Building an App to Explore Cancer and Nitrate

Project 1 Report

2020-02-23

Cory Leigh Rahman, University of Wisconsin-Madison, Master's in GIS & Web Map Programming, GEOG 777: Capstone in GIS Development

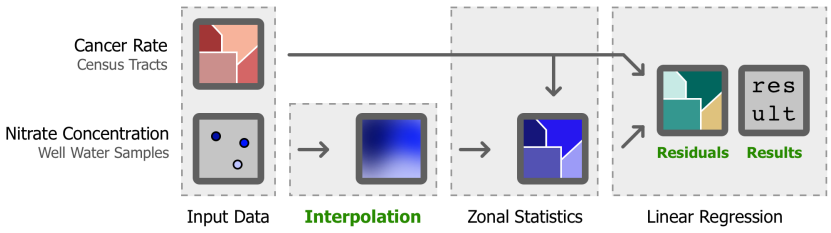
Introduction

Nitrate is a contaminant which can get into your drinking water supply because of fertilizers and other waste. Data from Wisconsin can be used to help figure out what the relationship is between nitrate in drinking water and cancer.

Methodology

To see how nitrate and cancer are related, linear regression was run on state-wide data from Wisconsin. Results of the regression could reveal how much of the cancer rate can be explained by nitrate concentration in drinking water.

- Independent (Explanatory) Variable: **Nitrate Concentration** from water well samples (parts per million)
- Dependent Variable: **Cancer Rate** from census tracts (percentage of population)



The green-colored steps (**Interpolation**, **Residuals**, and **Results**) are the steps the end-user can interact with. The rest is done automatically by the back-end server in Python. End-users can set the Interpolation parameters, run the full analysis, and explore the results.

The flow chart visualizes the process of the analysis. The analysis starts with two inputs from 2016: cancer rates from census tracts, and nitrate concentrations from well water samples. In order to explore the relationship between these factors they must be in the same form. To get the data in the same form, the well samples were interpolated and zonal statistics were used to get the average nitrate concentration per census tract. Once both cancer rates and average nitrate concentration were averaged per census tract, finally linear regression could be run on the data.

Implementation

A web application was built to automate the analysis and allow users to explore the results.

Technologies Used

Back-end (Server-Side) Technologies Used:

Technology	Description	Used For
Flask	Python Web Framework	Server code & API
Anaconda	Data Science Platform, Package Manager	Managing tools / dependencies
GDAL	Geospatial Data Abstraction Library	IDW Interpolation of water wells
rasterstats	Raster Statistics Library	Zonal statistics
GeoPandas	Geospatial Data Python Library	Working with GeoJSON, organizing data
statsmodels	Statistics Library	Ordinary least-squares (OLS) regression

Front-end (Client-side) Technologies Used:

Technology	Description	Used For
Angular	JavaScript Front-End Web Framework	Code organization, TypeScript support, data binding, and built-in tools
NPM	Package Manager	Managing tools / dependencies
Leaflet	JavaScript Mapping Library	Displaying regression residuals on a map
Bootstrap	Front-end UI Component Library	Page layout, input and button styling

Other Technologies

Technology	Description	Used For
VS Code	Microsoft Visual Studio Code IDE / Text Editor	Writing all code
colorbrewer2.org	Map Symbology Color Picker	Map color scheme
Sketch	Vector Graphics Design Software	Creating the flow chart graphic and IDW parameters graphic

Back-end Analysis Automation

The back-end is entirely written in Python and can be found in `app.py`. Through Flask, a simple API was written which accepts an HTTP request to `/analyze` with parameters.

```
...
@app.route('/', strict_slashes=False)
@app.route('/static', strict_slashes=False)
def main():
    return render_template('index.html')

@app.route('/analyze', methods=['POST'])
def analyze():
    ...
```

Once the request is received, the following actions are taken:

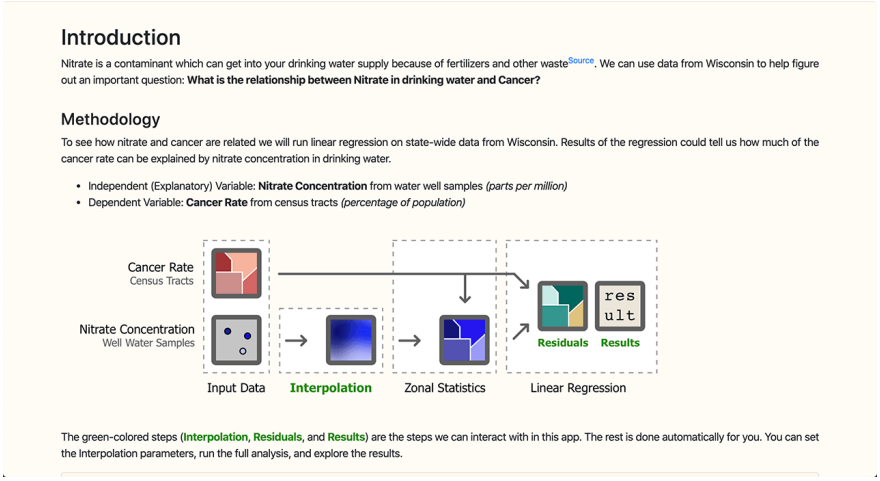
1. Get user inputs (Power and Smoothing for IDW)
2. Perform the IDW interpolation on water well data
3. Get the average nitrate concentration per census tract using Zonal Statistics
4. Organize the appropriate attribute data for regression (cancer rate and nitrate concentration)
5. Perform the regression analysis
6. Merge the residuals back into the dataset
7. Return two items: a) Regression results summary text, b) GeoJSON of residuals

Front-end Interaction and Design

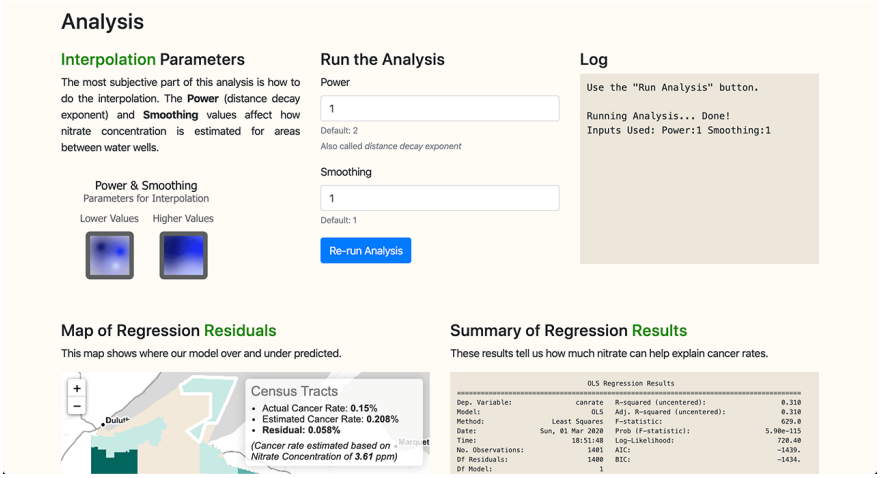
The end-user is guided through the problem and shown how to explore it. The interface is split up into two main sections:

1. Introduction
2. Analysis

The Introduction section explains how nitrate gets in water and prompts the end-user with the question of how nitrate and cancer are related. The previously seen Analysis Flow Chart graphic is displayed for the end-user here, along with optional supplementary information.



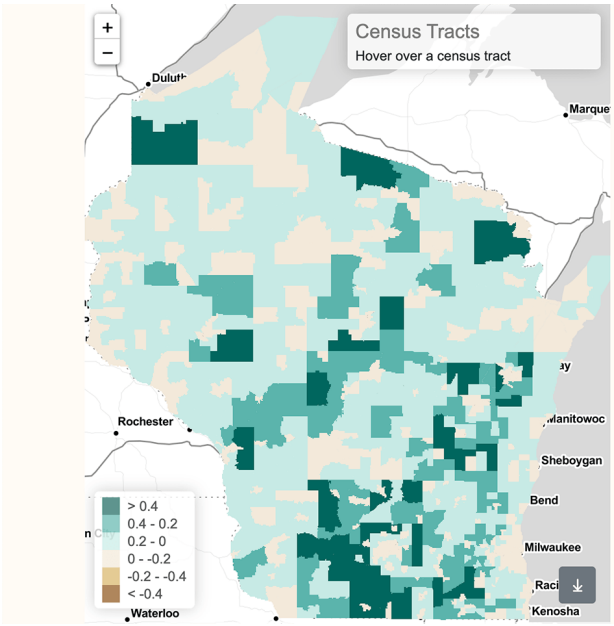
The Analysis section provides interfaces for the inputs and results. Because the most subjective part of this analysis is the IDW (inverse distance weighted) interpolation, the end-user is allowed to input their own values for the Power (distance decay exponent) and Smoothing values. After the end-user selects their IDW inputs and runs the tool, they are presented with a map of residuals from the regression alongside a full regression summary.



The code for the front-end can be found in the folder `front-end-exploring-cancer-nitrate/`

Conclusions from Analysis

After experimenting with multiple IDW inputs, the highest R-Squared value received was using the default values of "2" for the distance decay exponent and "1" for the Smoothing value. These inputs yielded the following results:



OLS Regression Results

Dep. Variable:	canrate	R-squared (uncentered):	0.320
Model:	OLS	Adj. R-squared (uncentered):	0.320
Method:	Least Squares	F-statistic:	659.2
Date:	Sun, 01 Mar 2020	Prob (F-statistic):	1.86e-119
Time:	20:49:39	Log-Likelihood:	730.75
No. Observations:	1401	AIC:	-1460.
Df Residuals:	1400	BIC:	-1454.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
mean	0.0258	0.001	25.676	0.000	0.024	0.028

Omnibus:	744.066	Durbin-Watson:	1.350
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4554.015
Skew:	2.489	Prob(JB):	0.00
Kurtosis:	10.297	Cond. No.	1.00

Map of Residuals

Using linear regression, the cancer rate can be estimated per census tract based only on the average nitrate concentration in water there. The map displays the difference (i.e. residual) between regression estimations and the actual cancer rates in each tract.

The residuals seem to be somewhat clustered around urban areas, but otherwise relatively random. With only one explanatory variable (nitrate concentration), it is not expected for the residuals to be completely random, as other explanations for cancer rates are certainly being missed.

Regression Results

The regression summary tells us two important things:

1. What the R-Squared value is
2. If we can trust the R-Squared value

The R-Squared value this analysis received was 0.32, suggesting that nitrate concentration in water can explain 32% of the cancer rate per census tract as long as the R-Squared value is trustworthy.

The other factors in the regression summary reveal that the R-Squared value is trustworthy. Because the P value (Prob F Statistic) is close to zero, and the F-statistic value is really large, we can reject the null hypothesis; this proves that there is a linear relationship between the independent and dependant variables. For the "mean" variable (mean nitrate concentration), given a large t-value (25.676) and a P-value close to zero, the null hypothesis can again be rejected, proving that nitrate concentration is significant in the prediction of cancer rate.

Given the statistical significance of the regression model, the results show that **nitrate concentration in well water can explain 32% of the cancer rate in Wisconsin**.