OPTIMALLY LOCATING OPIOID TREATMENT CENTERS IN UNDER SERVED AREAS USING R AND ALTERYX

Dan Putler and Ben Burkholder, Alteryx

alteryx | The Thrill of Solving

THE ROADMAP

- *Fast forward* to the final product, an app for locating opioid treatment facilities in underserved areas
- What the app requires and the challenges involved
- A three-step approach for developing census tract level estimates of the number of adults who abuse or dependent on opioids
- An evolutionary optimization algorithm for locating new treatment centers given the location of opioid abusers and existing treatment facilities

DEMO OF THE APP

WHAT DOES THE APP REQUIRE?

- A “demand surface,” which is the expected level of demand for a set of small geographic areas
  - The smaller the better since demand will be aggregated across areas touching or contained in a polygon
- The lat/lon coordinates of existing treatment facilities
- The coordinates of a set of potential new locations
- The number of desired new locations
  - The number in the set of potential new locations needs to be equal to or (preferably) greater than the desired number of new locations
- An optimization algorithm
  - This problem does not lend itself to traditional mathematical programming methods, so an approximate optimization method is needed

THE DEMAND SURFACE CHALLENGE

- Data on the number of individuals who abuse or are dependent on opioids in small (sub-county level) geographic areas is simply not collected
  - The closest available data is the number of individuals who died of a drug overdose at the county level, but it is only available for roughly half the counties in the US
  - The good news? The data needed to predict the number of opioid abusers can be predicted down to the census tract level is available
- The data available for making predictions:
  - Individual level, annual survey data from the National Survey on Drug Use and Health from the US Department of Health and Human Service
  - Census tract level demographic and socioeconomic data from the American Community Survey 5-Year Summary data
  - American Community Survey 5-Year PUMS data (a sample of individual and household records)

alteryx | The Thrill of Solving

PREDICTING OPIOID ABUSERS

- Maintained hypotheses of our approach
  - Socioeconomic and demographic factors can be used to the predict the probability that an adult is dependent on, or abuses opioids
  - Unobserved spatial effects do not overwhelm the observed socioeconomic and demographic effects
- The approach itself
  - Estimate an individual level, binomial probability model of opioid abuse/dependence among adults age 18 and older using demographic and socioeconomic categories as predictors
  - For each census tract in the 50 states and the District of Columbia, estimate the joint distribution of the demographic and socioeconomic factors using iterative proportional fitting to estimate the number of adults in each joint demographic and socioeconomic category in a census tract
  - Multiply the probability that an individual in a given demographic and socioeconomic category will abuse or be dependent upon opioids, and the sum these values across all groups

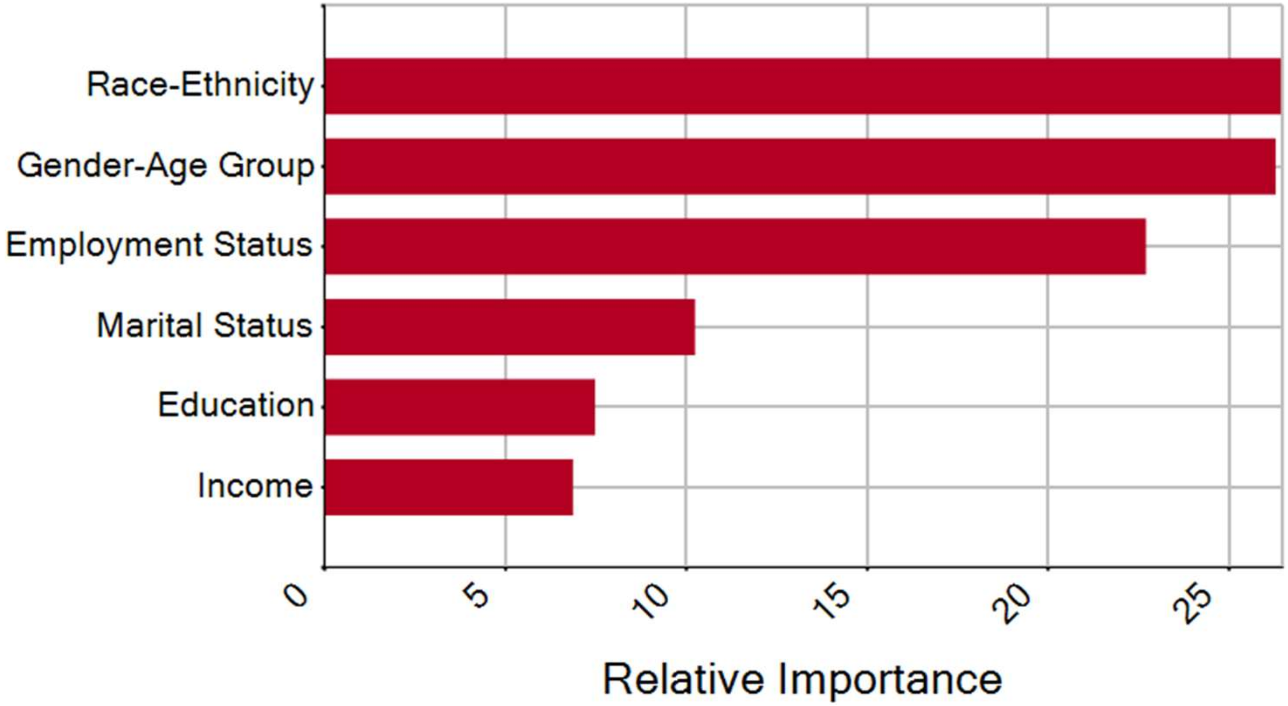
OPIOID ABUSE/DEPENDENCE MODEL

- The National Survey on Drug Use and Health
  - The 2016 survey (the most recent available) was used for this analysis
  - Special methods are used in the survey to elicit sensitive information from respondents, who are paid an incentive of $30 for participating
  - The 2016 survey had a reported 67,942 respondents age 12 years and above, but only 56,897 respondents are included in the public use file that was used in this analysis
- Data preparation
  - Appropriate data for projecting those who abuse or are dependent on opioids between the ages of 12 and 17 years is not available in the American Community Survey summary data, so respondents in this age range were removed (14,272 records), as were respondents who did not report a work status (436 records), leaving 42,189 respondents for model development
  - The sample is highly unbalanced with just over 1% of the sample abusing or dependent on opioids
  - A number of socioeconomic and demographic variables were recoded in order to conform with data reported in the American Community Survey summary data

OPIOID ABUSE/DEPENDENCE MODEL

- A traditional train/test methodology was used to develop a model
  - 65% of the records were in the training set, 35% in the test set
- Modeling algorithms considered
  - Gradient boosted models (gbm package)
  - Random forest models (randomForest package)
  - Neural network models (nnet package)
  - Binomial GLM using a complimentary log-log link function (stats package)
- The model with the "best" test sample performance based on the F1 (0.9944) and AUC (0.7688) statistics is a gradient boosted model with four-way interactions
- The "best" model was re-estimated using all of the records

OPIOID ABUSE/DEPENDENCE MODEL



alteryx | The Thrill of Solving

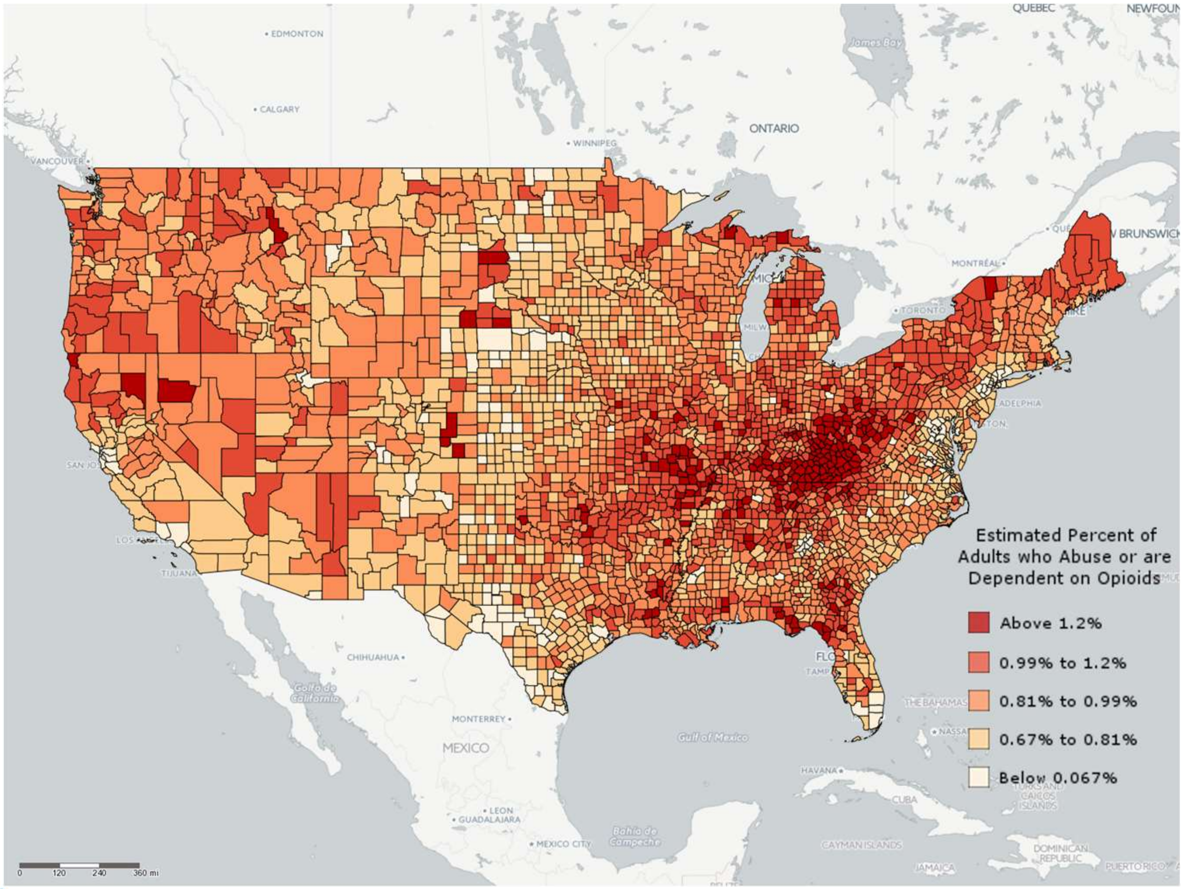
ITERATIVE PROPORTIONAL FITTING

- Iterative proportional fitting (or IPF) is an algorithm that has been used since at least the late 1930s to estimate the interior cells of a contingency table given a set of "seed" cells and a new set of marginal distributions so that the newly estimated interior cells produce the new set of marginal distributions
  - If the seed cells are all set to a constant value (typically 1) then the resulting dimensions are independent of one another
  - If the seed cells are informative, then the estimated cells will display the same pattern of interactions between dimensions as the seed cells
- The algorithm starts by scaling the cells along one dimension to match the marginal distribution of that dimension and then iterates over the dimensions of the table using the same scaling approach until the changes in the estimated interior cells are below a specified threshold

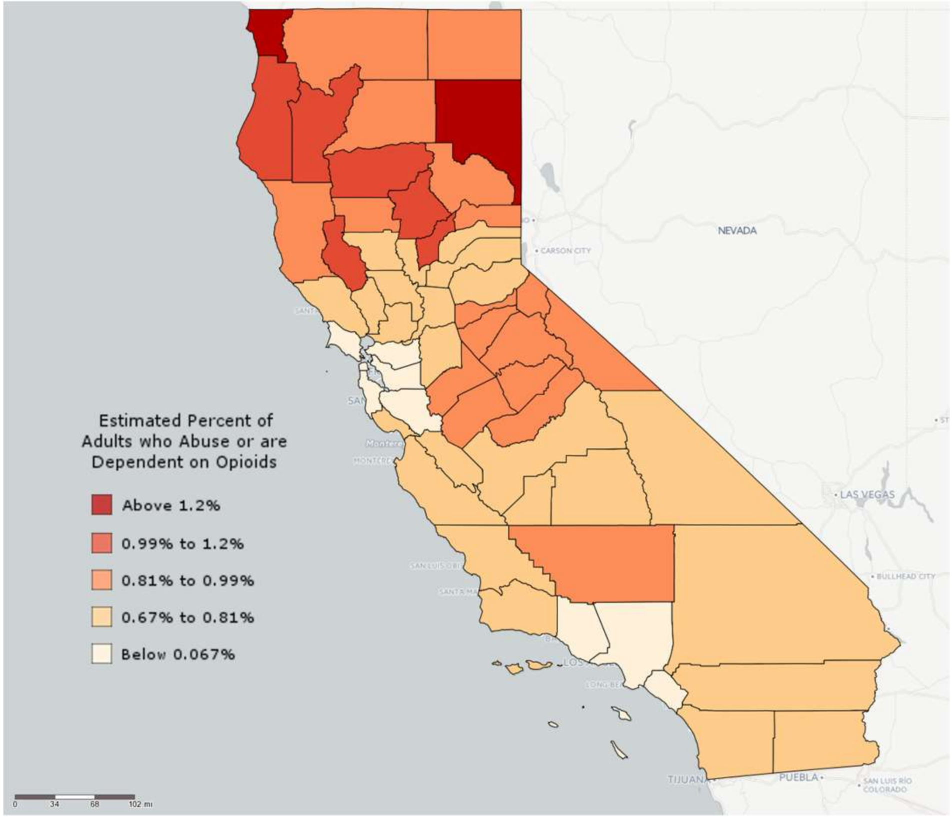
IMPLEMENTING IPF FOR THE APP

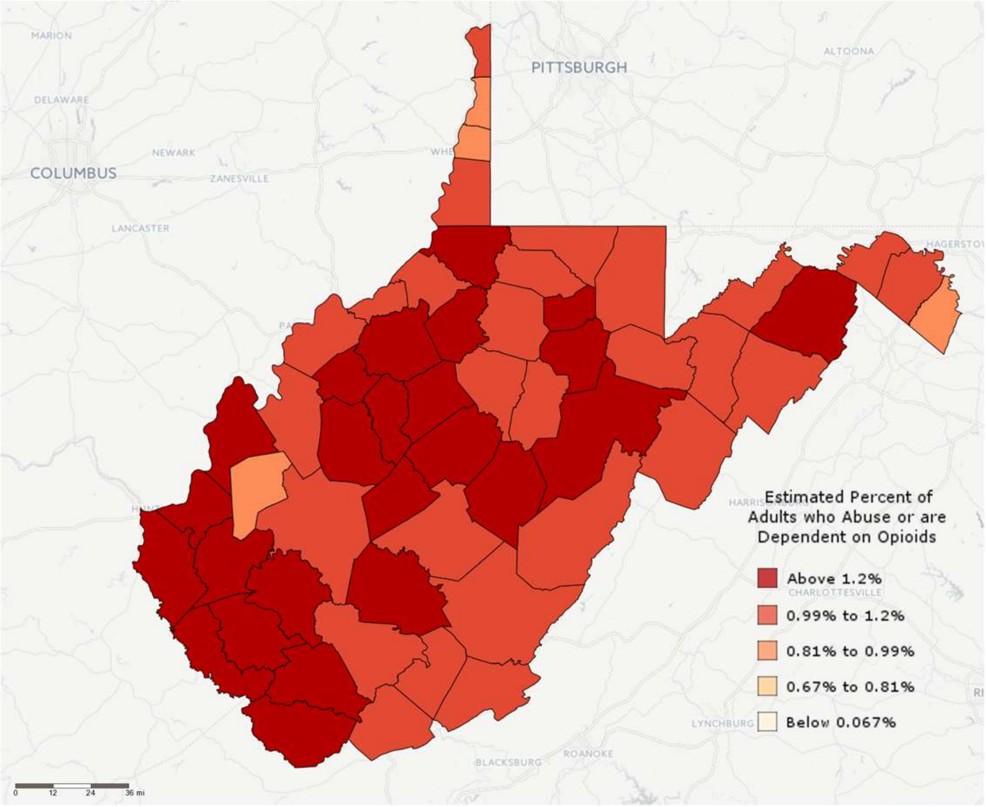
- The census tract level marginal distributions come from the summary data of the American Community Survey 2012-2016 Five Year Estimates for each of the 72,240 census tracts in the US
- A single informative seed table is constructed from the Public Usage Microdata Sample of the American Community Survey 2012-2016 Five Year Estimates, which is a subset of responses (15.7 M) from the full survey, with identifying information removed
- For each census tract this creates an estimated contingency table with 26,880 cells, with each cell representing the expected number of adults with a specific set of socioeconomic and demographic characteristics
- The mipfp package implements the algorithm in R, but a custom function was used in this application

PREDICTED OPIOID ABUSE RATES



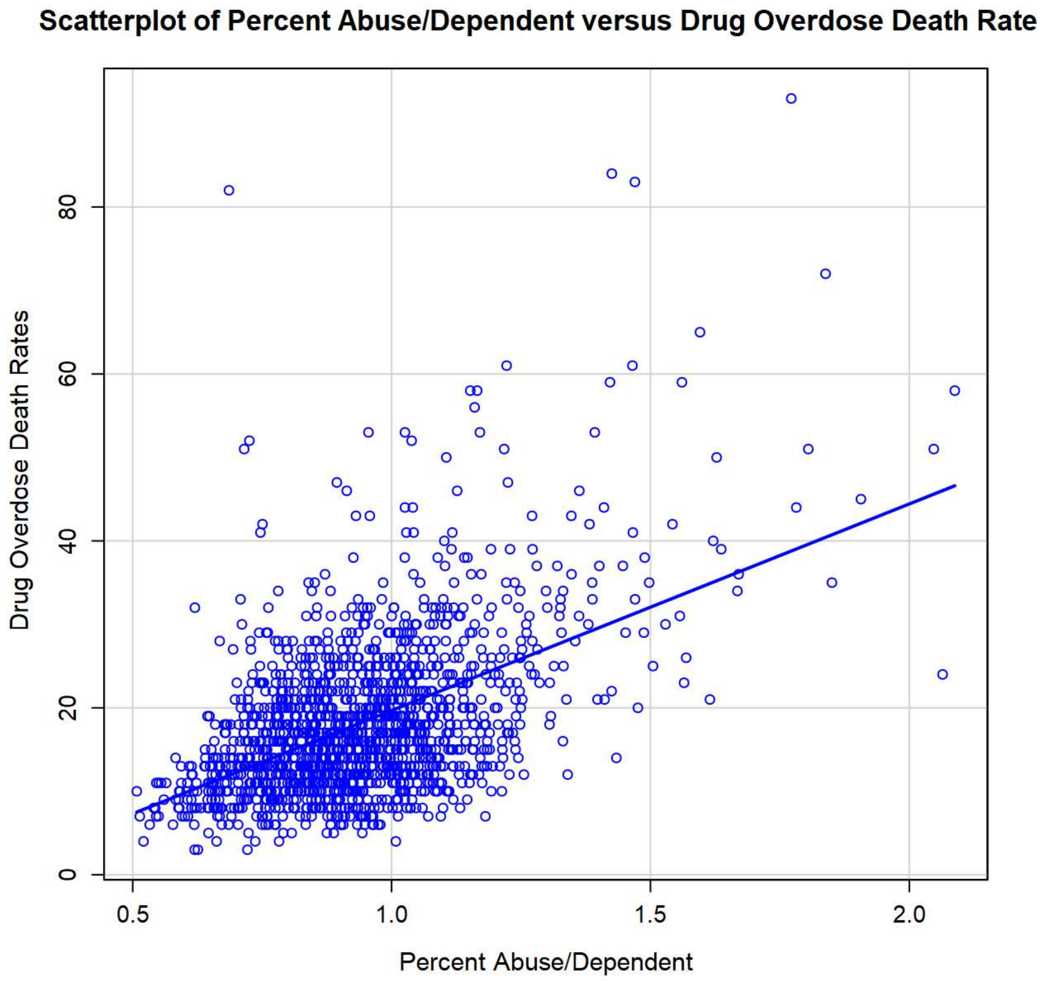
PREDICTED OPIOID ABUSE RATES





THE APPROACH'S VALIDITY

- We are using the three-step approach since the data we want is not available, so a direct examination of the approaches validity is not possible
- However, data is available on drug overdose death rates for just over half of the counties in the US is available from countyhealthrankings.org, and opioid use is the leading cause of overdose deaths
  - If the approach is valid, then the estimated percentage of adults who abuse or are dependent opioids should be strongly correlated with the drug overdose death rates
- There are some issues with drug overdose death rates that influence this analysis
  - Inconsistencies in reporting deaths across jurisdictions
  - Drugs other than opioids can lead to overdose deaths
  - There are spatial patterns in both the types of heroin available and the presence of fentanyl
  - Non-resident overdose deaths in a county



Pearson correlation coefficient: 0.51

THE OPTIMIZATION PROBLEM

- **Objective function**: Locate a desired number of treatment facilities such that the set of locations maximizes the expected number of adults who abuse or are dependent on opioids that are within a 10 mile radius of the new sites, and who are currently further than 5 miles from an existing treatment facility
  - Allowing for radius sizes other than 5/10 miles can easily be implemented, and drive time polygon could be used instead
- To avoid having two new sites locate adjacent to one another, a penalty function is used to discount the expected number of opioid abusers served by each facility
- The list of possible locations are taken to be set of census tract centroids, but a list of actual addresses could be used instead
- The optimization is implemented using an approximate evolutionary algorithm via a "location optimizer" macro within Alteryx

POTENTIAL IMPROVEMENTS

- Developing measures of uncertainty via simulation
- Investigate the benefits of newer methods for estimating the joint distribution of the relevant socioeconomic and demographic predictors
- Improvements in the “distance decay” component of the optimization model to address proximity to existing treatment facilities