Exploring the Relationship between Nitrate Levels and Cancer Occurrences in Wisconsin through an Interactive Application

By Aspen Neville

I. Introduction

Due to concerns that high nitrate levels in drinking water could be linked to cancer cases, the Wisconsin Department of Natural Resources were interested in evaluating that concern further and examining the magnitude of the cancer risk that nitrate levels pose. The agency collected nitrate levels from water wells drilled throughout the state as well as recorded the location of cancer occurrences over a ten-year period. Using that data, the department needed a way to run spatial analysis methods and view the results in order to evaluate the relationship between the nitrate levels and cancer occurrences.

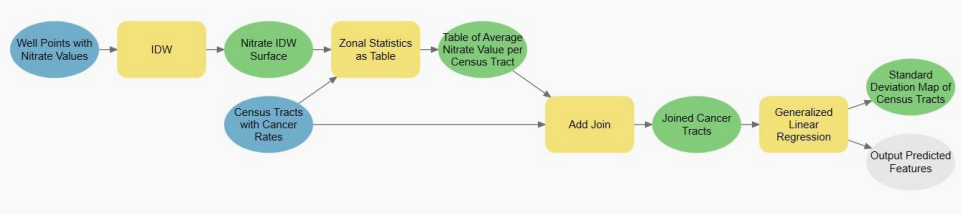
An application was built to run several spatial analysis methods based on user input for a certain value in an analysis method in order to end up with a linear regression equation of the relationship between average nitrate levels and cancer rates for census tracts in the state of Wisconsin. The application displayed the results in both a text output and a map for the user to see and evaluate the results with the option to save the results for further comparison to analysis runs with different input values.

II. Implementation

One of the first steps involved in implementing the application was to evaluate the data and determine the steps of analysis that would be used for exploring the relationship between the nitrate levels collected at well sample points and cancer occurrences in the state of Wisconsin. Since the data for the nitrate levels existed at discrete sample locations, the first step of the analysis needed to involve creating an interpolated surface of the values. With many spatial interpolation methods available, due to the assumption of spatial autocorrelation in the data (points closer together were more likely to have the similar nitrate levels in the water than those farther away) and for its ease of implementation, the inverse distance weighted method was chosen. Since no universal method existed for determining the distance-decay coefficient and the value of the distance-decay coefficient was a vital part of that method, the application needed to be designed to allow the user to change values for the distance-decay coefficient in order to explore which value was the best fit for the data.

The cancer occurrences data were previously aggregated to the level of census tract by taking the number of occurrences in a tract and dividing it by the tract's population. As such, for a comparison to be made between the nitrate values and the cancer rate using regression, the interpolated surface of nitrate values needed to be aggregated to the same spatial unit (census tract) as the cancer rate. It was decided that taking the average nitrate value in each census tract would best accomplish that goal. From there, linear regression was going to be used to explore the relationship. Linear regression was chosen as it was a simpler way to implement regression into the analysis and provided a good starting point to explore the relationship. A global linear regression model rather than a local model such as GWR, proved more effective for the data given the scale of the full state of Wisconsin. Additionally, a continuous linear regression model was used since the values for the cancer rate and nitrate levels could take on any range of numbers (not restricted to whole integers).

With the decisions made for the basics of the model, the spatial analysis was tested to determine the specific ArcPy methods to run the analysis. The methods were as follows (see diagram below):

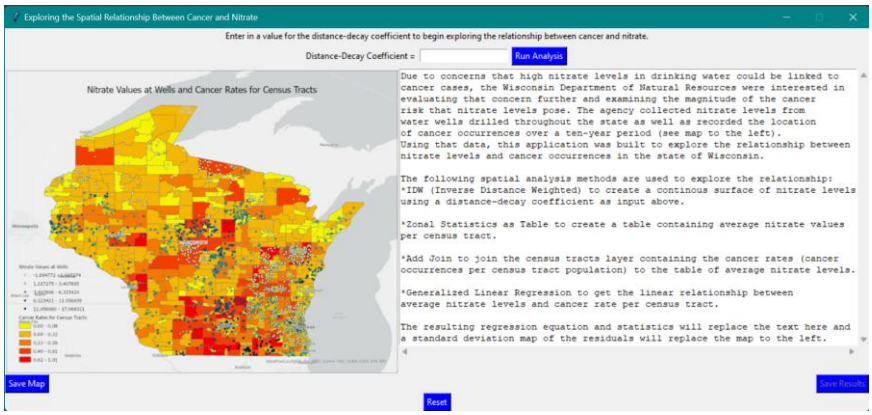


1. IDW (Inverse-Distance Weighted): Generated a continuous surface of nitrate values from the well points shapefile using the inverse-distance weighted method.[1]
2. Zonal Statistics as Table: Created a table of average nitrate values per census tract from the interpolated surface generated with the IDW method.[2]
3. Add Join: Joined the census tracts shapefile containing the cancer rates to the table of average nitrate values per census tract.[3]
4. Generalized Linear Regression: Determined the linear relationship between average nitrate value (exploratory/independent variable) and cancer rate (dependent variable) for each census tract, resulting in an output of a standard deviation map of residuals for each of the census tracts and analysis output text results containing the linear regression equation and statistic values to evaluate the statistical significance of the model.[4]

Another step of implementation involved determining how the application was to be built, including the tools leveraged to run the analysis. Python was chosen as the language for the application because of the spatial analysis capabilities available using the ArcPy library. To build the GUI and provide a front-end to present the analysis results as well as allow the user to run the model using different distance-decay coefficients for the inverse distance weighted portion of the analysis, the library of Tkinter was leveraged.

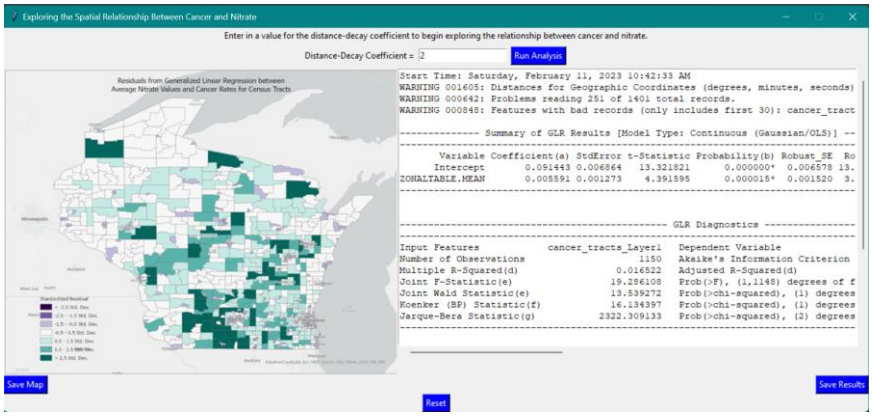
To execute the implementation and develop the application, the following steps were taken using Pycharm as the IDE:

1. Built GUI framework that provided a place for the user to input a value for the distance-decay coefficient (Tkinter entry widget with label) and run the analysis (Tkinter button widget) with places to display the maps (Tkinter label widget) and results (Tkinter text widget). The initial screen was set up to display an overview map and introduction text in the label and text widgets respectively (see below for initial state of GUI).



2. Imported ArcPy library and combined the Python code from each of the spatial analysis methods to run the whole analysis. Added in additional Tkinter window that appeared during analysis and closed when analysis was complete to keep the user informed of when the analysis was running.

3. Adjusted analysis so that it ran after the user clicked the appropriate button and used the user input into the entry widget in the GUI for the distance-decay coefficient. Also added in a way to check that the user input a proper value (number) into the entry widget (if not, the results text widget was set up to inform the user to input a number).
4. Added the map image into the label widget in the GUI and set it up so that the map image that was displayed updated based on the analysis results.
5. Set up the application so that the model output results wrote to the text widget in the GUI and updated the text widget to have scrollbars to fully view the results (see below for updated GUI after analysis was completed).



6. Added in buttons with functions for the user to save the map image and results text from each analysis run to their computer for further comparison outside of the application, as well as a button to return the GUI to its original state with the introduction text and overview map. The button to save the results text was disabled until the analysis was complete as it was not necessary with the introduction text.

III. Results

The application proved to be effective for running the spatial analysis and presenting the results in visual ways for the user to evaluate the linear relationship between cancer occurrences and average nitrate levels at the census tract level in the state of Wisconsin. To illustrate the use cases of the application in exploring the relationship between the variables of interest, the application was run by inputting the values of 1.5, 2, and 2.5 into the entry widget for the distance-decay coefficient and the resulting maps and results were saved to files after each run for comparison. As a rule of thumb, a reasonable distance-decay coefficient should be between 1.5 and 2.5 depending on the range of spatial autocorrelation in the data.[5] As such, those three values were chosen to show the results of the analysis for a range of reasonable values for the distance-decay coefficient.

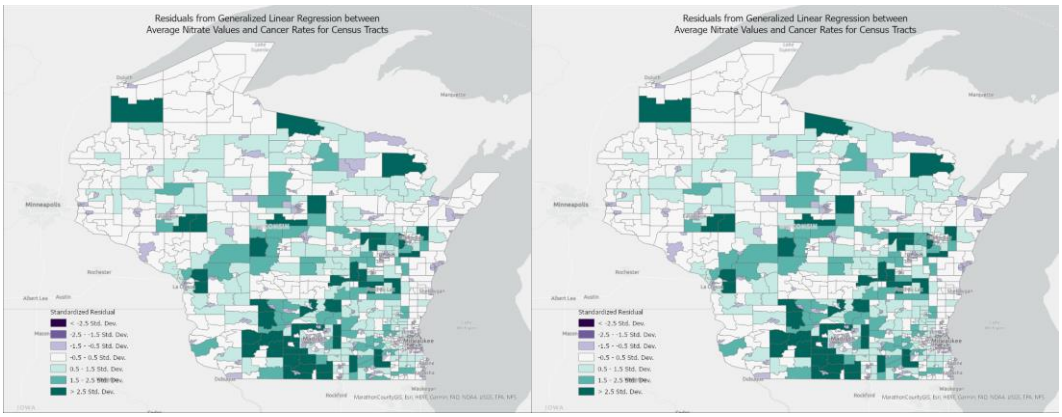
The model that used the distance-decay coefficient of 2 output the following linear regression equation representing the relationship between average nitrate value per census tract (x) and census tract cancer rate (y): $y = 0.005591x + 0.091443$. That equation stated that there was a positive relationship between the two variables with the census tract cancer rate expecting to increase by 0.005591 for every 1 unit increase in average nitrate value. Additionally, it implied that if the average nitrate value was zero, the cancer rate would be expected to be 0.091443. The equation for the other models with the coefficients of

1.5 and 2.5 also showed a positive relationship with values near the same and were respectively as follows: y = 0.005621x + 0.091309 and y = 0.005562x + 0.091570.

The statistical significance of the models and results were expanded upon through several statistical tests and were displayed in the application as part of the results output. All statistics were evaluated at the 99% confidence level to determine significance.

Statistic	**1.5 Model Value**	**2 Model Value**	**2.5 Model Value**
Coefficient t-Statistic	4.379927	4.391595	4.396106
R-squared	0.016436	0.016522	0.016556
Joint Wald	13.449050	13.539272	13.591450
Koenker (BP)	16.325631	16.134397	15.967977
Jarque-Bera	2318.720860	2322.309133	2325.353222

The r-squared value, which examined how fit the model was to the data and had a range of 0 to 1, was near 0.02 for all 3 models. Since the r-squared value was so close to zero, that suggested that the model was not a good fit for the data. Because the Koenker (BP) Statistic was significant, the Joint Wald rather than the Joint F-Statistic was consulted, along with the robust probabilities for the coefficients using the t-Statistic. The Joint Wald Statistic for all three models was significant, suggesting that the model was statistically significant and was unlikely to have occurred by chance. The robust probabilities for the coefficients were likewise significant, which implied that the average nitrate value was an effective variable in the model. The Koenker (BP) Statistic were all significant, suggesting that there was statistically significant heteroscedasticity (i.e., the effect of average nitrate value on cancer rate varied based on magnitude of average nitrate value) and/or nonstationarity (i.e., the spatial processes involved in the average nitrate value varied based on spatial location) in the model. The Jarque-Bera Statistics were also significant, implying that the residuals were not normally distributed and were biased. The bias in the residuals could have occurred due to missing a key variable in the model, the relationship being nonlinear between the variables, influential outliers, or strong heteroscedasticity.[6] Further tests would be required to determine which of those caused the bias in the model. The maps showing the standard deviations for each census tract were likewise nearly identical with only slight changes for a few tracts changing to the next range of standard deviation values for their residuals (see below for the maps with the model using coefficient of 1.5 on the left and 2 on the right).



Based on those results, there was not strong evidence for a linear relationship between average nitrate value per census tract and cancer rate per census tract in the state of Wisconsin. Though the model was

statistically significant (unlikely to have occurred by chance which suggested that there was some relationship between the nitrate value and cancer rate), the model output from the regression analysis was not a good fit for the data (r-squared value near zero). In addition, the statistics indicated that heteroscedasticity, nonstationarity, and bias existed in the model.

To examine the relationship further, it would be best to consult other analysis models to find one that is a better fit for the data. For example, a different interpolation method such as spline could be used in the analysis to create the interpolated surface of nitrate values. Even without using a different interpolation method, other variables could be adjusted in the IDW model to see those effects on the resulting interpolation surface. One good variable to adjust would be the sample window. For the application, the sample window remained consistent: variable distance but with a fixed number of sample points (12). That guaranteed that the model would always have enough sample points to estimate the value for each location; however, in areas of less dense sample points, the sample points included in the estimation process could have been at a distance that was outside the range of the spatial autocorrelation in the data. Instead of using the variable window size, one could explore the range of spatial autocorrelation in the data and then use a fixed distance to guarantee that no sample points would be included in the estimation outside of the range of spatial autocorrelation. To achieve that, the application could be adjusted to allow the user to choose between a fixed distance method (and provide the fixed distance) or a variable distance method (and provide the fixed number of sample points). In addition to changes in the interpolation process, different regression models could be explored. Because the linear regression equation was a poor fit to the model and the residuals were not normally distributed, a nonlinear regression model might be a better fit to explore the relationship between the variables. Other regression models could be replaced in the analysis portion of the application.

IV. Conclusion

Overall, the application that leveraged the Python libraries of ArcPy and Tkinter provided an easy-to-use and effective interface to explore the relationship between average nitrate levels and cancer occurrences at the census tract level in the state of Wisconsin. The application allowed the user to input values for the distance-decay coefficient, run the analysis, see the results of the analysis in both a map and text format, and save those results (map and text results) to compare to subsequent model runs. Though there was not a strong linear relationship between the two variables since the linear models were not good fits for the data, the application could be adjusted to fit other analysis methods and allow further exploration of the relationship between average nitrate value and cancer rate at the census tract level for the state of Wisconsin.

V. References

1. *IDW (Spatial Analyst)*. (n.d.). Esri. https://pro.arcgis.com/en/pro-app/2.9/tool-reference/spatial-analyst/idw.htm.
2. *Zonal Statistics as Table (Spatial Analyst).* (n.d.). Esri. https://pro.arcgis.com/en/pro-app/2.9/tool-reference/spatial-analyst/zonal-statistics-as-table.htm.
3. *Add Join (Data Management)*. (n.d.). Esri. https://pro.arcgis.com/en/pro-app/2.9/tool-reference/data-management/add-join.htm.
4. *Generalized Linear Regression (GLR) (Spatial Statistics)*. (n.d.). Esri. https://pro.arcgis.com/en/pro-app/2.9/tool-reference/spatial-statistics/generalized-linear-regression.htm.
5. Zhu, A. X. (n.d.). Module 5, Lesson 6: Inverse Distance Weighted Method. In *UW—Madison GEOG 579: GIS and Spatial Analysis: Fall 2022*. [Canvas Lesson].
6. *How Generalized Linear Regression Works*. (n.d.). Esri. https://pro.arcgis.com/en/pro-app/2.9/tool-reference/spatial-statistics/how-glr-works.htm.