Exploring the Relationship Between Nitrate Levels and Cancer Incidence: A Spatial Analysis for the Wisconsin Department of Natural Resources

October 03, 2023

Dominic Cwalinski

Table of Contents

Introduction

High concentrations of nitrate in Wisconsin's drinking water have raised concerns over a potential link to cancer. However, the magnitude of this health risk remains ambiguous. This study aims to elucidate the relationship between nitrate levels and the spatial distribution of cancer rates in Wisconsin. Utilizing data from test wells and cancer occurrences, I employed Inverse Distance Weighting and linear regression to spatially analyze the correlation. By making this relationship clearer, I hope to provide a data-driven foundation for public health recommendations and policy.

Implementation Plan

To meet the project's objectives, which focus on exploring the spatial relationship between nitrate levels in well water and cancer rates across specific tracts, a robust Python script was developed in conjunction with ArcGIS Pro. This script integrates various geospatial analyses including Inverse Distance Weighting (IDW) interpolation, zonal statistics, and Ordinary Least

Squares (OLS) regression. Moreover, the script incorporates a user-friendly interface developed using Tkinter to enable dynamic selection of some analytical parameters.

This implementation plan outlines the technical aspects of the project, providing a step-by-step breakdown of the methodology encapsulated in the Python script. It will detail the tools and technologies leveraged, data preprocessing steps, primary analytical methods, and the user interface design. By doing so, this document aims to offer a transparent roadmap that underpins the logic and functionality of the script, thereby ensuring replicability and scalability of the analysis.

Tools and Technologies Used

- ArcGIS Pro
- Python
- arcpy library
- Tkinter for GUI

Preprocessing Steps

1. **Data Import:** Import the shapefiles for cancer rates (cancer_tracts.shp) and well nitrate levels (well_nitrate.shp) using ArcPy's Describe function.
2. **Environment Setup**: Enable spatial analysis extension and set environment variables for output overwrites.

Main Steps

1. **IDW Interpolation**
    - Explanation: Perform Inverse Distance Weighting (IDW) on the well nitrate data to interpolate nitrate concentration levels across Wisconsin.
    - Parameters: Cell size, power (user-defined through GUI)
    - Output: idw_nitrate.tif
2. **Zonal Statistics**
    - Explanation: Calculate the average nitrate concentration in each cancer tract using the IDW output.
    - Parameters: MEAN
    - Output: zonal_stats.dbf

3. **Data Join**
    - Explanation: Join the zonal statistics table to the original cancer tracts shapefile.
    - Fields: GEOID10, MEAN
4. **Ordinary Least Squares (OLS) Regression**
    - Explanation: Conduct OLS regression to understand the relationship between nitrate concentration and cancer rates.
    - Variables: Dependent - canrate, Independent - MEAN
    - Output: ols_analysis.shp and ols_report.pdf
5. **Spatial Autocorrelation (Moran's I)**
    - Explanation: Assess the spatial autocorrelation of the OLS residuals.
    - Output: Moran's I HTML report
6. **Layout Export**
    - Explanation: Export the IDW and OLS layouts to PDF for further analysis and presentation.
    - Output: idw_layout.pdf and ols_layout.pdf

User Interface

- A Tkinter GUI with a slider to allow the user to define the k-value for IDW interpolation.
- Execution button to run the analysis.

Testing and Debugging

- Console messages are printed at various stages for debugging purposes.
- Conditional checks are in place to ensure field existence and type.

Outputs

The Python script generates multiple outputs, each serving a unique analytical purpose:

1. I**DW Interpolation Raster**: A raster dataset is generated depicting the interpolated nitrate levels across the study area.
2. **Zonal Statistics Table**: This table summarizes the mean nitrate levels for each cancer tract, aggregating the raster values within its boundaries.

3. **Spatial Join Output**: A new shapefile is created, enriching the cancer tracts with their corresponding mean nitrate levels.
4. **OLS Regression Analysis**: A new shapefile and a PDF report are generated, presenting the statistical relationship between cancer rates and mean nitrate levels.
5. **Spatial Autocorrelation Report**: An HTML report detailing Moran's I statistic for the residuals of the OLS regression.
6. **PDF Layouts**: Exported layouts from ArcGIS Pro that visualize the IDW and OLS results.
7. **Tkinter User Interface**: A graphical user interface allowing the user to specify parameters for IDW interpolation.

These outputs collectively provide a holistic view of the spatial relationship between nitrate levels and cancer rates, assisting in further research or policy formulation.

Results and Analysis

Inverse Distance Weighting (IDW) Interpolation

IDW was chosen because it's a reliable method for interpolating spatially correlated data, making it well-suited for estimating nitrate levels based on point samples from wells.

- **Choice of k-value:** In this analysis, a k-value of 2 was selected for the distance decay exponent in the IDW interpolation.
- **Rationale for k=2:** The k-value dictates how quickly the influence of a point diminishes with distance. A k-value of 2 was chosen because it provided a balance between local and more distant influences, producing a more realistic representation of nitrate distribution. Higher k-values were tested but led to over-smoothing, thereby diluting local variations in nitrate concentrations.
- **Implications:** The choice of k-value has a direct impact on the interpolated surface, which in turn influences subsequent analyses like zonal statistics and OLS regression. Therefore, the k-value of 2 was deemed optimal for this specific analysis.

Ordinary Least Squares (OLS) Regression

The initial part of our analysis involved performing an Ordinary Least Squares (OLS) regression to investigate the relationship between nitrate concentrations in well water and cancer rates across specific tracts in Wisconsin.

- **Coefficients**: Our OLS model yielded an Intercept coefficient of 0.071968 and a coefficient of 0.006084 for the variable "MEAN," which represents the mean nitrate level.
- **Statistical Significance:** The t-statistics and p-values for both variables were close to zero, confirming their statistical significance. This implies that nitrate levels are a significant predictor of cancer rates.
- **Goodness of Fit:** However, it's crucial to note that the R-squared value was as low as 0.020665. This suggests that the model accounts for only a small portion of the variability in cancer rates, indicating a weak predictive power.

Moran's I - Spatial Autocorrelation

To further scrutinize our OLS model, we performed a Moran's I analysis on its residuals. This helps to assess whether the model has accounted for the spatial distribution of the data.

- **Moran's Index:** Our Moran's I analysis yielded an index value of 0.204435, indicative of a clustered pattern.
- **Statistical Significance:** With a z-score of 21.558783 and a p-value close to zero, the spatial patterns detected are highly statistically significant.
- **Implications:** This significant Moran's I index suggests that there are spatial clusters where nitrate levels and cancer rates are similar. This is a critical aspect that the initial OLS model could not capture.

Conclusions

Upon synthesizing the results from both the OLS and Moran's I analyses, a more nuanced picture emerges. While the OLS model points toward a statistically significant but weak relationship between nitrate levels and cancer rates, the Moran's I results reveal significant spatial patterns.

This leads us to consider that other factors, possibly spatial in nature, could be influencing the relationship between nitrate levels and cancer rates. Thus, the analysis suggests that a more complex spatial regression model is warranted to better capture these intricacies.