The polygon overlay problem in Flint Water Crisis

May 5, 2020

Abstract

This project mainly focuses on the change of support problem (COSP), which deals with the problem of how the spatial variation in one variable associated with a given support relates to that of the other variable with a different support. Specifically, we are looking at the Polygon overlay problem, one of the cases of COSP that occurs when we have incompatible area to area data. Our main goal is to reproduce the results generated by Hanna-Atisha and Salder in 2016 in R. Therefore, this paper aims to reallocate the data from ZIP code level to the municipal level to allow for more accurate results by performing areal interpolation.

1 Introduction

The Flint Water Crisis is an ongoing public health concern occurring in Flint, Michigan. This issue began in April 2014, when Flint changed their water supply to the Flint River. The water was treated with chemicals to remove any contaminants in order to make the water safe for use and consumption. However, inadequate water treatment and monitoring caused the water pipes to corrode and the lead from the pipes leaked into the water. This water was consumed by people residing in Flint and became a health concern once there were reports on lead being present in their blood. Lead is a chemical, specifically a type of natural metal, that may take weeks to leave the body or in some cases, it may not at all leave the body, depending on the location it is stored in the body. High levels of lead might cause lead poisoning,which results in brain and kidney damage, and in severe cases, death. Children are more susceptible than adults to the adverse health problems caused by lead, since their physical, behavioral, and cognitive health are at risk during their developmental stages. Currently, there is no specified safe level of lead within children. In order to verify whether the lead in the new water supply posed a health risk to the people of Flint, officials directed health professionals to collect blood samples from individuals living within the ZIP codes of Flint and to measure the lead levels within each sample. If there was an increase in the number of Flint residents with lead in their blood or if lead levels increased significantly enough to be a health risk as a result of consuming the water, then the water supply from the Flint River would prove a concern for further investigation. Officials reported that the blood lead levels remained steady before and after the change in Flint's water supply. This means the lead levels in blood samples obtained after the water supply changed to the river did not significantly differ from the blood lead levels due to the previous water supply. This caused officials to conclude that change in the water supply is not a main concern.

However, Dr. Mona Hanna-Attisha, director at one of the pediatric units in Flint, Michigan, was not convinced by such results because they were not consistent with the results from her study analyzing lead levels in infants and children ages 5 and younger.She reported in September 2015, that at least "twice as many children had elevated levels of lead in their blood" since Flint's water supply changed in 2014. She reached out to her collaborator, Richard Sadler, a geographer at Michigan State University, and requested him to analyze the lead levels within the blood to re-address whether the current Flint water supply was unsafe for use and consumption. Together, they conducted a study determining whether there was a significant increase in the number of children with elevated lead levels**?**. Sadler analyzed and aggregrated data that came from blood samples obtained from a local children's hospital and he applied spatial analysis at the neighborhood-level; that is, he analyzed data from children who lived within the neighborhoods of Flint city. Contrary to what the officials reported, Sadler and his collaborators found that found that there was a significant increase in the number of children with elevated Blood levels of lead since the water supply has changed.

Sadler, initially didn't understand the discrepancy between the results and then later found that the results reported by the officials were based on data collected and lead levels analyzed at the ZIP code level, rather than the neighborhood level. In simple terms, the blood samples were collected from individuals who lived within the ZIP codes that overlapped but incongruent with the city of Flint.Since the analyses was conducted at ZIP code level, he found that at least 50% of the data consisted of individuals who did not live in the city of Flint, but within the Flint ZIP codes. Therefore, the individuals possibly who did not consume the water from Flint River would not add information but rather "mask the Flint Water Crisis problem;". Since, their lead levels would be zero,they lowered the average number of people with elevated lead levels. To address whether the change in water supply is a major concern,the only people who should have been in the analysis were those who consumed Flint's water. Analyses conducted at ZIP-code level made the results show that there was no increase in the number of people with elevated lead levels before and after the change of water supply leading to the conclusion that the change in water supply did not pose as a public health issue. The sample design and analysis of selecting all individuals residing at ZIP code level (instead of the municipal level) led to misrepresentation of the data and inaccurate results, which masked the harmfulness of Flint River new water supply. He stated that it is common knowledge for geographers to know that ZIP code is not a reliable spatial unit to use for analysis since the ZIP code borders are created for the sole purpose of the post office delivering mail to houses and buildings most efficiently. It is rare that regions formed by the ZIP code borders is aligned with or best represents the distribution of the data (to be analyzed) within the context of the main investigation. Hence, when officials reported insignificant results about the water supply not being a main concern it is because the analysis was based on Flint ZIP codes. Such events revolving around the Flint Water Crisis is one of the (many) examples that is centered around Modifiable Areal Unit Problem(MAUP). It is clear that the spatial unit initially chosen at ZIP code level led to misleading results and how these results differed from when Sadler analyzed lead levels at the neighborhood-level, which in this latter case provided a more accurate representation of the data and phenomena of interest. This discrepancy due to MAUP and how it masked the Flint Water Crisis problem motivates the current project in this report. Instead of analyzing the data at ZIP code level, the aim of this project is to transform that spatial data to the municipal level to reproduce similar results using R.

The term “spatial data transformations” refer to situations in which the spatial process of interest is inherently of one form but the data observed are of another form, resulting in a “transformation” of the original process of interest. For example, Assume a situation in which we have the data for population by county but don’t have data for population by zip code, which is our desired scale of interest . This situation and all of spatial data transformations are special cases of what is called the change of support problem (COSP) in geostatistics. The term “support” has come to mean simply the size or volume associated with each data value, but the complete specification of this term also includes the geometrical size, shape, and spatial orientation of the regions associated with the measurements. Changing the support of a variable , which is done typically by averaging or aggregation creates a new variable. This new variable is related to the original one, but has different statistical and spatial properties. COSP deals with the problem of how the spatial variation in one variable associated with a given support relates to that of the other variable with a different support. Polygon overlay problem is one of the cases of COSP, which occurs when we have incompatible area to area spatial data. Being confronted with this problem, we want to reallocate data from a source data set to a target data set, or in other words from the areal units with available data to the areal units of interest (for our concerns from zipcode level to city level), by using areal interpolation methods.

2 Methods

2.1 Areal Interpolation

Areal Interpolation is the process of making estimates from a source set of polygons to an overlapping but incongruent set of target polygons. This is required if, for example, a researcher wants to derive population estimates for neighborhoods in a U.S. city from the Census Bureau’s census tracts. Neighborhoods do not typically align with census tract boundaries, and areal interpolation can be used to produce estimates in this situation. Using Areal Interpolation, reaggregating polygonal data (for example, downscaling population counts) is a two-step process. First, a smooth prediction surface for individual points is created from the source polygons (this surface can often be interpreted as a density or risk surface); then, the prediction surface is aggregated back to the target polygons. Areal interpolation methods model spatial distribution based on strong assumptions such as homogeneity (evenly spatial distribution within the areal unit), isotropy (same spatial distribution in every direction) and stationarity (smooth variation according to distance).

2.2 Kriging

Kriging is the process to model the spatial data across an area, when there is spatially correlated distance or directional bias in the data. It assumes that the distance or direction between sample points reflects a spatial correlation that can be used to explain variation in the surface. The kriging tool fits a mathematical function to a specified number of points, or all points within a specified radius, to determine the output value for each location. This method is a multistep process; it includes exploratory statistical analysis of the data, variogram modeling, creating the surface, and (optionally) exploring a variance surface.

2.3 R packages

Functions from R packages such as “areal”, “tidyverse”, “sf”, “rgdal”, “curl”, “sp” and “gstat” were used.

2.3.1 areal

This package aims to reduce the barriers that prevent a spatial researcher from performing Areal Interpolation Prener and Revord [2019]. We used this package to perform areal interpolation to generate polygon data at the city level from polygon data aggregrated at the zipcode level. Functions such as `ar_validate` and `aw_interpolate` were used to perform the interpolation. The function `ar_validate` ensures that our data was ready for areal interpolation by checking five conditions. These conditions include:

1. Are both objects sf objects?
2. Do both objects share the same coordinate system?
3. Is that coordinate system planar?
4. Do the given variables exist in the source data?
5. Will interpolation overwrite columns in the target data?

The function `aw_interpolate` was then used to perform areal interpolation on the now validated source and target datasets.

2.3.2 gstat

This package is used for geostatistical modeling, prediction and simulation. We mainly used this package for variogram modeling, plotting and kriging. Function gstat::vgm generates a variogram model, takinglattice layers as input. We used function gstat::variogram to calculate the sample variogram from the data. After storing our variogram object, we utilized the gstats::fit.variogram function to fit a simple or nested variogram model to a sample variogram. As a result, we could call function krige0, which is a user-supplied covariance function for simple, ordinary, or universal kriging.

2.3.3 sf

This package supports for “simple features”, a standardized way to encode spatial vector data**?**. Binds to ”GDAL’ for reading and writing data, to ”GEOS’ for geometrical operations, and to ”PROJ’ for projection conversions and datum transformations. This was the most used package in the project, as we are dealing with shapefiles. We used functions like `st_read`, `st_transform`, and `st_as_sf` for data wrangling. The function `st_make_grid` was used to make a grid out of the shapefile.

2.3.4 sp

This package provides classes and methods for spatial data; the classes document where the spatial location information resides within the dataset, for 2D or 3D data. Utility functions are provided, e.g. for plotting data as maps, spatial selection, as well as methods for retrieving coordinates, for subsetting, print, summary, etc. In our project, this package was mainly used when performing kriging as sf objects need to be converted to spdf objects prior to the process.

2.3.5 tidyverse

This is a single “meta”-package that installs a collection of packages with a simple command **?**. In this project, we used this package to load packages like ggplot2, dplyr and readr with one R command library(tidyverse).

1. ggplot2: This package helps us to graphically show our resulting polygon data using ggplot, geom_sf, geom_polygon and scale_fill_gradient functions.

2. dplyr: This package helps us to combine two different datasets using inner_join command.

3. readr: This package was primarily used to import data from csv files.

3 Data

3.1 Data collection

In this project, we worked with three primary datasets, Flint, BLLZIPCODE and Sfzipcodegrid, using which our goal was to find percent of the lead in population under 6 and compare them at Flint city and zipcode levels. The Flint dataset was originally extracted from a city shapefile that is imported from Michigan state department boundaries. Since we focused on Flint, we filtered the shapefile using the city name “Flint”. The resulting dataset contains the following primary variables:

```
## Simple feature collection with 1 feature and 14 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 1252718 ymin: 4796523 xmax: 1264068 ymax: 4811960
## epsg (SRID):    26915
## proj4string:    +proj=utm +zone=15 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_def
##   OBJECTID FIPSCODE FIPSNUM  NAME         LABEL TYPE   SQKM SQMILES    ACRES
## 1       57    29000   29000 Flint City of Flint City 88.283  34.086 21814.99
##   VER   LAYOUT PENINSULA ShapeSTAre ShapeSTLen                       geometry
## 1 17A portrait     lower   88283092   74566.12 MULTIPOLYGON (((1256937 480...
```

- ObjectID: A unique identifer of the current city.
- FIPSCODE: A code to uniquely identify geographic areas.
- Name: The name of the city.
- Type: Specifies the type of the object (whether it is a city or county).
- geometry: This variable gives us the city's boundaries/shape in the form of polygon outline.

Like the Flint dataset, the BLLZIPCODE dataset was also extracted from a shapefile called "BLL" that contains the data of Michigan Childhood Blood Lead Levels data reported by The Michigan Department of Health and Human Resources. This comma separated value(CSV) file was imported from Dr.McNamara's Github repository titled "BLLunder6zip2016" and shows results for blood tests done on children under the age of 6, sorted by zipcode. It contains 935 rows with 6 variables. Since our primary focus is Flint, we only selected rows whose zipcodes are covered by Flint and ended up storing the result as an object named "BLLflint". Since BLLflint is a dataframe, to convert this into a shapefile, we performed a left join on BLLflint and FlintZP, which resulted in the BLLZIPCODE sf object. BLLZIPCODE contains the following primary variables:

```
## Simple feature collection with 6 features and 10 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: 1249332 ymin: 4793261 xmax: 1268198 ymax: 4816458
## epsg (SRID):    26915
## proj4string:    +proj=utm +zone=15 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_def
## # A tibble: 6 x 11
##   ZCTA5CE10 Pop_under_age_s~ `Children_teste~ Children_tested~ any_samples_n
##   <chr>     <chr>                       <dbl> <chr>                    <dbl>
## 1 48502     20                             15 75                           0
## 2 48503     2295                          637 27.8                        36
## 3 48504     2711                          701 25.9                        28
## 4 48505     1776                          596 33.6                        16
## 5 48506     2426                          600 24.7                        18
## 6 48507     2915                          836 28.7                        13
## # ... with 6 more variables: any_samples_percent <dbl>, AFFGEOID10 <fct>,
## #   GEOID10 <fct>, ALAND10 <dbl>, AWATER10 <dbl>, geometry <MULTIPOLYGON [m]>
```

- ZCTA5CE10: Unique identifer for the area.

- Pop_under_age_six: The total count of the population under age 6.
- any_samples_percent:The percent of children with Blood Lead Levels greater than five micrograms of lead per decilitor of blood.
- geometry: This variables gives us the zipcode's boundaries/shape in the form of list.

Sfzipcodegrid(grid) is the grid version of BLLZIPCODE(sfobject). While BLLZIPCODE only contains 7 out of 12 zipcodes of Flint, Sfzipcodegrid contains all the 12 zipcodes. This grid is made from the FlintZP code datafile, which contains the geometric data of all the zipcodes in Flint. The number of rows this dataset contains would be equivalent to the specified cell size of the grid. For example, if we specify the cell size to be 500, there will be 500 rows in the dataset. Since we are breaking the zipcodes into cells to form a zipcode, each cell contains its own boundaries (geometric attribute: lat and long) along with its id that acts as an unique identifier of the cell.

3.1.1 Data Wrangling:

After extracting and cleaning the datasets, we performed data wrangling by converting BLLZIPCODE and Flint to sf objects. In addition to this, we also verified all of the datasets used the same coordinate system to prevent any errors when validating our dataset in the process of performing areal interpolation.

4 Validating Areal Interpolation:

Before we started our interpolation, we wanted to validate whether areal interpolation is an effective way to explore MAUP. In order to achieve this, we performed interpolation on population data aggregrated at congressional district (congressDist) level to aggregrate the population at the zipcode level and reallocating data from zipcode level to congressDist level. Based on this, we were able to verify Areal Interpolation was one of the valid methods for COSP. Also, we were able to find that the areal interpolation is not pynchophalactic, but the aggregrated average value generated by `aw_interpolate` is closer to the observed mean value compared to the mean value produced `st_interpolate_aw` method of sf object.

5 The polygon overlay problem in Flint water crisis data:

After making sure, Areal Interpolation is a valid process for reallocating the polygonal data. We performed areal interpolation, as explained in the Data, we performed data cleaning and data wrangling to ensure we can do interpolation without any interruptions. We performed interpolation twice, once to aggregate BLL values from the zipcode level to the city level and angain from the few flint zipcodes to all the zipcodes. The main purpose of the second interpolation to aggregate the data we dont have data for 5 out of 12 zipcodes in Flint.

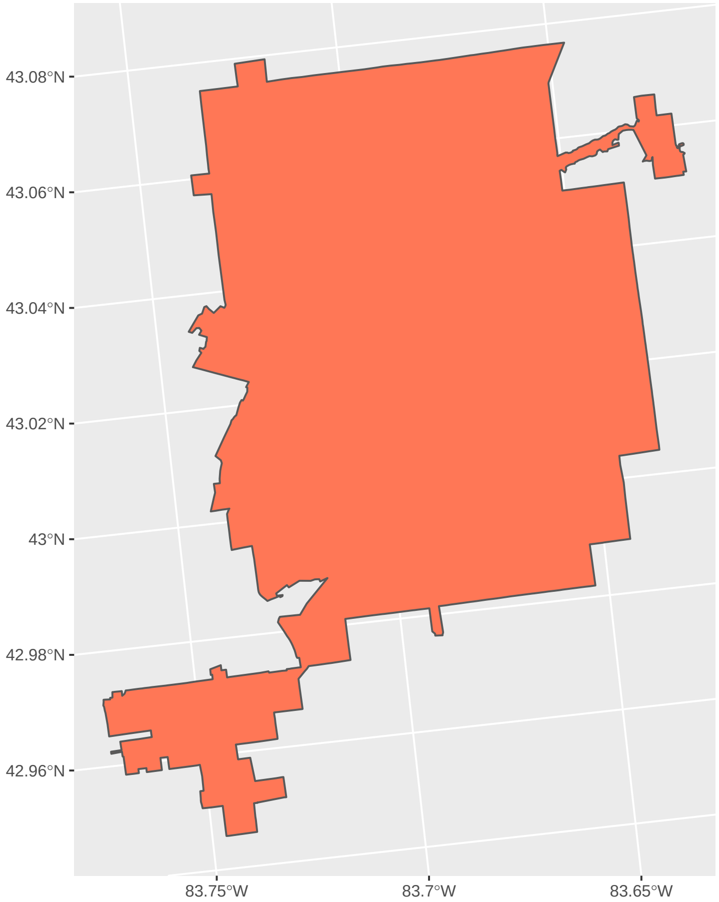

Figure 1: This fig shows the average BLL lead levels in population under age 6 to be 3.4211

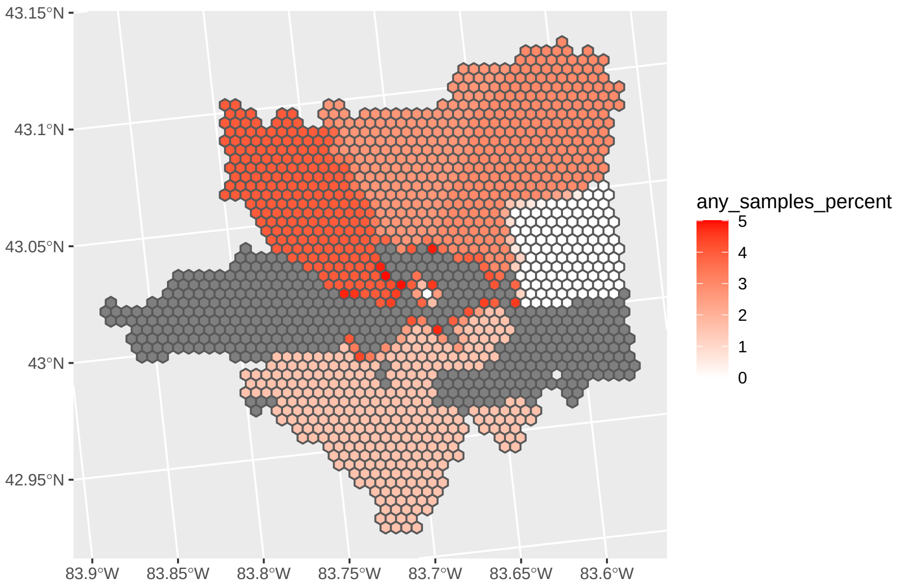

Figure 2: This fig shows the average BLL lead levels in different zipcode areas on the layer of Flint zipcode grid on SFZIPCODE grid

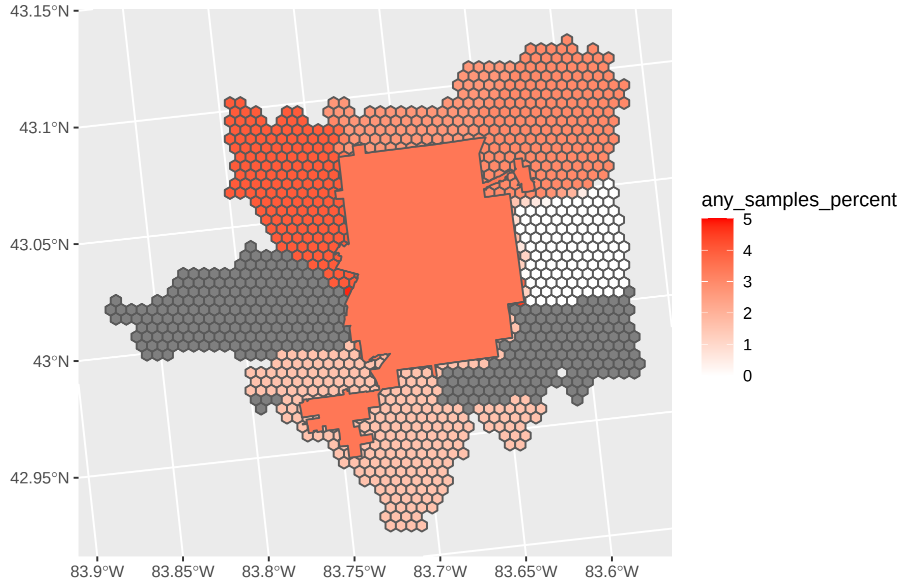

Figure 3: This fig shows the average BLL lead levels in the Flint city on the layers of SFZIPCODE grid, Flint zipcode grid and Flint city grid

By performing interpolation, we can model the missing data for these 5 zipcodes. Before performing interpolation, we validated our datasets by using the function `ar_validate` on the source(BLLZIPCODE) and target(Sfzipcodegrid (in first interpolation) and flint (in second interpolation)) datasets. Later, we used the function `ar_interpolation` on the source and target datasets, which resulted in the two interpolated datasets. The visualizations of these datasets can be seen below. When we interpolated from zipcode level to city level, we got a aggregated value of 3.4211 for BLL in pop_under_age_6 in flint city , which is smaller than 5mgdl.

6 Exploring Modifiable Areal Unit Problem with Kriging interpolatio

Before we started kriging, we converted all our sf objects to spdf. Since we don't have any have method in sf package or any other packages to perform kriging directly on sf objects. The process of kriging starts with creating variograms and covariance functions to analyze the statistical relationship with known data. Typical models we often use are Spherical, Exponential, Circular models. We attempted to find the best variogram model for the BLL data. We tried with several models(Spherical, Circular and Exponential). We used a fit.variogram function in R to find the most suitable variogram for this dataset. We found that Spherical is the best model to use.

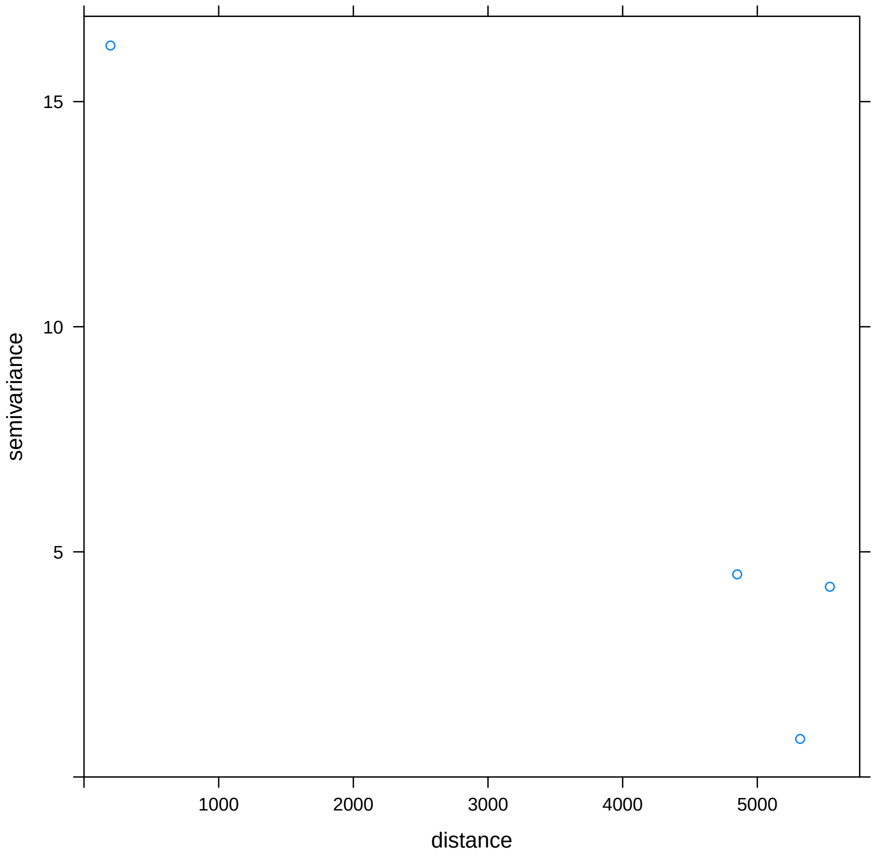

Figure 4: Variogram for BLLZIPCODE

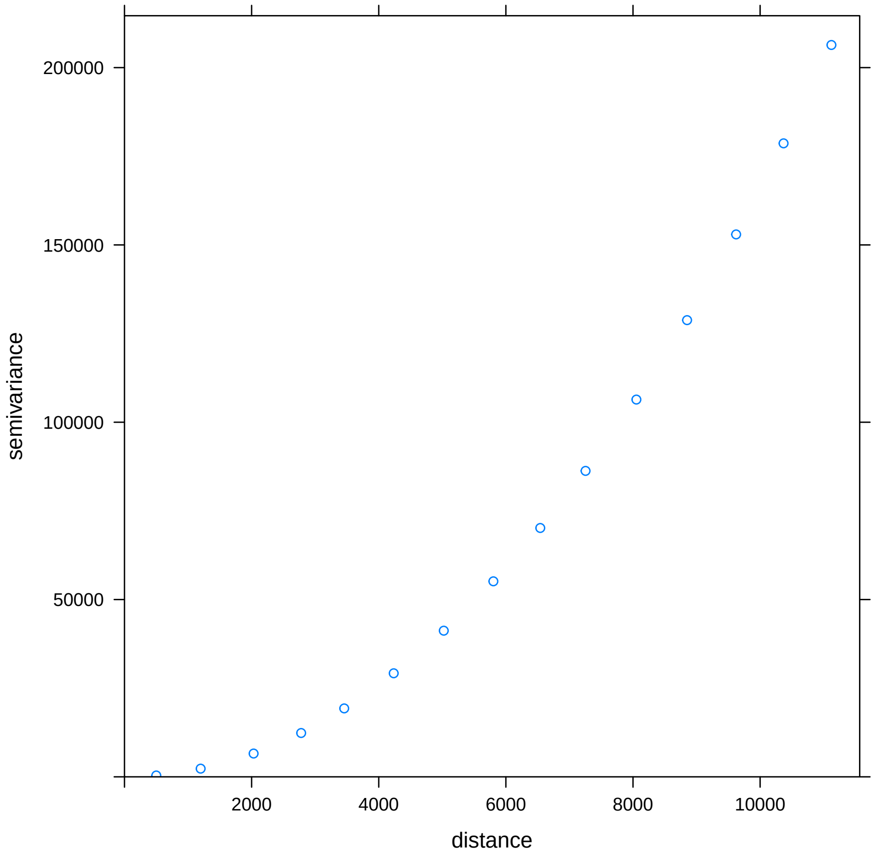

Figure 5: Variogram for sfZipcodeGrid

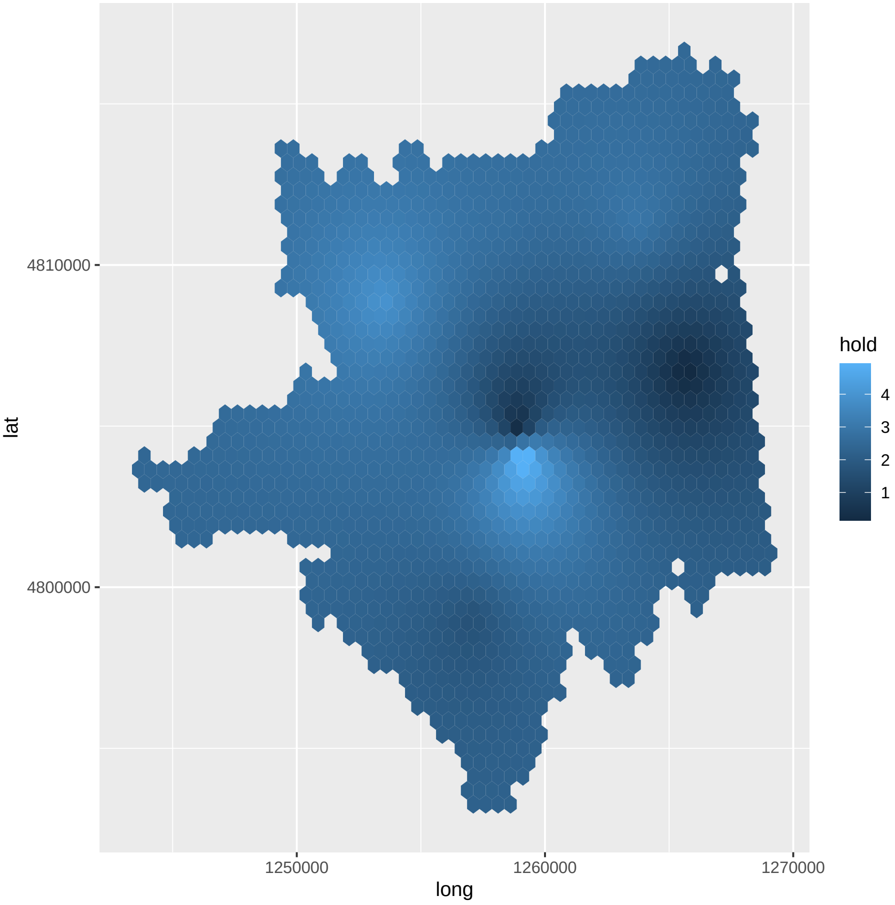

Figure 6: This fig shows different BLL levels in different zipcodes

When we performed kriging from BLLZIPCODE to flint city, we got the BLL of 2.371, which is smaller than the value we got when we performed areal interpolation. We performed kriging interpolation three times, with the same dataset BLLZPspdf, but with different target datasets (flintspdf, zipcodespdf(sfzipcodegrid spdf object) and BLLALLzipcodes). So, to perform krigging from BLLZIPCODE to Zipcodespdf, we repeated the specified above process and got the below the visualization, the lighter the blue, the more seriously the children from that area are affected by the lead.

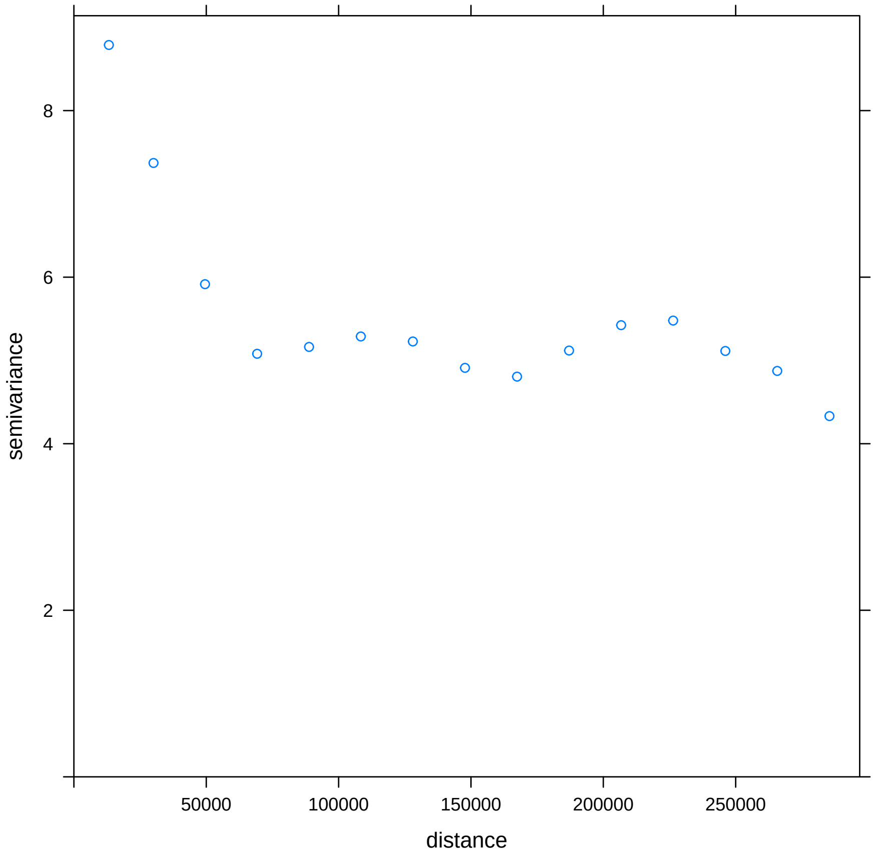

Figure 7: Variogram of all zipcodes in Michigan

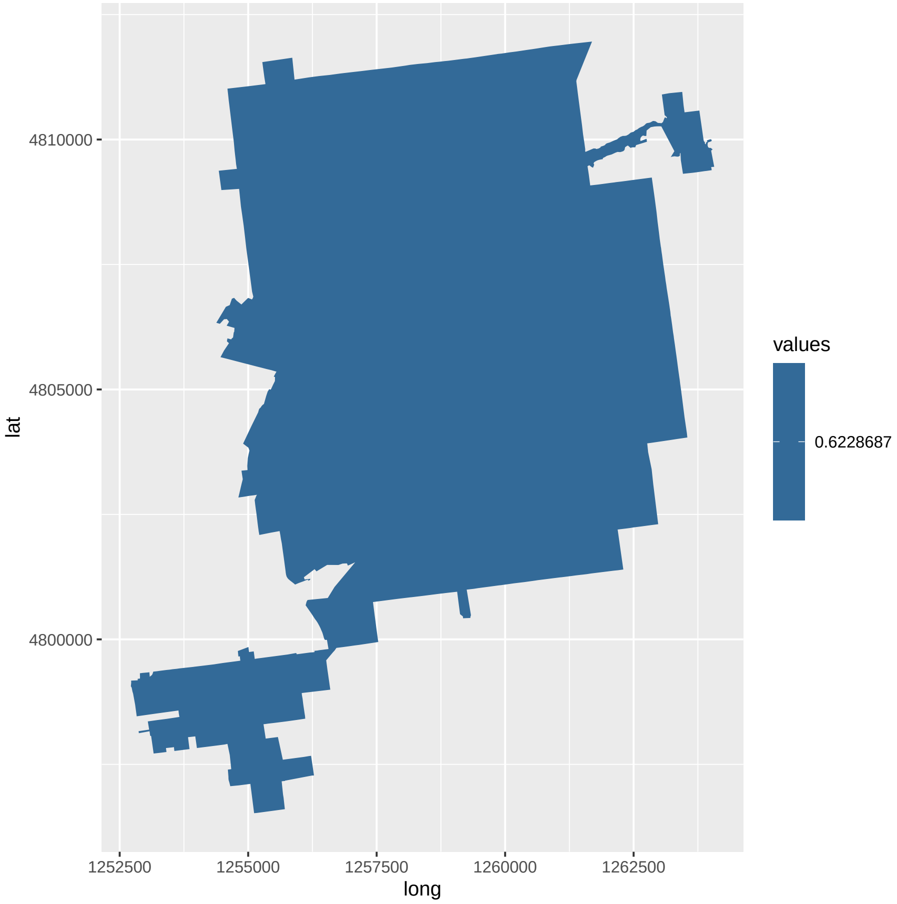

Figure 8: Outline of flint city filled with its average BLL

We tried to krige from all of the Michigan's zip codes to the flint city and got a value of 0.623, which is way smaller than all the values we got before.

References

MPH Jenny LaChance MS Richard Casey Sadler PhD Mona Hanna-Attisha, MD and MD Allison Champney Schnepp. Elevated blood lead levels in children associated with the flint drinking water crisis: A spatial analysis of risk and public health response. *American Journal of Public Health*, 106(2):283–290, February 2016.

Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi: 10.32614/RJ-2018-009. URL `https://doi.org/10.32614/RJ-2018-009`.

Christopher Prener and Charles Revord. areal: An r package for areal weighted interpolation. *Journal of Open Source Software*, 4(37):1221, 2019. doi: 10.21105/joss.01221. URL `https://doi.org/10.21105/joss.01221`.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.