Large-Scale Spatial Prediction by Scalable Geographically Weighted Regression: Comparative Study

Daisuke Murakami[1]
Institute of Statistical Mathematics, Tokyo, Japan

Narumasa Tsutsumida
Saitama University, Japan

Takahiro Yoshida
The University of Tokyo, Japan

Tomoki Nakaya
Tohoku University, Seindai, Japan

Abstract

Although the scalable geographically weighted regression (GWR) has been developed as a fast regression approach modeling non-stationarity, its potential on spatial prediction is largely unexplored. Given that, this study applies the scalable GWR technique for large-scale spatial prediction, and compares its prediction accuracy with modern geostatistical methods including the nearest-neighbor Gaussian process, and machine learning algorithms including light gradient boosting machine. The result suggests accuracy of our scalable GWR-based prediction.

2012 ACM Subject Classification Computing methodologies → Model development and analysis

Keywords and phrases Spatial prediction, Scalable geographically weighted regression, Large data, Housing price

Digital Object Identifier 10.4230/LIPIcs.COSIT.2022.12

Category Short Paper

Funding This research was funded by the Joint Support Center for Data Science Research at Research Organization of Information and Systems (ROIS-DS-JOINT) under Grant 006RP2018, 004RP2019, 003RP2020, and 005RP2021.

1 Introduction

Geostatistical Gaussian process (GP) models have been used for spatial prediction in geology, environmental science, and other fields (see [2]). Although GP-based spatial prediction is known to be accurate as demonstrated in [9], the computational complexity inflates in an order of N^3 where N is the sample size due to a matrix inversion. The classical GP model is unavailable for large samples (e.g., $N > 10,000$). To address the drawback, fast GP approximations have been developed in geostatistics (see [7]) and machine learning areas (see [11]). For example, nearest-neighbor Gaussian process (NNGP [3]) is widely accepted as a fast approximate GP in geostatistics. NNGP and other scalable GPs achieve linear-time computational complexity (i.e., the computational complexity increases in an order of N). They are available for very large samples.

[1] corresponding author

Scalable GPs usually model stationary spatial process assuming model parameters including regression coefficients as constant over space. However, [4, 13] among others have demonstrated that regression coefficients can vary over geographical space. Nevertheless, fast prediction technique considering such spatially varying coefficients (SVCs) is quite limited.

Geographically weighted regression (GWR [1]) is a popular SVC modeling technique that has been used for spatial prediction (e.g., [6, 5]). It is hard to apply the classical GWR for very large data in terms of the computational complexity and memory usage. To overcome the limitation, [10] and [12] developed algorithms for estimating the GWR model computationally efficiently. In particular, the latter developed the scalable GWR technique achieving a quasi-linear computational complexity with very small approximation error. The scalable GWR is potentially useful for spatial prediction. However, it has never been used for spatial prediction.

The objective of this study is to examine the usefulness of the scalable GWR in terms of spatial prediction for large samples through a comparison with modern prediction methods in geostatistics and machine learning areas.

2 GWR model

2.1 Basic GWR model

GWR describes the explained variable y_i at i-th sample site on a two-dimensional space using the following model:

$$y_i = \sum_{k=1}^{K} x_{i,k}\beta_{i,k} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

where $x_{i,k}$ is the k-th explanatory variable and σ^2 is the variance parameter. $\beta_{i,k}$ is the k-th regression coefficient at the i-th location. The model is estimated by a weighted least squares (WLS) method assuming greater weights for nearby samples using a distance-decaying kernel whose decay-speed is dependent on a bandwidth parameter w. Later, we will use an exponential kernel $k(d_{i,j}; w) = \exp(d_{i,j}/w)$, where $d_{i,j}$ is the Euclidean distance between locations i and j. The bandwidth parameter w is typically estimated by a leave-one-out cross-validation (LOOCV). In each iteration of the LOOCV, regression coefficients must be estimated for all the N sample sites. This property makes GWR slow for very large samples.

2.2 Scalable GWR model

Scalable GWR estimates the same local model (Eq. 1). To lighten the computational cost, the kernel function $k(d_{i,j}; w)$ with unknown b is replaced with a linear combination of kernel functions with known b values.

$$k^*(d_{i,j}; a, b) = a + \sum_{l=1}^{L} b^l k(d_{i,j}; \tilde{w})^{4/2^l}, \tag{2}$$

where $\tilde{w}$ is a known bandwidth, which is specified based on the median of the 100-nearest neighbor distance. a and b are parameters being estimated through the LOOCV. The first term represents a global weight assigning a constant weight across samples while the second term represents a local weight assigning greater weight on nearby samples. The l-th kernel $k(d_{i,j}; \tilde{w})^{4/2^l}$ has a faster-decay for small l while slower-decay for large l. If $b > 1$, the weight b^l for faster-decay kernels are larger than those for slower-decay kernel, while the opposite is true $b < 1$. Thus, Eq. 2 estimates the decay speed of the kernel by estimating the b parameter.

Unlike $k(d_{i,j};w)$, which is used in the ordinary GWR, $k^*(d_{i,j};a,b)$ is just a linear function with respect to the parameters a and b^l. Thus, a quasi-linear time algorithm, which is explained in [12] is available for the model estimation.

2.3 Spatial prediction using the scalable GWR model

The basic GWR is readily applicable for spatial prediction by assuming a spatial kernel centered on the prediction site. Similarly, once the a and b parameters are estimated by the LOOCV, the regression coefficient at the prediction site s_0 is estimated by a WLS in which the samples are weighted by using the following kernel function:

$$k^*(d_{0,j};a,b) = a + \sum_{l=1}^{L} b^l k(d_{0,j};\tilde{w})^{4/2^l}, \tag{3}$$

where $d_{0,j}$ is the distance between the prediction site and j-th sample site. Eq. 3 assigns large weights on observations nearby the site s_0. Thus, the regression coefficients are estimated to reflect the local property nearby the prediction site. Spatial prediction at site s_0 is performed by substituting the estimated local coefficient $\hat{\beta}_{i,k}$ into the following model:

$$\hat{y}_0 = \sum_{k=1}^{K} x_{0,k}\hat{\beta}_{0,k} \tag{4}$$

Thus, the scalale GWR is easily employed for spatial prediction. Importantly, the spatial interpolation achieves a (quasi-)linear computation cost that is considered as desirable for large-scale spatial predictions in geostatistics. Nevertheless, the scalable GWR has never been applied fro spatial prediction.

3 Application

This section examines the performance of our proposed prediction method through a comparison of prediction accuracy with modern prediction techniques including (conjugate) NNGP and the light gradient boosting machine (LightGBM), which is known as an accurate and computationally efficient gradient boosting algorithm [8]. We also consider the linear regression (LM) for reference.

The Lucas housing data, which consists of the data of 25,357 single family houses sold in Lucas, Ohio in 1993-1998 ($N = 25,357$), is used for the comparison. The conventional GWR is hard to apply because of the computational burden. The explained variable is logged housing price (see Figure 1) and the explanatory variables are total living area in square feet (TLA), garage area in square feet (garagesqft), and building age (AGE).

Prediction accuracy is compared though a 5-fold cross-validation (CV). The root mean squared errors (RMSEs) are as follows: 0.462 (LM); 0.360 (LightGBM); 0.362 (NNGP); 0.309 (Scalable GWR). Surprisingly, scalable GWR has outperforms NNGP and LightGBM, which are modern geostatistical method and machine learning method respectively. Figure 2 plots the observed price in the x-axis and the price predicted during the CV in the y-axis. Based on this figure, the scalable GWR tends to have smaller prediction error relative to alternatives. The result suggests the potential of the scalable GWR as a spatial predictor. From Figure 2, it is also observed that several predicted values of the scalable GWR have large error. Further stabilization might be required to improve the prediction accuracy.

Finally, Figure 3 compares interpolated housing price map. Here, logged price in each 250 m grid covering the study area is predicted. Because explanatory variables are unavailable on the grids, we compare NNGP with constant mean and the scalable GWR with spatially varying intercept. Based on the result, the scalable GWR tends to have smoother prediction result relative to NNGP.

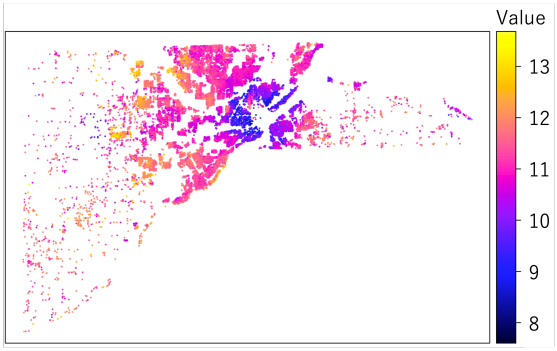

Figure 1 Logged housing price in Lucas county, Ohio.

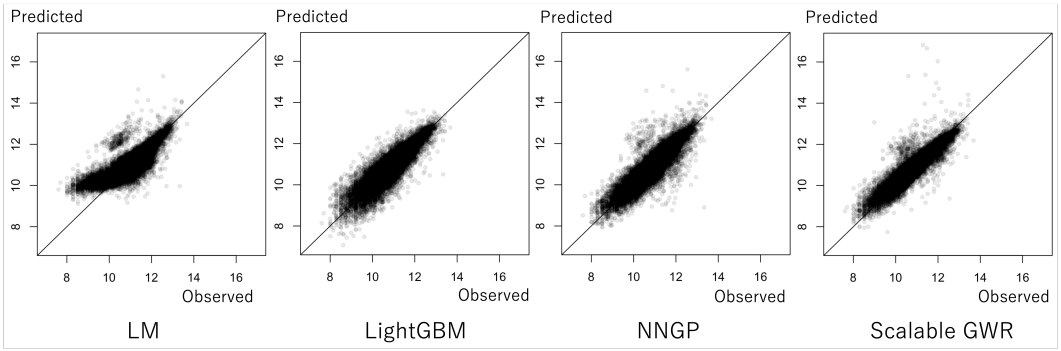

Figure 2 Comparison of observed (x-axis) and predicted housing prices (log-scale; y-axis).

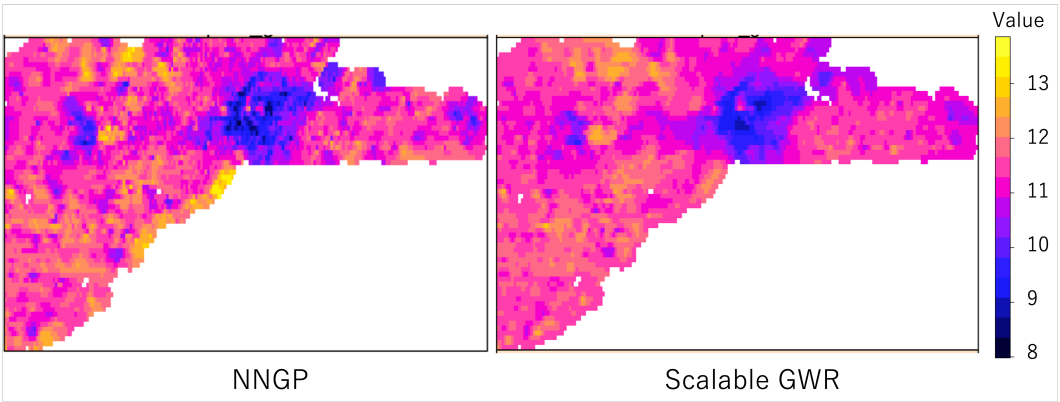

Figure 3 Comparison of observed and predicted housing prices (log-scale). Intercept-only model is used for comparison.

References

1 Chris Brunsdon, A Stewart Fotheringham, and Martin Charlton. Some notes on parametric significance tests for geographically weighted regression. *Journal of regional science*, 39(3):497–524, 1999.

2 Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

3 Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

4 A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

5 Melanie S Hammer, Aaron van Donkelaar, Chi Li, Alexei Lyapustin, Andrew M Sayer, N Christina Hsu, Robert C Levy, Michael J Garay, Olga V Kalashnikova, Ralph A Kahn, et al. Global estimates and long-term trends of fine particulate matter concentrations (1998–2018). *Environmental Science & Technology*, 54(13):7879–7890, 2020.

6 Paul Harris, AS Fotheringham, R Crespo, and Martin Charlton. The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. *Mathematical Geosciences*, 42(6):657–680, 2010.

7 Matthew J Heaton, Abhirup Datta, Andrew Finley, Reinhard Furrer, Rajarshi Guhaniyogi, Florian Gerber, Robert B Gramacy, Dorit Hammerling, Matthias Katzfuss, Finn Lindgren, et al. Methods for analyzing large spatial data: A review and comparison. *arXiv preprint*, 22, 2017. `arXiv:1710.05013`.

8 Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

9 Jin Li and Andrew D Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, 6(3-4):228–241, 2011.

10 Ziqi Li, A Stewart Fotheringham, Wenwen Li, and Taylor Oshan. Fast geographically weighted regression (fastgwr): a scalable algorithm to investigate spatial process heterogeneity in millions of observations. *International Journal of Geographical Information Science*, 33(1):155–175, 2019.

11 Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423, 2020.

12 Daisuke Murakami, Narumasa Tsutsumida, Takahiro Yoshida, Tomoki Nakaya, and Binbin Lu. Scalable gwr: A linear-time algorithm for large-scale geographically weighted regression with polynomial kernels. *Annals of the American Association of Geographers*, 111(2):459–480, 2020.

13 Tomoki Nakaya, Alexander S Fotheringham, Chris Brunsdon, and Martin Charlton. Geographically weighted poisson regression for disease association mapping. *Statistics in medicine*, 24(17):2695–2717, 2005.