**STAT 7020 Final Project – Spatial Analysis of Ohio Radon Concentrations**

Kyle Mann

Dec 7, 2021

**Table of Contents**

[1 Background 2](#_Toc89816491)

[1.1 Radon 2](#_Toc89816492)

[1.2 Data Source 2](#_Toc89816493)

[1.3 Literature Review 2](#_Toc89816494)

[2 Exploratory Analysis 3](#_Toc89816495)

[3 Modeling 3](#_Toc89816496)

[3.1 Modeling Concerns 3](#_Toc89816497)

[3.2 Kriging 3](#_Toc89816498)

[3.2.1 Ordinary Kriging 3](#_Toc89816499)

[3.2.2 Lognormal Kriging 4](#_Toc89816500)

[3.2.3 Kriging with Adjustment 4](#_Toc89816501)

[3.3 Local Approximate Gaussian Process 4](#_Toc89816502)

[3.4 Predictive Process 5](#_Toc89816503)

[3.5 Markov Random Field 5](#_Toc89816504)

[4 Conclusion 5](#_Toc89816505)

[5 References 5](#_Toc89816506)

# Background

## Radon

Radon is an element that occurs naturally as a radioactive gas, and can pose a harm to human health when it collects in large concentrations inside of buildings. The gas is colorless and odorless, and can thus can go unnoticed without targeted testing. It is a part of the radioactive decay chain of uranium, a radioactive element that is found in underground soils and rocks. When it is formed underground, it can seep into housing, and is especially common in basements. When concentrated inside buildings, radon gas can decay into radioactive airborne particles that can cause damage to the lungs if inhaled. Radon has been found to be the second leading cause of lung cancer after smoking. Radon is measured in pCi/L, a unit that characterizes radioactivity in terms of the number of radioactive decays per second in a given volume of air. The EPA estimates the average outdoor radon level to be 0.4 pCi/L and the average indoor radon level to be 1.3 pCi/L, and recommends taking action to make building modifications if the radon level in a building is above 4 pCi/L (EPA 2016).

Radon levels are affected by the local geology. Though its precursor uranium is found to some degree in nearly all soils, it is found in greater quantities near certain rock deposits (Harrell 1993). In addition to the geology of an area, the radon concentration in any building is also related to the building’s structure. Houses with basements are associated with higher radon levels. In multistory buildings, radon concentrations tend to be lower in the higher levels, leading to lower radon concentrations in urban areas. Indoor radon concentration has also found to be associated with the porosity of building materials and the setup of ventilation systems (Dai 2019). Further variation in radon measurement can come from inter- and intraday variation in true radon concentrations as well as measurement error and inconsistency in the handling of testing equipment (Grafton 1988).

## Data Source

In 1990, Dr. Ashok Kumar and colleagues at the University of Toledo created a database of about 50,000 indoor home radon concentration measurements in Ohio that was complied with data collected from private sector companies, governmental agencies, and universities (Heydinger 1991). Dr. Kumar’s Group maintained and added to this database throughout the 1990s and 2000s and developed a website front-end providing public access to the data, helped by grants from agencies including the Ohio Air Quality Development Authority, the ODH, and the EPA. During this time, the project went under the name, the “Ohio Radon Information System (ORIS)”. By 2007, the database had grown to about 130,000 records (Maroju 2007), and today the database stands at over 600,000 records. The ORIS website is at now deprecated. On my contact, Dr. Kumar indicated that the database has been transferred away from the University of Toledo and is now maintained by the ODH. On the ODH website are published choropleth maps, as well as pdf files containing the summary statistics for each county and zip code in the database. In these summary tables, for each area the of the number of records, minimum, maximum, mean, standard deviation, and geometric mean are provided. I used this as the data source for my analysis.

## Literature Review

In the 1990s, Dr. Kumar collaborated with Toledo geologist Dr. James Harrell and the Ohio Geological Survey to study the association between indoor concentrations and geologic formations. It was found that in Ohio, radon was associated with both the Ohio Shale outcrops and the Miami and Scioto glacial lobes of the Wisconsinan glacial deposits (Harrell 1991, and Harrell 1993).

In 2007, Suman Maroju, a PhD student advised by Dr. Kumar submitted a thesis analyzing the ORIS data with the goal of estimating indoor radon concentrations in zip codes that had no database measurements (Maroju 2007). For this analysis, Maroju took the geometric mean of each zip code in the database, converted it to a point feature at the zip code centroid, applied a log transformation, performed first order detrending, and then tried five different geostatistical interpolation techniques using ArcGIS Geostatistical Analyst. Maroju tried ordinary kriging, as well as the deterministic methods of inverse distance weighted, global polynomial, local polynomial, and radial basis functions. Maroju selected kriging as the best modeling method through comparing the methods via metrics such as Normalized Mean Square Error and Fractional Bias from train/test validation as well as visually inspecting maps of the predicted values.

In 2010, Kumar and colleagues compared the method of kriging following Maroju’s approach with cokriging using underground uranium levels as a covariate, and found that cokriging performed slightly better under train/test validation (Kumar 2010).

In 2010, Akkala, a PhD student at Toledo submitted a thesis trying out the use of neural networks to predict zip code GM radon concentrations based on spatial coordinates. Akkala tried three-layer neural networks with various numbers of nodes in the hidden layer, and also experimented with several network architectures designed to incorporate uranium concentration as a covariate. Using train/test set validation, Akkala found that the neural network methods could achieve higher accuracy than the kriging and cokriging methods used by Maroju and Kumar (Akkala 2010).

In a 2013 PhD thesis, Gummandi tested the methods of Maroju, Kumar, and Akkala, and compared them against Support Vector Regression (SVR) and Random Forest Regression (RFR) and found that SVR and RFR both had lower accuracy than the neural network methods. Gummandi argued that despite the lower accuracy, SVR may be less prone to overfitting if the sample size were smaller (Gummandi 2013).

In a 2014 PhD thesis, Bandreddy repeated the analysis of Gummandi, additionally using a Quantile Regression Forest model, and found that the model performed favorably compared with all of the prior models including Akkala’s neural networks.

In 2019, geographer Dr. Yanqing Xu worked with Dr. Kumar to study the spatial correlation between indoor radon concentrations and fracking wells in Ohio. Taking a different approach than most of the previous analyses, the authors obtained a dataset of about 112,000 of the home-level records from the ORIS database, and set up a multilevel model using zip code-level variables in addition to the individual home-level variables. The home-level predictors included season and type of radon measurement device and the zip code-level variables included the distance from the zip code centroid to a fracking well, the population density, and whether the zip code was urban, rural or suburban. The results of the analysis indicated that the distance from zip code centroid to the nearest fracking well was negatively correlated with the indoor radon concentrations. That is, zip codes closer to fracking wells tended to have lower higher radon concentrations (Xu 2019).

In summary, the majority of the analyses of the ORIS data have focused on zip code-level geostatistical analysis and interpolation, using the geometric mean of each zip code, often log-transformed. Ordinary kriging and cokriging with underground uranium levels as a covariate were found to be more accurate than deterministic methods. Neural network and quantile regression forest methods were found to be more accurate than kriging and cokriging.

# Exploratory Analysis

## Raw Data

The PDF report of the zip code level radon concentration summary statistics was accessed at the ODH website at the link provided in the references section. The table contained in the PDF was copy-pasted to a Notepad file, then copied into Excel and converted to .csv format. The first ten records of the CSV are displayed in Table 1. Zip code is the postal zip code, N is the number of home-level radon measurements present in the database, and min, max, mean, SD (standard deviation), and GM (geometric mean) are summary statistics for the radon measurements for the named zip code. Units for radon measurements are pCi/L. There are 1746 rows present in the raw data, each with a unique 5-digit zip code.

Table 1

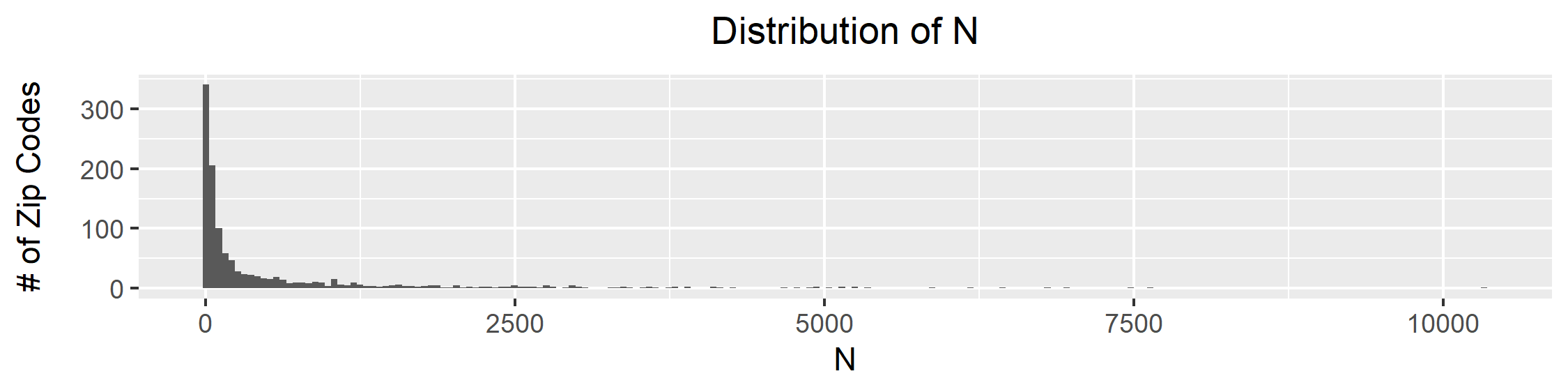
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Zip Code | N | Min | Max | Mean | SD | GM |
| 43001 | 335 | 0.1 | 317.7 | 15.7 | 40.5 | 6 |
| 43002 | 11 | 0.1 | 40.1 | 12.2 | 12.4 | 7.8 |
| 43003 | 86 | 0.1 | 56.1 | 9 | 11.3 | 4.4 |
| 43004 | 2925 | 0.1 | 126.5 | 7.6 | 9.9 | 4.4 |
| 43005 | 10 | 0.1 | 27.6 | 10.8 | 8.1 | 7.4 |
| 43006 | 16 | 0.1 | 15.9 | 7.5 | 5.3 | 5.3 |
| 43007 | 7 | 0.1 | 14.7 | 7.2 | 4.6 | 5.6 |
| 43008 | 59 | 0.1 | 14.3 | 2.7 | 3.1 | 1.6 |
| 43009 | 121 | 0.1 | 106.5 | 13.7 | 19.4 | 6.3 |

The US census publishes shape files for US Zip Code Tabulation Areas (ZCTAs), areas corresponding to US postal zip codes. 2010 census populations are provided for ZCTAs in the in the census-published zip code-to-county correspondence file. Merging the radon data with the shape files and keeping only zip codes that are present in both sources leaves 1155 of the original 1746 zip codes. After further merging with the 2010 census populations, we are down to 1134 zip codes. Though 612 (35%) zip codes were dropped, inspecting N, we find the dropped zip codes account for only 2958/612,119 (0.5%) of the home-level sample size. That is most of the dropped zip codes had very small N. Perhaps many of the dropped zip codes were transcription errors or deprecated zip codes. For subsequent analysis, we use the radon dataset post shape file merge. For analyses using the 2010 census population, we use the dataset post shape file and population merges.

## Univariate Summaries

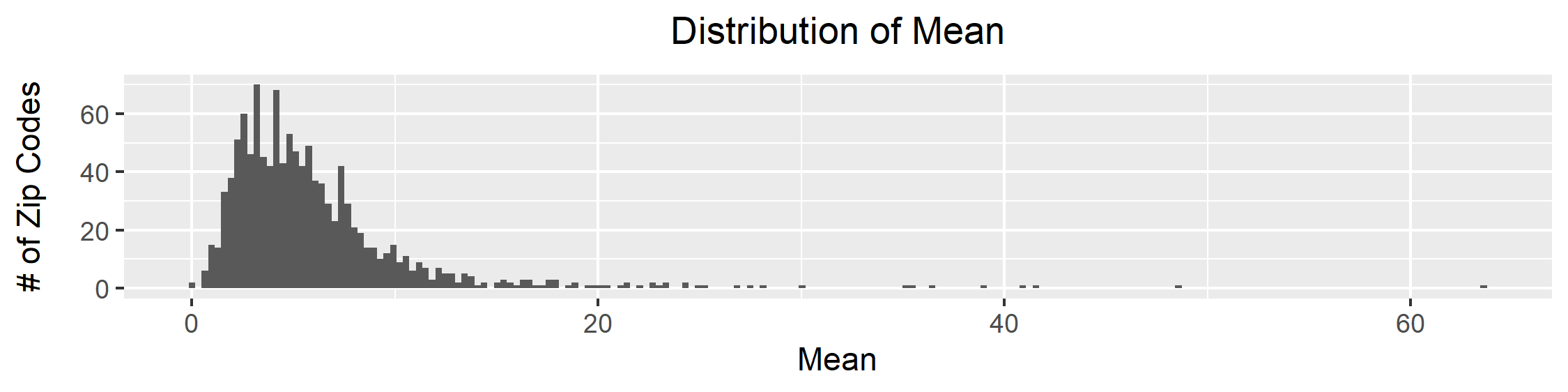
### N

While the number of observations per zip code ranges from 1 to 10,331, most zip codes have a relatively small number of observations. The first quartile is 20, the median is 90, and the third quartile is 501. The distribution of values is plotted in Figure 1.

Figure 1 

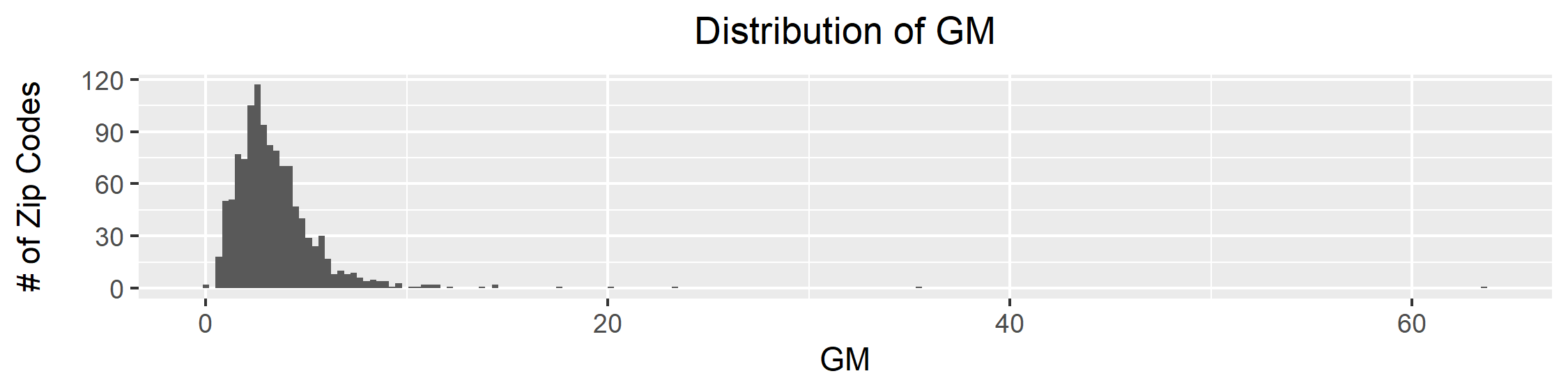
### Mean

The mean radon concentrations range from 0.1 to 63.7 with median 4.9 and quartiles 3.1 and 7.3 The distribution of the mean is plotted in figure 2. The distribution of the mean is far from symmetrical and highly skewed. We will not use the mean in the analysis, and will instead use the geometric mean, as it is a more reliable central statistic for radon concentrations due to the skewed distribution of the home-level measurements that is approximately lognormal.

Figure 2 

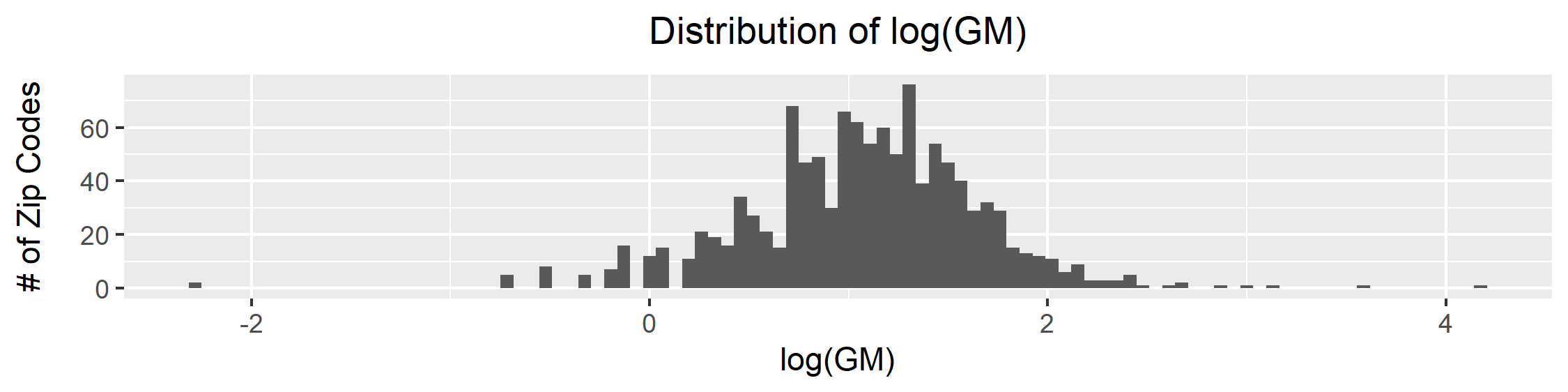
### Geometric Mean (GM)

The GM radon concentrations range from 0.1 to 63.7 with median 3.0 and quartiles 2.1 and 4.2 The distribution of the mean is plotted in figure 3. As for mean, the distribution of GM is far from symmetrical and highly skewed, with some extreme outliers on the upper end.

Figure 3

Taking the natural logarithm of the geometric mean, the distribution of the transformed variable is plotted in Figure 4. The distribution of is fairly symmetric and is appears approximately consistent with data drawn from a normal distribution, indicating that is approximately lognormal.

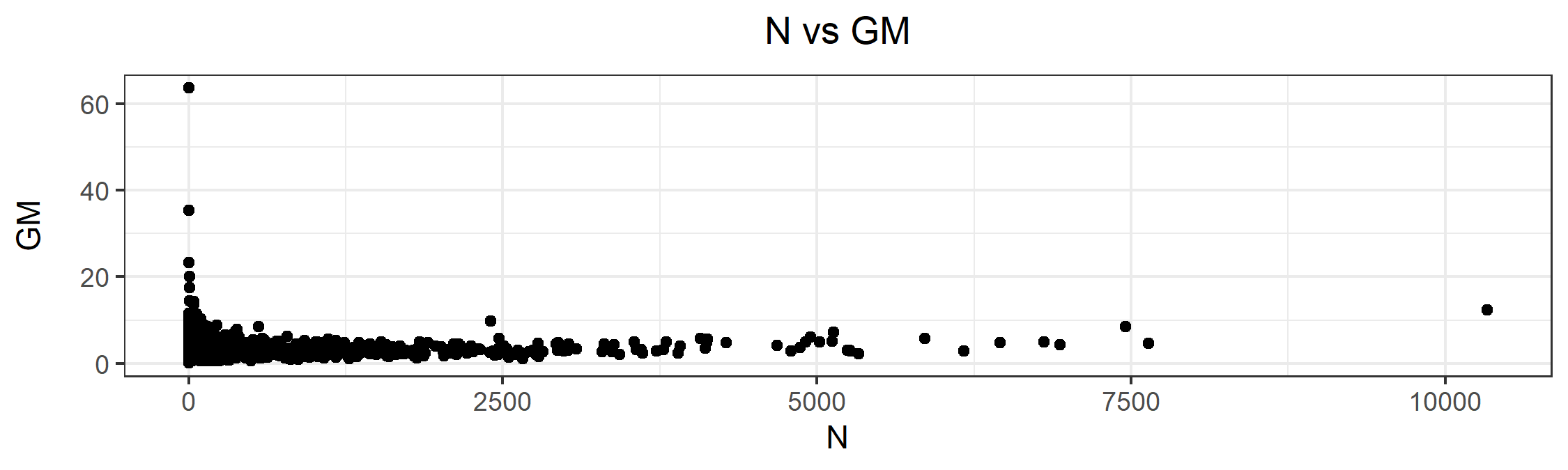
Figure 4



## Extreme Values

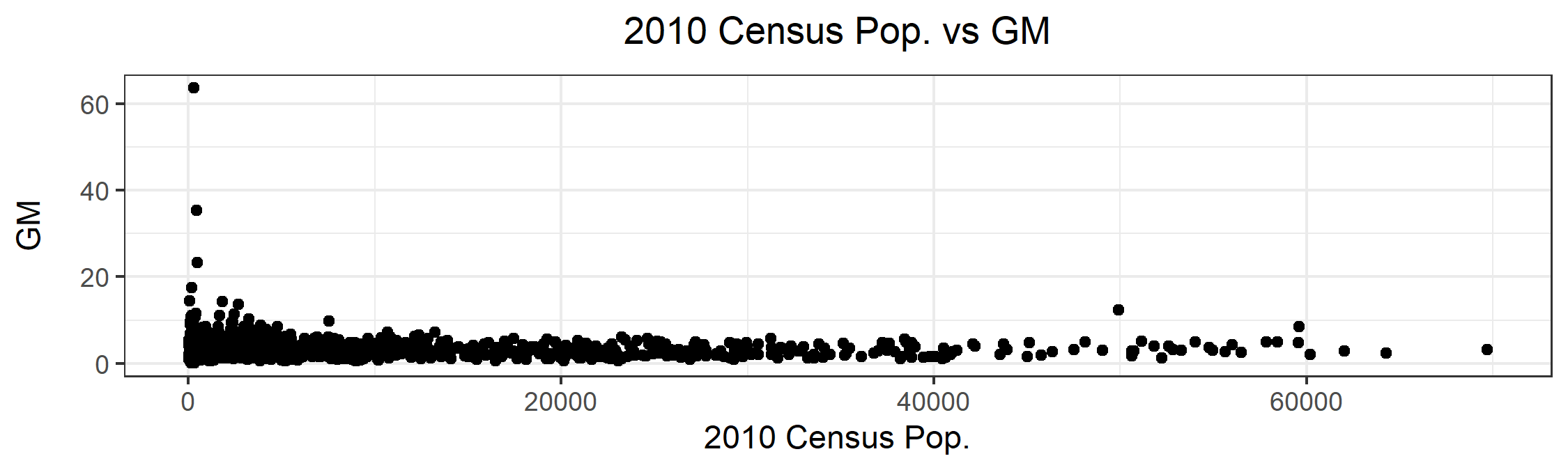
As seen in sections 2.2.2 and 2.2.3, the geometric mean has some extreme outliers on its right tail. To illustrate the cause of these outliers, the samples size (N) for each zip code is plotted against the GM. There is greater variability in the GM for zip codes with lower samples sizes. This is naturally expected, and illustrates that the high outliers may not be indicative of truly larger radon levels, but may just be due to the high GM variability for low-N samples.

Figure 5



In Figure 5, we can also see that to the right of N=250, there appears to be a positive correlation between the samples size and geometric mean. Interestingly, when we recreate this plot using the 2010 census population instead of N, this correlation disappears. This illustrates the fact that in zip codes with higher radon levels, there may be a higher amount of radon testing done per capita.

Figure 6



Having noted the extreme values, as well as the reason for their occurrence, we present the six zip codes with highest geometric means in Table 2.

Table 2

|  |  |  |  |
| --- | --- | --- | --- |
| Zip Code | GM | N | 2010 Census Population |
| 45816 | 63.7 | 1 | 291 |
| 45650 | 35.4 | 2 | 435 |
| 43144 | 23.3 | 1 | 477 |
| 44639 | 20.1 | 6 | - |
| 43317 | 17.5 | 6 | 168 |
| 43925 | 14.4 | 5 | 82 |

## Mapping

With the shape file merged in with the radon data, we can create choropleth maps of the radon concentrations in Ohio

Figure 7

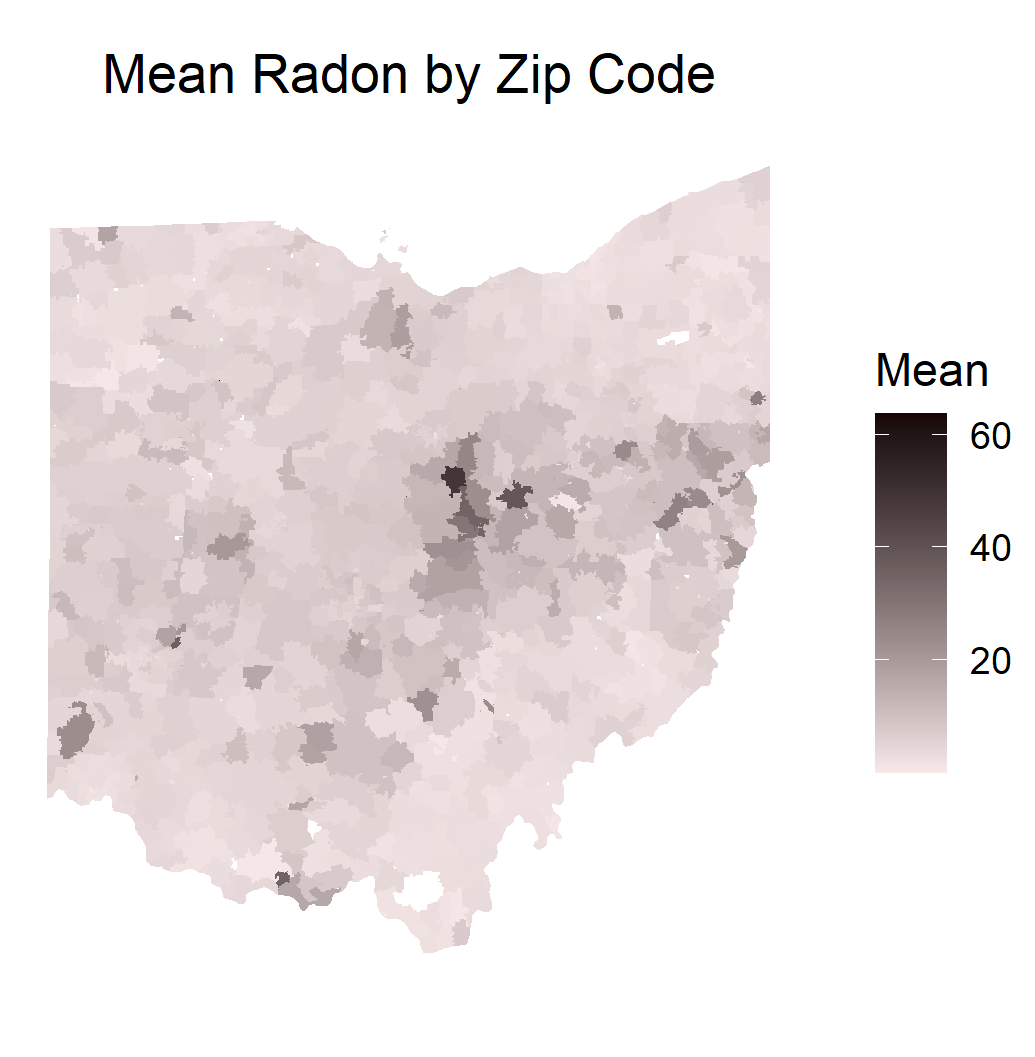


Figure 8

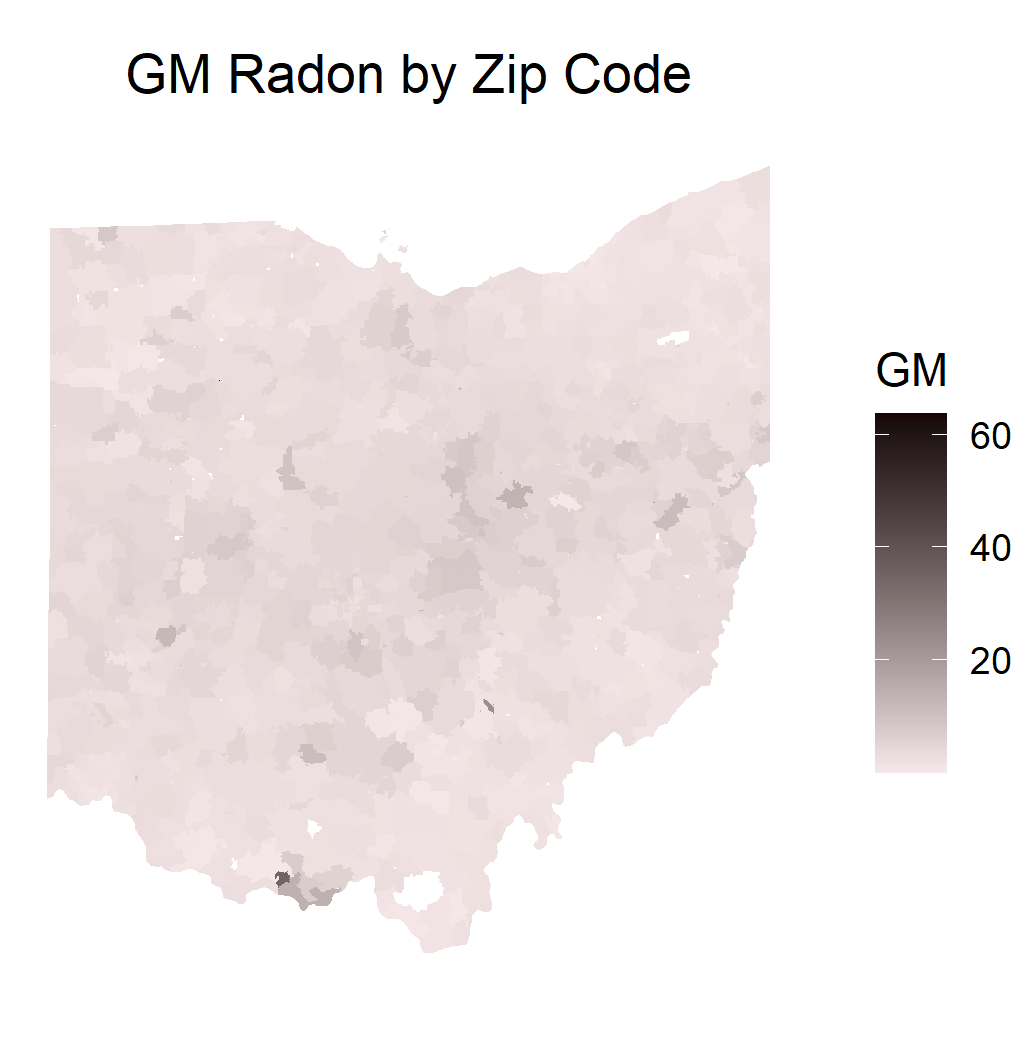
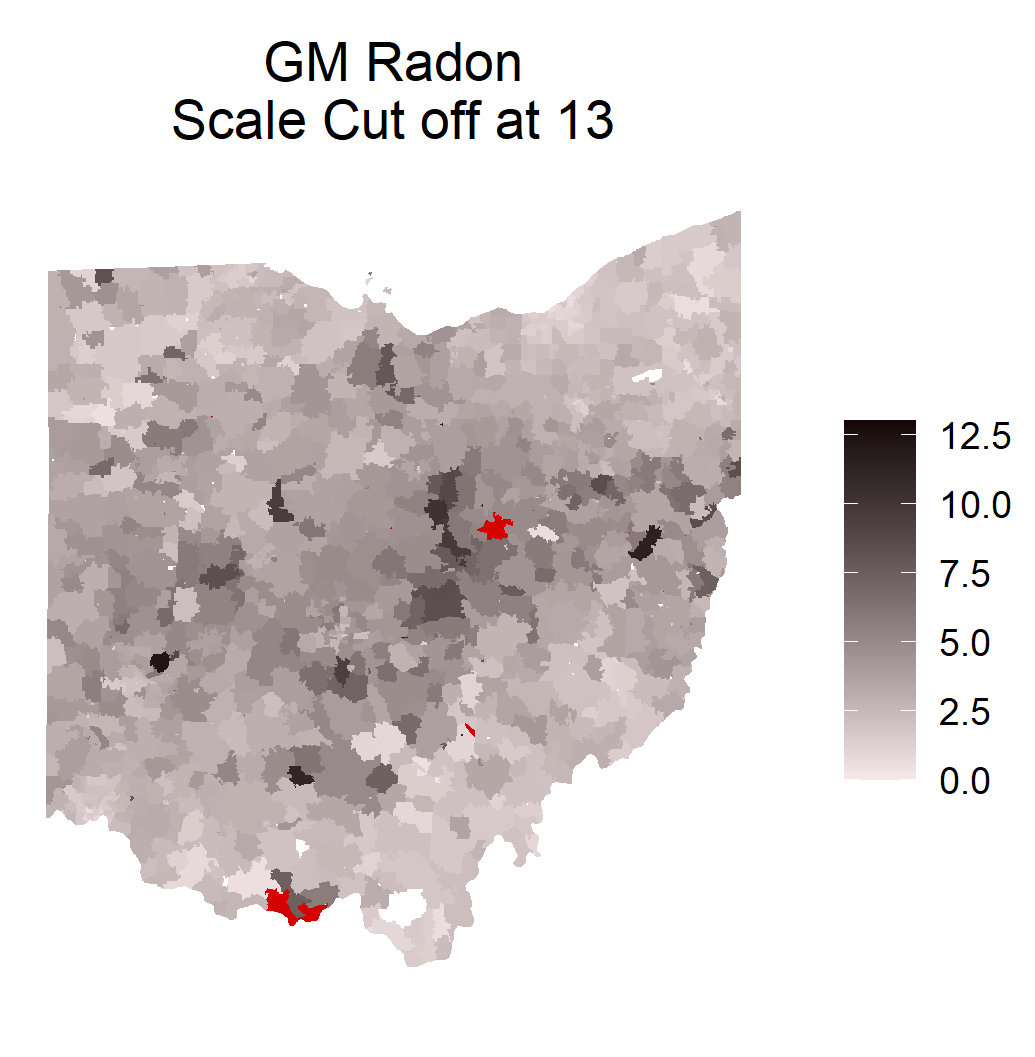


Figure 9



In Figure 8, the outliers throw off the scale such that the map of the geometric mean is hard to read. In Figure 9, all values above 13 are cut off of the scale and colored red.

Figure 10

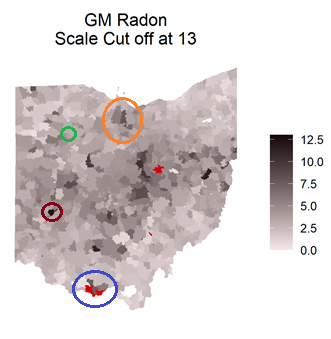


Figure 10 is Figure 9 with several circles superimposed to highlight portions of the map. We will refer to these areas by the following names from here on out:

**Benton Ridge** - The green circle indicates the zip code that has the highest GM of 63.7 pCi/L. This zip code is so small in area that it can hardly be seen on the map. It is a tiny spec of red! The N for this zip code is only 1. This zip code falls on the town of Benton Ridge.

**Rockville** - The blue circle falls on a cluster of three zip codes, that have high GMs and a small-medium sample size. The sample size is larger than Benton Ridge, but still small in comparison with some other parts of the state.

**Norwalk** – The orange circle highlights several zip codes around the town of Norwalk that have higher than average GMs and also a substantial sample size.

**Dayton** -The maroon circle highlights a zip code in Dayton, which has a higher than average GM. This zip code also has one of the largest sample sizes.

Based on our observations of the small-sample variability in Figure 6, examining these four areas, we have the highest confidence that Dayton has truly high radon levels, followed by Norwalk, then Rockville, then Benton Ridge, in that order. As we observe the modeling predictions in section 3, we can keep these four areas in mind.

In Figures 11 and 12 the GM radon is plotted by zip code with the scale cut off at 13, similar to the choropleth map in Figure 10, however the zip codes are plotted with scatter plot markers with the marker size based on the samples size N, and the 2010 census population ZPOP, respectively. This view of the data can provide both a picture of how the GM varies across the state, and also how the sample sizes vary. We see the largest samples sizes in urban areas such as Columbus, Cincinnati, and Dayton.

Figure 11

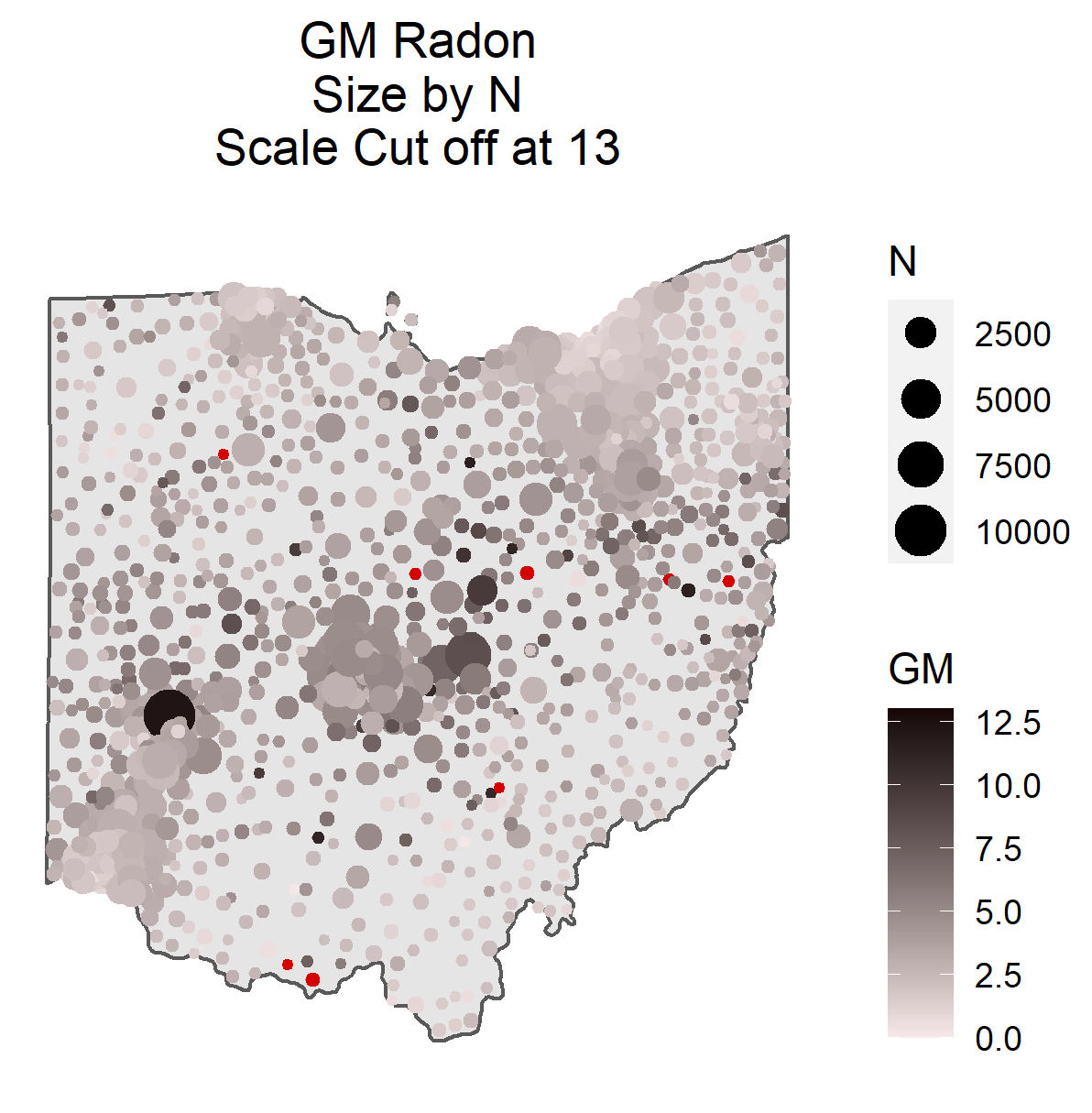
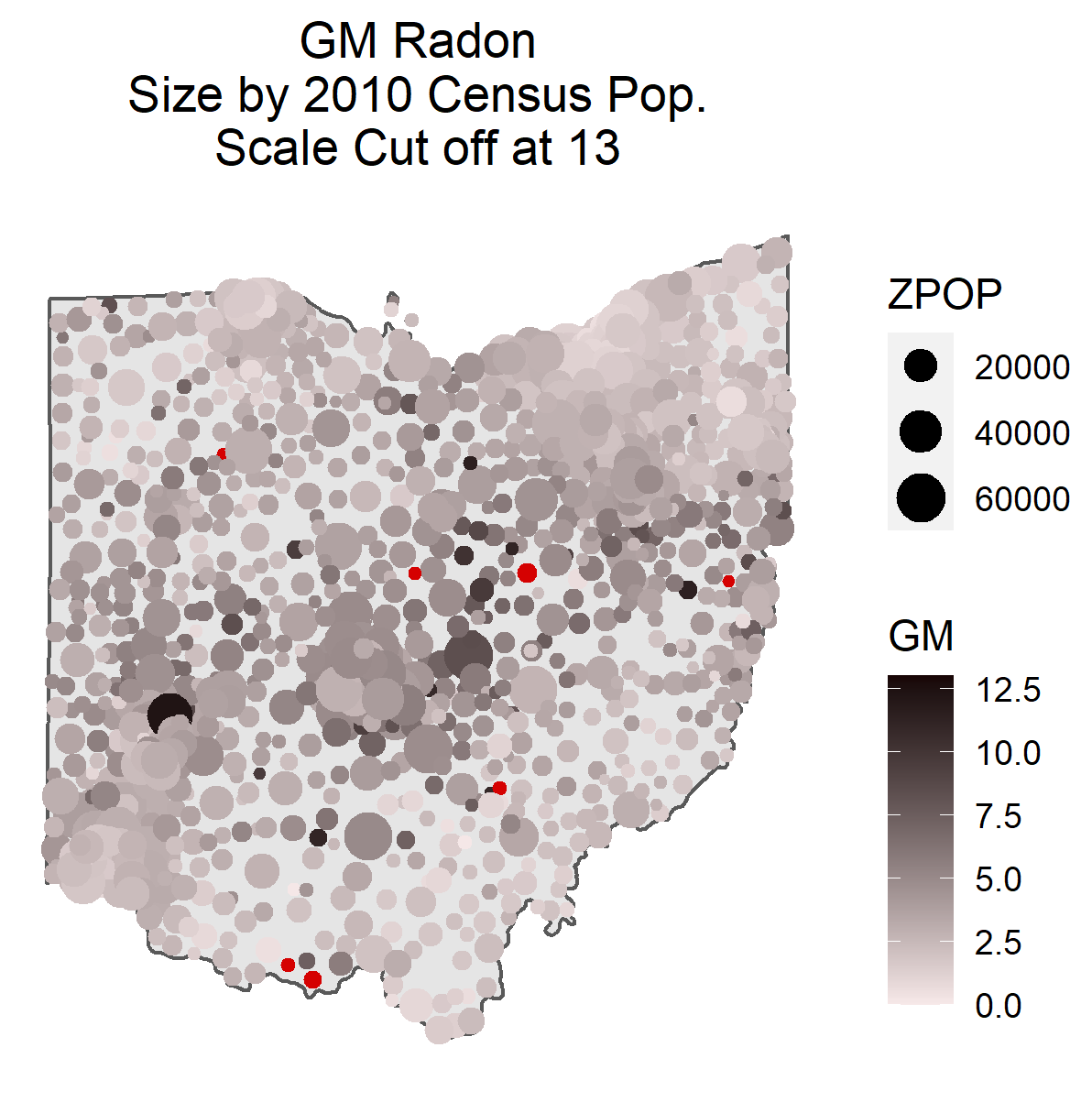


Figure 12



# Modeling

## Overview

For our investigation, we start off with a similar approach to the bulk of the analyses conducted by Dr. Kumar’s research group, as described in section 1.3. We will first consider the zip code geometric means as point features located at the centroid of each zip code ZCTA, and attempt to build a model that can be used to predict the geometric mean indoor radon concentration at any point in the state of Ohio. We will try out the techniques of ordinary kriging, lognormal kriging, local approximate Gaussian processes, and predictive processes. After considering the geostatistical mentioned above, we will try a Markov random field model that considers the lattice structure of the data.

Several points to keep in mind throughout the analysis:

* The distribution of GMs is non-symmetric. In section 2.2.3, we found it to be approximately lognormal.
* Each zip code has a different sample size N. There is greater variability in the GMs for the zip codes with smaller N. This correlation was illustrated in section 2.3.
* In figure 10, we identified 4 regions that all have higher than average GM, but varying sample sizes. In order of increasing sample size, these regions were Benton Ridge, Rockville, Norwalk, and Dayton.

Whenever converting from lattice data to point features, we need to consider whether the centroid will well represent the locations from which the samples were taken. If the areal features were large, such as counties, then there would be the concern that most of the samples could have been taken from some far-flung corner of the county that is not well represented by the centroid. Because zip codes in Ohio cover fairly small areas, this concern is mitigated, though not eliminated.

## Kriging

### Ordinary Kriging

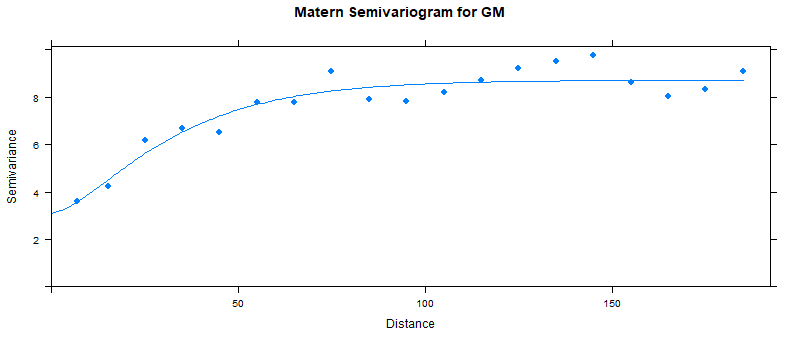
Ordinary kriging assumes that the data is generated from a second-order stationary and isotropic spatial random process with unknown mean. This can be written as follows, where is the spatial process, is a spatial point, is the unknown mean, and is a mean-0 isotropic process.

The ordinary kriging predictor for point is set up as follows where are the observed points

The predictor is derived by finding that minimize the mean squared prediction error (MSPE) subject to the constraint .

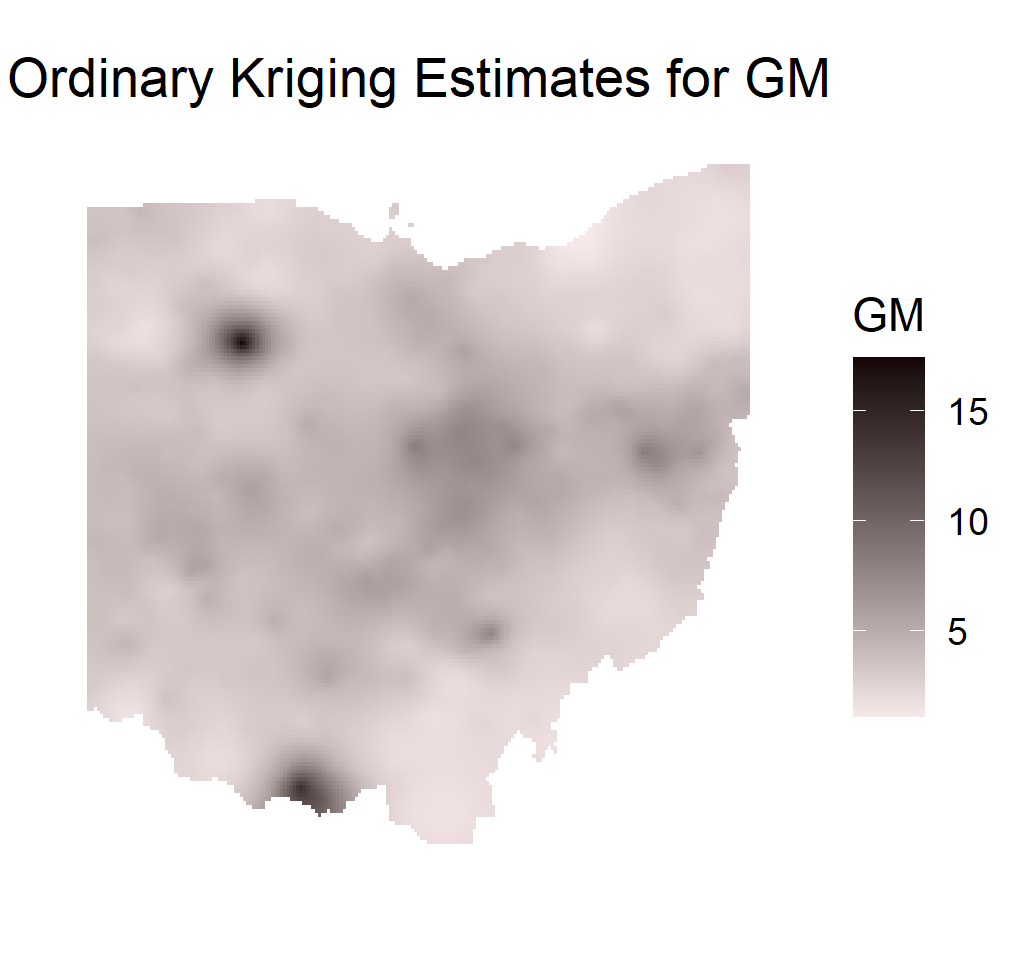
Ordinary Kriging is applied to point features derived as the GM of each zip code, located at the zip code centroid. First, the covariance structure is estimated by fitting an empirical semivariogram, and then fitting a Matern model including nugget to the empirical semivariogram points, displayed in Figure 13.

Figure 13



Supplying the estimated semivariogram to as the covariance structure, ordinary kriging predictions are computed across a fine-grain grid covering the state of Ohio, and plotted in Figure 14.

Figure 14



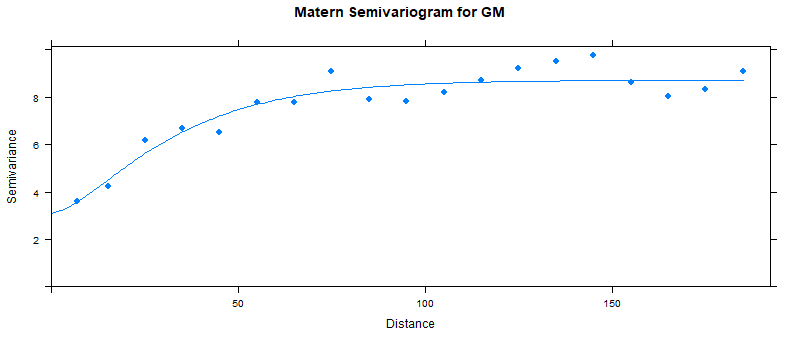
On first glance, we can see some considerable problems with the kriging predictions. Namely, the outlier at Benton Ridge in the northwest has a large affect on the predicted values. Remember, the zip code of Benton ridge had a sample size of 1, and so it is not desirable for this value to exert so much influence on the predicted values. What could be causing this problem? Although ordinary kriging does not rely on any assumptions about the shape of the random process’s PDF beyond it’s isotropic stationarity, in practice, a skewed distribution with extreme outliers can throw off the predictions. Now remember that the GM was seen to be approximately lognormal. With this in mind, we try a second approach called Lognormal Kriging.

### Lognormal Kriging

With lognormal kriging for GM, we first perform a log transformation on GM and then perform ordinary kriging. Now if we were interested in predicted values of log(GM), then we could stop here. However, in order to interpret the results, we want to transform the predictions back into the original units. If we were to backtransform by simply taking the exponential of the predicted values, this would unfortunately provide biased estimates of the GM (Yamamoto 2007). Yamamoto provides several adjusted backtransformation methods that are designed to reduce this bias. We apply the backtransformation from equation (3) of this paper, such that the equation is given as follows, where is the ordinary kriging estimate on the log transformed data, is the associated variance, and is a factor computed such that the mean of the backtransformed values match the mean of the original data:

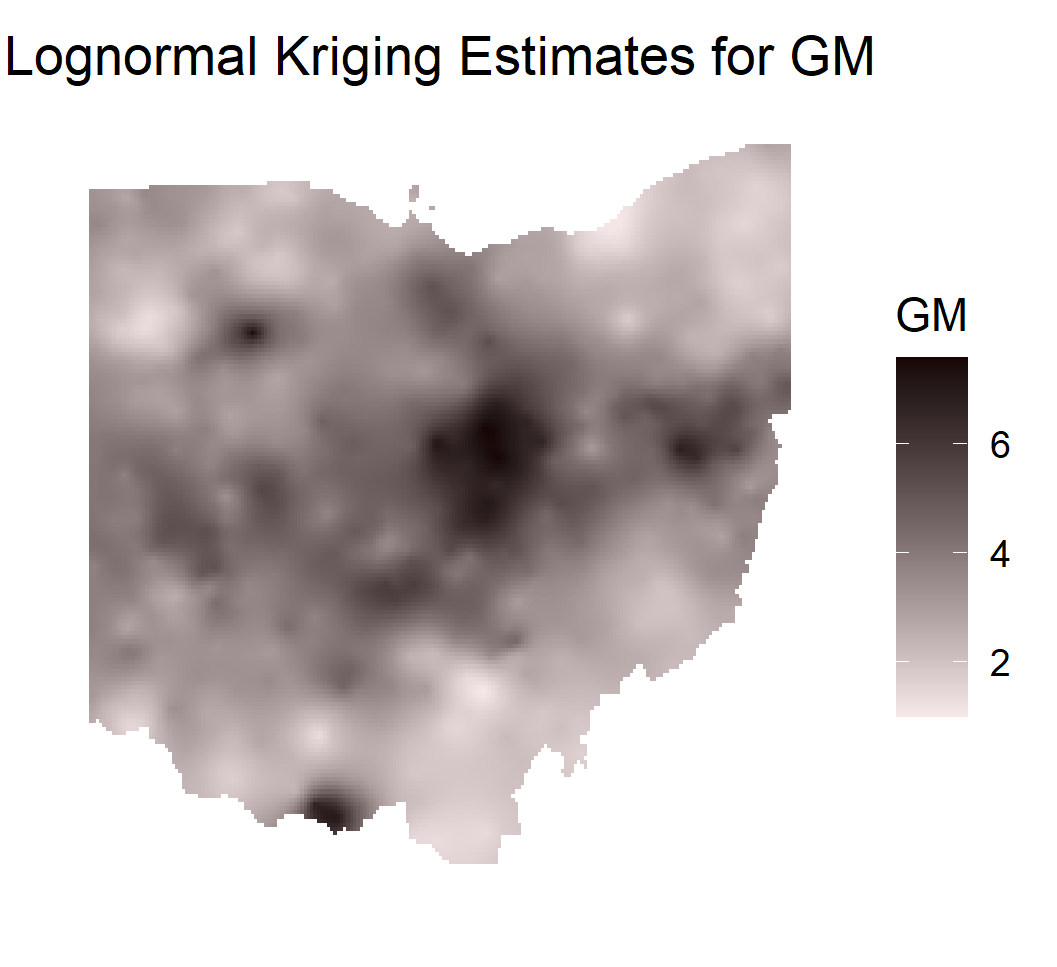
As for ordinary kriging, we estimate the covariance structure by fitting an empirical semivariogram, and then fitting a Matern model including nugget to the empirical semivariogram points, displayed in Figure 15.

Figure 15



After obtaining the ordinary kriging estimates for the log(GM), and applying the backtransform, the lognormal kriging predicted values are plotted in figure 16.

Figure 16



The predicted values for the lognormal kriging look preferable to those for ordinary kriging. The predicted values near Benton Ridge and Rockville are no longer dwarfing those for the rest of the state. However, it appears that Benton Ridge in the northwest still has undue influence on the predicted values. Remember that Benton Ridge is a zip code that accounts for just a single one of the 600,000+ radon measurements in the database. What is causing this undue influence, even with the log transformation? One apparent problem with the lognormal kriging approach is that like ordinary kriging, it assumes that the data are being generated by an isotropic stationary spatial process. In our case, we can say for certain that the process is not stationary. Because the variability of the GMs is correlated with the zip code samples size, this means that the variance is not stationary.

## Local Gaussian Process Approximation

In a method called local Gaussian process approximation, Gramacy and Apley proposed a method of predicting values in spatial processes as follows: For each point to be predicted, a local Gaussian Process is fit on a selection of nearest neighbors, in which the mean squared prediction error is minimized.

Predictions are made using the laGP package in R using the MSPE fit method, and plotted in Figures 17 and 18. In figure 17, the local Gaussian process approximation is applied directly to GM. In Figure 18, the local Gaussian process approximation is applied to log(GM) and then the predicted values are backtransformed. The same backtransform adjustment as in lognormal kriging is performed, admittedly without theoretical justification.

Figure 17

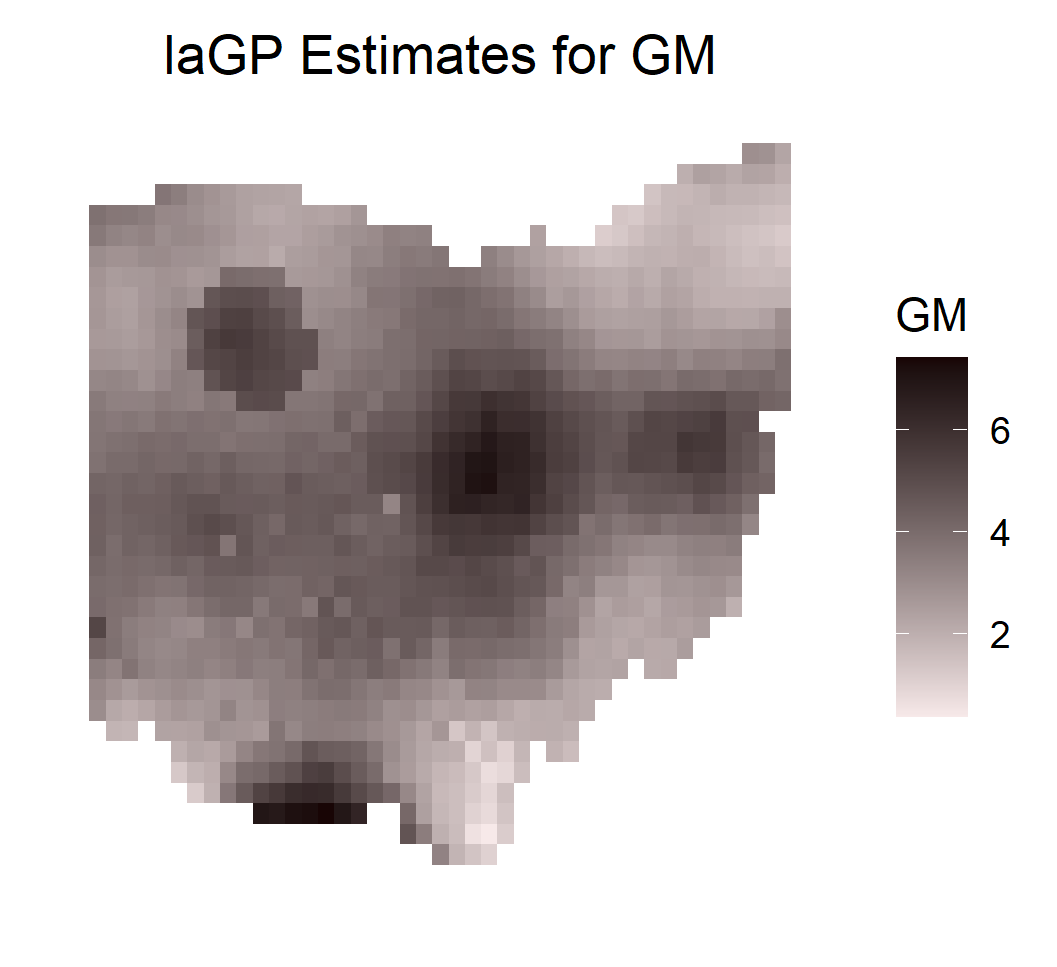
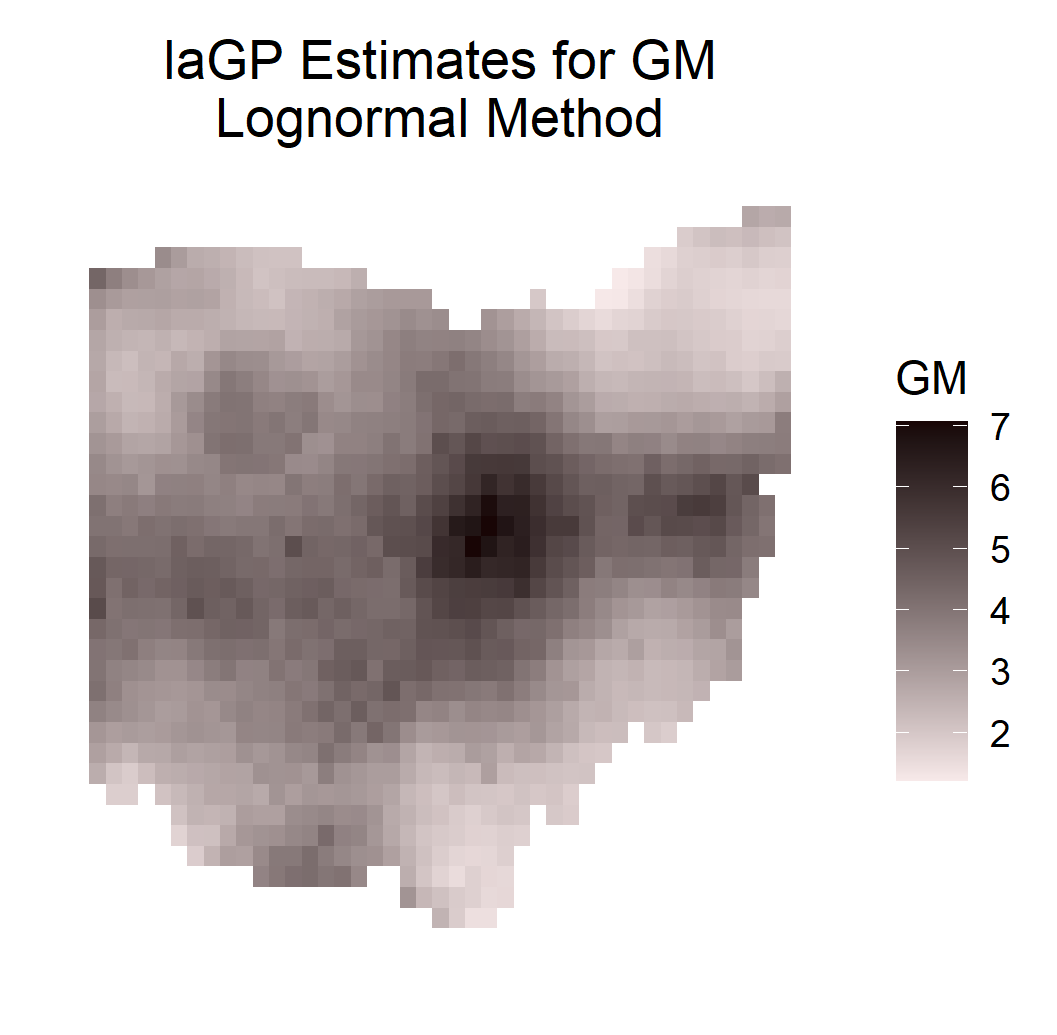


Figure 18



Inspecting the predicted values, the local Gaussian process approximation performs poorly. Benton Ridge appears to have even wider-ranging influence compared with the ordinary and lognormal kriging approaches. A strengths of local Gaussian process approximation is it’s allowance for nonstationary covariance structures, due to the local nature of its estimation. However, our data not only has regional nonstationarity (i.e. urban areas have higher sample sizes and lower variance), but also considerable point-to-point non-stationarity due to the variation in sample size from one zip codes to another. It is possible that the local Gaussian process approximation method cannot account for this point-to-point nonstationarity even if it can account for the regional nonstationarity.

## Predictive Process

The predictive process approach is a hierarchical Bayesian model. First, “knot” points are specified in the spatial domain, for example as a low-resolution grid of points, though they are not required to be evenly spaced. Write the knot locations and let . Then the spatial random process is modeled with a hierarchical structure where its value is related to both the point-level effects and knot-level effects.

Predictions using the predictive process method are made using the spBayes package. In figure 17, the predictive process approach is applied directly to GM. In Figure 18, the predictive process approach is applied to log(GM) and then the predicted values are backtransformed. The same backtransform adjustment as in lognormal kriging is performed, admittedly without theoretical justification.

Figure 19

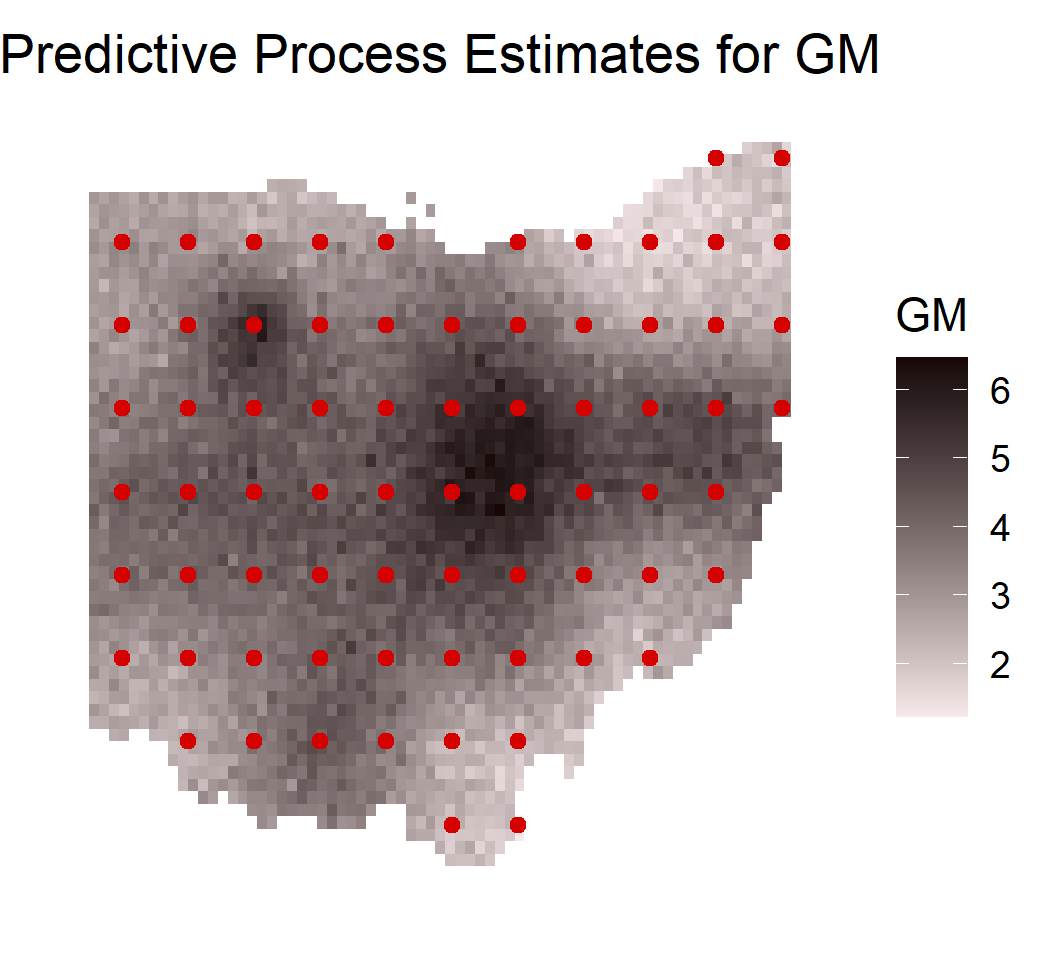
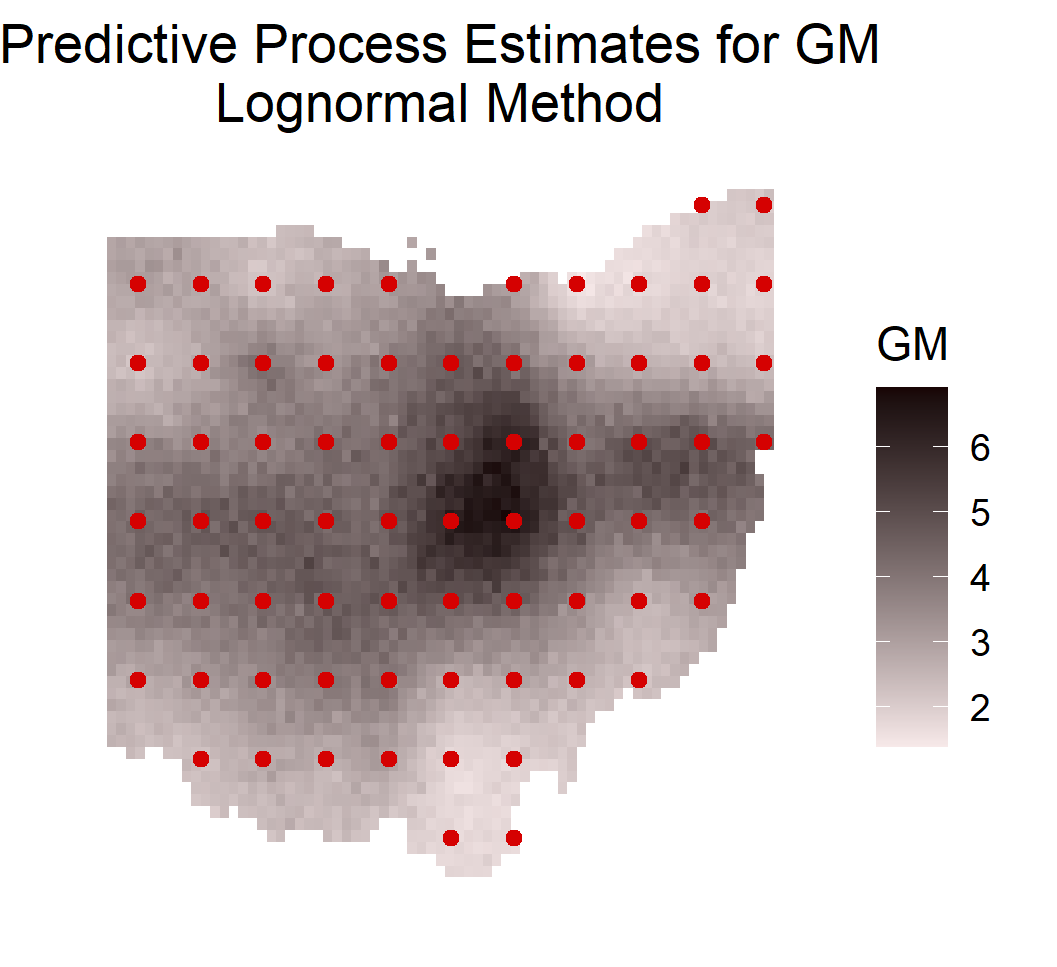


Figure 20



The predicted values display two weaknesses of the predictive process model for our application. First of all, the predictions appear to be oversmoothed, a known weakness of the model. For example, there is nothing that distinguishes Dayton from it’s surrounding zip codes. It is possible that the oversmoothing could be improved by experimenting with different configurations of knot points. A further weakness is the fact that Benton Ridge still has undue influence on the predictions. While the predictive process method can account for non-stationary random processes, it is possible, as for the local Gaussian process approximation, that it may handle regional nonstationarity better than point-to-point nonstationarity. In order to make predictions that are not affected by the low samples size outliers, we need to find a method that incorporates the samples sizes as input into the model.

## Gaussian Markov Random Field

Having tried a number of geostatistical modeling approaches and finding that all of them gave undue influence to the outliers coming from low samples size zip codes, we turn to a method that both uses the lattice structure of the original data and can account for the varying zip code sample sizes. This method is modeling the GMs using a Markov random field. When lattice data is modeled using a Markov random field, each section of the lattice is considered as a node in a network. In our application, two zip codes (nodes) are connected if they are geographically adjacent. The Markov property is assumed to be true, which means that each node may be correlated with its neighbors, but is conditionally independent of the other nodes in the network, conditioning on the values of its neighbors.

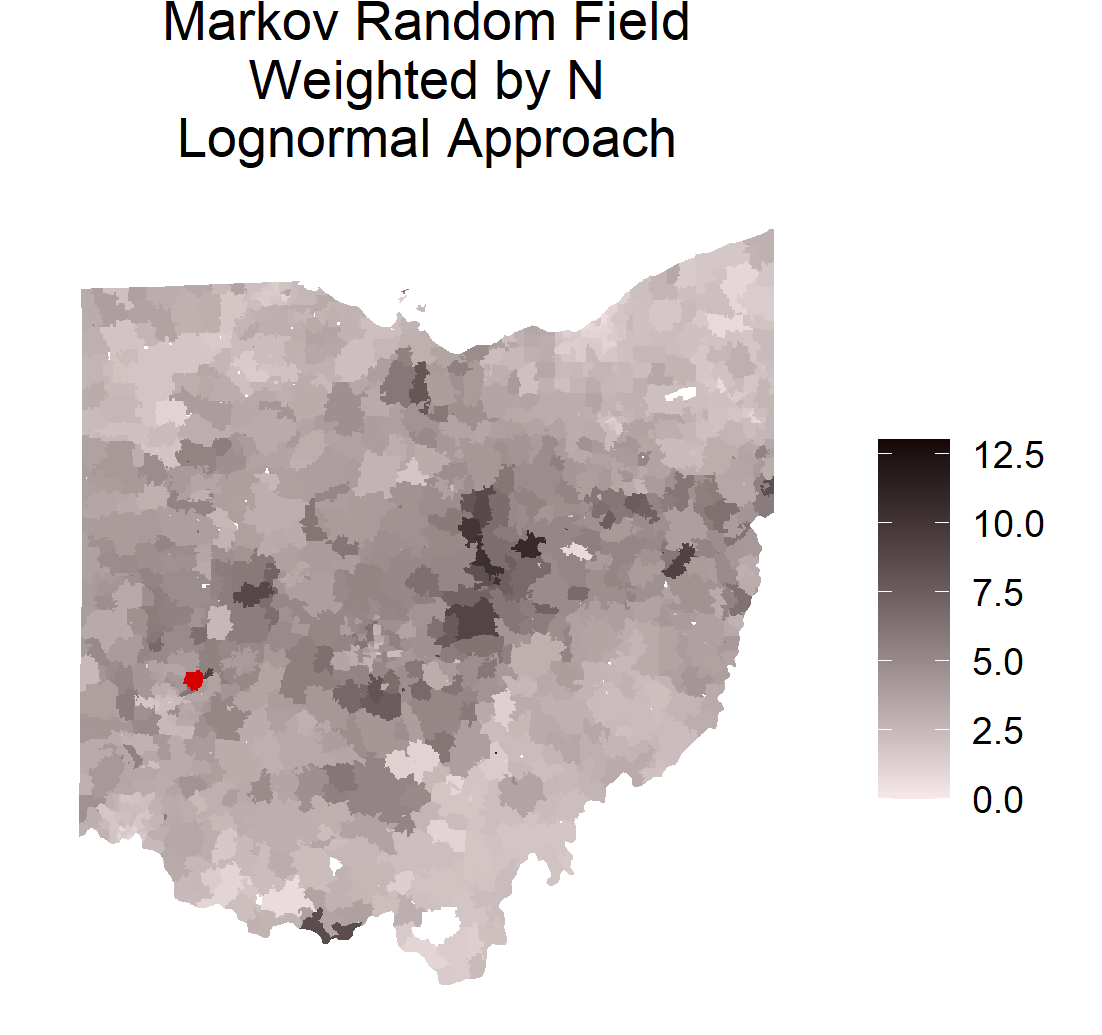
For our case, we will fit a Gaussian Markov random field using the gamlss.spatial package. Following the documentation present in (Bastiani 2018), our model can be written,

is a nx1 random effects vector (one effect for each observed zip code) that is assumed to be multivariate Gaussian, is the supplied precision matrix that can be computed from the neighbor structure of the lattice data, and is a diagonal matrix of weights. If we have sample sizes associated with each zip code, we can supply the sample sizes as the matrix. When is inverted and multiplied by the error variance parameter , it will decrease variance for the larger samples size zip codes and increase the variance for the larger sample size zip codes. This this model can account for our expectation that smaller sample size zip codes should have a larger variance.

I’ll note that one aspect of the Bastiani 2018 paper that I don’t understand at the time of writing is that in the paper’s model specification, the overall mean is not included. I did not center the data at zero in preprocessing, so if the function MRFA() in gamlss.spatial cannot account for a non-zero population mean, then I may be implementing this part of the analysis incorrectly.

The Markov random field parameters are estimated using the gamlss.spatial package. Because we are modeling approximately lognormal data using a Gaussian model, we use the lognormal/backtransform method. The resulting predictions are displayed in Figure 21.

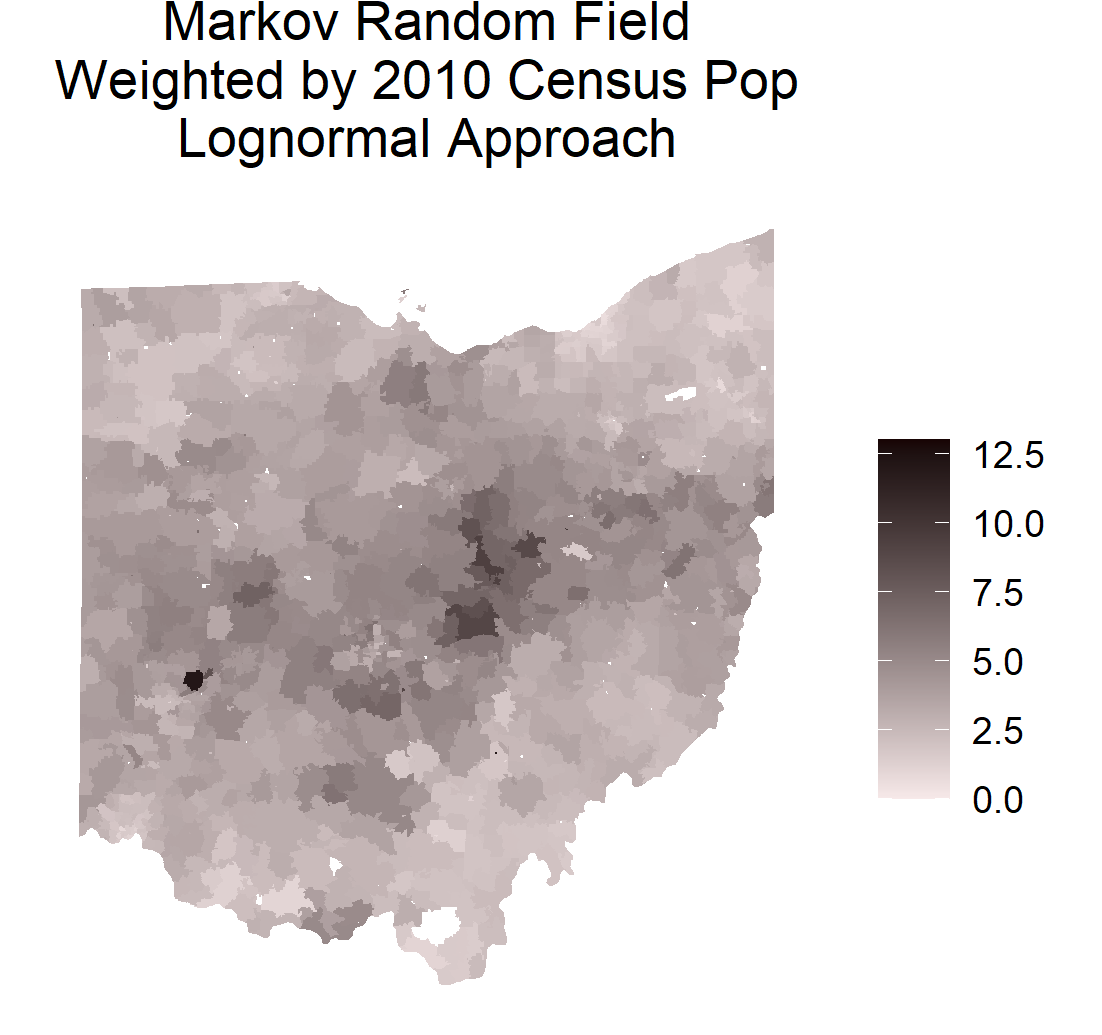
Figure 21



When we inspect the predictions for the regions we identified in Figure 10, the predictions look more realistic than any of the geostatistical methods. The small sample-size outlier at Benton Ridge has little influence over the predictions. Rockville and Norwalk are both still higher than average, but Norwalk increased vis-à-vis Rockville owing to its larger sample size. Dayton now has one of the highest predicted values (colored in red above the 13 pCi/L cutoff), due to the combination of it’s high GM and it’s very large sample size. Having made these observations, the Gaussian Markov random field approach seems to balance GM value and sample size better than any geostatistical method we tried, none of which were able to consider sample size as a direct input.

In Figures 5 and 6, we noted that there appeared to be some degree of correlation between the mean GM and the sample size, which could be explained by samples being taken more frequently in areas known to have higher levels of radon. The 2010 census population displayed no such discernible correlation with GM. Note that the correlation between 2010 census population and zip code sample size is 0.75. We try fitting the model again weighting based on population instead of weighting based on sample size, and plot the predicted values in Figure 22.

Figure 22



The predictions from the population-weighted model are generally similar to the predictions from the sample size-weighted model, though the population-weighted model display greater smoothing/shrinkage.

As a demonstration of the differences in smoothing between the models, we plot the predicted values passed through a binary indicator function that flags all values above 4 pCi/L, the EPA’s action threshold for indoor radon levels, as well as similar plots for both the raw data and for the lognormal kriging predictions.

Figure 23

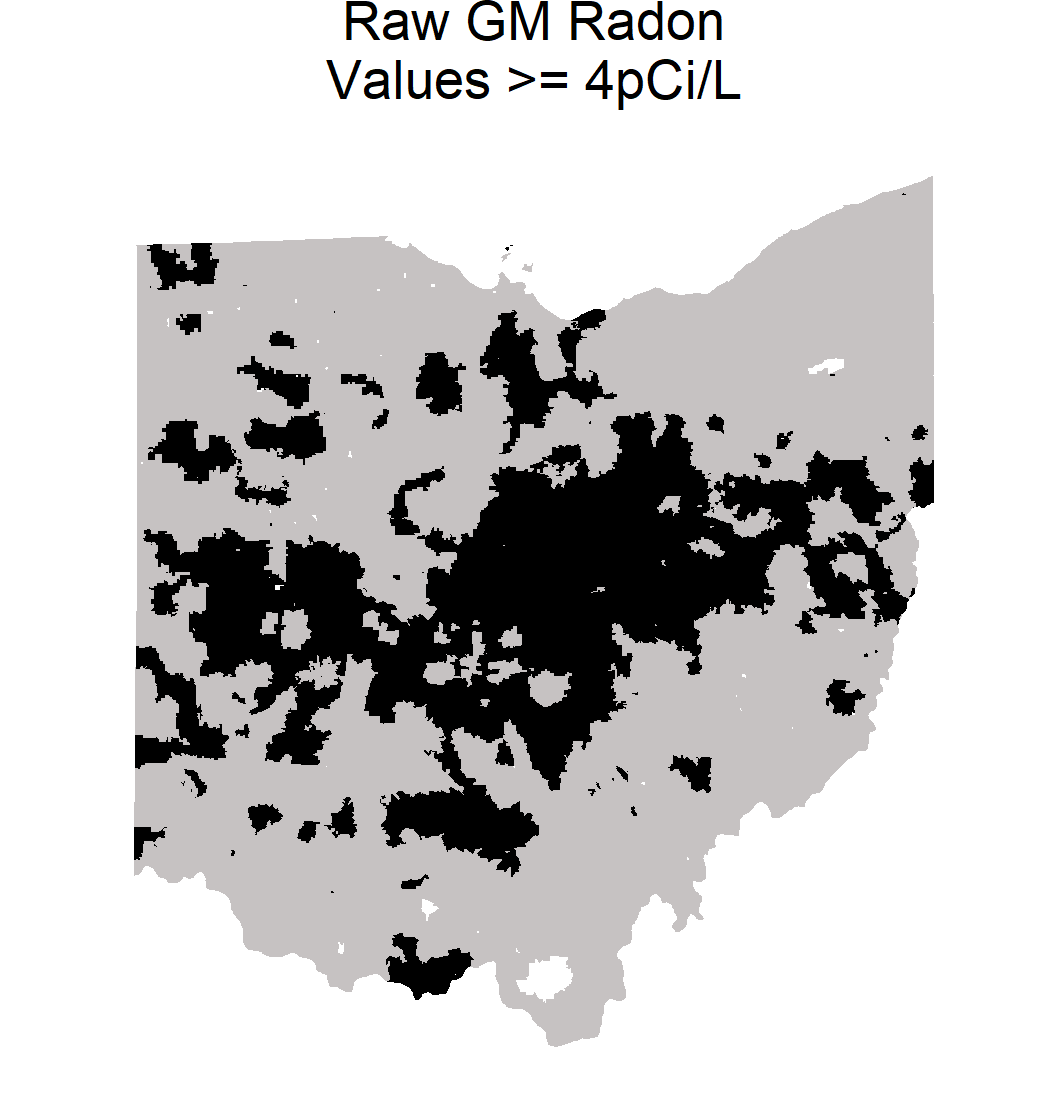


Figure 24

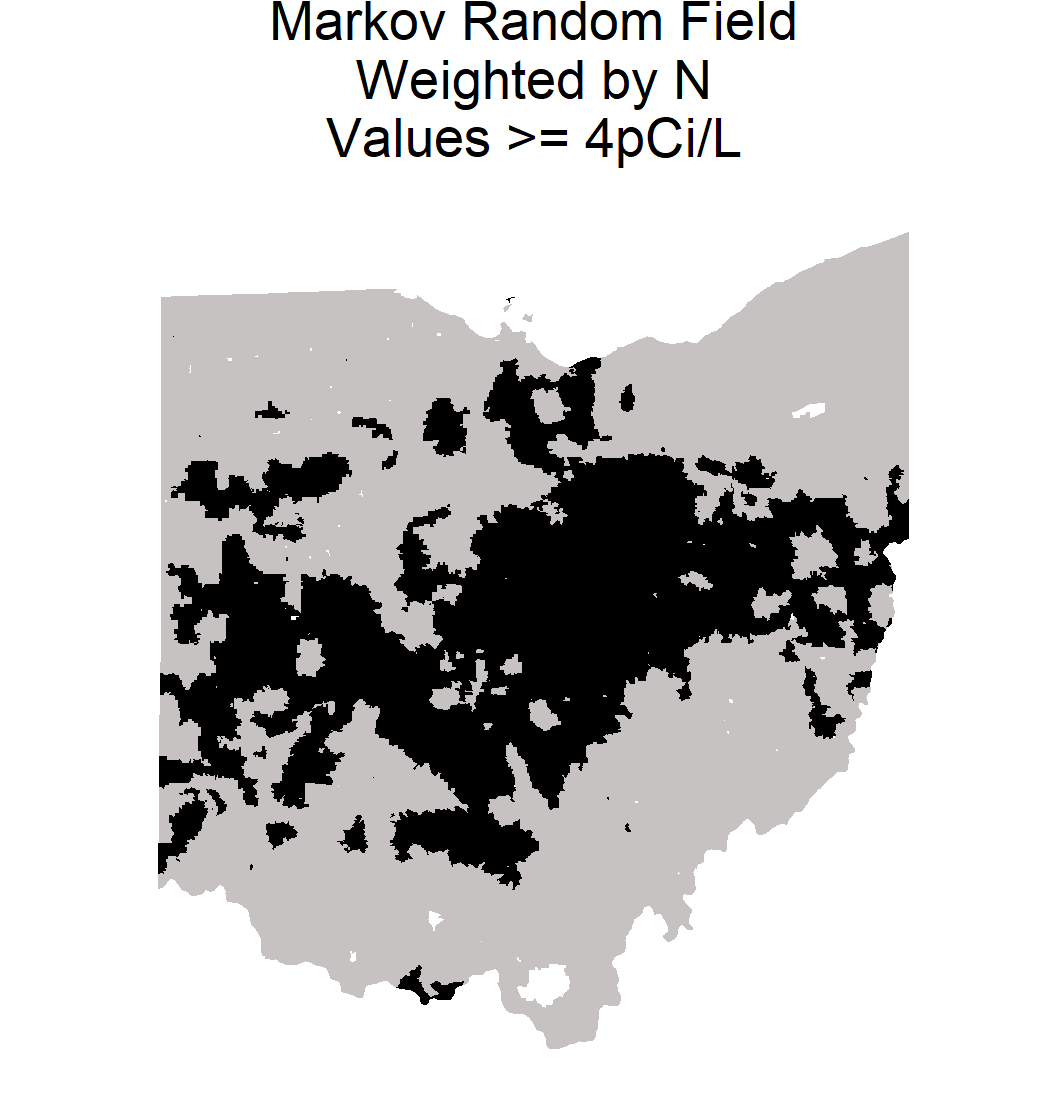


Figure 25

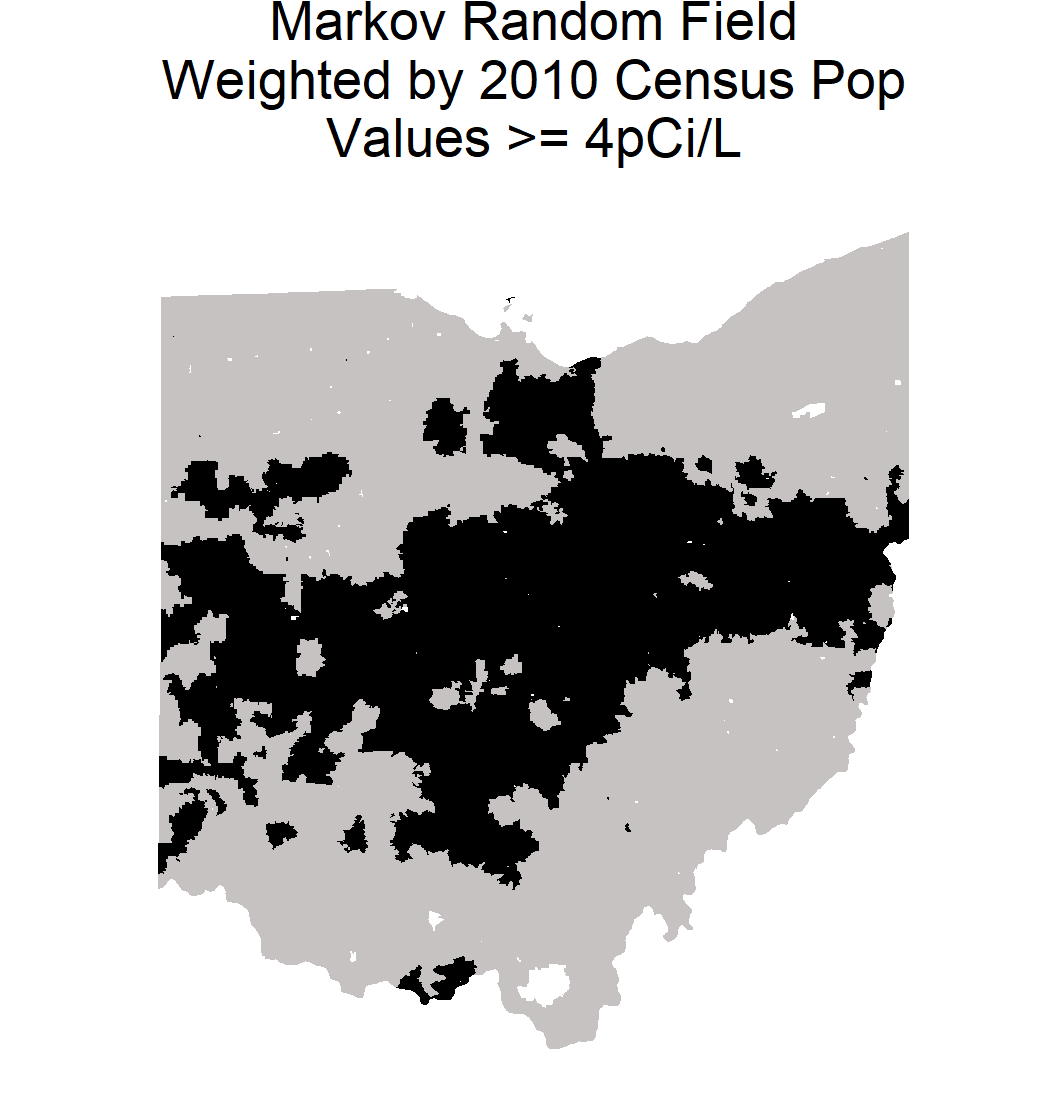
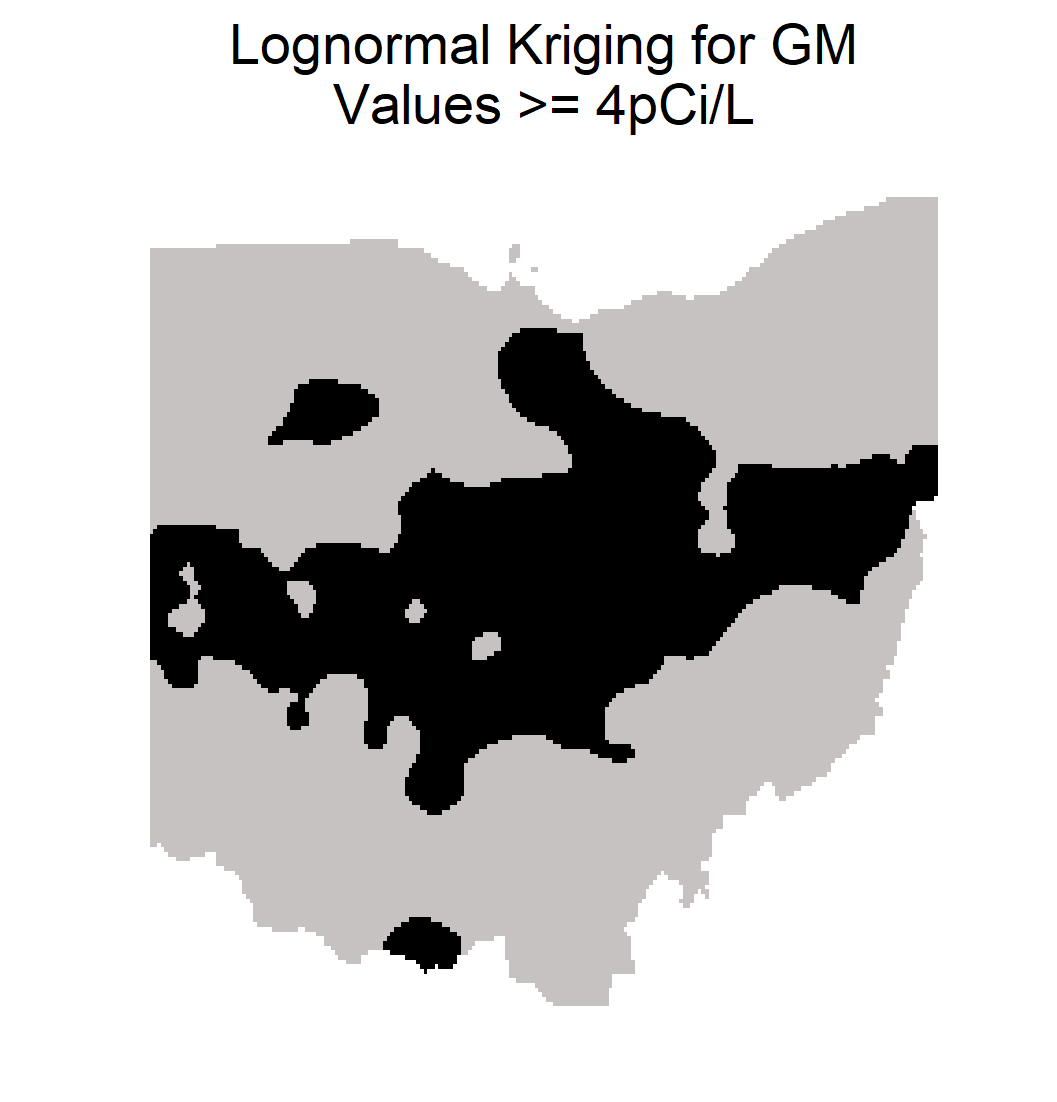


Figure 26



# Conclusion

## Modeling Results

After trying the geostatistical approaches of ordinary kriging, local Gaussian process approximation, and predictive process, and the lattice data approach of a Gaussian Markov random field, in order to generate predicted radon levels across Ohio, given the geometric means of 1155 zip codes, and visually inspecting the predicted values, the Gaussian Markov Random field was found to be the best model as it was able to provide a balance between the observed GM and the sample size while also providing smoothing based on the spatial correlation structure of the data. In fact, none of the geostatistical methods were able to incorporate the zip code samples size as model input, and thus were all vulnerable to the extreme outliers caused by high variability in the GMs at lower sample size zip codes. Out of the geostatistical methods, the lognormal Kriging provided the most realistic estimates, despite this major weakness.

Comparing our predictions with the radon maps and predictions in the literature such as those in Harrell 1993 and Maroju 2007, we see a similar pattern in the overall distribution of high radon values.

## Future Directions

Though it was apparent from inspection of the predicted values that the Gaussian Markov Random field was the preferable modeling approach, a more rigorous comparison could be done via train/test validation and model comparison using metrics such as MSPE on the test set to assess the quality of the predictions.

During the modeling process I searched for geostatistical methods that could incorporate the zip code sample size as modeling input to account for the high variability of GM in smaller sample size zip codes. After my class presentation, Dr. Kang suggested that this could be approached by performing a transformation on the GMs before kriging using a transformation that would shrink the observed GMs toward the center of the population, with more shrinkage being applied to the lower sample size zip codes.

The leading methods for zip code-level geostatistical analysis that came out of Dr. Kumar’s group are the neural network methods of Akkala and the Quantile Regression Forest method of Bandreddy. These two methods could be replicated and added to the analysis.

As far as my current understanding, I see several weaknesses in the analysis approach used by Xu, Sajja, and Kumar in the analysis of the relationship between hydraulic fracking and indoor radon concentration. I think that the multilevel model may not account for the spatial correlation of the radon measurements, due to the fact that the proximity to fracking sites is derived as a zip code-level variable rather than a home-level variable. I think that it could be worthwhile to replicate this work, while also testing out other modeling approaches that do consider the spatial correlation. Another point to consider is that like indoor radon, commercial fracking activity is highly correlated with local geological features, being a method of hydrocarbon extraction that is viable in select shale deposits. In any replication, it would be worth considering how to account for the correlation both variables have with geologic formations.

# References

Link to raw radon data: <https://odh.ohio.gov/wps/wcm/connect/gov/aab7c8e9-6835-49a1-856a-61f957849f70/Ohio+Zip+Codes+up+to+September+2020.pdf?MOD=AJPERES&CONVERT_TO=url&CACHEID=ROOTWORKSPACE.Z18_K9I401S01H7F40QBNJU3SO1F56-aab7c8e9-6835-49a1-856a-61f957849f70-nF3AkhK>

Link to ZCTA shapefiles: <https://www2.census.gov/geo/tiger/TIGER2020/ZCTA520/>

Link to ZCTA to place dataset (source of 2010 population by zip code): <https://www.census.gov/geographies/reference-files/2010/geo/relationship-files.html#par_textimage_674173622>

Akkala, “Development of Artificial Neural Networks Based Interpolation Techniques for the Modeling and Estimation of Radon Concentrations in Ohio”, The University of Toledo, 2010

Bastiani, Rigby, Stasinopoulos, Cysneiros, Uribe-Opazo, “Gaussian Markov random field spatial models in GAMLSS”, Journal of Applied Statistics, 2018

Bastiani, Stasinopoulos, Rigby, “The GMRF implementation in GAMLSS”, 2017

Dai, Neal, Diem, Deocampo, Stauber, Dignam, “Confluent impact of housing and geology on indoor radon concentrations in Atlanta, Georgia, United States”, 2019

EPA (United States Environmental Protection Agency), “A Citizen’s Guide to Radon - The Guide to Protecting Yourself and Your Family from Radon”, 2016

Grafton, “Local Radon Research: Meaningful and Cost Effective”, 1988

Gramacy, Apley , “Local Gaussian Process Approximation for Large Computer Experiments”, Journal of Computational and Graphical Statistics, 2015

Gummadi, “A Comparison of Various Interpolation Techniques for Modeling and Estimation of Radon Concentrations in Ohio”, The University of Toledo, 2013

Harrell, Belsito, Kumar, “Radon Hazards Associated with Outcrops of Ohio Shale in Ohio”, 1991

Heydinger, Kumar, Harrell, “Development of an Indoor Radon Information System”, 1991

Kumar, Kadiyala, Devabhaktuni, Akkala, Manthena “Examination of Ohio Indoor Radon Data”, 2010

Maroju, “Evaluation of Five GIS Based Interpolation Techniques for Estimating the Radon Concentration for Unmeasured Zip Codes in the State of Ohio”, The University of Toledo, 2007

Yamamoto, “On unbiased backtransform of lognormal kriging estimates”, Science of the Total Environment, 2007