The Data Open

Analysis of US demographic segmentation and spatial pattern regarding tobacco usage and COPD mortality

Chuqi Bian, Shutong Li, Yuantao Shi, Minglun Pan

November 21, 2021

1 Topic Selection

Tobacco is a direct threat to people's health, as it not only leads to addiction and causes life-threatening diseases to its direct smokers, but also harnesses the health of people who are exposed to second-hand smoke. According to the World Health Organization, around 8 million people die prematurely every year from smoking. Although this topic should bring up more and more public awareness, people still treat it as a common thing. To address such concern within the United States, it is important for the government and citizens to acknowledge the trend of tobacco usage and the tobacco-related premature death at different demographic levels. With this in mind, we decide to investigate:

The intrinsic patterns and group segmentation regarding tobacco usage and tobacco related premature disease/premature death within the United States.

Within the scope of our analysis, we primarily use the *tobacco_use_us* and *us_chronic_resp_disease* dataset. External data are utilized to find the geographical autocorrelation of proportion of reported smokers. In our analysis, We will break down our insights into 4 angle of aspects:

1. Tobacco usage among age groups
2. Tobacco usage among social classes (by income and education)
3. Quantifying smoking culture using tobacco usage and COPD mortality rate
4. Statistically analyze spatial clustering significance of COPD mortality and tobacco usage and times-series modeling COPD mortality rate.

2 Executive Summary

We perform demographic analysis of smoking habit horizontally across break out groups and COPD mortality rate vertically across county, state and nation level.

From our analysis of smoking habit, we first investigated tobacco usage across age groups in US. We found that 1) new generation is getting more resistant towards smoking, 2) middle age demographic is the one least likely to quit smoking and 3) smoking prevalence seems to cluster spatially (around tobacco producing states).

Additionally, we learned that population with lower educational level and/or lower household income has higher smoking prevalence.

Moving on to COPD mortality rate, we observed that it also seemingly clusters in the spatial dimension. We also explored ways to explain current smoking culture of every state through bivariate analysis of mortality rate and smoking prevalence among young adults.

As a result of our exploration of tobacco usage and COPD mortality rate. We devised two direction for modeling and statistical analysis.

1. To verify the spatial autocorrelation of COPD mortality rate and smoking prevalence at state level (univariate and bivariate). i.e. to verity these two factors as indicators for regional smoking cultures. Our results verifies the presence of correlation across state and portrays the shift of clusters across time.

2. To model and forecast COPD mortality factoring state-level smoking prevalence. Our forecasting predicts the COPD mortality rate in the year of 2015 to 2019. The rate is slightly increasing at the beginning, then slightly decreasing in the end.

3 Technical Exposition

3.1 Data Wrangling

Tobacco Usage data

Column for sample size consists of all standard string representation of integers with comma separators at thousands. We type cast it into integers without comma to better impute for the column data_values. For the column data_value, we find out that, for every (year, state, question, response, breakout) subgroup, data_value is approximately calculated as the ratio between sample size of the row and the summation of sample size of all responses from the same subgroup (factored with degree of freedom). We use the same method for imputing missing data values inside the dataframe. When zero-division is encountered, we treat the result data-value as 0.

Additionally, we performed several feature engineering on the dataset. We are encouraged to see if any yearly patterns exist within different response buckets and data_values. To do that, we calculate the delta δ and rate of change of the data_values. To check the data at all levels, we create a region(levels are West, Midwest, Northeast, South, and Other US territories) factor and categorize each datapoint into different region buckets based on their state.* To measure the addictiveness of the citizens, we applied weights to the sample sizes according to their different levels of smoking status. That is, under the Smoking Status break out category, we quantify former smoker as -1, never smoke as 0, smoke some days as 1, and smoke everyday as 2. In the end, addictive_score $= \frac{\sum_i W_i * Sum\_of\_Sample_i}{\sum_i Sum\_of\_Sample_i}$, where W_i is the weight, Sum_of_Sample is the sum of sample size within each demographic break out.

When distributions of addictive score and count percentage follow the same distribution, we showcase then side by side to distinguish different insights. Otherwise, we only show the count percentage.

* Note: When aggregating state statistics into larger regions (Northern vs Eastern etc.), we encountered Simpson's Paradox when calculating smoking prevalence of various age-groups (most noticeably 18-25) due to our method of calculation (recalculating data value by summing together sample sizes). These results are promptly discovered and left out.

US Chronic Respiratory Disease Dataset

The dataset for chronic respiratory disease is structurally robust with minimal missing values. The mortality rate is standardized by age group therefore is fitting for comparison among states. For our analytical purposes we only subset the mortality rate of gender='both' (i.e. we are agnostic of gender vs mortality rate).

3.2 Exploratory Analysis

US demographic is a diverse entity. When exploring data, we examine the data by state, region, country level separately, draw insights depending on different sample levels and different break out category. During exploratory analysis, we have had several takeaways that provide critical insights regarding the distribution of smokers in the US.

3.2.1 Age group

Age serves as our first layer to segment the U.S. tobacco usage. We first investigate how different age groups behave at the country level.

Country-level Analysis

When surveying the smoker status among different age groups on the country level(aggregated across state level data), the addictive score of each age group differs a lot. 65 + has a negative addictive score, which is reasonable given the majority had a habit of smoking but already quit. The middle-aged group have a smooth addictive score curve, while that of the younger generation speedily decreases across years. Similarly, we observe the percent young adult non-smoker (18-25) is increasing at a distinctly fast rate, and that of the elderly(65+) remains steadily high(Fig 1).

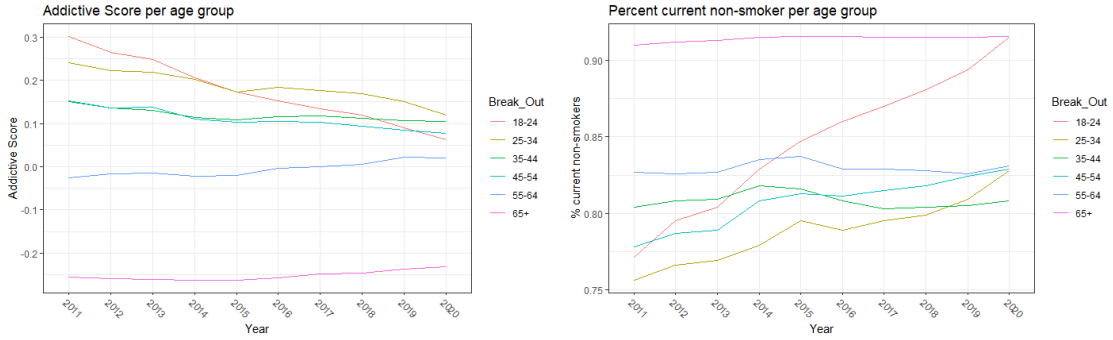


Figure 1: Addictive score and percent non-smoker per Age Group

The observation is partially surprising since the young adults appear to be smoking less despite the advent of E-cig. We suspect the sharp decrease has to do with the raising literacy on the harms of tobacco.

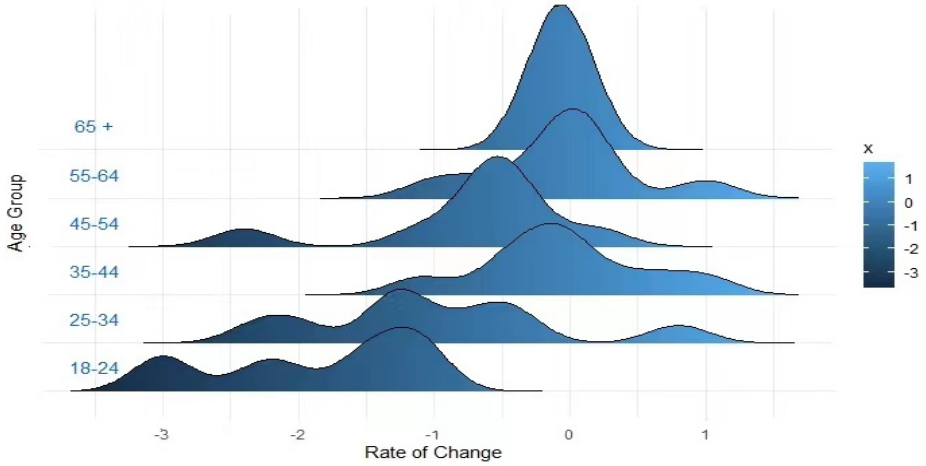


Figure 2: Density plot of rate of change by percentage of smokers in U.S. across age groups

A more intuitive way to visualize the change in trend is to plot the density of them across years (Figure 2). Aggregated on the country level, the distribution of changes provides us with three main insights. 1) Consistent with our above-mentioned observation, young adults consistently exhibits large decrease in smoker prevalence, 2) The age group of 65+ being already very low on smoker population, shows the lowest rate of decrease across time, 3) Combining fig1 and fig2, middle age is the age group where smoker population is relatively large and effort to quit smoking is relatively low.

One major takeaway is that as people age, they are more reluctant to change. This holds true for their tobacco usage. Plus tobacco is addictive, making it harder to quit as they have more years' of smoking history. To advocate for changing the smoking habit, we need to either pay special attention to the middle-aged group, as they are more reluctant to make changes; or to prevent/intervene young adult smoking habits.

Decrease and convergence of percent young adult smokers

We further breaks down the granularity of smoking prevalence of young adults from country-level towards state level. We discovered that the decrease is consistent across all states and the deviation from average among the values is growing less (i.e. values are converging).

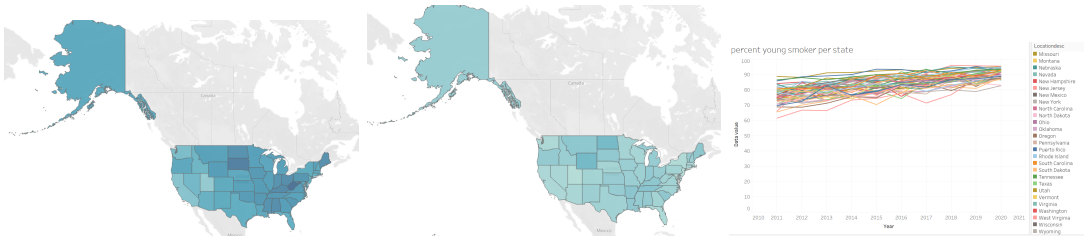


Figure 3: Percent current young adult smoker at 2011 vs 2020 vs overall trend

Additionally, smoking prevalence among young adults was high in 2011 centering around KY, WV and NC. Kentucky, West Virginia (or Virginia) and North Carolina are the top three tobacco production states in the US. We will see this in a more detailed manner at section 3.2.3.

3.2.2 Social Dynamic

As people with different backgrounds may treat tobacco with different levels of seriousness, we conduct analysis on the social aspect of citizens with tobacco usage at the region level.

Education

We analyze the Education indicator at the region level(aggregated by states), given that we are not sure if people at different regions will have similar educational backgrounds, which, by our assumption, may affect their smoking habits. Later in section 4, we justify that spatial autocorrelation in regards of data value exists. This validates our idea to analyze the following factors at the regional level.

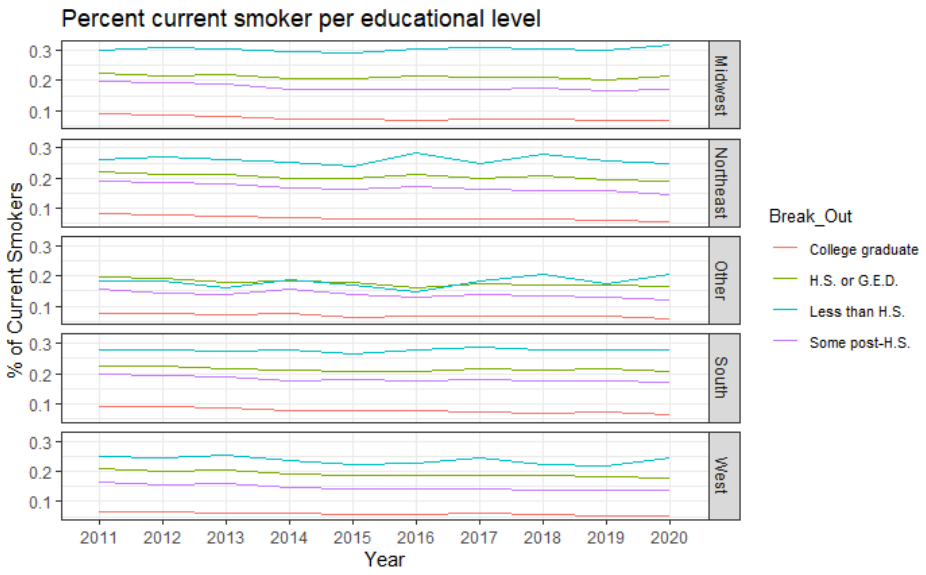


Figure 4: percentage current smoker across educational level per year by region

Citizens whose education history are less than high school have the highest percentage count of current smokes across all regions. As people have higher education, the percentage becomes smaller. From the above analysis, we suspect a negative correlation between citizens educational level and percentage count of current smokers.

Income

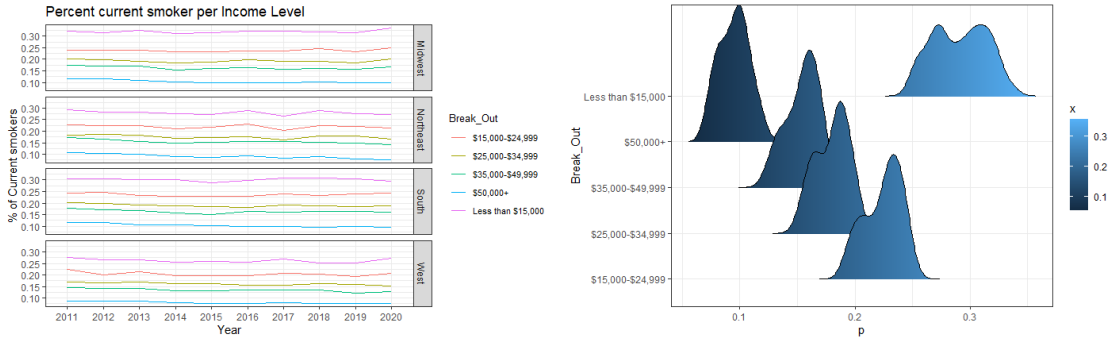


Figure 5: Line and density plot of percentage current smokers per income group by region

Interestingly, from figure 8, we also observe a negative correlation between income level and percentage count of current smokes across all region. Two side-by-side plots share the same insight from different angles. The line plot breaks down the percent current smokes by income group and regions and gives the most straightforward comparison; the density plot focuses on comparing the distribution of percentage count across income levels, year and region agnostic. The high residuals between line plots, and the clear trend of density plot both indicates a clear boundary on the smoking habit per income level. In addition, our observation of similar trend of smoking prevalence across income and educational level is consistent with our preliminary knowledge that education level and income status themselves are highly (positively) correlated.

From the above two segments, we are confident to conclude that social class plays a salient role in smoking habits and/or health literacy of people in U.S.

3.2.3 Charting the cigarette culture

Culture through diseases

We believe the mortality rate of COPD diseases to be an indicator of the nation's past smoking culture - we know that 1) COPD ≈ smoking habit due to it being the main cause of the disease and 2) chronic diseases like COPD takes time to manifest their often lethal symptoms. Therefore, mortality rate of COPD can be considered a time-lagged data that reflects the smoking culture in the past that is available to us at a state level granularity (as opposed to sales data). Figure 4 hopefully gives some intuition on the relationship between the two factors.

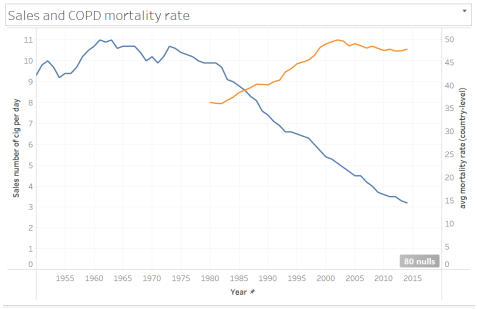


Figure 6: The COPD mortality rate at country-level and Tobacco sales data

Comparing mortality rate of COPD from 1980 to 2014, we discover formation of a hot-spot for COPD mortality rate concentrating around Kentucky, West Virginia, and North Carolina (Figure 5).

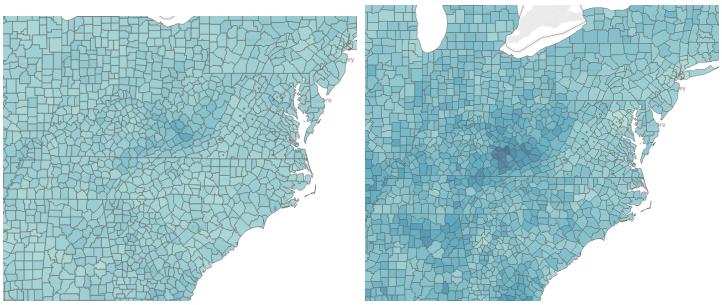


Figure 7: Mortality rate hot spot at intersection of Kentucky, West Virginia and North Carolina

This pattern according to our assumption implies a heavy presence of the top three tobacco production states. The similar behavior exhibited by these three states also leads us to inquire more on the relationship between a state's current COPD mortality rate and prevalence of current smoker.

Last but not least, the two observations shines light on the intrinsic spatial correlation among our data points. We will further analyze spatial autocorrelation in our modeling section. For now we have to remember that - any time we want to compare values of states among one another - we have to be cautious about using statistical analysis that assumes i.i.d. of data points.

Connecting disease to current smoking behavior

Similar (or converse) to what we believe about the implication of mortality rate, we postulate that estimate of young adults reporting to be non-smokers every year to be an indicator of the presence (or lack thereof) of smoking culture into the future. Plotting the two factors together at any fixed year that overlaps the two data set (2011-2014), we obtain visualizations of Figure 8.

Even though we are limited by the absence of i.i.d assumption to statistically prove correlation between

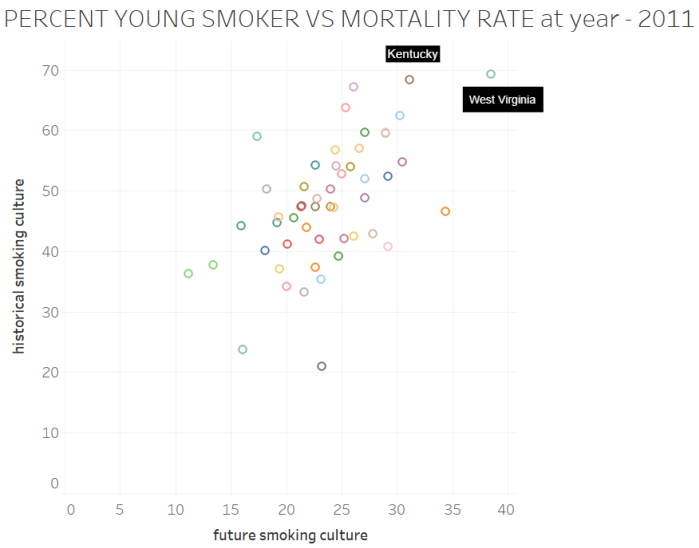


Figure 8: The scatter plot of percent of smoker between (18-24) vs mortality rate of COPD for 2011. Our version of past smoking culture vs future smoking culture

the two variable, the two variables none-the-less allows us to compare the indicators of past and future smoking culture descriptively. Kentucky and West Virginia again appear on the top right of the scatter plot having the most people died from smoking induced diseases while having the most of their new generation picking up smoking habit. On the other hand, states with stricter smoking laws such as California appears at the very bottom left of the chart.

4 Modeling

4.1 Spatial Autocorrelation Analysis

Reverberating with the previously recurring notion that mortality rate and smoking prevalence both correlates with the physical location of states. We decide to perform hypothesis testing on the auto-correlation of the two variables in spatial dimensions.

Metric: The metric for quantifying spatial autocorrelation is Moran's I. Moran's I is a correlation index aimed to capture global autocorrelation. This fits our purpose because we want to examine whether certain features group at a national scale. Mathematically, Moran's I is a function of pairwise similarity score of the feature in question (denoted as $Sim(y_i, y_j)$) and the weight of the two regions (denoted as W_{ij}). For the weight of two regions, it is common to binarize it as 1(TRUE) or 0(FALSE) depending on whether the two regions are neighbors. The similarity score is often formulated as $(y_i - \overline{y})(y_j - \overline{y})$. Combined together, formulation for Moran's I is

$$I = \frac{1}{s^2} \frac{\sum_i \sum_j (y_i - \overline{y})(y_j - \overline{y})}{\sum_i \sum_j w_{ij}}$$

(s^2 = sample variance). We will be performing hypothesis test on our Moran's I with the null hypothesis that feature value of states are randomly scattered across space.

Limitation in Data: The polygon shape files for state boundaries that we have available to us contain built-in error margins in shape coordinates, rendering a few supposedly neighboring states isolated 'islands' (having a weight of 0 rather than 1 in entries of weight matrix). That being said, the results still largely coincides with our observations and it is logical for us to follow through on our premise.

Tobacco Usage Dataset: Since a spatial instance (a feature layer) is a snapshot of smoking prevalence of a fixed demographic breakout and year, we plot a matrix of Moran's i where rows are different years of survey data and columns different demographic breakouts (Figure 9).

(a) Moran's I is largely statistically significant

(b) all Moran's I are positive, indicating clusters

(c) example of map with the least significant I

Figure 9: p-value and value of moran's I, as well as an example of when Moran's I is justifiably insignificant

From our testing we can confidently conclude that the smoking prevalence among different age groups (as well as in general) bounds with the geographic locations. As we previously mentioned, this shouldn't be a surprising discovery because geographic location largely forms regional culture, among which likely be one that defines the culture for smoking.

COPD Mortality rate: Similarly for COPD mortality rate data, we need to plot the result of statistical testing across Years. In Figure 10 we plot both the Moran's I and its statistical significance as a scatter plot where the grey line is our rejection threshold of 0.05 (Figure 10).

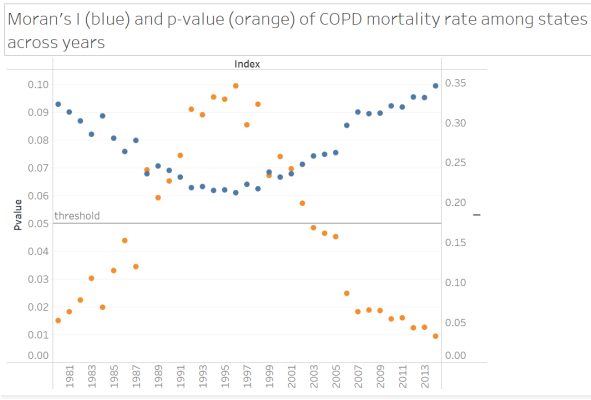


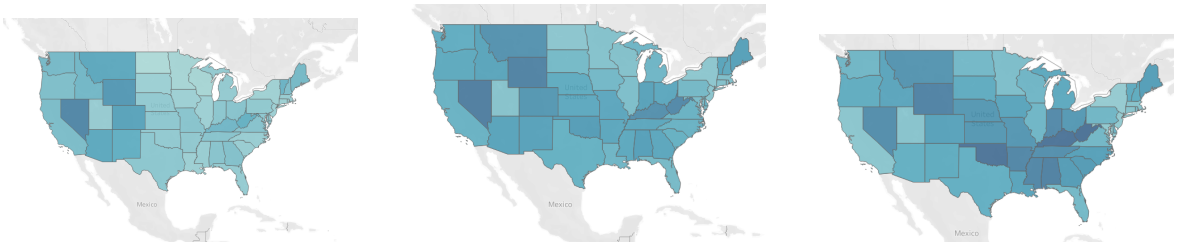
Figure 10: Blue: Moran's I, Orange: p-value for I, and a reference line for our p-value threshold

We can see that across-the-board (global) clustering of COPD mortality rate experiences a fall and rise throughout the years. We plot out the choropleth map of one year for each three periods (Figure 11) for better understanding of this phenomenon.

Even though Moran's I calculates significance for across the board clustering, the shift in local hot-spot is nonetheless captured by the trend in I's value and p-value. We can observe that 1) an overall increase in COPD mortality in every state and 2) a faster rate of increase among eastern regions that leads to a shift in hot-spot.

Bivariate Mortality vs Young Smoker Prevalence

Lastly we decide to examine our intuition about the correlation between COPD mortality and smok-



(a) Mortality initially clusters in the west

(b) Mortality in other areas caught up

(c) Mortality in the South-east surpasses west

Figure 11: a shift in COPD mortality hot-spot explains the change in Moran's I across time

ing prevalence among young adults by testing the bivariate spatial correlation of the two. Just like univariate Moran's I, bivariate I looks at the spatially-lagged correlation of a regional feature not by itself but by another feature. The result is presented in Figure 12.

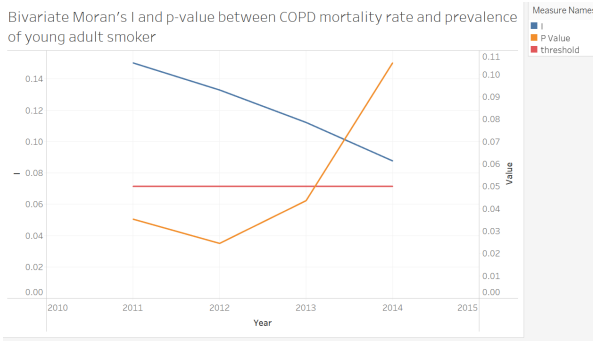


Figure 12: Blue: Moran's I, Orange: p-value for I, and a reference line for our p-value threshold 0.05

From Figure 12 we see that the spatial correlation between the two variables decreases both in value and its statistical significance. From our EDA in young smoker prevalence we understand that the decrease in the bivariate correlation is likely due to the aggressive decrease in both average and variance of the percent young smoker. Nevertheless, statistics regarding year 2011 and 2012 provides support to our intuition that, when young adult prevalence is sufficiently high, it clusters in a similar fashion with COPD mortality rate, implying that both quantifies something spatial/regional about smoking behaviors in US.

* Caveat: Just like how spatial correlation affects traditional correlation, the opposite is equally true. However, the trend in Moran's I and its respective p-value is still informative *even if* the value itself is optimistically biased.

4.2 COPD mortality rate prediction

In this subsection we introduce a regression method to predict Chronic obstructive pulmonary disease mortality in US, which takes the majority of Chronic respiratory diseases mortality. Like Figure 8 in EDA suggests, we note that the proportion of people who are current smokers is highly correlated with COPD mortality rate , therefore we consider a auto-regressive model with exogenous information and intercept:

$$Y_{t+1} = \alpha Y_t + \beta X_{t+1} + \gamma + \epsilon_t \tag{1}$$

where Y is the vector of COPD mortality rate in different states, X is the vector of proportion of people who are current smokers in different states and ϵ_t is random noise with zero mean. Here α, β, γ

are unknown constants. We estimate the coefficients α, β, γ by minimizing the following loss:

$$\sum_t (Y_{t+1} - \alpha Y_t - \beta X_{t+1} - \gamma)^2,$$

After getting the coefficients, we use equation (1) to get the prediction of COPD mortality rate in each state in the year 2015-2020. Then we average over states to get an overall estimate of COPD mortality rate in US, which can be seen in plot 13. Also, the average of our fitted value is 47.09, 47.66, 47.91,

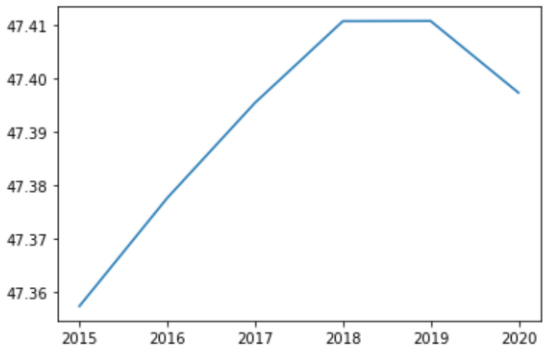


Figure 13: Estimate of US COPD mortality rate in the year 2015-2019.

47.51, whereas the true average value is 47.84, 47.44, 47.48, 47.83, the average error is 0.23.

5 Conclusion

Intrinsic patterns and group segmentation

Younger generation has less habit of smoking than the middle aged group, with the amount of difference almost at 10% during 2020. Elders have the lowest smoking prevalence and are usually the time when people decide to quit smoking. Social Classes also stays as an influential factor. People with higher income and/or education level tend not to smoke. The level of influence for the two indicators are almost the same, with about 20% of different between the highest social and the lowest social group.

Spacial Analysis

Presence of regional smoking culture becomes salient from our visualization as well as univariate and bivariate spatial correlation analysis of COPD mortality rate and present smoking prevalence. This aligns with geo-dependent features such as average income, educational level (as we just discussed) as well as features like yearly tobacco production (the top 3 states always have the most people smoking in recent years).

Future Campaigns

As of time series prediction shows, the US COPD mortality rate did not get better in recent years. A slight increase number of people still die due to tobacco-driven diseases. More attention should be brought to the group segments that have higher percentage of current-smokers, like middle-aged groups and lower social class groups. Younger generation has received good education on the harm of tobacco, but they should be encouraged to be consistent in staying away from it. For states that have a prevailing smoking culture, the government should involve to establish straight laws, and its citizens should be guided to have their own awareness of the harness of tobacco regardless of their surroundings.

6 Future Consideration

Policy influence

With the spatial correlation of smoking culture existing across states, we can further investigate detailed policy for different states regarding tobacco usage and production, including the tax and average price of tobacco per states by year. This can be done through geoenrichment for census areas (state/county level) and examining potential influential factors for the spatial correlation of smoking culture.

Usage of e-cigarette

The BRFSS combines usage of e-cigarette in the *tobacco_use_us* data. However, we suspect there being a different trend between tobacco vs. e-cigarette. While currently there are not enough data to show the premature death rate caused by direct usage of e-cigarette, it is still important to analyze it separately and get the trend in advance.