POISSON REGRESSION FOR SMOOTH GEOGRAPHIC STRATIFICATION OF RISK

A PREPRINT

Delphine Boursicot
Département de sciences de la décision
HEC Montréal
Montréal, Québec
delphine.boursicot@hec.ca

Maxime Comeau
MRCC
Desjardins General Insurance Group
Lévis, Québec
maxime.comeau@dgag.ca

Bastien Ferland-Raymond
MRCC
Desjardins General Insurance Group
Lévis, Québec
bastien.ferland-raymond@dgag.ca

Philippe Gagnon
Department of Statistics
University of Oxford
Oxford, United Kingdom
philippe.gagnon@stats.ox.ac.uk

Charles Gauvin
MRCC
Desjardins General Insurance Group
Lévis, Québec
charles.gauvin@dgag.ca

Rachel Han
Department of Mathematics
University of British Columbia
Vancouver, British Columbia
hanrach@math.ubc.ca

Michael Lindstrom
Department of Mathematics
University of California, Los Angeles
Los Angeles, California
mikel@math.ucla.edu

Nassim Razaaly
DeFI Team (INRIA SIF, École Polytechnique)
Palaiseau, FRANCE
nassim.razaaly@inria.fr

Juliana Schulz
Département de sciences de la décision
HEC Montréal
Montréal, Québec
juliana.schulz@hec.ca

Junwei Shen
Department of Statistical and Actuarial Sciences
University of Western Ontario
London, Ontario
jshen255@uwo.ca

Tony Wong
Department of Mathematics
University of British Columbia
Vancouver, British Columbia
tonyw@math.ubc.ca

September 28, 2019

ABSTRACT

Segmentation of risk over spatial location is important in the insurance industry, particularly for home insurance, as each region has its own innate level of risk based on features of the location and its surroundings. It is also important that risk segmentation be spatially smoothed for business considerations. That is, models should not predict risk levels that vary rapidly in space to avoid unfair pricing differences for two clients in similar living conditions separated by a short distance. In this report, we outline the approaches we took to address this problem. Notably, we applied the methods of Geographically Weighted Regression, Poisson Kriging, and Fused Lasso, to insurance claim counts data from Desjardins. The models were applied to aggregated data on a postal code basis in order to predict spatial risk levels.

Keywords fused lasso, geographically weighted regression, kriging regression, spatial smoothing, spatial statistics, risk segmentation

1 Introduction

In the 2019 Industrial Problem Solving Workshop in Montréal, Québec, Canada, hosted by the Centre de recherches mathématiques (CRM) and the Institute for Data Valorization (IVADO), the Desjardins Groupe d'assurances générales presented a problem. The problem was to investigate and develop statistical methodologies capable of characterizing the geospatial elements of the risk of household theft, while balancing both prediction accuracy and ensuring slowly varying geographic estimators. At the time of the workshop, the company used a four step process: (i) a piecewise (over each region) generalized linear model (GLM), (ii) supplemented by gradient boosting with Xgboost, (iii) a subsequent smoothing via Markov Random Field, and (iv) finally making a prediction with the resulting smoothed GLM. This methodology was somewhat convoluted, involving many steps, and lacked robustness against small changes in the data. The challenge posed was to identify alternative methods to serve the problem specifications.

We investigated three alternatives including Geographically Weighted Regression (GWR) [5], the Kriging Method [2], and Fused Lasso [7], each modified suitably to model Poisson random counts. Geographically Weighted Regression is a technique that provides a local regression for a response variable at every point in space, with nearby observations having a greater influence in the regression, similar to a locally weighted scatterplot smoothing (LOWESS) approach. The Kriging method bears similarities to GWR but in making a prediction, empirical spatial autocorrelations in the response variable are taken into account to develop the weightings. With Fused Lasso, a regression is obtained for the data in conjunction with two penalty terms, the first being used for model selection (reducing parameters) and the second for enforcing spatial smoothness of estimators.

Having described the methods, the remainder of this report is organized as follows: in Sections 2 and 3, we introduce our data and notation conventions; in Section 4, we explain our techniques; our results are presented in Section 5 and we provide a conclusion and discuss future work in Section 6.

2 Data

The data provided were individual records of insured individuals and their resulting number of claims of household theft. Several hundred features were included in each record, including personal characteristics like age and geographic features such as local crime. For privacy, no names were included and all data were provided at the postal code level. Several record features were of a categorical nature, such as the type of roof. Close to 900,000 records were provided for analysis.

Due to the enormity of the dataset in both records and number of variables, we chose to employ various reductions and simplifications. These included:

- *One-hot encoding* [3]: transforming categorical data into a binary representation. This greatly increases the number of variables but allows for categorical data to be treated numerically. There is otherwise not a natural way to combine numerical values with categorical values.
- *Feature Selection*: 50 features were selected as being the most important and influential in the models and we worked with only these 50 features. This was done with Xgboost [4].
- *Subsampling*: we worked with a dataset of approximately 47,000 records rather than 900,000.
- *PCA*: principal component analysis [1] allows for high dimensional data to be represented, at least approximately, in a lower dimensional linear subspace. See Figure 1.

For simplicity in geographically weighted regression, we did not convert the latitude and longitude to distances and we treated each degree change in latitude as equal in length to a degree change in longitude. For making actual predictions, proper conversions would be in order.

As for the Kriging approach, the used the R package **geoRglm** to do the analysis, which did not allow incorporation of a weight variable for each location, which affects the level and the validity of the predictions. This would need to be addressed.

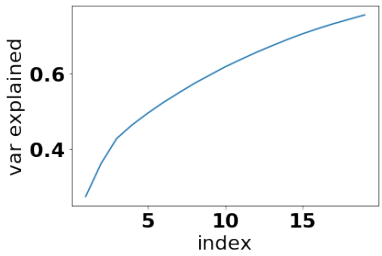

Figure 1: Plot of cumulative variance explained versus the number of principal components used after the subsampling of features was performed and one-hot encoding was applied. When PCA was used, we used the first 10 principal components.

3 Notation

Here we present the notation used in describing the models. In general, we assume a features matrix

$$\Theta \in \mathbb{R}^{N\times(1+d_c+d_s)}$$

where there are N records of clients, with d_c client-specific features (age, etc.) and d_s geography-specific features (local crime rate, etc.). The extra 1 is for the intercept. Thus, the first column of Θ is an all ones vector, columns 2 through $1+d_c$ describe the client data, and columns $2+d_c$ through $1+d_c+d_s$ describe the spatial data. We shall adopt the notation that a subscript of $i:$ is the ith row of a given matrix. We also denote

$$X \in \mathbb{R}^{N\times 2}$$

to be a matrix of N corresponding locations (centroids of postal codes in our data). We then denote a *column* vector

$$y \in \mathbb{R}^{N}$$

to be the realizations (claim counts).

Some models encoded spatial location as one of the d_s spatial variables. At times in this report, we will denote

$$T \subset \{1, 2, ...N\}$$

to be set of records used for training and

$$V \subset \{1, 2, .., N\}$$

to be a set of records used for validation. We often required that the centroids contained in the two sets are disjoint, i.e.,

$$(\cup_{i\in T} X_{i:}) \cap (\cup_{i\in V} X_{i:}) = \emptyset. \tag{1}$$

This is relevant because proper validation and parameter selection require a fair test of the models' predictive powers without biasing the models with samples of regions it needs to predict. However, for the Fused Lasso method, locations are categorical data and the disjointedness condition in (1) is removed.

As the objective of the models is to be predictive, hypothetically a new client (not part of the known records) could be assigned a record index $i \notin T \cup V$ and our models can make predictions as to their claim rate using additional data $\Theta_{i:}$ and a position $X_{i:}$.

We interpret Y_i, the number of insurance claims made by client i, as a Poisson random variable where $\Pr(Y_i = y_i)$ can be explicitly computed from a known Poisson rate λ_i :

$$\Pr(Y_i = y_i|\lambda_i) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!}. \tag{2}$$

We apply a GLM to describe the Poisson rate so that client i is modelled as having $Y_i \sim \text{Pois}(\lambda_i)$ where

$$\log \lambda_i = \Theta_{i:}\hat{\beta}_i + \epsilon_i. \tag{3}$$

In the GWR and Lasso methods, $\epsilon_i = 0$, but $\epsilon_i \neq 0$ for the Poisson Kriging method. The vector $\hat{\beta}_i$ is a column vector of estimators to be found at location $X_{i:}$.

The means by which the $\hat{\beta}_i$'s are found depend upon the method.

4 Techniques

4.1 Geographically Weighted Poisson Regression

4.1.1 Method

To make a prediction for a client index i (possibly not in T), with a known feature row-vector $\Theta_{i:} \in \mathbb{R}^{1+d_c+d_s}$ and location $X_{i:}$, we compute

$$\hat{\beta}_i = \text{argmin}_\beta \Big(-\sum_{j \in T} w(X_{i:}, X_{j:}) \log \Pr(Y_j = y_j | \lambda_j = \exp(\Theta_{j:}\beta)) \Big) \tag{4}$$

which is a weighted log-likelihood of observing all of the data. In general, a w function is chosen to be a decreasing function of the distance between its arguments. For our work, we chose the Gaussian

$$w(x, y) = \exp(-\frac{||x - y||^2}{2\alpha^2}), \tag{5}$$

where x and y are two positions. Other choices can be made. The value of α in (5) is a hyper-parameter, which we chose by cross-validation. For GWR, we used the top 10 principal components in developing the regressions.

4.1.2 Evaluation

To perform cross-validation and assess the model fits, we considered two objective functions $J(\alpha)$. If V denotes the set of points in the validation set, i.e., all clients at a point with a Poisson rate to be estimated, we used

$$J(\alpha) = L(\alpha) = -\frac{1}{|V|} \sum_{i \in V} \log \Pr(Y_i = y_i | \lambda_i), \tag{6}$$

the average negative log likelihood of observing the validation set (see (2)), and

$$J(\alpha) = D(\alpha) = \frac{1}{|V|} \sum_{i \in V} \Big(y_i \log(y_i/\lambda_i) - (y_i - \lambda_i) \Big), \tag{7}$$

the average deviance. It is understood that if $y_i = 0$, the logarithm term is not included as part of the summand in (7). The optimal α, α^*, is found from

$$\alpha^* = \text{argmin}_\alpha J(\alpha) \tag{8}$$

In Figure 2, the results of the objective functions for different values of α are plotted.

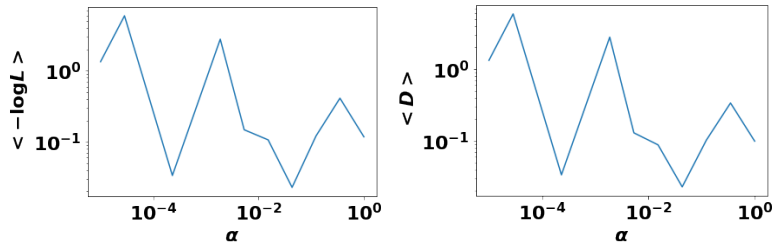

Figure 2: A parameter sweep over α for the objective functions (6) and (7) to estimate the best hyper-parameter α. The values plotted are the mean values over the validation set. The optimal value is $\alpha \approx 4.3 \times 10^{-2^\circ}$.

4.2 Poisson Kriging

To implement Poisson Kriging, we made the assumption that the log of the Poisson rate could be decoupled into the sum of an average rate that can be predicted from the features and a Gaussian random field. This resulted in

$$\log \lambda_i = f(X_{i:})\hat{\beta} + \varepsilon(X_{i:}) \equiv s_i. \tag{9}$$

where $f(X_{i:})$ is a row vector of geovariables for the ith location plus the intercept term of 1. Above, we have that $\hat{\beta}$ is an estimator to be found by maximizing the marginal likelihood and ε is a random field with mean $\mu = -c(0)/2$ and covariance function

$$C(x, y) = c(||x - y||) = \sigma^2 \exp(-||x - y||^2/\phi)$$

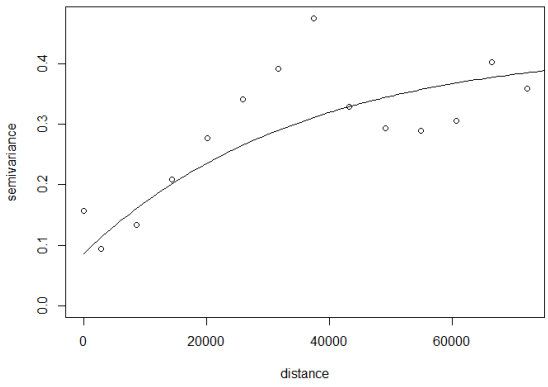

Figure 3: Empirical semivariogram plot based on all data available. The sill (σ^2), range (ϕ) and nugget (τ^2) parameters can be estimated approximately from this plot around 0.37, 38000 and 0.08 respectively.

with σ and ϕ as parameters. So when $|c(0)|$ is small, from Taylor expansion we have

$$\begin{aligned} E(Y_i) &= E[E(Y_i|\epsilon(X_{i:}))] \\ &= e^{\Theta_{i:}\hat{\beta}} E(e^{\epsilon(X_{i:})}) \\ &\approx e^{\Theta_{i:}\hat{\beta}}(e^{\mu} + \frac{1}{2}e^{\mu}\text{Var}(\epsilon_i)) \\ &= e^{\Theta_{i:}\hat{\beta}} e^{-c(0)/2}[1 + \frac{1}{2}c(0)] \\ &\approx e^{\Theta_{i:}\hat{\beta}} e^{-c(0)/2} e^{c(0)/2} = e^{\Theta_{i:}\hat{\beta}} \end{aligned}$$

To estimate σ^2 and ϕ, we estimated the spatial covariance of the response variables Y_i by empirically estimating the semivariogram function (the relationship between the semivariance of a response variable and the distance between points)

$$\gamma(r) = \frac{1}{2}\text{Var}(\{Y(u) - Y(v) \text{ s.t. } |u - v| = r\}). \tag{10}$$

Equation (10) describes the variance in the difference between the response variable at all points a distance r apart. In theory, it should vanish at $r = 0$ and increase as r increases, possibly saturating at some point. In practice, the continuous values of r are replaced by buckets of finite width. This curve is related to the covariance function by $c(r) + \gamma(r) = \gamma(\infty)$ so that having one function allows the other to be found as well. The fit we obtained for the semivariogram is illustrated in Figure 3. In practice, a nugget is present (nonzero y-intercept) and the semivariance is nonzero at zero distance. This is due to possible measurement errors in the process of data collection.

With the semivariogram computed, we were able to estimate σ^2 and ϕ in the covariance function. Suppose $Y = (y_1, y_2, \ldots, y_n)^T$ is the response for training data and $Y^* = (y_1^*, y_2^*, \ldots, y_t^*)^T$ is the response for the test data. Similarly, we defined S, S^*, ϵ, and ϵ^* as the vectors $(s_1, ..., s_n)^T$, $(s_1^*, ..., s_t^*)^T$, $(\epsilon_1, ..., \epsilon_n)^T$, and $(\epsilon_1^*, ..., \epsilon_t^*)^T$, respectively. Here we make the prediction of y^* through the prediction of S^* and ϵ^*. We can predict ϵ^*, the ϵ^*'s at positions $X_{1:}^*, ...X_{t:}^*$, based on information from ϵ. Then predictions can be made via MCMC simulation stated as below.

Step 1: Simulate ϵ from the distribution of ε conditioned on y

$$P(\varepsilon|Y) \propto P(Y|\varepsilon)P(\varepsilon)$$

where

$$P(Y|\varepsilon) = \prod_{k=1}^{n} \frac{\exp(s_k y_k)}{y_k!} e^{-\exp(s_k)},$$

$$s_k = \hat{\beta} f(X_{k:}) + \epsilon(X_{k:})$$

and ε comes from the Gaussian random field with estimated parameters $\hat{\sigma}^2 = \sigma^2$ and $\hat{\phi} = \phi$ from the semivariogram.

Step 2: Simulate ϵ^* from the distribution of ε^* conditioned on ε (since they come from the same Gaussian random field, the conditional distribution is also Gaussian with associated parameters $\hat{\sigma}^2$ and $\hat{\phi}$)

Step 3: $s_k^* = f(X_{k:}^*)\hat{\beta} + \epsilon^*(X_{k:}^*)$ and $y_k^* = e^{s_k^*}$

Step 4: Repeat steps 1-3 m times, where m is the number of simulations. Then the final predicted response for the t locations of interest is the average value of the m predictions.

4.3 Fused Poisson Lasso

With the Fused Poisson Lasso regression, spatial location (postal code) represents a categorical variable with d_ℓ values that was encoded as d_ℓ binary covariates. Denote by $\mathcal{L}$ the set of column indices containing these covariates. We seek a single vector of estimators $\hat{\beta}$ found from the minimization problem

$$\hat{\beta} = \text{argmin}_\beta \left\{ -\frac{1}{|T|} \sum_{i \in T} \log \Pr(Y_i = y_i | \lambda_i = \exp(\Theta_i \beta)) \right\} \quad \text{subject to} \quad \sum_j |\beta_j| \leq t_1 \text{ and } \sum_{j \in \mathcal{L}} |\beta_j - \beta_{j^*}| \leq t_2 \tag{11}$$

where $t_1, t_2 > 0$ are hyperparameters that require tuning and j^* is the index of the "nearest" (in a sense defined below) neighbor of the j^{th} location. Note that the penalization above represents a special case of a broader penalization framework, where for instance the differences between all coefficients of the neighbors of the j^{th} location would be explicitly penalized.

The ℓ^1-constraint with t_1 is a model selection mechanism, working to make the regression coefficients in $\hat{\beta}$ to be sparse. The ℓ^1-constraint with t_2 works to make the (categorical) coefficients describing similar locations identical (if it makes sense according to the optimisation problem): if a client moves from region 1 to region 2, where both regions have similar geographic features (median income, local crime, etc.) and their two successive houses are in similar condition, there should not be a significant change in their risk factor.

In practice, given a location indexed by j, we defined j^* to be the index of the "nearest" location in terms of distance with respect to the features. For lack of time, we used Euclidean distance in feature space.

Unlike GWR, this method explicitly assumes that regardless of spatial location (for all values of the categorical variable *location*), the other covariates all have the same effect upon the response variable. Regardless of whether this is true, it very directly satisfies the problem specifications where client-specific factors cannot have varying influence over space.

5 Results

Here we present the results of the different techniques.

5.1 Geographically Weighted Poisson Regression Results

In Figure 4, with the optimal α, we plot the prediction of the Poisson rate λ at a subset of locations. There is a degree of smoothness in the risk present.

5.2 Poisson Kriging Results

Using Poisson Kriging as described, we can generate a prediction for the frequency of claims being made over space as seen in Figure 5.

5.3 Fused Poisson Lasso Results

Using the Fused Lasso method, we depict in Figure 6 the spatial distribution of the risk with various values for t_1 and t_2.

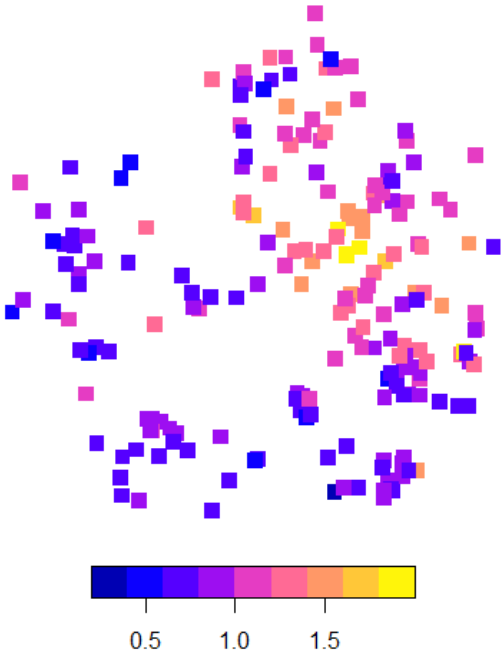

Figure 4: Over a small validation set, the average predicted λ for GWR was 0.00644 per year and the values ploted are relative to this average value.

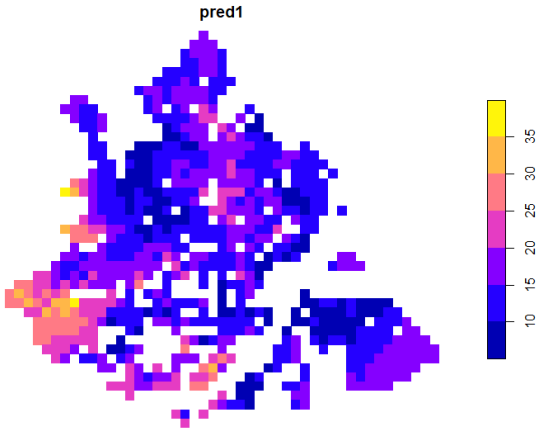

Figure 5: Plot of risk over space using Poisson Kriging. The response variable is smooth. It is important to note that the limitation in the R package **geoRglm** not able to manage weights in observations, actual frequencies are therefore not accurate.

6 Conclusions and Future Work

Through the workshop, we built upon and used three alternative techniques for the prediction of risk over space. All techniques were able to provide predictions of risk over space, and each method has its own pros and cons. We identify them below.

For Geographically Weighted Regression:

- One single model serves to describe the entire dataset and make predictions.
- The resulting spatial distribution of predicted risks appears to be smooth.

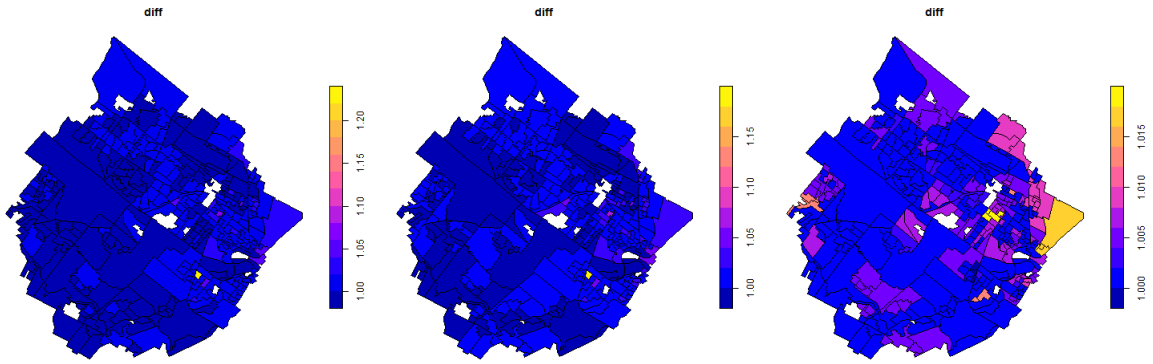

Figure 6: Plot of risk over space using Fused Lasso. The different plots show the results for different choices of t_1 and t_2 (the values increase as we move to the right).

- Predictions can be made at locations where there were no previous data.
- The algorithm is flexible in that different forms of weighting functions could be used, but this also adds a level of arbitrariness to the fitting.
- The algorithm is not fast to run. It took more than 2 hours to run a sweep of 15 $\alpha-$values in cross validation using only 1% of the unique postal codes in the smaller dataset of 47,000 records.

For Poisson Kriging:

- The basic model is very simple and easy to interpret.
- Smoothness emerges from using the model.
- Unfortunately, the model is difficult to apply for Poisson and other distributions (such as Gamma and Tweedie which are often used in insurance to model individual claim amounts and total losses, respectively.)
- The algorithm is very memory intensive and slow: it took more than 2 hours and 25G of ram to run on a set of 5000 records.

For Fused Lasso:

- The method is "global", fitting everything all at once, and the results are intuitive.
- It meets the problem specifications.
- It cannot directly make predictions at locations where no client data are available since that would introduce a new categorical value; however, for a new location, heuristics based in finding the most similar location from known observed points, similar to how GWR uses nearby locations to make predictions, could be used.
- It can be very memory intensive: to make a prediction, at least as many dummy variables as there are unique client locations are required. However, we can *fuse* areas together (the j-th location with its nearest neighbor, and then consider this as a new area with its own nearest neighbor) prior running the optimisation until the number of covariates is reasonable.

For future work, it would be worthwhile to:

- modify the objective function in the GWR to perform model selection (by adding an ℓ^1-penalty to the minimization problem (4));
- better manage observation weights in Poisson Kriging and increase the amount of data that can be analyse;
- use the Mahalanobis distance [6] to select nearest neighbours for Fused Lasso.

Author Contributions

All authors contributed to the scientific investigations, wrote code, and interpreted and discussed results. The report was written by ML with technical assistance from PG and JSh. All authors read and reviewed the final manuscript.

Acknowledgements

We would like to acknowledge the CRM and IVADO for hosting this event and generous compensation of travel expenses.

References

[1] H. Abdi and L. J. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[2] M. Ali, P. Goovaerts, N. Nazia, M. Z. Haq, M. Yunus, and M. Emch. Application of poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of bangladesh. *International journal of health geographics*, 5(1):45, 2006.

[3] J. E. Beck and B. P. Woolf. High-level student modeling with machine learning. In *International Conference on Intelligent Tutoring Systems*, pages 584–593. Springer, 2000.

[4] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[5] A. S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

[6] P. C. Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[7] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.