Project 2

Eli Horner, Jack Andrew, Rina Deka

2025-04-04

Introduction

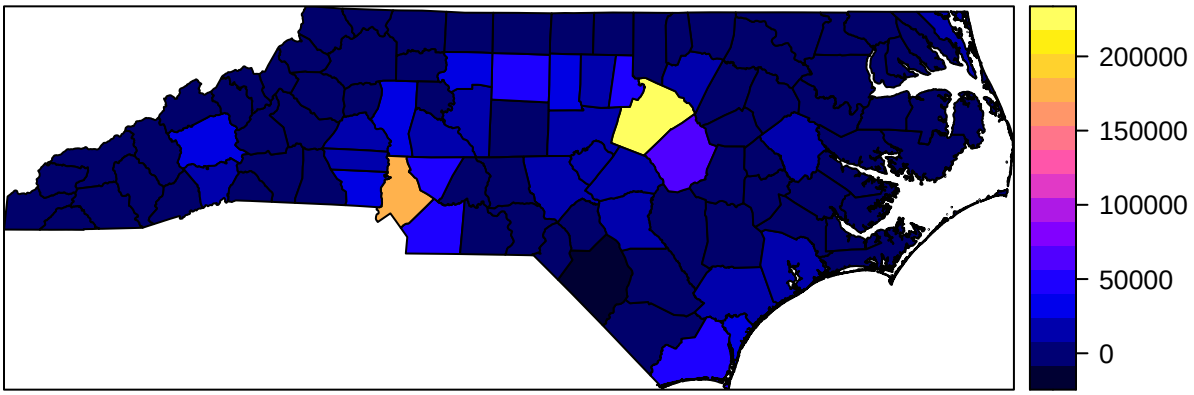
In this project, we first analyze spatial dependence in 3 variables of interest and then make predictions of a 4th variable. For this analysis, we chose Overdose Deaths per 100,000 as the response variable due to its direct relevance to public health concerns, particularly in the context of the opioid crisis. Overdose deaths are a pressing issue in many regions, and understanding the factors that contribute to them is critical for effective policy and intervention strategies

Part 1

The state of North Carolina collects a lot of data at the county level. Here, we will examine 3 of these variables, namely: Population Change since 2014 (2024), Median Age, and Opportunity Youth (2022).

Population Change since 2014

First, let's look at population change since 2014 and analyze is there is spatial dependence.



Visually, we see evidence of some clear spatial patterns. Let's calculate a Moran's I and Geary's C to test for spatial dependence. First, we define our neighborhood. We will use queen neighborhoods, considering corner borders to be borders. We use this neighborhood to define our proximity matrix, which we construct as a W matrix (rows normalized).

Now, we can perform our Moran's I test with this neighborhood.

```
##
##  Moran I test under normality
##
## data:  nc_counties$pop_change
## weights: Wlist
##
## Moran I statistic standard deviate = 3.685, p-value = 0.0001143
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.230280366      -0.010101010       0.004255248
```

We find our Moran's I statistic to be 0.2303, which is significantly different than our expectation under the assumption of normality of -0.0101 with a p-value of 0.0001. This result indicates positive spatial dependence, indicating that nearby counties have more similar population changes over the past decade. Now, let's check this results by calculating Geary's C.

```
##
##  Geary C test under normality
```

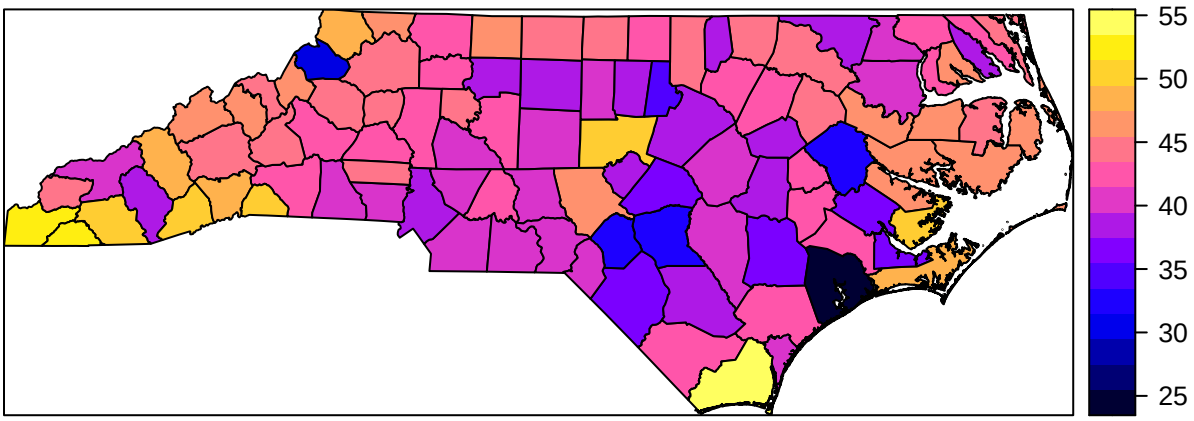
```
##
## data:  nc_counties$pop_change
## weights: Wlist
##
## Geary C statistic standard deviate = 3.0434, p-value = 0.00117
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##       0.791205335       1.000000000       0.004706728
```

We find the Geary's C statistic to be 0.7912, which is significantly different than our expectation under normality of 1 with a p-value of 0.0012. This result also indicates positive spatial dependence.

Overall, we can conclude that there is positive spatial dependence in the population change in NC counties since 2014.

Median Age

Let's look at median age. Here is a plot of median age by county in 2022.



Now, we can check for spatial dependence using the Moran's I test and the Geary's C test.

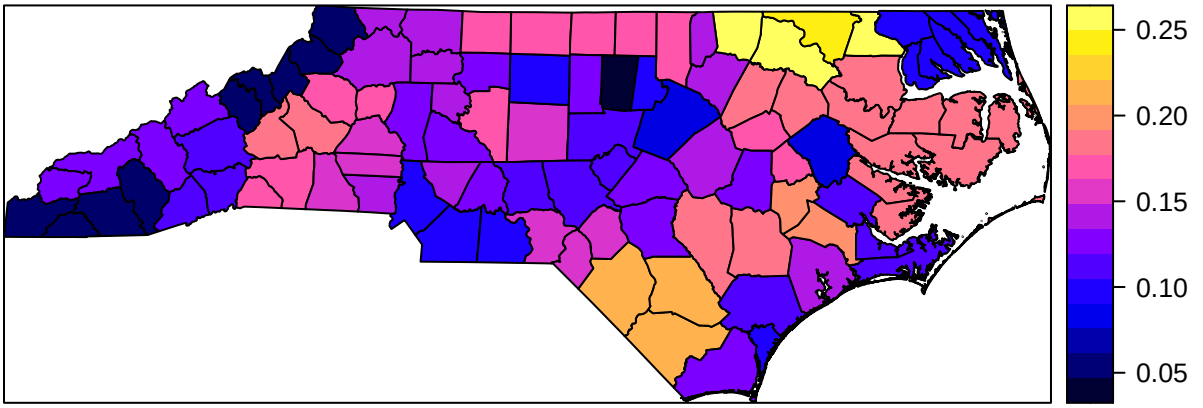
```
##
##  Moran I test under normality
##
## data:  nc_counties$age
```

```
## weights: Wlist
##
## Moran I statistic standard deviate = 2.3628, p-value = 0.009068
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.144032443      -0.010101010       0.004255248
```

We find our Moran's I statistic to be 0.1440, which is significantly different than our expectation under the assumption of normality of -0.0101 with a p-value of 0.0091. This result indicates positive spatial dependence, indicating that nearby counties have more similar median ages. Now, let's check this result by calculating Geary's C.

Opportunity Youth

Let's look at opportunity youth in 2022. Here is a plot of opportunity youth by county in 2022.



Now, we can check for spatial dependence using the Moran's I test and the Geary's C test.

```
##
##  Moran I test under normality
##
## data:  nc_counties$opp22
## weights: Wlist
##
## Moran I statistic standard deviate = 6.6639, p-value = 1.333e-11
```

```
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.424599269      -0.010101010       0.004255248
```

We find our Moran's I statistic to be 0.4246, which is significantly different than our expectation under the assumption of normality of -0.0101 with a p-value of 1.3e-11. This result indicates positive spatial dependence, indicating that nearby counties have more similar percentage of opportunity youth in 2022. Now, let's check this result by calculating Geary's C.

```
##
##  Geary C test under normality
##
## data:  nc_counties$age
## weights: Wlist
##
## Geary C statistic standard deviate = 2.6419, p-value = 0.004123
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##       0.818753085       1.000000000       0.004706728
```

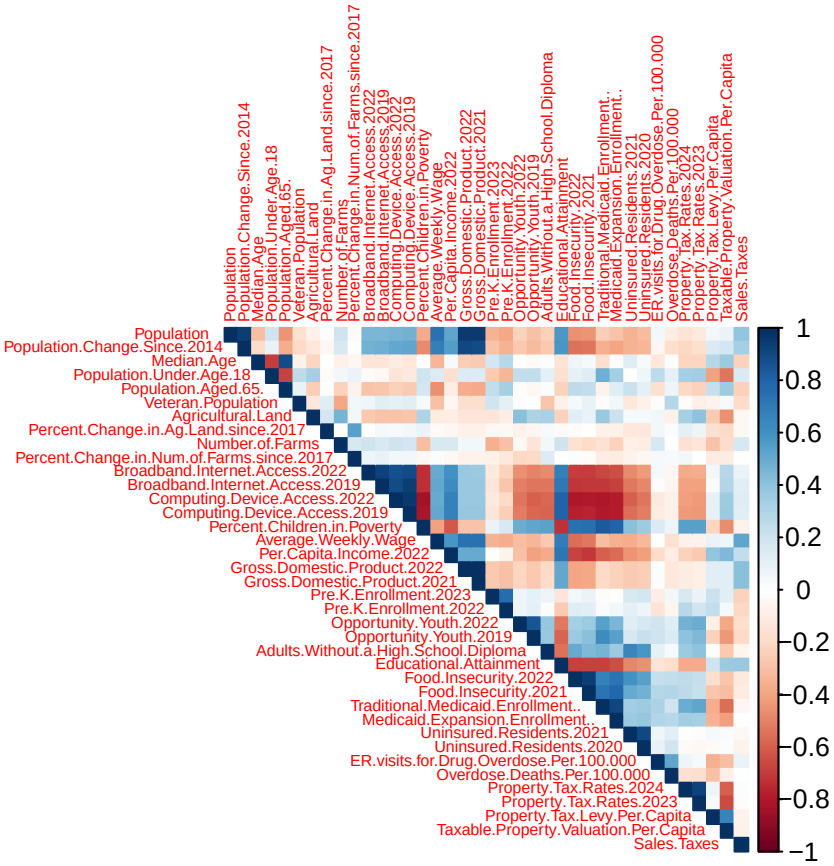
We find the Geary's C statistic to be 0.8188, which is significantly different than our expectation under normality of 1 with a p-value of 0.0041. This result also indicates positive spatial dependence.

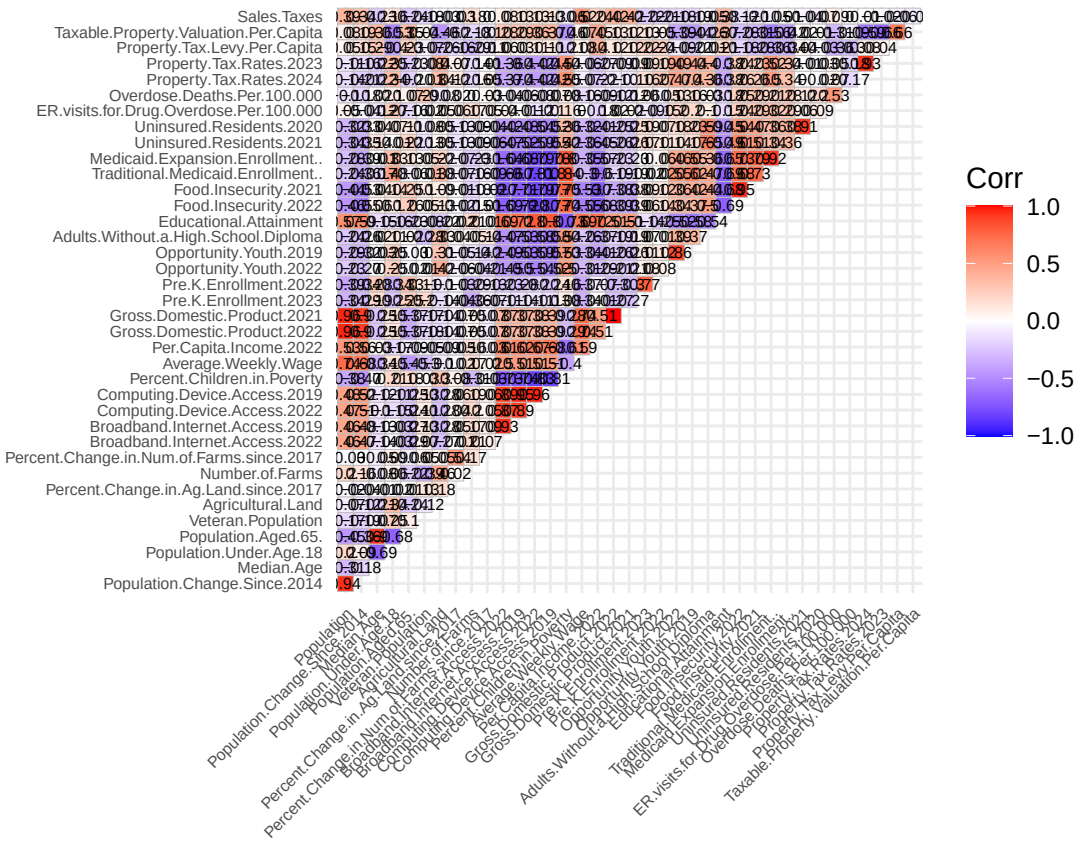
Overall, we can conclude that there is positive spatial dependence in median age in NC counties.

Part 2

For this section, we will regress one variable on some of the other variables in the data set.

Let's start by choosing our variable.


Population
Population.Change.Since.2014
Median.Age
Population.Under.Age.18
Population.Aged.65.
Veteran.Population
Agricultural.Land
Percent.Change.in.Ag.Land.since.2017
Number.of.Farms
Percent.Change.in.Num.of.Farms.since.2017
Broadband.Internet.Access.2022
Broadband.Internet.Access.2019
Computing.Device.Access.2022
Computing.Device.Access.2019
Percent.Children.in.Poverty
Average.Weekly.Wage
Per.Capita.Income.2022
Gross.Domestic.Product.2022
Gross.Domestic.Product.2021
Pre.K.Enrollment.2023
Pre.K.Enrollment.2022
Opportunity.Youth.2022
Opportunity.Youth.2019
Adults.Without.a.High.School.Diploma
Educational.Attainment
Food.Insecurity.2022
Food.Insecurity.2021
Traditional.Medicaid.Enrollment..
Medicaid.Expansion.Enrollment..
Uninsured.Residents.2021
Uninsured.Residents.2020
ER.visits.for.Drug.Overdose.Per.100.000
Overdose.Deaths.Per.100.000
Property.Tax.Rates.2024
Property.Tax.Rates.2023
Property.Tax.Levy.Per.Capita
Taxable.Property.Valuation.Per.Capita
Sales.Taxes
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



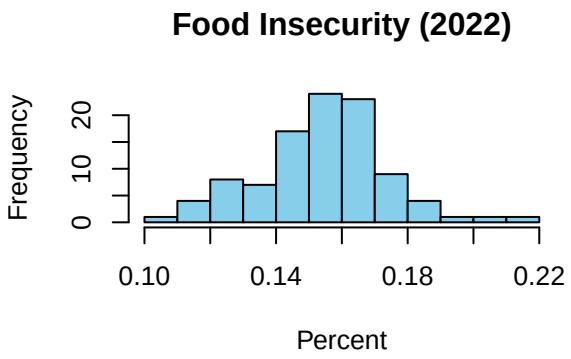
Based on the heat map, some interesting variables might be (for select topics): - Population.Change.Since.2014 (urbanization, migration, policy) - Food.Insecurity.2022 (health equity & access) - Overdose.Deaths.Per.100.000 (public health) - Percent.Children.in.Poverty (long-term socioeconomic outcomes)

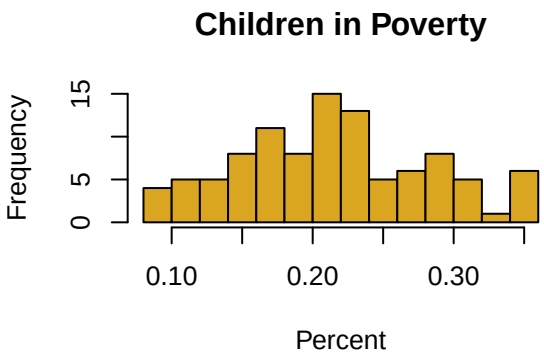
We want to pick predictors that are: - Correlated (but not too correlated) with our outcome - Interpretable and likely to explain the response - Not multicollinear (check pairwise correlation < 0.8 if possible)

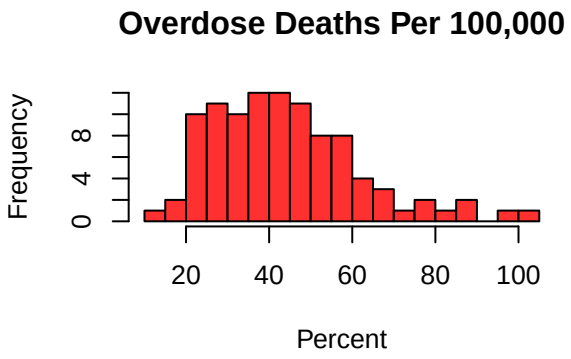
Again using the heatmap, it looks like the best 3 predictors are: - If we use Population.Change.Since.2014 as our target, then predictors could be Median.Age, Broadband.Internet.Access, Poverty.Rate, Educational.Attainment (Since young, educated areas with internet might grow more).

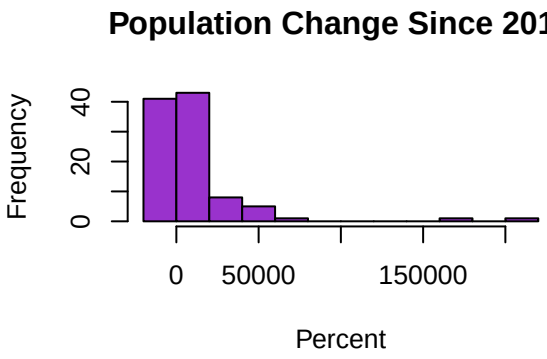
- For Food.Insecurity.2022 as the target, we could look at Percent.Children.in.Poverty, Adults.Without.a.High.School.Diplo Uninsured.Residents.2020 etc. (due to the link between poverty and food insecurity, education and food scarcity, and lack of healthcare and food insecurity).
- If we use Overdose.Deaths.Per.100.000 as the target, we could look at things like Population Change since 2014, Educational Attainment, Veteran Population (structural vulnerabilities).
- If we look at Percent.Children.in.Poverty, we might look at things like: Unemployment.Rate, Food.Insecurity, Educational.Attainment, Average.Weekly.Wage (since these speak to parental resources and economic stability).

Let's take a look at histograms since understanding the shape of our distribution will help us understand what type of model we should use (linear or poisson, etc), help us avoid too many outliers dominating the model, decide if a transformation of the target variable is needed, and to visually asses normality and skewness.









Which seems to be the best for the target variable?

For the Food Insecurity target, we have:

- Distribution: Fairly normal (bell-shaped), centered around ~15–18%.
- Pros: Good for regression — a nice, symmetric target.
- Cons: If you're interested in identifying outliers or extremes, this might not be ideal since it's quite balanced.
- Verdict: Solid choice for standard regression, but not particularly surprising or extreme.

For Children in Poverty, we have: - Distribution: Roughly uniform or slightly bimodal, with a wide spread.

- Pros: Could be interesting for classification (e.g., high vs. low poverty), or exploring clusters.
- Cons: Might be more difficult for regression due to less clear central tendency.
- Verdict: Good for classification or policy-based segmentation.

For Overdose Deaths, we have:

- Distribution: Right-skewed, long tail — many counties with low deaths, few with very high rates.
- Pros: Super interesting for prediction. That skew suggests some counties are outliers worth understanding. Great for modeling rare but serious outcomes.
- Cons: Requires care in handling outliers (e.g., log transformation).

- Verdict: Most interesting if your goal is identifying high-risk areas or understanding disparities.

For - Distribution: Highly skewed right, many near zero or negative, with a few extreme positives.

- Pros: Could be fascinating to model migration or growth dynamics.
- Cons: Huge range, lots of zeros or small changes — may require normalization or feature engineering.
- Verdict: Also very interesting, but potentially noisier.

My personal opinion is that we should go with Overdose Deaths.

Overdose deaths are heavily spatially structured and certain regions (like Appalachia, parts of the South, or the Rust Belt) have had historically higher rates. That makes it a perfect candidate for a spatial regression model.

The right skew also gives us the challenge of transforming or modeling a non-normal target — often the more statistically interesting and policy-relevant scenario. You might log-transform the response or try a Poisson or negative binomial model depending on how it's measured.

Furthermore, modeling overdose deaths allows us to make a policy case — identifying predictors like poverty, unemployment, healthcare access, or education could point to real solutions. It's a meaningful and timely topic.

Even after controlling for variables like poverty or population change, we'll probably have leftover spatial structure — so your spatial dependence test (e.g., Moran's I on residuals) will likely be fruitful and interesting.

Model Building

Let's start by fitting a simple linear regression model.

We find the following model: predicted overdose deaths per 100,000 = 3.874 - 2.825e-5 * population change since 2014 - 2.137 * educational attainment + 22.41 * veteran population. This implies that as veteran population increases in a county, so does overdose rate. It also implies that as educational attainment decreases, overdoes rate increases. This effect is smaller than that of veteran population. Finally, we find a tiny effect of population growth since 2014. This effect is also negative, meaning increased population growth implies lower overdose rate. All of these findings make sense conceptually. Our model is not super predictive, and only Veteran Population has a significant effect. Let's check for spatial dependence in our residuals, using the Moran's I test, using the same neighborhood from part 1.

```
##
##  Moran I test under normality
##
## data:  nc_counties$residuals
## weights: Wlist
##
## Moran I statistic standard deviate = 2.7749, p-value = 0.002761
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.170909782      -0.010101010       0.004255248
```

We find our Moran's I statistic to be 0.1709, which is significantly different than our expectation under the assumption of normality of -0.0101 with a p-value of 0.0028. This result indicates positive spatial dependence in our residuals, suggesting we need a model that accounts for spatial dependence.

We will fit both a SAR and CAR model for our spatial dependence. First, let's examine the CAR model.

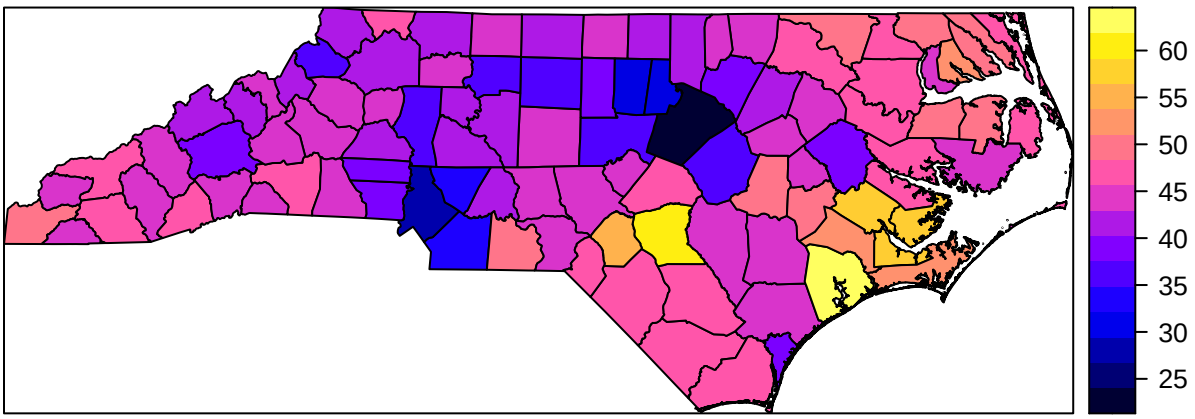
We find the following model: predicted overdose deaths per 100,000 = 3.829 - 5.202e-5 * population change since 2014 - 1.709 * educational attainment + 20.52 * veteran population. This implies that as veteran population increases in a county, so does overdose rate. It also implies that as educational attainment decreases, overdoes rate increases. This effect is smaller than that of veteran population. Finally, we find a tiny effect of population growth since 2014. This effect is also negative, meaning increased population growth implies lower overdose rate. Now, we will fit using SAR model.

We find the following model: predicted overdose deaths per 100,000 = 4.008 - 4.367e-5 * population change since 2014 - 2.158 * educational attainment + 21.09 * veteran population. This implies that as veteran population increases in a county, so does overdose rate. It also implies that as educational attainment decreases, overdoes rate increases. This effect is smaller than that of veteran population. Finally, we find a tiny effect of population growth since 2014. This effect is also negative, meaning increased population growth implies lower overdose rate.

Since the pseudo-R^2 of 0.088 for the SAR model is higher than the CAR model's pseudo-R^2 of 0.086, we will use the SAR model going forward.

Spatial Map

Here is a map of our model's prediction for overdose death rate in North Carolina counties.



Our model is not very good at predicting drug overdoses, but this seems to be a challenging thing to predict from available data.

Discussion

The key aim of our analysis was to identify socio-economic and demographic factors that might influence overdose death rates at the county level. We selected the following predictor variables based on their potential connections to health outcomes, specifically overdose mortality.

Population Change: Population growth or decline can be an important indicator of economic and social changes within a region. Population decline could signal economic instability, lower access to healthcare, or a loss of social support structures, potentially contributing to higher overdose rates. We found this variable not to be significant in our model.

Veteran Population: Veterans are known to face unique health challenges, including mental health disorders, substance use issues, and limited access to healthcare. By including the veteran population as a predictor, we aimed to capture this vulnerable group's potential influence on overdose rates. We found this variable to be the only significant predictor in our model.

Education (Percentage of Adults 25-44 with a secondary degree by 2030): Education is a well-established determinant of health. Areas with higher levels of education tend to have lower rates of substance abuse and overdose deaths. Additionally, more educated populations might have stronger community networks, which can provide social support and reduce the risk of overdose. We found this variable not to be significant in our model.