A Spatial Analysis of the Frequency of Oil Wells on Cancer and Asthma Rates in Pennsylvania

Parker Johnson

Abstract: Across the US, the desire to divest from fossil fuels is growing, but Pennsylvania continues to be home to many of the oil wells in the US. The public outcry against the fossil fuel industry has been heard around the world, with demands for the industry to be held accountable for the environmental and health damages they have caused. Previous literature has identified proximity to oil wells being linked to increases in asthma exacerbations and cancer diagnoses, as well as oil wells being located more often in poorer neighborhoods. This paper builds upon this previous research to build spatial models for both cancer and asthma rates in Pennsylvania counties based on socioeconomic characteristics. Four spatial models were considered, all using 5 different neighborhood matrices, to find the best model and neighborhood matrix combination to model the disease rates. The spatial Durbin model was the best model for modeling cancer rates, while the spatial error model was best for asthma rates, and the row-standardized neighborhood matrix based on spatial adjacency was the best matrix for both models. Additionally, the percent of white people in a county was found to be a statistically discernible predictor for both cancer and asthma rates.

Introduction

For individuals concerned with the climate crisis, 2030 is an important year by which carbon emissions need to be reduced to limit the effects of global warming (IPCC 2018). As a result, the desire to diverge from fossil fuel use is becoming ever present in public discussion. However, many portions of the US still have large numbers of fracking and drilling locations open for oil and gas extraction. Since the fossil fuel industry is a dominant force in certain regions of the country, studies have been conducted regarding the health impacts of these locations, finding health impacts such as the pollution from fossil fuel generation increasing the risk for respiratory diseases such as asthma (Hendryx 2019), and contamination from oil production increasing the risk for cancer in males (Sebastián 2001). However, information regarding these health impacts is not often widely disseminated to the vulnerable individuals the fossil fuel industries affect the most. Additionally, a 2015 study found that the poor in Pennsylvania are unequally exposed to pollution from unconventional gas wells (Ogneva-Himmelberger 2015), and while communities with certain socioeconomic characteristics house the most wells, the significance of burdening these is not often discussed. This paper aims to investigate how socioeconomic factors and proximity to oil and gas drilling sites contribute to asthma and cancer rates, and if these factors contribute differently between asthma and cancer rates.

This paper will build off of previous research surrounding the health impacts that oil and gas facilities have on its communities. A 2016 study conducted by the Johns Hopkins Bloomberg School of Public Health found that those living near natural gas fracking wells in Pennsylvania were more likely to have asthma exacerbations than those living farther from these sights (Rasmussen 2016). Additionally, a 2020 study on oil refinery proximity and cancer rates found a statistically discernible increased risk in cancer diagnosis rates across all types of cancer included in the study (Williams 2020). Finally, a 2021 paper from the Energy Research & Social Sciences journal found that more oil and gas related complaints were filed for higher median household income areas, and the odds of a complaint yielding a positive determination was lower in more marginalized communities (Clark 2021). Therefore, this could be motivation for there being more oil wells in poorer areas that are less likely to file complaints, and there being more adverse health effects for wells in marginalized communities. With these strong findings, this paper seeks to further investigate these previous studies.

Data

The key variables in this study are the number of oil wells at the county level, the cancer and asthma emergency department visit rates at the county level, and the census variables being analyzed at the county level. The census variables being analyzed are percent white, percent unemployed, median household income, and total population from the 2021 US Census, as motivated by the Clark 2021 study (U.S. Census Bureau 2022). The number of oil wells per county were collected by FracTracker in collaboration with Earthworks; this 2021 update on the number of oil wells includes all conventional and historical oil wells, totalling 100,956 wells in Pennsylvania (Jackson 2021). It is not uncommon for these wells to be located on farms, in forests, or in people’s backyards, and most of these wells are low-producing wells that have been neglected to be plugged (Jackson 2021). A log transformation for the number of oil wells per county will be used, as there are some counties in western Pennsylvania that have very high numbers of wells. The cancer rates were collected by State Cancer Profiles, a collaboration between the National Cancer Institute and the CDC, and measures all reported incidences of cancer per 10,000 residents within each county (State Cancer Profiles 2023). The asthma rates were collected by the Pennsylvania Asthma Surveillance System, and measures the number of asthma related emergency department visits per 10,000 residents within each county (Department of Health 2023). Finally, the census variables were collected at the county level, since the asthma and cancer rates were recorded at this level (U.S. Census Bureau 2022).

Methods

For this analysis, data was first aggregated to the county level. A log transformation for the number of oil wells per county will be used, as there are some counties in western Pennsylvania that have very high numbers of wells. It’s worth noting that many counties, especially in eastern Pennsylvania, have no oil wells, so in order to use a log transformation on this variable, I added 1 to the number of oil wells in each county before performing the transformation. An indicator variable was also added based on whether a county is part of eastern or western Pennsylvania, given the stark differences in the number of oil wells between the two regions of the state. Also, an interaction term was added between the log number of wells and the indicator variable, as there appears to be a large relation between the number of oil wells in a county and whether the county is in western or eastern Pennsylvania. Additionally, the variance inflation factor (VIF) was checked between each variable to ensure that variables didn’t have high collinearity between one another. For this measure of collinearity, one accepted cutoff for too high of collinearity is for VIF values greater than 10 (Ferré 2009), which would indicate that this variable should be removed from analysis. Since all of the variables included had VIF values less than 10, no variables demonstrated too high of collinearity between one another, so no variables had to be removed from this analysis.

We then needed to conduct a Moran’s I test on the residuals of a linear regression model to determine if there is statistically discernible clustering of the residuals for the disease rates per Pennsylvania counties, after modeling them using the log number of oil wells and socioeconomic characteristics. This process was first conducted to model cancer rates, then the same process was followed to model asthma rates. The value for our Moran’s I test is the slope between the residual for the disease rates of a given Pennsylvania county and the average of residuals for the disease rates per neighboring Pennsylvania counties (Bivand 2013, p. 276); therefore, in this case, it’s the slope between these 67 pairs of values for each Pennsylvania county.

To test for spatial dependence, we use a Monte Carlo test in which the disease rate values are randomly assigned to counties for the purpose of simulating complete spatial randomness, and the Moran’s I value is calculated from this many times, being 5000 for this analysis (Bivand 2013, p. 281). We then compare our observed Moran’s I value to these many permutations to compute a p-value to determine the probability that our observed Moran’s I value occurred by chance, where the null hypothesis we’re testing is that there’s no spatial structure in the residuals (I = 0), and the alternative hypothesis is that there is a spatial structure in the residuals (I > 0). For p-values below .05, we reject the null hypothesis that there is no spatial structure in the residuals, so we conclude that there is statistically discernible evidence of clustering in the residuals of cancer or asthma rates in Pennsylvania counties, which gives us evidence to consider different spatial models to apply to our data. For p-values between .05 and .1, we will still reject the null hypothesis, but it’s worth noting that these models do not give as strong of evidence to consider different spatial models compared to models that produce p-values below .05. This gives us evidence to consider different spatial models because if we have statistically discernible evidence for a spatial structure to the error terms in our linear model, the assumption that error terms are independent and identically distributed is violated, so we have evidence for spatial autocorrelation present in the data.

In order to use spatial regression, a neighborhood matrix needs to be used. To find the most optimal spatial model, multiple neighborhood matrices were tested, including a row-standardized matrix and binary matrix based on spatial adjacency, and different k-nearest neighbors matrices. Instead of being based off of all of the surrounding counties from a given county, which are what the first two neighborhood matrices are based off of, k-nearest neighbors is based off of the k-nearest neighbors from a given county (Varga 2017), and in this analysis, we will use a row-standardized matrix based off of different k-nearest neighbors. Since oil emissions contain both the particles PM2.5 and PM10 (California Air Resources Board) which can travel very different distances in the air (EPA 2021), we will consider 3 different k-nearest neighbors matrices to represent this possible range of the impact of emissions, specifically a 3-nearest neighbors, 5-nearest neighbors, and 10-nearest neighbors matrix.

With our neighborhood matrices specified, we can now consider different spatial models to explore the spatial variability in our data. The four models that are explained below were investigated, as they are all common spatial regression models used for data that is aggregated. The first spatial model considered is the spatial lag model, in which spatial dependence is based on the neighboring disease rates (Bivand 2013, p. 305), and takes the form

where y is the disease rate, ⍴ is the quantified dependence on the neighboring disease rates, W is the neighborhood matrix, X is matrix notation for the census variables and variable for the log number of wells, β is matrix notation for the coefficients of our variables, and e is random error.

The second spatial model considered is the spatial Durbin model, which is similar to the spatial lag model, but now spatial dependence is also accounted for through the explanatory census variables and log number of wells in each neighboring county (Bivand 2013, p. 305). The form it takes is

where 𝛾 is the quantified dependence on the neighboring explanatory variables. Since we have multiple explanatory variables, 𝛾 represents a value for each explanatory variable.

The third spatial model considered is the spatial error model, in which spatial dependence is accounted for through the effect of residuals on surrounding counties on the disease rates (Bivand 2013, p. 305). The form it takes is

where λ is the quantified dependence of the effect of residuals on surrounding counties on the disease rates.

The final spatial model considered is the conditional autoregressive model, in which spatial dependence is accounted for through the disease rate residuals of the immediate neighbors of a given county (Bivand 2013, p. 298). The form of the mean function is

where yj~i is the disease rate of the immediate neighboring counties, and the mean of the function is essentially a spatial error model. While the details of the conditional autoregressive model are beyond the scope of this analysis, the error term of this model is not independent, and allows for this memoryless property of only the immediate neighboring counties having an effect on the model, as opposed to the other models in which counties that are further away simply have a weaker effect on the disease rate of a given county.

With these four specified models and five specified neighborhood matrices, we can test the Moran’s I for each neighborhood matrix, and then use all statistically discernible neighborhood matrices to test each spatial model. To compare each model to one another, the Akaike’s Information Criterion (AIC) value for each model will be compared. The AIC value is the weighted sum of the log-likelihood of the model and the number of fitted coefficients (Bivand 2013, p. 298), where the lower the AIC value is, the better fit a model is to the data. The model and associated neighborhood matrix with the best AIC value will be the final model chosen, where the statistical discernibility of each explanatory variable will be assessed with p-values. This process was conducted for first cancer, and then asthma rates.

Results

In order to assess any spatial distribution of the log number of oil wells in Pennsylvania counties, a choropleth can be used as initial motivation for graphical analysis. A choropleth of the log number of oil wells and median household income can be seen in Figure 1, in which it appears that there is a higher number of wells in Western Pennsylvania, compared to Eastern Pennsylvania that has almost no wells, motivating a numerical test for spatial correlation. Additionally, western Pennsylvania appears to have lower median household income values compared to most eastern Pennsylvania counties.

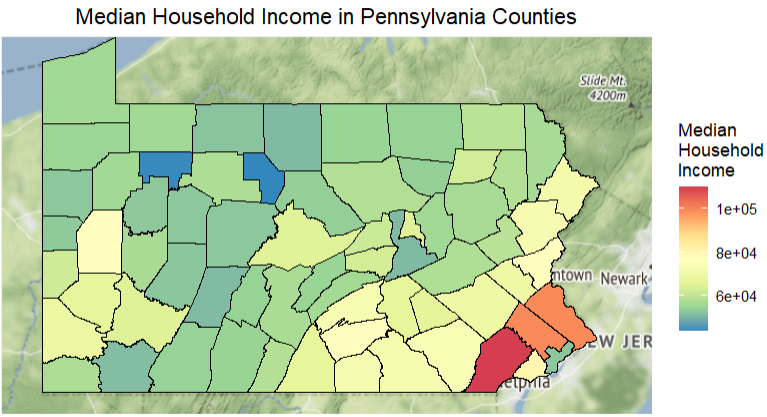
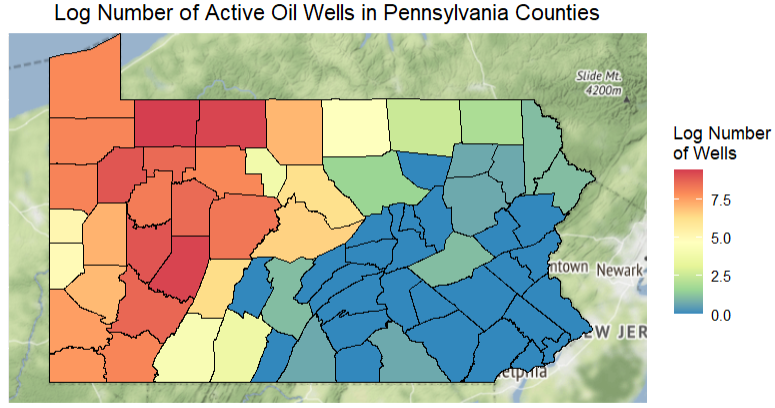


Figure 1: Number of oil wells and median household income in Pennsylvania counties

Next, we want to assess the spatial distribution of the number of cancer incidents and the number of asthma emergency department (ED) visits. Choropleths of these disease variables can be seen in Figure 2 below. It is not as clear from these choropleths whether or not there is spatial association between counties for cancer and asthma cases, motivating the need to perform some form of numerical analysis to test if there is spatial correlation present. There appears to be slightly higher numbers of asthma emergency department visits in eastern Pennsylvania counties, but more analysis is needed. From these initial choropleths, it’s hard to determine if there’s an association between the log number of oil wells in a county and socioeconomic variables to either of our disease variables, so further analysis is necessary.

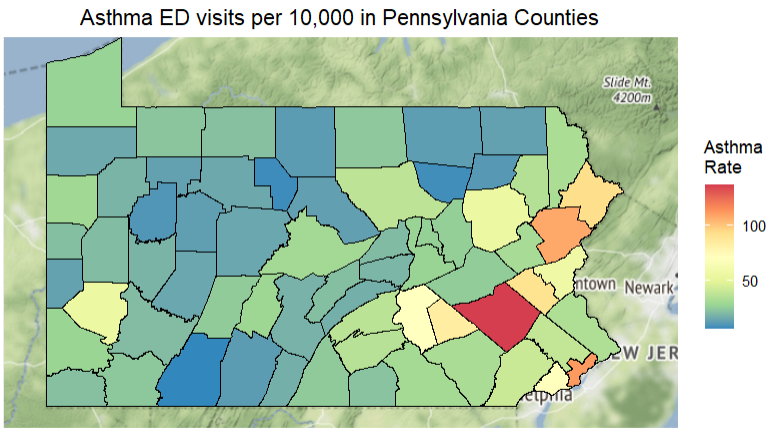
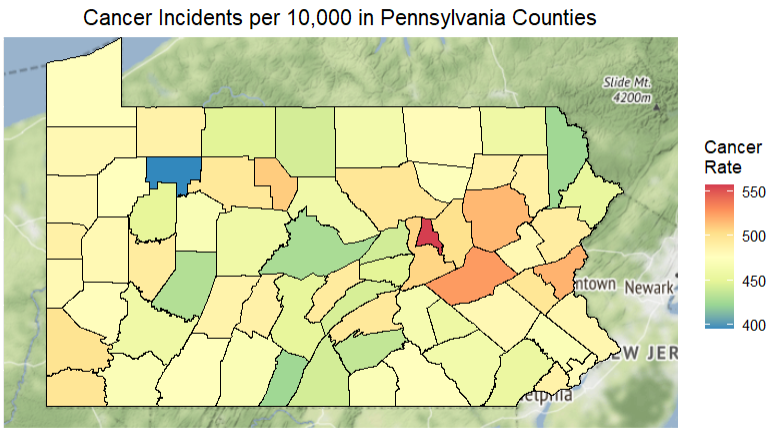


Figure 2: Cancer incidents and asthma ED visits per 10,000 residents in Pennsylvania counties

Now that we have visualized the spatial distribution of some of our variables and our disease rates, we can consider our numerical analysis for testing for spatial autocorrelation. In order to assess the significance of the neighborhood matrices, the Moran’s I test on the residuals of a linear regression model were first tested, in which cancer rates were being modeled. The p-values for each Moran’s I test can be found in Table 1 below.

Table 1: Moran’s I test p-values for each neighborhood matrix

|  | Binary, spatial adjacency | Row-standardized, spatial adjacency | 3-nearest neighbors row-standardized | 5-nearest neighbors row-standardized | 10-nearest neighbors row-standardized |
| --- | --- | --- | --- | --- | --- |
| p-values | 0.0228 | 0.015 | 0.09778 | 0.1724 | 0.02919 |

From Table 1, we can see that with a 5-nearest neighbors row-standardized neighborhood matrix, there is not enough evidence to determine that there is statistically discernible clustering of the residuals for the cancer rates per Pennsylvania counties, so we will not consider this neighborhood matrix for the spatial models for cancer rates. Additionally, a 5-nearest neighbors row-standardized neighborhood matrix only provides slight evidence for statistically discernible clustering of the residuals for the cancer rates per Pennsylvania counties, but we will still consider this neighborhood matrix for our spatial models. The remaining neighborhood matrices do have enough evidence to determine that there is statistically discernible clustering of the residuals for the cancer rates per Pennsylvania counties, so each spatial model will be paired with each remaining four neighborhood matrices to find the best fitting model.

To find the best fitting model for cancer rates based on our census variables and the log number of oil wells in each county, the AIC values of each model can be seen in Table 2 below.

Table 2: AIC values for each spatial model and neighborhood matrix pairing

|  | Binary, spatial adjacency | Row-standardized, spatial adjacency | 3-nearest neighbors row-standardized | 10-nearest neighbors row-standardized |
| --- | --- | --- | --- | --- |
| Spatial lag model | 642.30 | 637.70 | 641.19 | 641.06 |
| Spatial Durbin model | 646.03 | 637.51 | 643.09 | 641.22 |
| Spatial error model | 639.36 | 638.06 | 641.37 | 640.81 |
| Conditional autoregressive model | 639.77 | 642.81 | 643.06 | 643.08 |

From Table 2, we can see that the spatial error model performs quite well compared to the other spatial models since its AIC values are generally lower compared to the other models, and the neighborhood matrix that performs the best is the row-standardized neighborhood matrix based on spatial adjacency compared to the other neighborhood matrices. We also see that the conditional autoregressive model generally performs worse compared to the other models, indicating that the residuals in the immediate surrounding counties do not model the cancer rate of a given county well. The best model here based on the AIC values is the spatial Durbin model based on the row-standardized neighborhood matrix based on spatial adjacency. However, the AIC values are all very similar to one another, so no model greatly outperforms any of the other models. Regardless, the coefficients and associated p-values for our final model for cancer rates can be seen below in Table 3:

Table 3: Coefficients and p-values for the best equation for modeling cancer rates

| Variable | Estimated value | p-value |
| --- | --- | --- |
| Intercept term | -228.63 | 0.372 |
| Percent white | 171.36 | 0.00539 |
| Percent unemployed | 499.19 | 0.173 |
| Median household income | -0.000207 | 0.624 |
| Total population standardized | 55.000 | 0.0796 |
| Log number of wells | -6.029 | 0.00760 |
| Eastern PA Indicator | 9.663 | 0.524 |
| I(Log number of wells \* Eastern PA Indicator) | -7.730 | 0.514 |
| Gamma: percent white | 349.37 | 0.0534 |
| Gamma: percent unemployed | 1796.4 | 0.0150 |
| Gamma median hh income | -0.000505 | 0.603 |
| Gamma total pop. standardized | 118.56 | 0.157 |
| Gamma log number of wells | 4.418 | 0.179 |
| Gamma eastern PA indicator | 57.613 | 0.0842 |
| Gamma I(Log number of wells \* Eastern PA Indicator) | -77.537 | 0.00158 |

Using the coefficients for each variable from Table 3, and using the value for ⍴ which was found to be 0.2610, our final model for cancer rates can be seen below:



where X is the variable, β is the coefficient for the variable, W is the row-standardized neighborhood matrix based on spatial adjacency, and γ is a vector of the gamma coefficients listed in Table 3 for each of the variables.

As we can see from Table 3 in the p-values below 0.05, the percent of white people in a county and the log number of wells in a county have statistically discernible effects on the rate of cancer in Pennsylvania counties. We see that the percent of white people in a county have a statistically discernible positive effect on the rate of cancer in Pennsylvania counties, while the log number of wells in a county have a statistically discernible negative effect on the cancer rates in Pennsylvania counties. It’s worth noting that the value for ⍴ is not statistically discernible, which means that the average of the neighboring cancer rates do not have a statistically discernible effect on the rate of cancer within the given Pennsylvania county. These results seem a bit counterintuitive; we would expect oil wells to have a positive effect on cancer rates, and for the neighboring cancer rates to have a positive discernible effect on the rate of cancer in a given county. While it might make sense that counties with a high percentage of white people are more likely to be located in rural areas, resulting in more oil wells to possibly increase the rate of cancer in the county, the odd results could be due to caveats with the data.

In order to compare models for cancer and asthma rates, we now need to create models for asthma rates. Once again, in order to assess the significance of the neighborhood matrices, the Moran’s I test on the residuals of a linear regression model were tested. The p-values for each Moran’s I test can be found in Table 1 below.

Table 4: Moran’s I test p-values for each neighborhood matrix for asthma rates

|  | Binary, spatial adjacency | Row-standardized, spatial adjacency | 3-nearest neighbors row-standardized | 5-nearest neighbors row-standardized | 10-nearest neighbors row-standardized |
| --- | --- | --- | --- | --- | --- |
| p-values | 0.183 | 0.07159 | 0.1948 | 0.06859 | 0.7271 |

From Table 4, we can see that with a binary spatial adjacency, 3-nearest neighbors row-standardized neighborhood matrix, and 10-nearest neighbors row-standardized neighborhood matrix, there is not enough evidence to determine that there is statistically discernible clustering of the residuals for the asthma emergency department visit rates per Pennsylvania counties, so we will not consider these neighborhood matrices for the spatial models for asthma rates. Additionally, the row-standardized neighborhood matrix by spatial adjacency and the 5-nearest neighbors row-standardized neighborhood matrix only provides slight evidence for statistically discernible clustering of the residuals for the asthma rates per Pennsylvania counties, but we will still consider these neighborhood matrices for our spatial models.

To find the best fitting model for asthma emergency department visit rates based on our census variables and the log number of oil wells in each county, the AIC values of each model can be seen in Table 5 below.

Table 5: AIC values for each spatial model and neighborhood matrix pairing for asthma rates

|  | Row-standardized, spatial adjacency | 5-nearest neighbors row-standardized |
| --- | --- | --- |
| Spatial lag model | 577.47 | 577.75 |
| Spatial Durbin model | 582.45 | 579.87 |
| Spatial error model | 576.5 | 576.59 |
| Conditional autoregressive model | 577.09 | 577.4 |

From Table 5, we once again see that the spatial error model performs quite well compared to the other spatial models, and the neighborhood matrix that performs the best is against the row-standardized neighborhood matrix based on spatial adjacency compared to the other neighborhood matrices. We also see that the spatial Durbin model generally performs worse compared to the other models, indicating that the census variables and log number of oil wells in surrounding counties do not model the asthma emergency department visit rate of a given county well. The best model here based on the AIC values is the spatial error model based on the row-standardized neighborhood matrix based on spatial adjacency. However, the AIC values are again all very similar to one another, so no model greatly outperforms any of the other models. Regardless, the coefficients and associated p-values for our final model for asthma rates can be seen below in Table 6:

Table 6: Coefficients and p-values for the best equation for modeling asthma rates

| Variable | Estimate | p-value |
| --- | --- | --- |
| Intercept term | 144.27 | 0.000564 |
| Percent white | -173.36 | 1.086\*10-6 |
| Percent unemployed | 446.16 | 0.0312 |
| Median household income | -0.000247 | 0.253 |
| Total population | -7.64\*10-6 | 0.514 |
| Log number of wells | -0.209 | 0.851 |
| Eastern PA Indicator | 14.218 | 0.119 |
| I(Log number of wells \* Eastern PA Indicator) | -9.639 | 0.199 |

Using the coefficients for each variable from Table 6, and using the value for λ which was found to be 0.3047, our final model for asthma emergency department visit rates can be seen below:



where X is the variable, β is the coefficient for the variable, and W is the row-standardized neighborhood matrix based on spatial adjacency.

As we can see from Table 6 in the p-values below 0.05, the percent of white people in a county and the percent of unemployed people in a county have statistically discernible effects on the rate of asthma emergency visits in Pennsylvania counties. We see that the percent of white people in a county have a statistically discernible negative effect on the rate of asthma emergency visits in Pennsylvania counties, while the percent of unemployed people in a county have a statistically discernible positive effect on the rate of asthma emergency visits in Pennsylvania counties. All of the other variables do not have statistically discernible effects on the rate of asthma emergency visits in Pennsylvania counties. It’s worth noting that the value for λ is not statistically discernible, which means that the average of the residuals of the surrounding asthma rates do not have a statistically discernible effect on the rate of asthma within the given Pennsylvania county. While it seems a bit counterintuitive that the number of oils wells and the average of the residuals of the surrounding asthma rates do not have statistically discernible effects on the rate of asthma in Pennsylvania counties, it seems to make sense that counties with high levels of unemployment experience higher rates of asthma ED visits, and counties with smaller percentages of white people also experience higher rates of asthma ED visits.

Comparing the models found for cancer and asthma rates, we see some notable differences. While the differences in AIC values were very small so the model choice was not extremely important, we see that a spatial Durbin model fit better for modeling cancer rates, while a spatial error model fit better for modeling asthma rates in Pennsylvania. However, a row-standardized neighborhood matrix based on spatial adjacency was the best neighborhood matrix to use for both modeling cancer and asthma rates. A key difference between the two models is that the percent of white people in a county has a statistically discernible positive effect on the rate of cancer, but a statistically discernible negative effect on the rate of asthma ED visits in Pennsylvania counties. This result does seem counterintuitive, as we would expect disease rates to be modeled similarly, but this could be due to the possibility of asthma being more present in urban areas while cancer is also fairly present in rural areas. Another interesting finding is that the models for asthma struggled to show significant evidence for spatial autocorrelation compared to the models for cancer, yet the lower AIC values within the asthma models suggests that the census variables included are better predictors for asthma rates than cancer rates in Pennsylvania counties. Finally, we see that median household income does not appear to be a good predictor of both cancer and asthma rates in Pennsylvania counties.

Discussion

From these results, we find that the Spatial Durbin model is the best fit for modeling cancer rates, while the spatial error model is the best fit for modeling asthma ED visit rates, but the row-standardized neighborhood matrix based on spatial adjacency was the best neighborhood matrix for both cases. We also find that the percent of white people in a county have differing spatial effects on the disease rates. The implications of these findings are that asthma and cancer rates, despite both being diseases that can arise from pollution from the fossil fuel industry, still need to be modeled differently from one another, and socioeconomic characteristics predict cancer rates differently from asthma ED visit rates. We also find that the number of oil wells in a county does not have a positive effect on either cancer or asthma rates, which is in contradiction to the previous literature. Additionally, we find that median household income is not a discernible predictor for either cancer or asthma rates, so while it could still be the case that there are more oil wells located in poorer Pennsylvania counties, the effects of these wells do not cause statistically discernible increases in cancer or asthma rates; however, these contradictory findings could be due to limitations in the data.

There are many limitations that impact the strength of this study. One possible limitation is that there might be variables that are unaccounted for that have a significant impact on cancer rates. For example, smoking is often linked to higher rates of cancer, especially lung cancer (Proctor 2001), so if smoking is a more important predictor for cancer than these variables, and smoking rates are different between Pennsylvania counties, this could be causing these models to poorly model cancer rates. Therefore, further research should be conducted to determine if there is data available on smoking rates within each county, to account for this variable that could be causing poor model fitting. Unaccounted variables could also be causing some correlation between the variables included in the model, which could explain why many of the variables do not show statistically discernible evidence, so addressing these unaccounted variables could reduce the correlation between variables and result in better fitting variables overall.

Another limitation is that areal data at the county level was used for this analysis. By aggregating the number of oil wells to each county, we lose information about where the oil wells are specifically located; for example, oil wells near the border of a county might have greater health impacts on the neighboring county than the county the well is in, and information is lost about if there are clusters of wells located anywhere within a county, which could lead to inaccurate results about the true health impacts oil wells have on individuals. Since the proximity to oil wells is likely more significant in predicting health impacts compared to the number of wells within a county, future work could attempt to fit either a non-homogeneous Poisson process model or a log-Gaussian Cox process model to this data, to account for the distance from an oil well. Additionally, future work could involve searching for asthma and cancer data at a more precise geographic level, such as at the census tract level, to allow for more granular values to produce more accurate estimates of these disease rates as they relate to oil wells. By using a more precise and granular model, this could help account for urban centers within each county that are currently not accounted for by using areal data at the county level.

Another limitation is that the presence of nearby oil wells might only impact a certain type of cancer, and therefore overall cancer rates might be too broad. Since bladder and colon cancer have been previously found to be linked to proximity to oil refineries (Williams 2020), perhaps cancer rate should be limited to just these types of cancers, as other cancers might not be as widely studied as these two types of cancers as they relate to proximity to oil refineries. Therefore, future work should conduct a more in-depth investigation into which types of cancers have been linked to proximity to oil refineries, and only create cancer rates based on these cancers. This could also help improve the accuracy of our fitted cancer models if some types of cancers are a better fit for our model compared to other cancer types.

One last limitation is that the dataset that collected oil locations might define what an oil well is too loosely, and very small operations that wouldn’t impact human health heavily might be included in the data, so looking more closely into how the oil well data was collected could be beneficial. Additionally, the data used for this analysis did not include unconventional wells such as fracking operations, and only included conventional oil wells. Since the Rasmussen 2016 study found an increase in asthma exacerbations for those living near natural gas fracking wells in Pennsylvania, finding data that includes these types of wells could be beneficial to finding a more accurate association between oil wells and disease rates.

Adding on to the future work listed within the limitations, future work could also consider only focusing on western Pennsylvania for the spatial domain, since there are much more oil wells in this region compared to eastern Pennsylvania. Additional variables related to census information could also be investigated, such as the percent of people in a given region that have health insurance, or if education level is associated with disease rates. Finally, the findings within Pennsylvania could be compared to other US states, to see if disease rates are modeled differently for different states across the country.

Works Cited

Bivand, Roger S., Edzer Pebesma, and Virgilio Gómez-Rubio. Applied Spatial Data Analysis with R. 2nd ed., Springer, 2013.

“California Air Resources Board.” *Inhalable Particulate Matter and Health (PM2.5 and PM10) | California Air Resources Board*, ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health. Accessed 18 May 2023.

Clark, C. J., Warren, J. L., Kadan-Lottick, N., Ma, X., Bell, M. L., Saiers, J. E., & Deziel, N. C. (2021). Community concern and government response: Identifying socio-economic and demographic predictors of oil and gas complaints and drinking water impairments in Pennsylvania. *Energy Research & Social Science*, *76*, 102070. https://doi.org/10.1016/j.erss.2021.102070

Ferré, J. “Regression Diagnostics.” *Comprehensive Chemometrics*, 2009, pp. 33–89, https://doi.org/10.1016/b978-044452701-1.00076-4.

Hendryx, Michael, et al. “Air Pollution Exposures from Multiple Point Sources and Risk of Incident Chronic Obstructive Pulmonary Disease (COPD) and Asthma.” *Environmental Research*, vol. 179, 2019, p. 108783, https://doi.org/10.1016/j.envres.2019.108783.

IPCC. “Summary for Policymakers of IPCC Special Report on Global Warming of 1.5°C Approved by Governments.” *IPCC Summary for Policymakers of IPCC Special Report on Global Warming of 15C Approved by Governments Comments*, 8 Oct. 2018, www.ipcc.ch/2018/10/08/summary-for-policymakers-of-ipcc-special-report-on-global-warming-of-1-5c-approved-by-governments/.

Jackson, Erica, et al. “Pennsylvania Conventional Well Map Update.” *FracTracker Alliance*, 28 May 2021, www.fractracker.org/2021/05/pennsylvania-conventional-well-map-update/.

Ogneva-Himmelberger, Yelena, and Liyao Huang. “Spatial Distribution of Unconventional Gas Wells and Human Populations in the Marcellus Shale in the United States: Vulnerability Analysis.” *Applied Geography*, vol. 60, 2015, pp. 165–174, https://doi.org/10.1016/j.apgeog.2015.03.011.

“Particulate Matter (PM) Emissions." Environmental Protection Agency, Report on the Environment, 2021, cfpub.epa.gov/roe/indicator\_pdf.cfm?i=19.

Proctor, Robert N. “Tobacco and the Global Lung Cancer Epidemic.” *Nature Reviews Cancer*, vol. 1, no. 1, 2001, pp. 82–86, https://doi.org/10.1038/35094091.

Rasmussen, Sara G et al. “Association Between Unconventional Natural Gas Development in the Marcellus Shale and Asthma Exacerbations.” *JAMA internal medicine* vol. 176,9 (2016): 1334-43. doi:10.1001/jamainternmed.2016.2436

Sebastián, M San, et al. “Exposures and Cancer Incidence near Oil Fields in the Amazon Basin of Ecuador.” *Occupational & Environmental Medicine*, 1 Aug. 2001, oem.bmj.com/content/58/8/517.short.

*State Cancer Profiles*. State Cancer Profiles > Incidence Rates Table. (n.d.). Retrieved May 2, 2023, from https://www.statecancerprofiles.cancer.gov/incidencerates/index.php?stateFIPS=42&areatype=county&cancer=001&race=00&sex=0&age=001&stage=999&year=0&type=incd&sortVariableName=rate&sortOrder=default&output=0#results

*Surveillance reports*. Department of Health. (n.d.). Retrieved May 2, 2023, from https://www.health.pa.gov/topics/programs/Asthma/Pages/Surveillance-Reports.aspx

U.S. Census Bureau. “American Community Survey 1-Year Data (2005-2021) [API].” U.S. Department of Commerce, Sept. 2022.

Varga, Viliam, Jan Hacka, and Jana Mullerova. "Using k-Nearest Neighbors for the Estimation of Forest Characteristics from LiDAR Data." Forests, vol. 8, no. 9, Sept. 2017, pp. 322.

Williams, Stephen B et al. “Proximity to Oil Refineries and Risk of Cancer: A Population-Based Analysis.” *JNCI cancer spectrum* vol. 4,6 pkaa088. 7 Oct. 2020, doi:10.1093/jncics/pkaa088