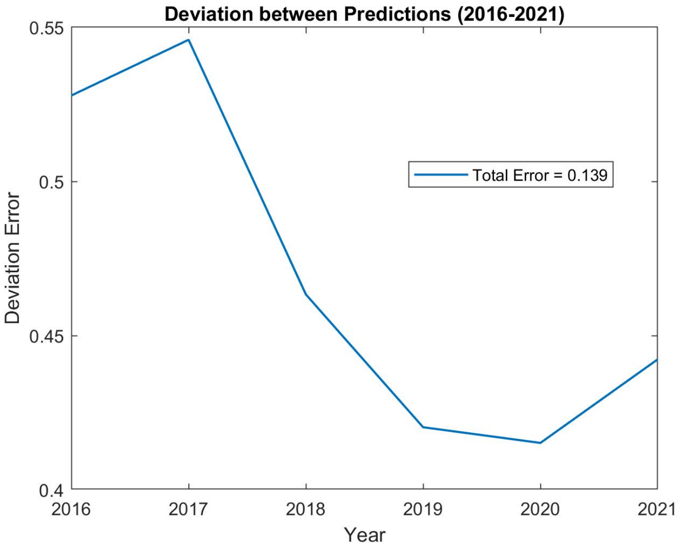

**Figure 9.** Error between random-walk and spatial-regression predictions.

## Drug Identification Thresholds

We classify a county as *problematic* if it is High and Stable, High and Increasing, or Low and Increasing for longer than 1.5 years (6 timesteps). **Figure 10** depicts the predicted number of problematic counties through 2028.

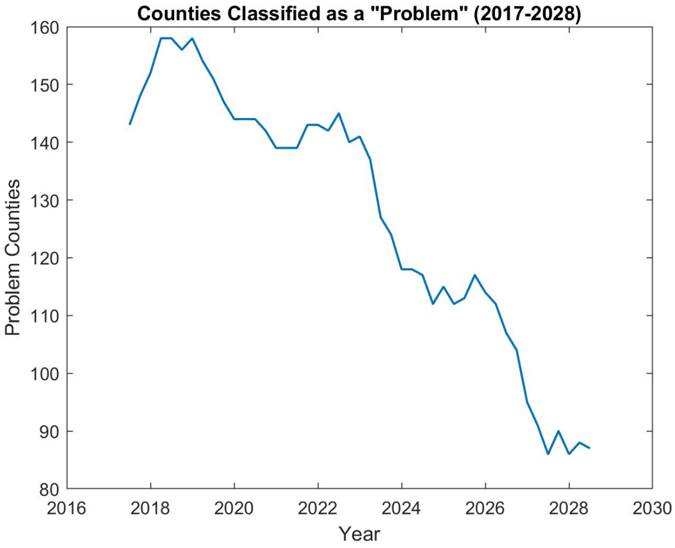

**Figure 10.** Counties classified as a problematic, 2017–2028.

Our model indicates that the number of problematic counties will decrease fairly steadily over the next 10 years. However, this may be due to the spillover effect: As the crisis diffuses to other areas of the country, the magnitude of the problem may decrease within our study region.

# Strategies to Combat the Opioid Epidemic

## Reducing Illicit Drug Use

Results from both our spatial regression and random forest models show that total illicit drug use rate is the most important predictor of opioid use rate in each county. The average results across the seven years of spatial regression models are that, all else constant, a 0.1% reduction in total illicit drug use would reduce total illicit opioid use by 0.05%, which is almost half of the median opioid abuse prevalence rate across counties in this region. Reducing total amount of drug use would reduce the amount of opioid use, because fewer people would progress through the drug hierarchy.

Recommended ways to address illicit drug use include investing in rehabilitation infrastructure [Chandler et al. 2009], reaching out and providing resources to communities willing to cooperate with police efforts [Moore and Kleiman 1989], and improving adolescent education about the addictive properties of illicit substances [Gerstein and Green 1993]. The results of our models with regard to ancestry and education indicate that aiming education initiatives at people of all races and socioeconomic statuses is necessary.

## Direct Opioid Interventions

In the early 1970s, the U.S. declared a “war on drugs,” increasing the size and presence of federal drug enforcement agencies while halting investigation into medical efficacy of these substances. Rates of criminal drug abuse have remained constant or increased in the decades since [Drug Policy Alliance 2018]. Launching another “war on drugs” will not address the opioid epidemic. Instead, we must treat it as a public health crisis, requiring science- and health-based solutions, rather than a combative approach. The first steps in this direction are the limitations on opioid prescriptions; the next steps in this process involve prevention, reducing overdose deaths, and improving addiction care in and beyond the criminal justice system [National Center on Addiction and Substance Abuse 2017].

- **Promote comprehensive public-awareness and education campaigns.** A series of campaigns should be addressed to those already abusing opioids to communicate urgent safety concerns, such as the dangers of blood-borne illnesses from sharing needles. Prevention measures directed at adolescents and young adults are particularly important because of their increased susceptibility to addiction [National Center on Addiction and Substance Abuse at Columbia University 2011]. Instruct educators in effective prevention and intervention strategies based in public health science rather than punishment that takes students away from resources, which can inadvertently reinforce the behavior. Use school websites to distribute information about opioid abuse to parents and children.

- **Reduce availability of and accessibility to opioids.** States should expand and encourage the use of Prescription Drug Monitoring Programs (PDMP) to track the prescribing and dispensing of controlled prescription drugs to identify suspected misuse, doctor shopping, or diversion. Programs should be put in place to inform medical professionals about safe prescribing practices for pain management and promote adherence to CDC prescribing guidelines. Around 61% of the drugs that are prescribed in the U.S. each year are not consumed [Kennedy-Hendricks et al. 2016]. Medicine take-back programs would decrease these drugs making it to the streets [National Center on Addiction and Substance Abuse 2017].

- **Curtail overdose deaths.** Get life-saving opioid overdose antidotes such as naloxone into the hands of first responders. Administration of the drug has already saved tens of thousands of lives [Wheeler et al. 2015]. In addition, expedite the distribution of new information about the crisis to stakeholders, including news about emerging synthetics and how to treat someone exposed to them.

- **Improve addiction treatment in and beyond the criminal justice system.** Increasing treatment capacity, expanding availability of medication-assisted treatment (MAT) programs, and extending insurance coverage for addiction treatment programs will help people outside of the criminal justice system get the help they need to overcome addiction. Providing adequate recovery support systems will encourage recovering addicts to remain clean. Improving treatment in the criminal justice system will decrease recidivism among addicts, three-quarters of whom are arrested for another crime within five years of release [Durose et al. 2014]. States are encouraged to educate law enforcement officers about addiction as a chronic health condition and to implement and support diversion programs such as Alternative to Incarceration (ATI), which gives individuals in the criminal justice system (CJS) greater access to treatment options. In addition, states should support treatment options for those in the CJS upon re-entry to civilian life.

# Limitations and Sources of Error

## Overall

- **Data restrictions:** A limitation of this report is the limitation of data to the ACS and NFLIS data provided.

- **Modifiable areal unit problem:** The MAUP describes discrepancies in spatial analyses performed at different scales arising because the scale at which one analyzes information or grouping schemes can produce different results [Emch et al. 2017]. Had we been provided with data at different scales (for instance, at the individual level), or aggregated at the census tract or ZIP-code levels, our investigation may have produced different results.

- **Adjacency vs. distance:** In both of our models, we incorporate spatial weighting based on county adjacency rather than on physical distance. To the models, it appears that all counties are equally spaced, despite some counties being dramatically different in size and shape. Changing the type of spatial weighting could alter our results [Waller and Gotway 2004].

- **Linear extrapolation of SES variables:** Forecasting opioid abuse rates for 2017–2021 requires extrapolation of socioeconomic indicators beyond the time frame of the given data. For simplicity, we assume a linear trend for each variable. In reality, these trends are likely to be more varied.

## Model I: The Random Walk

- **The Markov property:** The defining characteristic of a Markov process is future states' dependence only on the current state. We nominally assumed the Markov property to simulate change in each county's prevalence of opioid abuse. The "nominally" here applies because the selection of a county's class depends on the present class of both it and its neighbors, which in turn depend on the past state(s) of that county. In either case, making this semi-Markovian assumption may be unrealistic in the long term because longer-term changes, such as policy implementation, are very much indicative of past drug abuse issues in that region, and would drive a county to have more consistent drug-abuse prevalence in the long run. This assumption likely reduces some of the validity of our longer-term forecasts.

- **Linear interpolation of variables in Markov model:** Similarly to the extrapolation discussed above, linear interpolation of socioeconomic variables for each "three month" timestep is a convenient but likely inaccurate shortcut. However, the error introduced by the interpolation is much less significant than that introduced by the extrapolation, because we are dealing with timesteps between given data points, and because socioeconomic factors such as percentage of people with certain household structure or ancestry are unlikely to change drastically in the span of one year, whereas they may change noticeably over, say, five years.

- **Identifying potential origin locations:** For the year 2000, we initialized all counties to the Low Stable class. Given the randomness inherent in this model, however, there is a small likelihood of a Low Stable county jumping to a High Increasing, High Decreasing, or High Stable class in one timestep. This is somewhat unrealistic, and probably contributes to the error in our approximation of the 2010 spatial distribution.

### Model II: Spatial Regression

- **Prediction:** In forecasting opioid abuse rates for 2017–2021, we apply our spatial lag model from 2016. Although using the most recent model seems like the best choice, given that we had to run a separate spatial regression for each year, the 2016 model is not necessarily representative for 2010–2016. This means that our forecasts are almost certainly biased by the socioeconomic factor data from 2016.
- **Spatial model tests:** Spatial autoregressive and spatial error regressions are asymptotic, that is, they give approximately valid inference only for large number of regions [Waller and Gotway 2004]. We have data for approximately 460 counties in each year, which seems good enough.

## Conclusion

The opioid epidemic is a crisis of epic proportions that requires immediate attention from both the public and policymakers. We characterize the spread of this crisis in the Appalachian region of the U.S., where the epidemic has been most prevalent, using a random-walk model and a suite of spatial regression models. Our models perform well and complement each other. These models provide useful insights regarding socioeconomic variables associated with the opioid epidemic and future spatial dynamics in this region, which allow us to make informed recommendations for public-policy interventions.

# Memo to the Chief Administrator

**October 3, 2019**
**To: Uttam Dhillon — Acting Administrator, Drug Enforcement Agency**
**cc: Dr. Nora D. Volkow — Director, National Institute on Drug Abuse**
**From: MCM Team #1901679**
**Subject: Strategies to Combat the Opioid Epidemic**

---

The opioid epidemic claimed 47,000 lives in 2017 [Centers for Disease Control and Prevention 2019] and could take 500,000 more before 2027 if not addressed [Blau 2017]. Our team has analyzed the given data and conceived two models that accurately characterize and predict the spread of the opioid epidemic in Kentucky, Ohio, Pennsylvania, Virginia, and West Virginia. To better understand the crisis and any chances we have of diminishing its effects, we explored current research and investigated the impact of socioeconomic status on opioid abuse within our model.

## Results

We utilize two models to gain a more comprehensive picture of the opioid epidemic. Both characterize addiction as a contagion spread outward from an initial source. We determined that the likely point of origin in this region was Montgomery County, Kentucky, a rural area that does not have much treatment support. The problem grew and spilled over into neighboring counties. We predict that the epidemic will continue to spread across the nation, particularly affecting counties that already have a high incidence of illicit drug use. Slowing the spread of this epidemic is of the utmost importance and should be addressed immediately.

## Proposal

To mitigate the effects of the opioid epidemic and reduce the number of overdose deaths in the coming years, we propose the following policies:

1. Place restrictions on prescription of opioids for acute pain, without limiting access for those with disabilities or chronic severe pain. Educate medical professionals on safe prescribing practices and encourage adherence to the CDC's opioid prescription guidelines.

2. Implement targeted education and public-awareness campaigns. One set should be aimed at adolescents, who are among the most susceptible to addition, and executed through school systems and websites. Another should address those already misusing opioids, warning them of potential dangers of continued use and promoting treatment programs and clinics near them.

3. Distribute the opioid overdose antidote naloxone to first-responders and potentially to family and friends of those with addictions, and inform them on how to use it.

4. Extend insurance coverage and treatment capacity for those with addiction issues. Expanding the availability of medication-assisted treatment programs that utilize medications such as methadone and buprenorphine, especially in rural areas, will also allow struggling individuals to overcome their addiction.

5. Improve addiction treatment in the criminal justice system. This will decrease recidivism rates among addicts, who are more likely to get arrested and three-quarters of whom will be arrested for another crime within five years of their release [Durose et al. 2014]. Educate law enforcement officers about addiction as a chronic health condition, and support continued treatment upon return to civilian life.

# References

Alexander, Monica J., Mathew V. Kiang, and Magali Barbieri. 2018. Trends in black and white opioid mortality in the United States, 1979–2015. *Epidemiology* 29 (5): 707–715. journals.lww.com/epidem/Fulltext/2018/09000/Trends_in_Black_and_White_Opioid_Mortality_in_the.16.aspx.

Anselin, Luc. 2003. An introduction to spatial regression analysis in R. labs.bio.unc.edu/buckley/documents/anselinintrospatregres.pdf.

BAART Programs. 2018. Vermont's opioid addiction: A family crisis. baartprograms.com/vermonts-opioid-addiction-a-family-crisis/.

Berezow, Alex. 2018. White overdose deaths 50% higher than blacks, 167% higher than hispanics. acsh.org/news/2018/04/05/white-overdose-deaths-50-higher-blacks-167-higher-hispanics-12804.

Bivand, Roger. n.d. `predict.sarlm`: Prediction for spatial simultaneous autoregressive linear model objects. Documentation reproduced from package `spdep` version 0.8-1. https://www.rdocumentation.org/packages/spdep/versions/0.8-1/topics/predict.sarlm.

Blau, Max. 2017. Stat forecast: Opioids could kill nearly 500,000 Americans in the next decade. statnews.com/2017/06/27/opioid-deaths-forecast/.

Centers for Disease Control and Prevention. 2019. Overview of the drug overdose epidemic: Behind the numbers. cdc.gov/drugoverdose/data.

Chandler, Redonna K., Bennett W. Fletcher, and Nora D. Volkow. 2009. Treating drug abuse and addiction in the criminal justice system: Improving public health and safety. *Journal of the American Medical Association* 301 (2): 183–190. ncbi.nlm.nih.gov/pmc/articles/PMC2681083/.

Christensen, Ole F., and Paulo J. Ribeiro Jr. 2017. `geoRglm`: A package for generalised linear spatial models introductory session. cran.r-project.org/web/packages/geoRglm/vignettes/geoRglmintro.pdf.

Corbett, John. 2006. *Torsten Hägerstrand: Time geography.* is.muni.cz/el/1431/jaro2006/Z0147/time_geography.doc.

Drug Policy Alliance. 2018. A brief history of the drug war. drugpolicy.org/issues/brief-history-drug-war.

Durrett, Richard. 2016. *Essentials of Stochastic Processes*. 3rd ed. Switzerland: Springer International Publishing.

Durose, Matthew R., Alexia D. Cooper, and Howard N. Snyder. 2014. *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010*. bjs.gov/content/pub/pdf/rprts05p0510.pdf.

Emch, Michael, Elisabeth Dowling Root, and Margaret Carrel. 2017. *Health and Medical Geography*. 4th ed. New York: Guilford Press.

Gerstein, Dean R., and Lawrence W. Green (eds.) 1993. Preventing drug abuse: What do we know? Chapter 1 in *Preventing Drug Abuse*, edited by Dean R. Gerstein and Lawrence W. Green. Washington, DC: National Academies Press. ncbi.nlm.nih.gov/books/NBK234579/.

Health Resources & Services Administration. 2019. Opioid crisis. https://www.hrsa.gov/opioids.

Katz, Josh, and Margot Sanger-Katz. 2018. 'The numbers are so staggering.' Overdose deaths set a record last year. *New York Times* (29 November 2018). nytimes.com/interactive/2018/11/29/upshot/fentanyl-drug-overdose-deaths.html.

Kennedy-Hendricks, Alene, et al. 2016. Medication sharing, storage, and disposal practices for opioid medications among US adults. *Journal of the American Medical Association* 176 (7): 1027–1029. https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2527388/ild160028.pdf.

Keyes, Katherine M., et al. 2014. Understanding the rural-urban differences in nonmedical prescription opioid use and abuse in the United States. *American Journal of Public Health* 104 (2): 52–59. ncbi.nlm.nih.gov/pmc/articles/PMC3935688/.

Klinkenberg, Brian. 2019. Spatial autocorrelation. ibis.geog.ubc.ca/courses/geob479/notes/spatial_analysis/spatial_autocorrelation.htm.

Moore, Mark H., and Mark A.R. Kleiman. 1989. The police and drugs. ncjrs.gov/pdffiles1/nij/117447.pdf.

National Center on Addiction and Substance Abuse. 2017. Ending the opioid crisis: A practical guide for state policymakers. centeronaddiction.org/addiction-research/reports/ending-opioid-crisis-practical-guide-state-policymakers.

National Center on Addiction and Substance Abuse at Columbia University. 2011. Adolescent substance use: America's #1 public health problem. https://files.eric.ed.gov/fulltext/ED521379.pdf.

National Institute on Drug Abuse. 2018. Sex and gender differences in substance use. drugabuse.gov/publications/research-reports/substance-use-in-women/sex-gender-differences-in-substance-use.

National Institute on Drug Abuse. 2018. Understanding drug use and addiction. drugabuse.gov/publications/drugfacts/understanding-drug-use-addiction.

National Institute on Drug Abuse. 2018. Overdose death rates. drugabuse.gov/related-topics/trends-statistics/overdose-death-rates.

National Institute on Drug Abuse. 2018. Abuse of prescription (Rx) drugs affects young adults most. drugabuse.gov/related-topics/trends-statistics/infographics/abuse-prescription-rx-drugs-affects-young-adults-most.

National Institute on Drug Abuse. 2019. Opioid overdose crisis. drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis.

Pedregosa, Fabian, et al. 2011. *Scikit-learn*: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Recovery First. 2018. Why veterans are prone to opioid abuse? recoveryfirst.org/veterans/opioid-abuse/.

Rudd, Rose A., et al. 2016. Increases in drug and opioid overdose deaths — United States, 2000-2014. *Morbidity and Mortality Weekly Report (MMWR)* 64 (50): 1378–1382. cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm.

Rummans, Teresa A., M. Carolline Burton, and Nancy L. Dawson. 2018. How good intentions contributed to bad outcomes. *Mayo Clinic Proceedings* 93 (3): 344–350. https://doi.org/10.1016/j.mayocp.2017.12.020.

Scommegna, Paola. 2018. Opioid overdose epidemic hits hardest for the least educated. prb.org/people-and-places-hardest-hit-by-the-drug-overdose-epidemic.

Simon, Laura, Derek Young, and Iain Pardoe. 2018. Detecting multicollinearity using variance inflation factors. https://newonlinecourses.science.psu.edu/stat501/lesson/12/12.4.

Sparks, Corey S. 2015. Spatial regression models. rpubs.com/corey_sparks/109650.

Spielman, Seth. 2015. Spatial analysis and regression. sethspielman.org/courses/geog5023/lectures/Lecture_8_2015.pdf.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. 2014. *Age of substance use initiation among treatment admissions aged 18 to 30*. samhsa.gov/data/sites/default/files/WebFiles_TEDS_SR142_AgeatInit_07-10-14/TEDS-SR142-AgeatInit-2014.htm.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. 2017. *Results from the 2017 national survey on drug use and health: Detailed tables*. https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.pdf.

United States Census Bureau. 2017. Cartographic boundary — Shapefiles — County. cb_2018_us_cd116_20m.zip. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html.

Ver Hoef, Jay M., et al. 2017. Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs* 88 (1): 36–59.
https://eprints.qut.edu.au/115891/1/115891.pdf.

Viton, Philip A. 2010. Notes on spatial econometric models.
pdfs.semanticscholar.org/64ab/4ec3a6cb25cb191818c5d65400e6c3697082.pdf.

Waller, Lance A., and Carol A. Gotway. 2004. *Applied Spatial Statistics for Public Health Data*. New York: Wiley. onlinelibrary.wiley.com/doi/book/10.1002/0471662682.

Wheeler, Eliza, et al. 2015. Opioid overdose prevention programs providing naloxone to laypersons—United States, 2014. *Morbidity and Mortality Weekly Report* 64 (23): 631–635.
cdc.gov/mmwr/preview/mmwrhtml/mm6423a2.htm.

White, Gary C. 2017. Likelihood ratio tests. https://sites.warnercnr.colostate.edu/gwhite/wp-content/uploads/sites/73/2017/04/LikelihoodRatioTests.pdf.

# About the Authors

Team members Ellen Considine, Suyog Soti, and Emily Webb.