Conclusion Validity (Plan 6)

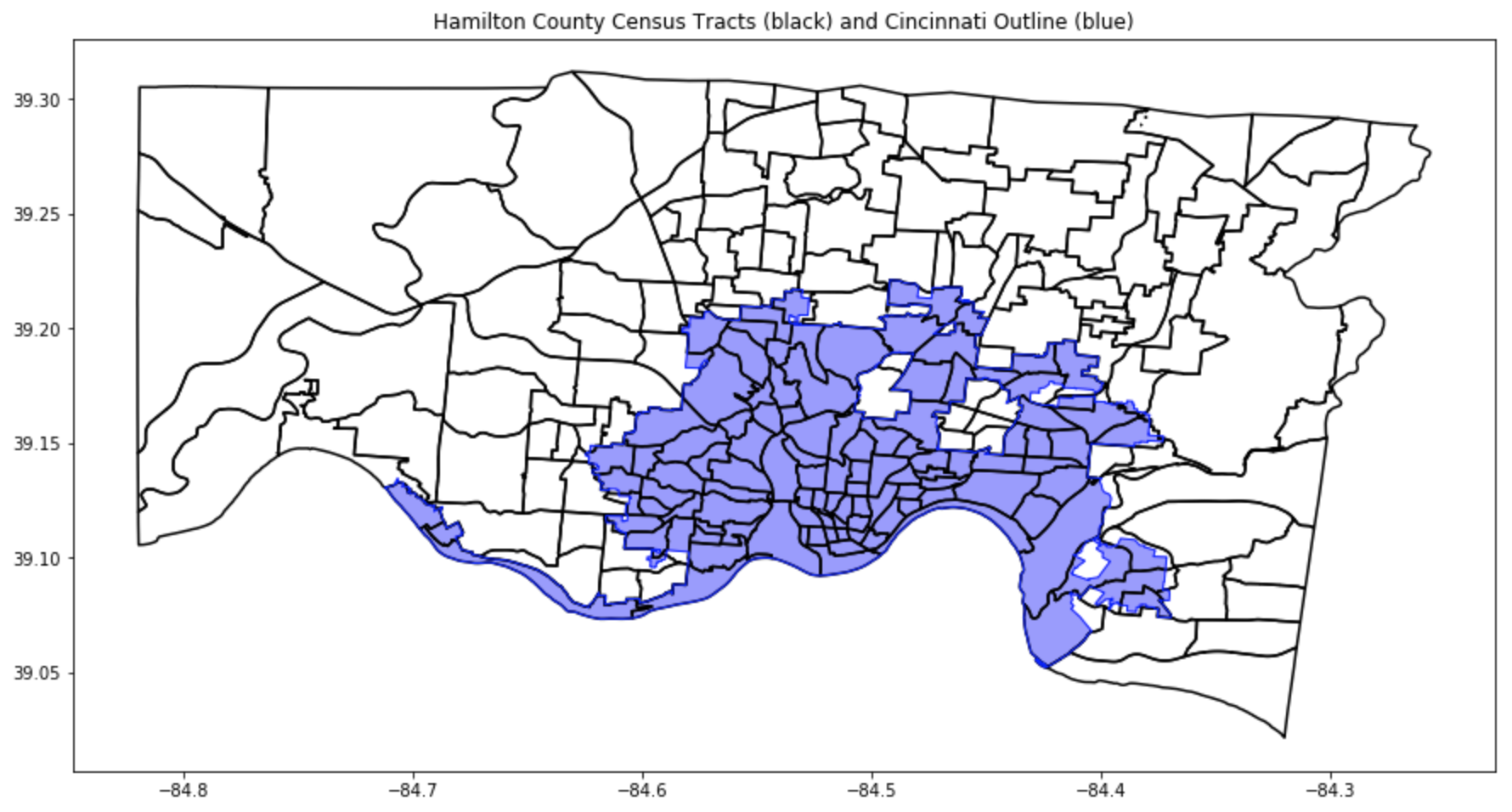
EMSE 6577

Drew Gobbi, Olivia Wang, Olatunji Akinbule, Maddie Warndorf

1. **Describe in detail the independent variables, dependent variables, and the relationship that you expect to find between them in your study.**

We select independent variables that represent the location of public health interventions across cities that could have an impact on the number of drug users in a given area. These include the number of medical facilities, pharmacies, drug dropoff locations, and naloxone distribution centers within two miles of each spatial unit in Tempe, Arizona and Cincinnati, Ohio.

We use census tracts as our spatial unit for Tempe and Statistical Neighborhood Approximations (SNA) for the spatial unit for Cincinnati. We used the SNA instead of the census tracts in this instance because the tracts do not overlay well with the border of the city of Cincinnati. Please see Figure 1.1 for reference. The white area is Hamilton County census tracts (black outlines) and the purple area is the union of the Cincinnati city and the census tracts. It shows that some census tracts within Cincinnati that also represent areas outside of Cincinnati.

Fig 1.1

We are planning to test two dependent variables in this study. First, we examine the relationship between our spatial predictors and the count of opioid-related EMS calls recorded per spatial unit (e.g. Census Tract or SNA) per month per year. Similarly, we also examine the relationship between the same spatial predictors and the population normalized mass load (PNML) of opioids in the wastewater of Tempe, Arizona. Because this unit is sampled directly by researchers at Arizona State University, we rely on the site collection areas for our spatial unit of analysis in the PNML regression. The collection areas used by ASU are smaller than a typical census tract in Tempe and the data is sampled daily. Likewise, because of the inter-day volatility in this measurement, we use the average PNML per site collection area per month per year as our dependent variable in this experiment. Both dependent variables – average monthly PNML and count of opioid-related EMS calls – represent indirect and flawed measurements of the number of drug users in our cities of interest.

We expect the spatial predictors we have included to have varying degrees of impact on our constructs of opioid use both in terms of magnitude and direction. For example, we expect to the number of medical facilities in an area to be negatively correlated with drug use because treatment is more readily accessible. However, because of the way our constructs are measured, the actual magnitude and sign of this relationship may not be as well represented. EMS calls could be negatively correlated with the number of medical facilities in an area if one believes that they act as depressants on overall drug use and represent treatment centers within an area. These same issues occur when hypothesizing the relationship between the medical facilities and the PNML measures. Medical waste that is dumped at a higher level near these areas may actually show a positive association with the PNML of the opioids in wastewater in an area, even though we expect these facilities to represent treatment.

Likewise, while we also expect naloxone distribution centers and drug dropoff centers to represent effective treatments that lower the amount of drug use in an area over time, there could be an association problem in our data. Areas with greater numbers of naloxone distribution centers or drug dropoffs may have already had a high number of drug users in them to begin with, which led city officials to install these interventions at a greater rate. For the purposes of this analysis, we assume these types of treatments are effective and expect our model to detect the hypothesized long term effects of lowering drug use as measured by both EMS calls and PNML measurements in an area in a point in time. If this assumption proves to not hold, we may have to turn to intervention or time-series based (or both) analyses of when these interventions were introduced to understand this relationship.

Finally, it is also difficult to guess how the number of pharmacies in an area will affect EMS calls. Because pharmacies may also produce medical waste, we believe they could have a positive effect on the PNML measurements of opioids in the area’s wastewater. The relationship is less clear for EMS calls. Pharmacies do act as treatment centers for opioids by administering naloxone and providing drug drop locations (depending on the city or state). But, pharmacies also prescribe and release more opioids into a given area to patients. Having more drugs in an area can be correlated with higher amounts of drug use.

1. **What statistical or machine learning analysis do you plan to conduct to determine whether this relationship exists or not? Justify this analysis relative to the types of variables (interval, ordinal, nominal, etc.) you are relating.**

We will conduct two types of analyses (one for the EMS calls and one for the PNML measurements) to accommodate the distribution of our dependent variables. Our analyses will attempt to determine the effect of the spatial predictors listed above on each drug use measure. All independent variables in this study are ratio measures, with the exception of the EMS calls, which we transform into a binary measure, as discussed below.

The count of EMS calls is power law distributed per census tract per year per month. This makes sense intuitively, since we can expect only a few calls a month, with long tails of high a number of calls happening in certain extreme cases. As we will discuss below, we transform this variable into a binary classifier for a typical number of calls in an area at a point in time and an atypical high number of calls (i.e. above average per the dataset). As such, we use the following logistic regression model:



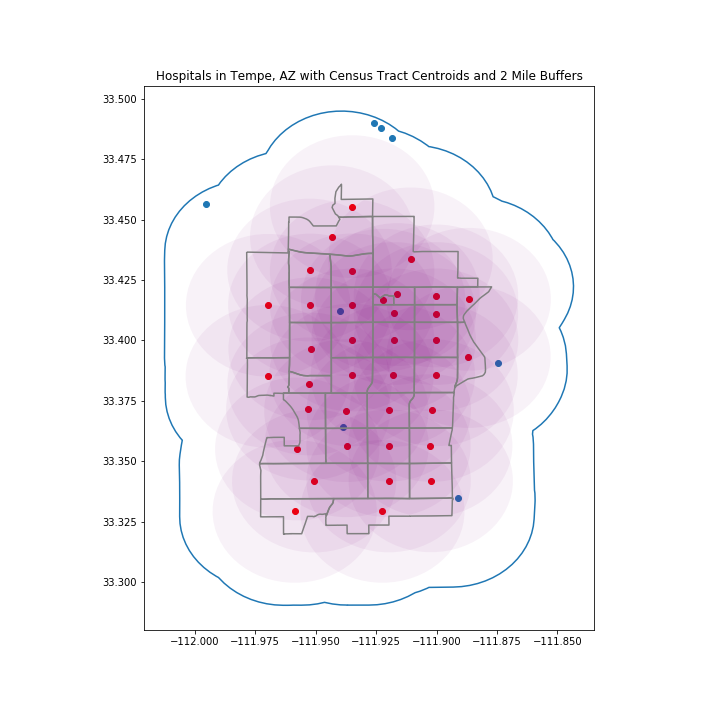
Secondly, we conduct a regression of the count of medical facilities, pharmacies, drug dropoff, and naloxone distribution variables in a given area and at a given point in time with the PNML measures. We add fixed effects to this model to account for the unobserved heterogeneity between site collection areas and points in time. These fixed effects are equal to one if a PNML measure occurs in time t or area j and 0 otherwise. Accordingly, our model follows the form:

.

1. **Generate and clean pilot data reflecting each of the variables identified above. Display descriptive statistics of these data, including histograms, measures of central tendency and dispersion.**

Cleaning of MedFacilitiesCount2mi, PharmCount2mi, DrugDropCount2mi, and NaloxoneDistribCount2mi:

* All the variables listed above were cleaned by:
  + Removing all facilities that were not within the 2 mile radius of the city.
    - For the Cincinnati and DC counts, we are still finding medical facilities for Kentucky, Maryland, and Virginia. Therefore these counts are not final.
  + A 2 mile radius (buffer) was created from each census tract/SNA centroid. An example is shown below.



* + From there, each facility that was in the 2 mile radius of the centroid was counted.

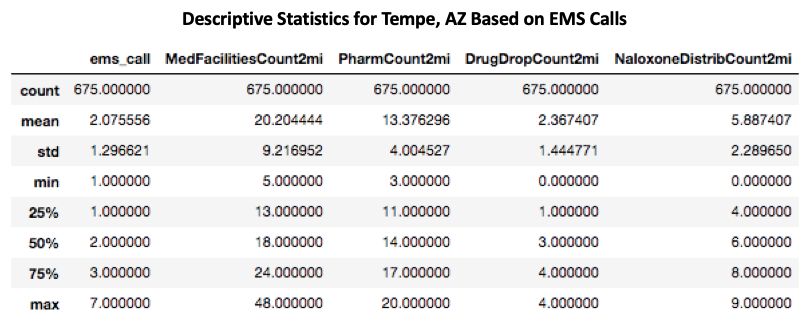
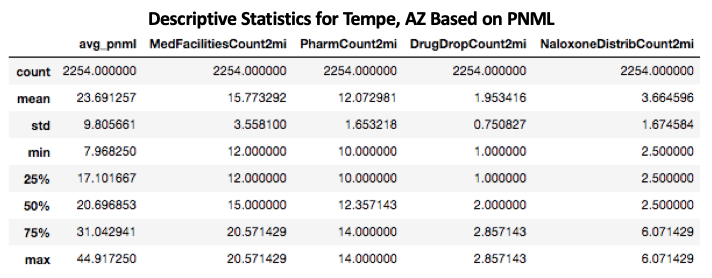
Cleaning of EMS Calls:

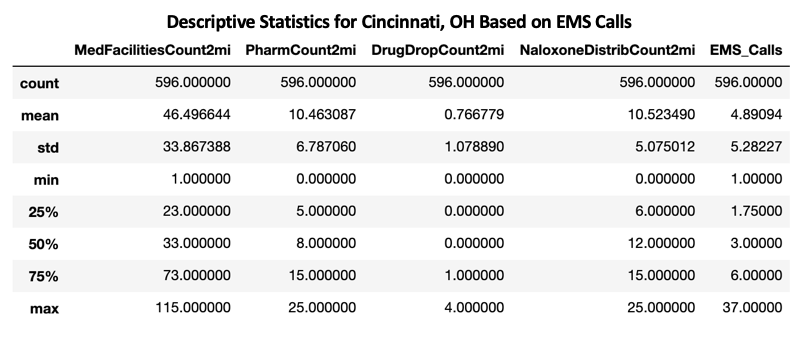
* The count of EMS Calls was cleaned by aggregating all of the call points by month year for each spatial unit (either census tract or SNA). For both Tempe and Cincinnati the dates range from 05/2017 to 08/2018.

Cleaning of PNML:

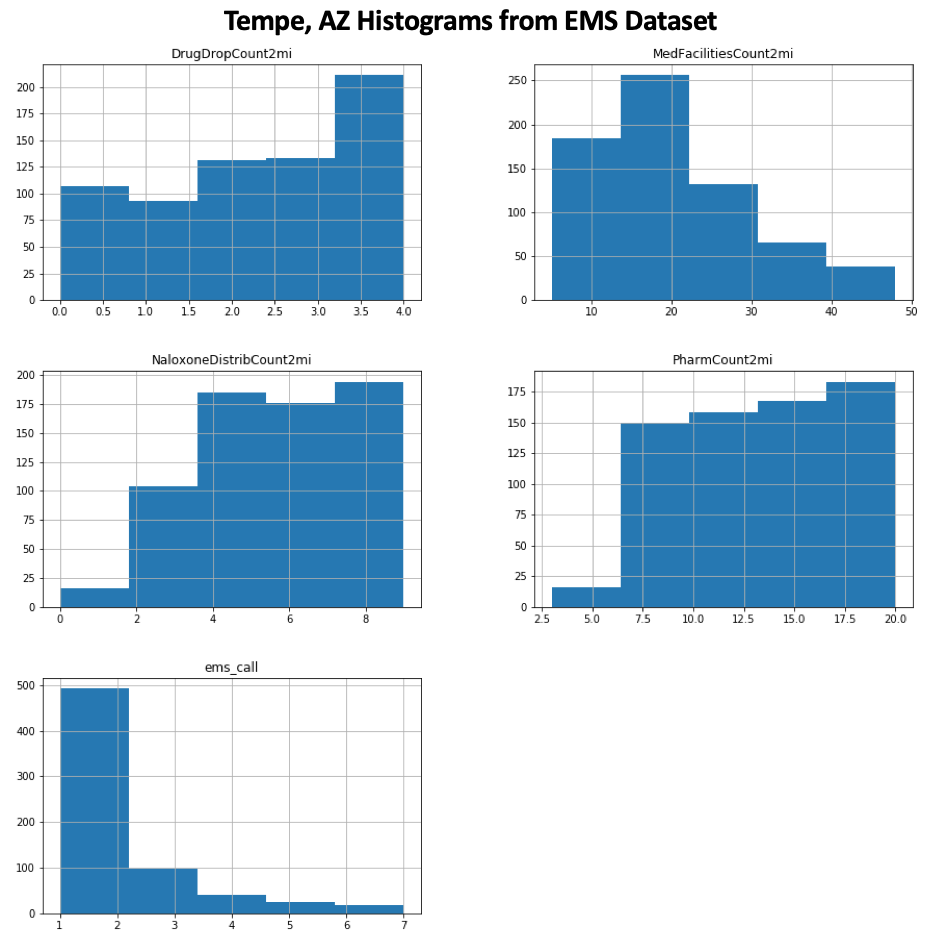
* PNML data is at a different spatial unit (site collection area) than the EMS calls. To get the counts of each spatial predictor in the collection areas, we took the mean of the census tract counts for that predictor for all census tracts that overlapped with the site collection area polygons.

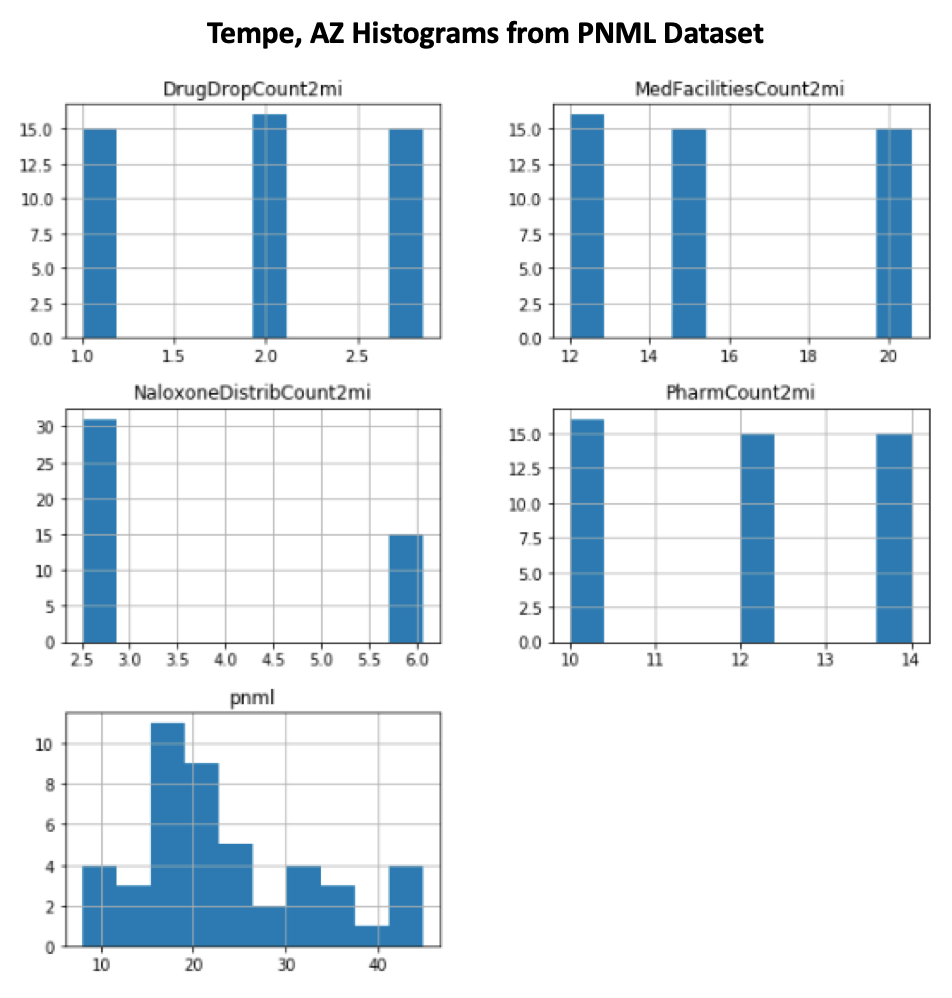
Descriptive Statistics:

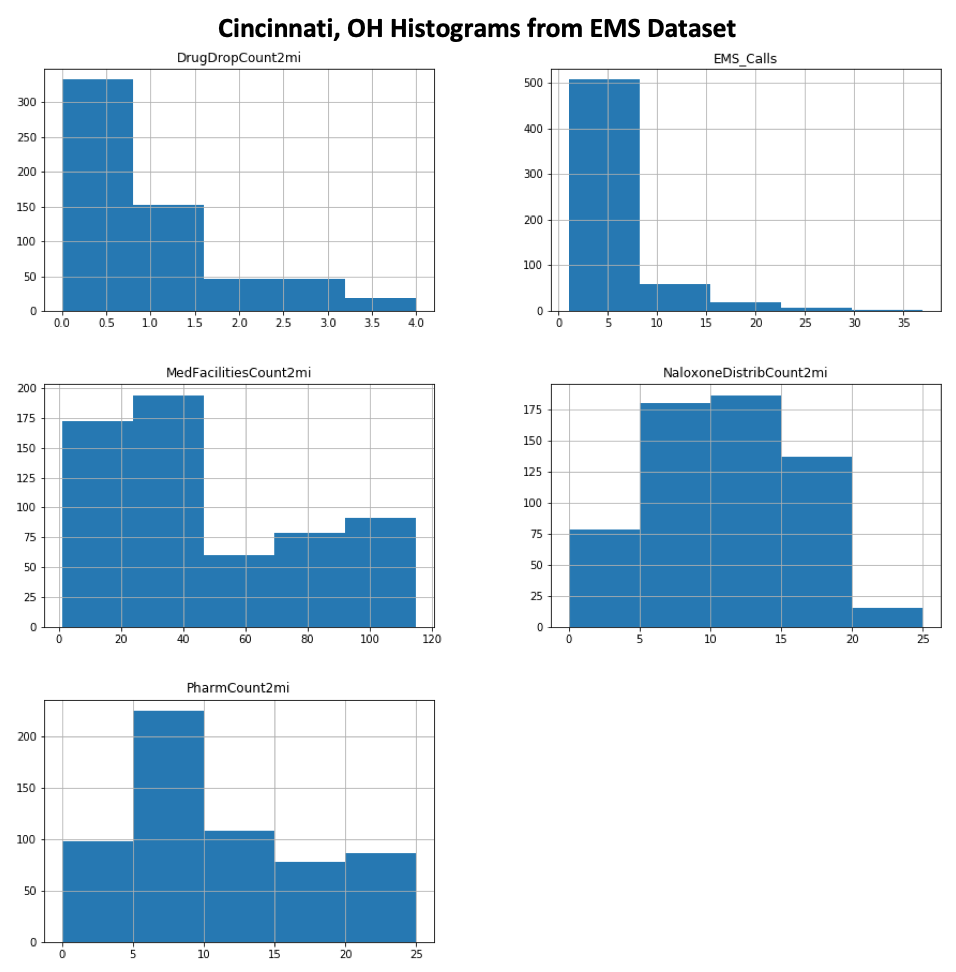




Histograms:







1. **For each variable, indicate what, if any, data transformation is necessary in order to conduct your intended analysis. Justify your choice of transformation relative to the distributional assumptions underlying your chosen statistical or ML model.**

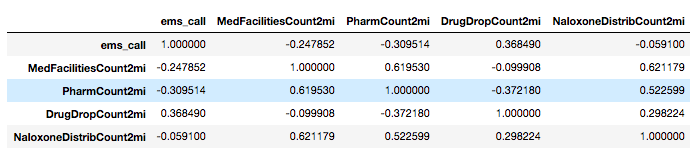
As discussed previously, we transform the count of EMS calls per area per year per month to a binary variable equal to one if the count of calls is above the series average and equal to zero otherwise. We do this to create two classes of outcomes in our dependent variable that can be analyzed in our logistic regression.

All other variables included in the study are ratio measurements. The PNML measures are somewhat skewed, them to be “Gaussian-like” enough to utilize as is. While the counts of facilities per spatial area are discrete, we expect their effect on the outcome of interest to vary proportionally to one another. As such, we do not perform any additional transformations at this time.

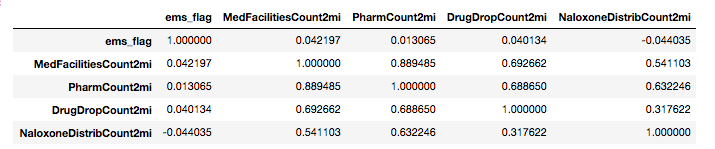
1. **Conduct a preliminary test of the relationships between your variables of interest using a correlational method. Justify your choice of this method given the distribution of the data.**

Given the power law distributions of the EMS data, we test the correlation between our dummy variables for a high number of EMS calls per spatial unit per month per year in the data set. Overall, we find weak to moderate associations between this binary variable and the spatial predictors of interest using a pearson correlation matrix.

*Tempe, AZ EMS Correlation Matrix*

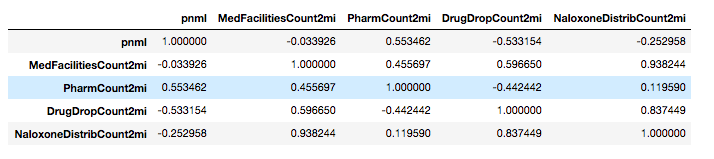


*Cincinnati, OH EMS Correlation Matrix*

**

Because the PNML data is approximately gaussian and it and all of the spatial predictors of interest we are looking at our ratio, we also leverage pearson correlation matrix for this dataset as well.

*Tempe, AZ PNML Correlation Matrix*



1. **How many analyses do you plan to conduct using your dataset? Make sure you control for multiple comparisons. Choose a Type I error rate using this information.**

We will conduct two analyses in this study. First, we will use a logistic regression of the count of EMS calls per area per month per year against the treatment facility spatial regressors. Then we will conduct a fixed-effects linear regression of the PNML values against these same predictors and a series of dummy variables for the given year-month and tract indicators in our data.

Because or dataset is limited, we are willing to sacrifice Type I error in order to maximize our detected effect size. We have a robust number of datapoints for the EMS calls and will utilize the standard 0.05 Type I error rate. We are far more limited with the PNML data and will accept a Type I error rate of 0.2 (i.e. risk a false positive every one in five experiments).

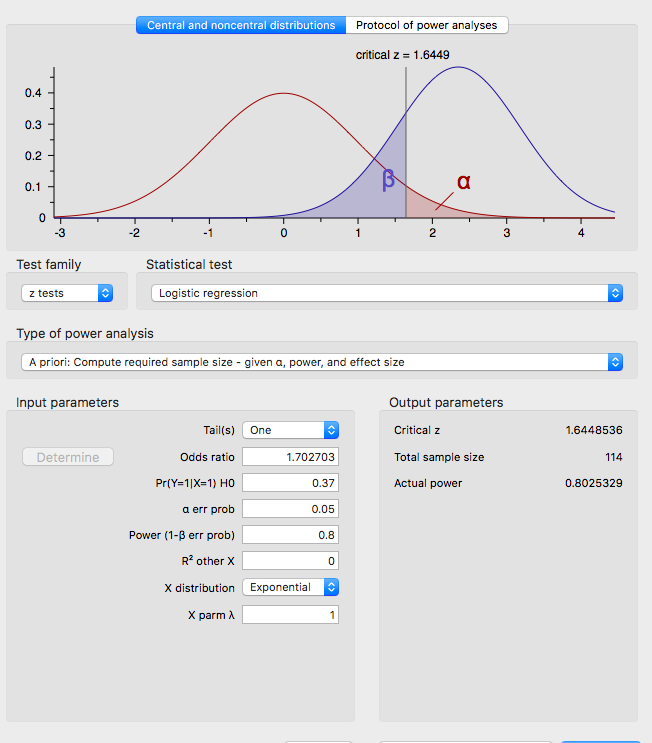
1. **Covariates should be included in your analysis if you expect confounds but cannot control for them because of limitations on your design. What covariates do you plan to control for statistically, if any? Justify your selection of these.**

The fixed effects included in the PNML analysis account for the mean difference in variation between spatial and temporal segments in our analysis. In doing this, we lose the ability to identify which specific confounds (such as differences in economic status for example between tracts or over time) lead to heterogeneity between spatial and temporal units. What the fixed effects do instead is control for the total effect of each confound in a spatial or temporal unit on our dependent variable. This applies the same level of control over variance of applying relevant confounds.

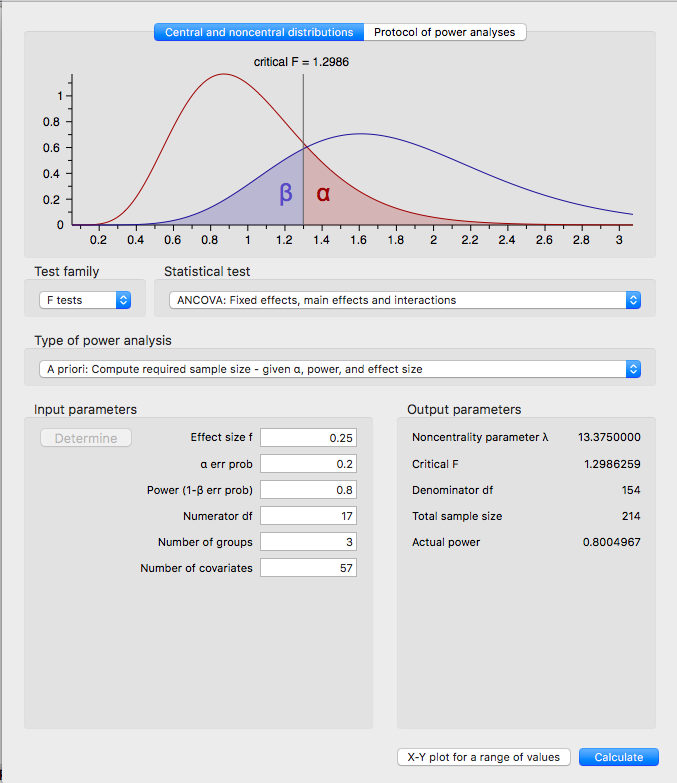
These same controls are not available to us in the logit model because of quasi-separation issues. While we have not implemented them at this time, we will leverage demographic controls from the census datasets to control for area to area heterogeneity.

1. **Conduct a power analysis to determine how much data are necessary to detect the hypothesized relationship. You may use off-the-shelf software. Your power analysis should be based on pilot data.**

We conduct two power analyses for these experiments. First, for the logistic regression against EMS calls, we assume that at their mean, the number of EMS calls per spatial area per month per year will follow the average proportion across the whole dataset. In the case of Tempe, this means that we observe a Pr(Y=1|X=1) for the null hypothesis of .37. Assuming, an effective proportion of .5 against our alternative hypothesis, we expect an odds ratio of 1.7027, a critical Z value of 1.64, and a minimum sample size of 114.

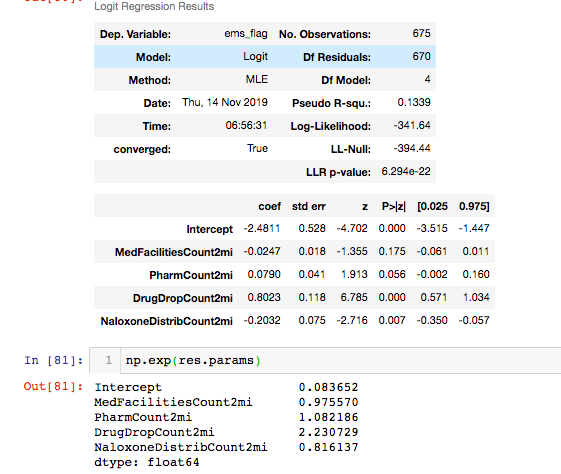


For the PNML regression, we rely on an analysis of covariance between predictors (ANCOVA). We have 37 county level fixed effects, 16 year-month fixed effects, and four regressors for a total of 57 covariates in our analysis. Using a .2 Type I error rate and .8 desired power, we find that we need a sample size of 154. This would equate to about two to about two more years of monthly PNML data collection under the current design.

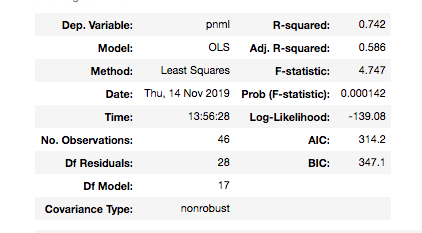


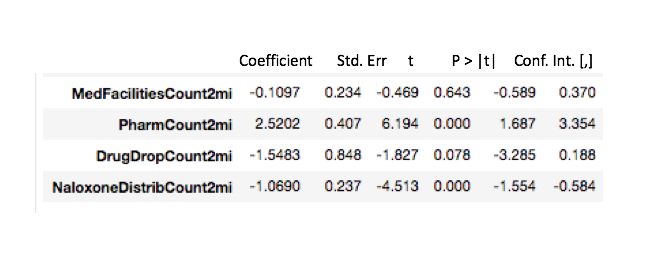
1. **Conduct your intended analysis using your pilot data. Specify and control for any covariates that you have identified. What does your analysis indicate regarding the hypothesized relationship in the data?**

Conducting a logistic regression of EMS calls on our pilot data yield somewhat confusing results. First, we observe a pseudo R-squared value of around 0.14, suggesting that the spatial opioid treatment predictors only explain about 14% of the variance in EMS calls. We find odds ratios close to 1 for both the number of pharmacies and medical facilities within two miles of a census tract which could imply that these regressors have no impact on a census tract having a high number of opioid related EMS calls made per month. For the naloxone distribution centers within two miles of a census tract, we find that the presence of these points reduces the likelihood of having a high concentration of opioid related EMS calls by around 20%. Strangely, we also find that the presence of drug dropoff centers increases this same likelihood by 130% over the mean with an odds ratio of 2.3. While this is an odd result, one initial hypothesis could be a reverse causality problem. High numbers of EMS calls related to opioids may actually cause a greater concentration of drug dropoff centers to be put in an area, which could explain the high association we observe in the data.



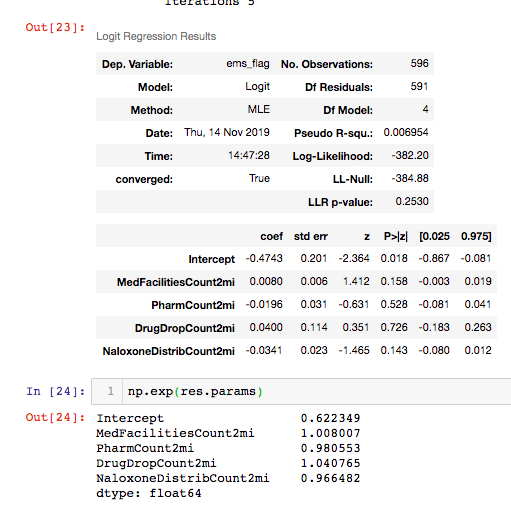
In the PNML regression, we find results that are more similar to our initial hypotheses and observe an R-squared value of 0.742 despite the small sample size used.





In this regression, we see that both naloxone and drug dropoff centers decrease PNML measurements by 1-1.5 units per month. The only positive predictor we observe is the count of pharmacies, which as discussed above, could be a result of more medical waste being produced around these centers. Medical facilities do not show any statistically significant effect on the PNML measure. One hypothesis could be that the medical waste effects in the water are offset by the facility itself’s role as treatment center and deterrent to overall drug use.

Extending our initial correlation test of the effect of treatment facility locations on the number of EMS calls per month per year per area to Cincinnati produces some initially discouraging results. We find a pseudo R-squared value of 0.006 which implies that essentially none of the variability in EMS calls is explained by this design. Moreover, we find odds ratios of approximately one for each spatial regressor included in the Cincinnati model. This means that the location of medical facilities, naloxone distributors, pharmacies, or drug dropoffs have no impact on increasing or decreasing the likelihood of having a high number of opioid related EMS calls in a given area at any point in time.



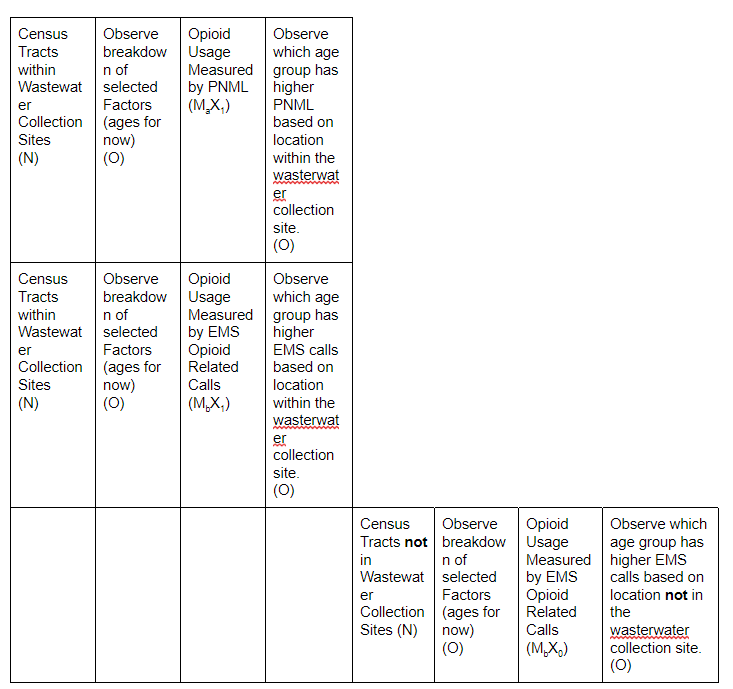
Comparing these results to the logistic regression conducted in Tempe, it appears that there is no association between these predictors and the number of opioid related EMS calls.

This is an interesting finding, given that the PNML regressions results diverge quite significantly from the EMS ones in Tempe. As we noted in our third presentation, even though both PNML and EMS are face valid measures of the same construct, opioid use, they only correlate with one another about 60% of the time. These divergent results may indicate that the PNML measures are picking up on different facets of the drug use construct we wish to study than the EMS calls. Testing to see if this divergence holds in future experiments where the interventions we study are significant for both the PNML and EMS measures will be a key focus going forward.

1. **Provide an overall summary of your plan thus far: question, sampling strategy, measurement strategy, design, and analysis plan. Indicate what you expect to find in the analysis, what the design allows you to say given these findings, what the implications are for your theory, how far you can generalize, and to what extent this answers your question.**

As a summary, our goal in this study is to establish a significant relationship between key public health interventions, demographics, behavioral and mental health predictors, and/or other variables of interest that could impact opioid usage within cities over time. We measure opioid by examining the monthly count of opioid-related EMS calls per census tract, and by looking at the monthly average PNML of opioids detected in urban wastewater. Because of this, we have used purposive sampling of two cities, Tempe and Cincinnati, that collect this data. Tempe, in particular is one of the only cities in the US to collect PNML data and is the only one that has made it publicly available.

Accordingly, we use a patched-up equivalent materials design that attempts to find consistent relationships between predictors we test based on theory and both of our measures of opioid use. Both PNML measures and EMS calls represent different aspects of drug use, and both the Cincinnati and Tempe data sources collect these measures via different methodologies, across different time frames and spatial units. We hypothesize that if we find a consistent relationship between a selected predictor (or group of predictor) in all cases, then this relationship may have a causal impact on opioid use.



So far, we have tested two of these relationships. First, we found that poverty status and being in the middle aged cohort of a population was not significantly correlated with drug use levels despite being cited by the Arizona Department of Public Health as a potential cause. Second, in both Tempe and Cincinnati, we used a logistic regression two test the effects of the count of medical treatment facilities, naloxone distribution centers, pharmacies and drug dropoff centers within two miles of a given area on an above average number of opioid related EMS calls. We also conducted a linear regression with fixed effects testing the effect of these same spatial predictors on the PNML measures in Tempe and found that the results diverge quite a bit between the two constructs of drug use.

Going forward, our goal will be to test more types of relationships that could have an impact on both of these opioid-use related confonds based on the public health literature and data available to us. While we stop short of claiming causality, we argue that any relationship we detect that is consistent across each experiment represents a significant enough association to be examined by future research and intervention studies.