Alex Zarley
Geog 777

Project 1 Report

A potential cancer risk to adults from nitrate and nitrite in drinking water has recently been identified. The purpose of this analysis is to explore the relationship between nitrate levels in drinking water and cancer occurrence in adults. The data used to perform this analysis is a point shapefile of well locations throughout Wisconsin with the sampled nitrate levels in parts per million (ppm), along with a Wisconsin census tract polygon shapefile with adult cancer rates already calculated. To explore the relationship between nitrate levels in well water and cancer rate in adult Wisconsinites, I aggregated the nitrate level ppm and cancer rate to the same spatial unit and performed Ordinary Least Squares (OLS) regression analysis. I created an application that allows the user to run the analysis using different k-values for Inverse Distance Weighting (IDW) spatial interpolation of nitrate levels from points and specify the output area of each hexbin in the output layer. Several statistics for the model and individual coefficients are calculated and displayed in the application. All shapefiles and statsitics generated from the analysis can be saved to a user-specified directory.

The goal of this project for me was to use only open source Python packages. The GUI was built with tkinter and maps are drawn with geoplot and matplotlib. Data and statistical analysis were performed with pandas and statsmodels packages, and geoprocessing was done with GDAL, geopandas, and shapely. All datasets generated during analysis are stored in memory and can be saved by the user after the analysis has run. I projected each dataset into UTM 16N before running the analysis.

The initial datasets for the project were in different spatial geometries. Nitrate levels were recorded as point locations throughout Wisconsin, while cancer rate was aggregated to census tract polygons. Each dataset needed to be abstracted into the same spatial geometry before performing regression analysis to determine what, if any, relationship exists between nitrate levels in drinking water and adult cancer occurrence at every location in Wisconsin. I chose to use a hexbin layer as the final aggregation unit. A user specifies the area of each hexbin, in square kilometers, from my app GUI before running the analysis. The hexbins are clipped to the census tract dataset extent before either variable is aggregated to them. For the well nitrate points, I used IDW

to create a raster layer estimating nitrate levels across the state. I then converted the raster layer to a point layer, with the point being at the center of each raster cell. To determine the nitrate level for each hexbin, I calculated the mean nitrate value for all IDW points within each hexbin after performing a spatial join between the nitrate IDW points and the hexbin polygons.

In IDW interpolation, samples closer to a given location will receive more weight than those farther away when calculating the estimated value. The relationship between distance from a sample and the weight given to that sample is non-linear in IDW, with the distance decay coefficient *k* specifying how quickly the weights given to sample points diminish with distance. A *k* value of 1 will produce a smoother interpolated surface because there will be a smaller difference in weighting between nearer and farther sample points. A higher *k* value (e.g., 3) will increase the influence of closer points on the interpolated surface and produce more localized "peaks". I used two methods to determine what the best value of *k* would be for IDW. The first method was a cross-validation of different *k* values in IDW. Cross-validation uses all supplied samples, removes each sample one at a time, and predicts what the value would be at that point's location based on the remaining samples. I could then determine which *k* value performed best by calculating the root mean square error (RMSE) for each interpolated surface and choosing the *k* value with the smallest RMSE. I performed IDW with *k* values of 1, 1.5, 2, 2.5, and 3. A *k* value of 2 produced the lowest RMSE, suggesting that is the appropriate value to use for IDW on the nitrate well point dataset.

The other method I used was to measure spatial autocorrelation in the input point dataset. Moran's I can be used to measure spatial autocorrelation in a dataset. The null hypothesis when testing for spatial autocorrelation with Moran's I is that there is no spatial autocorrelation in the observed values, or that the values in their locations occur due to random chance. The alternate hypothesis is that spatial autocorrelation does exist among the values. Moran's I values range from -1 to 1, with -1 meaning perfect negative spatial autocorrelation, 0 representing the absence of spatial autocorrelation, or a totally random distribution, and 1 representing perfect positive spatial autocorrelation. Moran's I for the nitrate level points is 0.748, which

represents a very positive spatial autocorrelation among values. The z-score is 89.76, and the p-value is 0.0. Since the z-score is very high and the p-value is very low, it indicates that it is very unlikely that the observed spatial pattern is caused by random chance and that the relatively high positive spatial autocorrelation is statistically significant. Since the positive spatial autocorrelation is so high for this dataset, a higher *k* value, such as 2 or 2.5 should be used. A higher *k* value will ensure nearer points are given a greater weight when interpolating values at unknown locations. Based on the results of these two exploratory methods, I determined a *k* value of 2 is appropriate for this analysis. Running the analysis with a *k* value of 2.5 produced similar results, and my conclusions were the same as with a k value of 2.

I used a weighted average by percent of hexbin area to calculate the cancer rate for each hexbin. I calculated the polygon intersection between the hexbin layer and the census tract cancer rate layer. I then calculated the area of each resulting intersection polygon. These intersection polygons have the unique ID of the hexbin and the unique ID of the census tract that created the intersection polygon. I divided this intersection area into the corresponding hexbin's total area to get the percent of the total area of the hexbin that the intersection represents. The total area of each hexbin was calculated after they were clipped to the census tract extent to make this calculation accurate for hexbins that were clipped on the edges of the state. I used this percentage as the weight to average the cancer rates of the census tracts intersecting each hexbin. I multiplied the census tracts' cancer rate by the percent of the total area of a given hexbin that census tract intersects to get the weighted cancer rate for that hexbin for that census tract, and then summed all the weighted cancer rates for each hexbin to determine the cancer rate for that hexbin. Using this area-weighted method provides a more accurate cancer rate for a hexbin than if I had calculated a simple average of census tract cancer rates that intersect a hexbin.

Once I had a cancer rate and a nitrate level calculated for each hexbin, I could perform regression analysis to determine what, if any, relationship exists between nitrate levels in well water and cancer rate. I chose to use a hexsize of 150 square kilometers for the analysis in this report. The values of the statistics

changed slightly with various hexsizes, but the significance and interpretation of the statistics and results did not change. I first attempted to use geographically weighted regression, but there was not enough variation in the dependent variable cancer rate to fit a model to the data. I confirmed this by calculating Moran's I for cancer rate for my hexbin layer and I = 0.68, with a z-score of 37.15 and a p-value of 0.0. This means I could reject the null hypothesis that there was no spatial autocorrelation, which indicates there is positive spatial autocorrelation, or clustering, among cancer rate in my hexbins. Next, I performed OLS regression analysis, with cancer rate as the dependent variable and nitrate level as the independent variable. Using a k value of 2 and hexbins of 150km2 produces an equation of:

*Canrate = 0.0239*nitrate level + 0.1411*

According to this model, for every 1 ppm increase of nitrate in well water, cancer rate increases by 0.0239, or the occurrence of cancer per 1,000 people increases by 23.9. The p-value for both the nitrate level coefficient (0.0239) and the intercept coefficient (0.1411) is 0, which means they are statistically significant, and I can reject the null hypothesis that each coefficient is not helping the model.

The RMSE is 0.1489, which means the average error for each predicted cancer rate is 0.1489, or 148.9 cancer occurrences per 1,000 people. Another measure of success of the fitted model in predicting cancer rates is the coefficient of determination (r^2). r^2=0.0922 for this model, which mean 9.22% of variance in cancer rates across the state of Wisconsin can be explained by nitrate levels in well water. This suggests that there are other important variables missing from this model that could help explain the variation in cancer rates across the state.

The null hypothesis for evaluating the model as a whole is that the explanatory variable (nitrate level) in this model is not effective. The Joint F-Statistic measures the statistical significance of the overall model. The Joint F-Statistic for this model is 110.5681, with a p-value of 0.0. This means I can reject the null hypothesis that the model is not effective.

Alex Zarley
Geog 777

I tested if the model predictions are biased by performing a hypothesis test with the Jarque-Bera Statistic, and by calculating Moran's I on the residuals to determine if spatial autocorrelation exists among the error in the model. The Jarque-Bera Statistic was 393.912 with a p-value of 0.0. When the Jarque-Bera Statistic is far from 0, as in this case, it indicates the residuals are not normally distributed and are skewed. The residual distribution is skewed right in this model, suggesting it overpredicts cancer rates at a higher rate than it under predicts them. This means I can reject the null hypothesis that the model predictions are not biased and infer that some bias exists in the model because the residuals are not normally distributed. Moran's I for the residuals is 0.63, which indicates a moderately strong positive spatial autocorrelation. The p-value for this statistic is 0.0, which again means I can reject the null hypothesis that there is no spatial autocorrelation in the residuals. This method confirms the inference from the Jarque-Bera-Statistic that there is positive spatial autocorrelation, or clustering, among residual values from this model.

I can conclude from the above statistics that this model is statistically significant, as are each individual coefficient in the regression equation. This model, which only uses nitrate levels as a predictor, does not fully explain the observed variance in cancer rates across the state though. Because this model only captures 9.22% of the variance in cancer rates, and because there are clusters across the state where it systematically over- and under-predicts cancer rates, it is fair to assume that there are one or more factors that should be included in a regression model to more fully and accurately predict cancer rates across the state of Wisconsin.