Andrew Pittman
Spatial Analysis Interactive Application
UW GIS WMP Master's Capstone Project 1
02/28/2020

It has been hypothesized that the consumption nitrate and nitrite has a potential correlation with higher cancer risk, however definitive evidence of this correlation has yet to be found in humans. Given this, an interactive geospatial tool has been commissioned to discover what if any link there is between these two variables via an aggregation method (Inverse Distance Weighting) and a linear regression method (Ordinary Least Squares). To emphasize clarity, nitrate levels will sometimes be referred to as the independent variable, and cancer rates will sometimes be referred to as the dependent variable. The goal of this development and consequent analysis effort is to enable environmental analysts, governmental bodies, and knowledgeable citizens to explore the raw data as well as drive analysis using reasonable inputs for the aggregation and regression methods chosen.

The creation of a web application available to the public internet was chosen to reach as large an audience as possible. This decision enables all users easy access to the tool for personal study and public presentation. In order to minimize barriers to usage, only open source technology options were selected when selecting the technology stack. Mapbox GL was chosen for displaying map based data, as it is low cost and highly effective at data driven styling of geospatial data. ReactJS was chosen as a JavaScript framework to support an efficient and fast user interface and user experience. Material UI was chosen to streamline the look and feel of the interface, enabling users to quickly or instantly understand how to use various features. TurfJS was used to facilitate geospatial calculations via free and open source methods. RegressionJS and SimpleStatisticsJS were chosen to facilitate regression calculations and standard deviation calculations, respectively. Workerize-LoaderJS was chosen to enable multi-threaded background processes to run, providing a more seamless user experience while resource intensive geospatial aggregations and statistical regressions were being performed. ShapefileJS was selected to ingest and process the raw data files into geojson.

The implementation plan began by researching open source methods to perform aggregation and regression analysis. The various spatial data manipulation libraries above were either known at the project start or discovered during this phase. Of particular note, the discovery of TurfJS's interpolate and tag functions allowed planning for the critical collocation of the data into a single geospatial layer and consequent aggregation using the Inverse Distance Weighting (IDW) algorithm. These steps were critical to plan for prior to setting up and implementing a relatively simple regression equation, as dissimilar spatial data can not be fed into a regression method. Upon understanding that TurfJS could handle aggregation tasks, the two statistics libraries above were identified as methods to perform various regression calculations. At this point, a rough outline for the development plan was generated to inform and guide the next phase of the project.

The initial development plan consisted of implementing a basic web application layout that included a Material UI header bar and map display from Mapbox GL. Upon completion of this basic framework, the raw data layers were added to the map on application load via ShapefileJS: well nitrate levels taken from wells in Wisconsin, as well as normalized cancer rates across Wisconsin various national census tracts. This data was loaded and styled accordingly to provide understanding of the data to be processed. Tool tips and data driven styling were added to enable user extraction of the data embedded in each raw data layer. Upon reviewing the raw data visually and re-familiarizing with TurfJS's functionality, an aggregation strategy was decided on.

Enabling repeated user initiated aggregation of the raw data required collocation of the raw data. In order to both collocate the data and provide a good user experience, a web worker is tasked with tagging all nitrate data points with their corresponding normalized cancer level (held in polygon form) via TurfJS's tag method. This is a highly resource intensive task due to the size of the raw data, taking about twelve seconds on average, and the web worker allows the job to be done in the background once immediately after the user loads the application. Once the job is complete, the nitrate data point layer has been infused with the corresponding cancer rates and re-ingested as a layer into the map. This allows all future aggregation and regression calls to utilized this information within the current application instantiation, however the user is unable to tell that anything has changed as the original point data has been available since initially loading the application.

To perform IDW aggregation, both a weight and an output polygon size are required. The decision was made to implement user controls allowing the adjustment of both weight value (k) and the polygon size (sqkm) to facilitate more user analysis control. Upon application load, the weight value defaults to 2, and the polygon size value defaults to 10 square kilometers in order to enable aggregation and regression. Utilizing the collocated data in the point layer, aggregation is kicked off immediately after collocation asynchronously using javascript's "async await" functionality via two more dedicated web workers- one for nitrate data aggregation to a hexagon polygon layer and one for cancer rate data aggregation to a matching hexagon polygon layer. This enables two simultaneous threads to perform the necessary geospatial aggregation work, shortening user wait times for the final analysis product.

Upon completion of both of these aggregations, a final web worker is tasked with performing the regression from the resulting aggregation data. The regression work returns a hexagon polygon layer embedded with various metadata associated with the regression calculation; residual values, standard residual values, standard deviation of residuals, slope and offset of the regression equation, R^2 value, and an absolute value of the standard residual values. The user is able to interact with the map and all current layers during all of this behind the scenes work due to the usage of web workers, creating a seamless user experience. When the regression hexagon layer is returned, the map simply drops the new layer in where the old regression layer was, if it existed previously. On first aggregation and regression run (application load), all analysis controls associated with aggregation input and the regression layer toggle are disabled to prevent errors from occurring.

Following the implementation of all statistics, thorough review was taken of the math and resulting visual map layers. This resulted in bug identification and fixes, and ensured that the final product was mathematically correct to the greatest extent possible while being visually pleasing and intuitive. Within this process, it was decided to implement a filter on the resulting spatial regression layer to only show those polygons that intersected with Wisconsin. This was necessary as TurfJS's interpolation output includes all polygons in a simple bounding box, and was accomplished via the import of an additional state boundary geojson that was utilized within TurfJS's intersect function to remove regression data on non-relevant hexagons. This process also involved numerous stylistic tweaking of color scales and geospatial visual display of metrics. It was decided during this process to add in numerous visual affordances to assist the user in identifying trends in data. The nitrate point data layer was originally styled by color from blue (low nitrate) to red (high nitrate), but during this process variation on the size of the point data was also added (small for low nitrate, large for high nitrate) to add extra reinforcement of the underlying data metric. Additionally, the spatial regression layer was implemented using both a diverging color scale to represent standard residual values as well as polygon extrusion height to represent the absolute value of standard residual values. These additional affordances assist each other in enabling users to quickly identify trends in the data layers.

Wrapping up the development included adding in a series of conditionally rendered legends that align with the various map data layers. These legends align visually with the layers themselves stylistically. An overlay card was added to assist users with a detailed explanation of the project's purpose and analysis methods. In order to convey that user input for aggregation and spatial regression was being processed, a non-intrusive loading bar was added to the base of the application to indicate work being performed in the background. Finally, an export capability was added to enable further analysis efforts to be conducted by the various user groups in other analytic tools.

Utilizing the application has yielded various but consistent results regarding the validity of the regression for predicting the correlation between the independent variable and the dependent variable. Various IDW weight (k) values were compared against varying hexagon size values to produce a field of results for review:

Hexagon size	k = 1.5	k = 2	k = 2.5	k = 3	k = 3.5	k = 4	k = 4.5
5 sqkm	1.82%	3.41%	4.33%	4.72%	4.64%	4.37%	3.79%
10 sqkm	2.85%	3.59%	5.13%	5.92%	5.87%	5.51%	5.11%
15 sqkm	0.72%	2.19%	3.29%	3.63%	3.58%	3.43%	3.26%
20 sqkm	1.18%	1.63%	2.37%	2.62%	2.55%	2.33%	2.10%

Table 1. R^2 values rendered from spatial analysis application

The results from Table 1 above show that the relationship between nitrate and cancer is likely to be poor due to low R^2 values, as seen above. In the best case scenario (k=3, size=10 sqkm), the R^2 value rises to about 6%, which explains about 2.5% of the standard deviation within the residual values. This figure drops upon variation of the input values in any way, indicating that there is not a strong correlation between the independent variable and the dependent variable. The map taken from the spatial analysis tool of the best case R2 scenario is shown below.

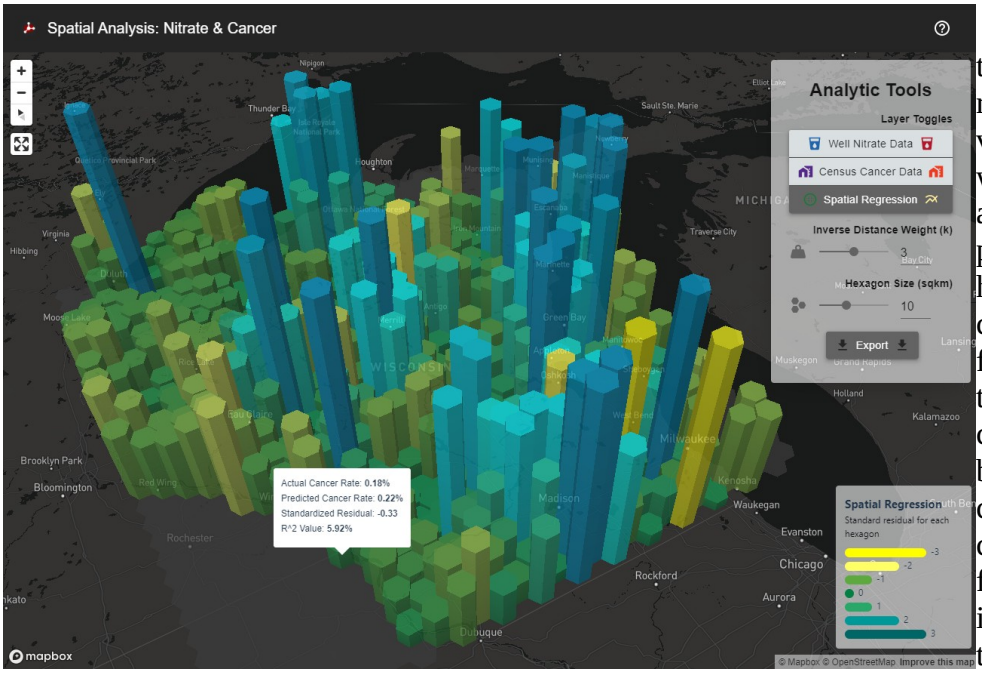

Figure 1: Spatial analysis with k of 3, size of 10 sqkm, coefficient of determination of 5.92%

While the tool enables thorough exploration of the relationship between nitrate values and cancer rates, it would seem that nitrate alone either is not a solitary predictor of cancer in humans or is one of many cancer indicators and further research is needed to determine if the more complex relationship between the two is more complex. Perhaps a more complex model is needed to fit the data, or more more independent variable need to be considered to draw more conclusive results.

Future plans for this application include public deployment to pittman.dev/spatialanalysis/ and socialization of the tool and results with peers and mentors for feedback.