Benchmarking Regression Models Under Spatial Heterogeneity

Nina Wiedemann[1]
Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Henry Martin
Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

René Westerholt
Department of Spatial Planning, TU Dortmund University, Germany

Abstract

Machine learning methods have recently found much application on spatial data, for example in weather forecasting, traffic prediction, and soil analysis. At the same time, methods from spatial statistics were developed over the past decades to explicitly account for spatial structuring in analytical and inference tasks. In the light of this duality of having both types of methods available, we explore the following question: Under what circumstances are local, spatially-explicit models preferable over machine learning models that do not incorporate spatial structure explicitly in their specification? Local models are typically used to capture spatial non-stationarity. Thus, we study the effect of strength and type of spatial heterogeneity, which may originate from non-stationarity of a process itself or from heterogeneous noise, on the performance of different linear and non-linear, local and global machine learning and regression models. The results suggest that it is necessary to assess the performance of linear local models on an independent hold-out dataset, since models may overfit under certain conditions. We further show that local models are advantageous in settings with small sample size and high degrees of spatial heterogeneity. Our findings allow deriving model selection criteria, which are validated in benchmarking experiments on five well-known spatial datasets.

2012 ACM Subject Classification Computing methodologies → Concurrent algorithms

Keywords and phrases spatial machine learning, spatial non-stationarity, Geographically Weighted Regression, local models, geostatistics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.11

Supplementary Material *Software (Source Code)*: `https://github.com/mie-lab/spatial_rf_python`; archived at `swh:1:dir:f7fc9b237a4e80f882b1785718ab884c2770ac66`

Acknowledgements We would like to thank Martin Raubal for the fruitful discussions about the project.

1 Introduction

The success of machine learning and artificial intelligence in recent years has sparked considerable interest in respective methods also in GIScience, and has led to a general proliferation of spatial data science [24]. While spatial statistics used to carefully address the special nature of spatial data, spatial data science often involves the direct application of (global) machine learning models to spatial data without explicitly modeling spatial properties. Nevertheless, these models oftentimes provide successful inferences on test data. Yet, spatial data may be subject to complex confounders including spatial heterogeneity, which is the focus of this paper. Currently, there is no comprehensive review available

[1] Corresponding author

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 11; pp. 11:1–11:15
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

that would show when global non-linear machine learning models can or should be used for interpolation or prediction tasks on spatially heterogeneous data without producing misleading or wrong results. While there exist analyses on the effectiveness of methods to deal with spatial autocorrelation [2], no such benchmarking has been done regarding spatial heterogeneity, either originating from non-stationarity of the actual process or from spatially heterogeneous exogenous noise. In this work, we benchmark the performance of global (machine learning) and local spatial regression models for the prediction of unseen test data that is subject to various kinds of heterogeneity. We simulate spatial heterogeneity with synthetic data in order to derive recommendations about the suitability of model types for specific heterogeneity-related scenarios. We finally validate our model selection criteria through experiments on several real-world datasets. The following two sub-sections briefly outline the state of the art as well as our contribution in more detail, before we present the experiments and our results.

1.1 Related work

Statistical learning methods have been adapted to geospatial data since a long time. A major step towards accounting for spatial heterogeneity has been the proposal of local models, such as Geographically Weighted Regression (GWR) [3, 9]. Next to variants of GWR [17], the idea also inspired adaptations of machine learning models, with spatial versions of Random Forests (RFs) [11, 28] or even Geographically Weighted Artificial Neural Networks [13, 7]. The proposed modifications of machine learning models such as Random Forests include 1) providing spatial coordinates as input [18], 2) deriving spatial features such as the distance from points of interest from the coordinates [14] in order to improve spatial generalization [5], 3) including the observations at nearby samples as covariates [28], and 4) fitting RFs on local subsets of data [11]. While these approaches have been shown advantageous in some situations, a recent study Zhou et al. [31] compared GWR with geographical RFs on health data and actually found that GWR provided better predictions than the more complex RF models, though the generalizability of the results is limited due to the very specific application context.

A common limitation of existing approaches is that the developed methods are usually evaluated on a single or few real dataset(s). The results may therefore be subject to unknown data properties. Synthetic data, in contrast, allows to benchmark methods in a controlled setting. While this solution is implemented, for example, by Beale et al. [2] and Santibanez et al. [26] for the purpose of assessing the effect of varying degrees of spatial autocorrelation, there is a lack of benchmarking with simulated spatial heterogeneity. Fotheringham et al. [10] and Hagenauer et al. [13] validate their methods on synthetic data that were designed to be non-stationary in space, and Finley et al. [8] compare GWR and SVC on non-stationary synthetic data, but they do not systematically vary the non-stationarity. The latter is our point of departure for the following sections.

1.2 Contribution

We evaluate the ability of different models to deal with varying degrees of spatial heterogeneity. Inspired by the work conducted by Comber et al. [4] presenting a route map when to use GWR and two of its variants, we derive model selection criteria from our results on synthetic data. We extend previous findings in three ways: first, in addition to GWR and other linear methods, we also consider Random Forests as non-linear models and compare their performance on non-linear tasks; second, we consider *predictive* performance instead of analysis in order to account for overfitting behavior; third, and most importantly, we provide a detailed analysis of model adequacy with respect to spatial non-stationarity and signal-to-noise ratio. To achieve this, we propose a synthetic data-generating process that allows to systematically vary the degree of spatial heterogeneity due to 1) the non-stationarity of the process, and 2) noise. We utilize this framework to compare seven models that are selected to reflect standard approaches that were, to varying degrees, developed to deal with spatial data. By analyzing the model performances in this controlled synthetic setting, we derive recommendations what model is appropriate dependent on the sample density, the spatial heterogeneity and the problem complexity. We validate our model selection criteria by benchmarking the models also on five real geospatial datasets.

2 Methods

We simulate a spatial regression problem with synthetically generated data that are subject to spatial heterogeneity. Spatial heterogeneity in our analysis stems from two effects; on the one hand, the dependence of the dependent variable[2] Y on the independent variables X may be non-stationary, i.e., the same input may lead to different outputs in different spatial regions. In previous work [10, 13], this was modeled by varying the coefficients β dependent on the coordinates (u, v); for example, Fotheringham et al. [10] set $\beta_1 = 1 + \frac{(u+v)}{12}$ and Hagenauer et al. [13] add coefficients with oscillating spatial distribution based on trigonometric functions. On the other hand, spatial heterogeneity may be caused by differences in the variance of the errors (and thus by noise). To understand the effect of the signal-to-noise ratio in spatial data subject to spatial heterogeneity of both types, we propose to vary the noise and the level of non-stationarity over space and to compare models on both linear and non-linear problems on test data.

2.1 Data-generating processes (DGPs)

One of our investigated DGPs represents a linear relationship of Y on k independent variables $x_j (j \in [1..k])$. It is given as

$$y_i = \sum_{j}^{k} \beta_j(u_i, v_i) \cdot x_{ij} + \epsilon(u_i, v_i) \ , \tag{1}$$

where x_{ij} is the j-th feature of the i-th sample, (u_i, v_i) are the coordinates of the i-th sample, and $\beta_j(u_i, v_i)$ is the location-dependent coefficient. $\epsilon(u_i, v_i)$ is the noise that may also be heterogeneous across space. The definition of β and ϵ will be given in detail in Section 2.1.1 and Section 2.1.2 respectively.

We also implement a non-linear DGP in order to analyze the model performances under the regime of a more complex phenomenon. The function is constructed such that there are interactions between variables and non-linear effects of single variables, and the terms are weighted with the non-stationary coefficients β:

$$\begin{aligned}\tilde{y}_i = \beta_1(u_i, v_i) \cdot x_{i1}^2 \cdot \sin(x_{i2}) + \beta_2(u_i, v_i) \cdot \sin(x_{i2}) \cdot x_{i4} \\ +\beta_3(u_i, v_i) \cdot x_{i5} \cdot \log(x_{i3}^2) + \beta_4(u_i, v_i) \cdot x_{i4}^2 \cdot \cos(x_{i2}) \\ +\beta_5(u_i, v_i) \cdot x_{i1}^2 \cdot x_{i4} \cdot x_{i5} + \epsilon(u_i, v_i).\end{aligned} \tag{2}$$

[2] Throughout this paper, we use capital characters for vectors and matrices and non-capitalized characters for referring to scalar terms.

In both scenarios, we construct n samples with pairs of geographic coordinates (u_i, v_i) and k attribute values x_{ij}. The coordinates are drawn from a uniform distribution $\mathcal{U}(-1, 1)$. In contrast to related work, we did not use coordinates on a regular grid in order to better mimic a realistic situation with irregular local clustering and dispersion patterns of observation sites. The independent variables X are assumed to be subject to spatial autocorrelation since we aim to simulate realistic spatial data. This is modeled by left-multiplying a vector of uniform random data X' by the so-called spatial autoregressive (SAR) generating operator[3] [16], that is, as $X = (I - \rho W)^{-1} X'$, where W is the weight matrix, computed as the inverse distances of the 20 nearest neighbors. After observing that the average spatial autocorrelation, measured using Moran's I, is around 0.3 in the considered real datasets, we calibrate the autoregressive parameter ρ such that the resulting values yield Moran's I values of around 0.3 accordingly ($\rho = 0.75$).

2.1.1 Non-stationary coefficients β

In contrast to previous work assuming a complete variation of the coefficients [10, 13], we argue that with many types of real-world processes, it would be more reasonable for the coefficients to vary around a constant value c_j. To simulate this, we frame spatial non-stationarity as an additive factor to the underlying coefficient c_j, and quantify its strength with a factor λ. The coefficients used are thus composed of the constant coefficient c_j and the spatial variation $\hat{\beta}_j(u_i, v_i)$:

$$\beta_j(u_i, v_i) = c_j + \lambda \cdot \hat{\beta}_j(u_i, v_i).$$

The spatial variation $\hat{\beta}$, in turn, is modeled based on trigonometric functions and thus in a similar fashion as presented in [10, 13]:

$$\hat{\beta}_j(u_i, v_i) = \sin(u_i \cdot 2\pi + j) + \cos(v_i \cdot 2\pi + j).$$

Since the coordinates are drawn from $\mathcal{U}(-1, 1)$, this definition of $\hat{\beta}$ leads to two cycles of the sine and cosine functions in x and y direction. Furthermore, the spatial variation is shifted by j for the j-th coefficient to ensure that the spatial heterogeneities attached to the coefficients are not all the same. The final coefficients $\beta_j(u_i, v_i)$ with weak ($\lambda = 0.2$) and strong ($\lambda = 0.5$) non-stationarity are shown in Figure 1.

2.1.2 Spatial heterogeneity of the errors ϵ

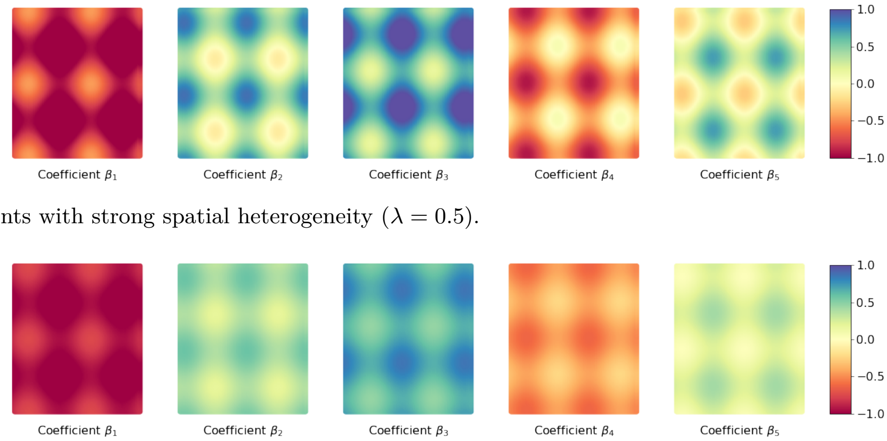
Not only β-coefficients but also the error terms can vary across space. A heterogeneous spatial distribution of the noise ϵ increases the difficulty of distinguishing signal from noise. The spatial distribution may thereby either be similar to one of the coefficients (i.e., also trigonometric) or different. Let σ be the average noise strength similar to the non-stationarity effect size λ as defined in Section 2.1.1. Using this, we consider three scenarios for varying the error terms:

$$\epsilon \sim \mathcal{N}(0, \sigma), \tag{3a}$$

$$\epsilon(u_i, v_i) \sim \mathcal{N}(0, \hat{\sigma}(u_i, v_i)) \quad \text{with} \quad \hat{\sigma}(u_i, v_i) = \sigma \cdot (\sin(u_i \cdot 2\pi) + \cos(v_i \cdot 2\pi) + 1), \tag{3b}$$

$$\epsilon(u_i, v_i) \sim \mathcal{N}(0, \hat{\sigma}(u_i, v_i)) \quad \text{with} \quad \hat{\sigma}(u_i, v_i) = \sigma \cdot (0.5 \cdot (u_i + v_i) + 1). \tag{3c}$$

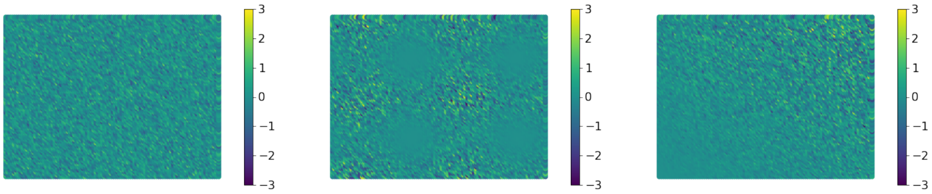
[3] See also `https://r-spatial.github.io/spatialreg/reference/invIrM.html`



(a) Coefficients with strong spatial heterogeneity ($\lambda = 0.5$).

(b) Coefficients with weak spatial heterogeneity ($\lambda = 0.2$).

Figure 1 Spatial non-stationarity is simulated as a trigonometric spatial variation of the coefficients β. The factor λ determines the overall strength of the non-stationarity.



(a) Uniformly distributed noise.

(b) Heterogeneous (trigonometric).

(c) Heterogeneous (linear).

Figure 2 Varying the spatial distribution of the variance of the errors ϵ. We simulate three scenarios: uniformly distributed noise ϵ, one that follows a similar distribution as the non-stationary process (i.e., trigonometric), and one that follows a different distribution (linear).

Equation 3a refers to a scenario with uniformly distributed noise. This scenario does not incorporate spatially varying errors. Equation 3b describes error terms that are heterogeneous in the sense that their variance oscillates trigonometrically around σ, depending on their spatial locations. The last scenario presented in equation 3c is also spatially varying but based on a diagonal linear trend over the map. Respective noise maps created under the scenarios outlined are illustrated in Figure 2.

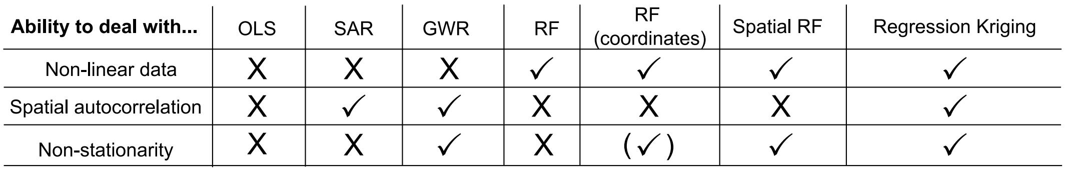
2.2 Regression models

We consider linear and non-linear, global and local models suitable for regression tasks. Figure 3 provides an overview of their properties. In the following, let $X \in \mathbb{R}^{n \times m}$ denote the m-dimensional feature matrix of n samples, and let $Y \in \mathbb{R}^n$ be the dependent variable that is to be predicted from X.

2.2.1 Ordinary Least Squares and a global spatial model

We employ two linear global types of regression models. One of these is the Ordinary Least Squares (OLS) model, which assumes a linear dependency of Y on X. It is given as

$$Y = X\beta + \epsilon,$$



Ability to deal with...	OLS	SAR	GWR	RF	RF (coordinates)	Spatial RF	Regression Kriging
Non-linear data	X	X	X	✓	✓	✓	✓
Spatial autocorrelation	X	✓	✓	X	X	X	✓
Non-stationarity	X	X	✓	X	(✓)	✓	✓

Figure 3 Overview of the compared models' abilities to handle non-linearity, their consideration of spatial autocorrelation, and their respective suitability for non-stationarity.

with ϵ being the error term and $\beta \in \mathbb{R}^m$ denoting the coefficients. In OLS, the coefficients can be estimated using matrix inverse and multiplication: $\beta = (X^T X)^{-1} X^T Y$. The intercept can be included in this model through a column vector of ones added to the feature matrix, which yields $X \in \mathbb{R}^{n \times m+1}$ and $\beta \in \mathbb{R}^{m+1}$. Note that applying OLS on spatial data is not generally advisable since it assumes that the samples are independent. This is not the case with (geo)spatial data because these are often taken from shared contexts, originate from processes with endogenous spatial dispersal mechanisms, or may be driven by spatially structured covariates. We nevertheless include OLS in our comparison as it is widely used as a yardstick against which to assess the usefulness of spatially explicit methods.

The second global linear method tested is the Spatial Lag in X model (SLX). This model takes into account spatially lagged independent variables and is given as

$$Y = \rho W X + X\beta + \epsilon,$$

where W is the spatial weights matrix that is computed as the inverse distance of the 20 nearest neighbors (see DGP), and ρ is the spatial coefficient. The estimation of β and ρ can be solved by adding the spatially-lagged X as additional covariates, and estimating two sets of coefficients for X and WX respectively.

2.2.2 Geographically Weighted Regression

Although Geographically Weighted Regression (GWR) was proposed for the analysis (not prediction) of spatial data, it is a suitable local model to account for non-stationarity in regression problems. GWR follows the standard linear regression framework but assumes that the coefficients β are dependent on locations (u_i, v_i). The model specification is given as

$$y_i = \sum_j \beta_j(u_i, v_i) X_{ij} + \epsilon.$$

In GWR, the local coefficients are estimated by building local models around each sample including only the spatial neighbors within a bandwidth. The latter can either be *fixed* (i.e., a pre-set distance) or *adaptive* (i.e., varying in space). The bandwidth is optimized by means of the golden-section search algorithm based on the Corrected Akaike Information Criterion (AICc) or with cross-validation (CV). Here, we tune a *fixed* bandwidth with the AICc criterion and use an exponential kernel. Our analysis aims to benchmark established local and global models on a synthetic (single-scale) task. We, therefore, use the original GWR specification but do not consider variants of the model such as multi-scale GWR [10].

2.2.3 Random Forest Regression models

Random Forests (RFs) are established machine learning models for regression tasks and have been shown to be very successful for a wide range of applications. We choose RFs as the main non-linear model in our experiments since it is arguably most prominent in spatial applications and does not require extensive parameter tuning. An RF is formed as an ensemble of decision trees that can learn arbitrary non-linear relations. The prediction of an RF is the average over the tree-wise outputs. We use the implementation provided through the `scikit-learn` [23] package.

To give RFs the ability to learn spatially non-stationary processes, a simple approach is to include the geographic coordinates as covariates [18]. We denote this RF-variant by *RF (coordinates)* in the following. In general, this approach is not recommended, since such a model is not applicable to other spatial regions [5]. However, we only regard regression within the same region here.

2.2.4 Spatial Random Forests

Aside from simply extending non-linear models by adding geographic coordinates or spatial features as covariates, another option is to fit them locally, as a non-linear counterpart to GWR. Similar to [11], we implement this approach for RFs. To provide a local yet efficient approach, we exploit the bootstrapping nature of RFs and fit a fixed number of spatially-disjoint decision trees. The decision trees are rooted in the cluster centers of K-Means clustering applied to the dataset. At test time, the prediction for a test sample is given by the weighted average of tree-wise predictions, where the weights are defined by the inverse distances of the test sample to the root of each tree respectively. While Georganos et al. [11] proposed a weighting of the spatial-RF and the global-RF predictions, we set the weight to 1 for a fair comparison between global and local models. Our version of spatial RFs is made available as an open-source package[4]. We validated that our spatial RF achieves similar performance as the implementation by Georganos et al. [11] and found that it is actually superior in 65.7% of all simulated scenarios and under ceteris paribus conditions.

2.2.5 Kriging

Another method that we employ is Kriging. This method is a well-known approach for interpolating geospatial data. A suitable variant for regression tasks is so-called Regression Kriging, which corresponds to universal Kriging with external drift. Regression Kriging essentially tackles (possibly non-linear) regression problems by fitting an arbitrary (global) regression model on the data and then applying Kriging on the *residuals*. Here, we use an RF as the base regressor in order to achieve maximum comparability to the global RF models, and employ the Kriging implementation offered in the `pykrige` package [20]. All Random Forest-based models are fitted with 100 base estimators and a maximum tree depth of 30. Increasing the number of estimators to 150 did not yield any significant improvements. We did not tune other parameters for a fair comparison. For GWR and spatial RFs, the bandwidth is tuned on validation data.

2.3 Experimental setup

We construct synthetic data following the DGPs described above, and evaluate the seven regression models in each scenario. The data is thereby randomly split and each model is trained on 90% of the data and tested on the remaining 10%. To study the effect of the sample size, we generate four datasets with $n = 100, n = 500, n = 1000$, and $n = 5000$ samples respectively, and $k = 5$ attributes for each sample. Our DGPs allow to compare model performances subject to varying degrees of non-stationarity (λ) and of the variance of

[4] `https://github.com/mie-lab/spatial_rf_python`

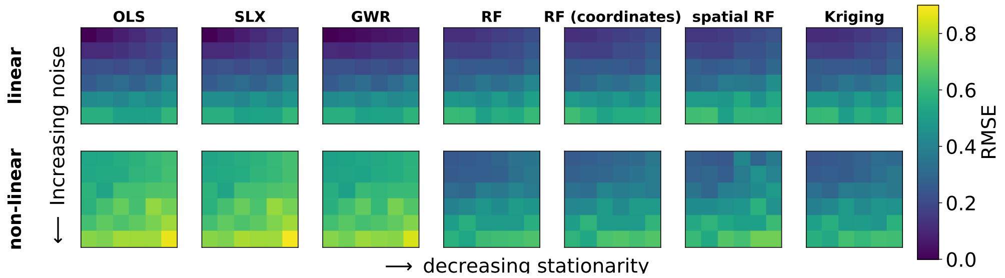


Figure 4 Results on the synthetic dataset (1000 samples). Performance in general decreases with noise and with the degree of non-stationarity (lowest performance in the bottom right of each plot). On linear data, GWR can account for non-stationarity, in contrast to other models. A random forest is better suited for non-linear phenomena, but spatial (locally fitted) RFs do not provide any benefits in these scenarios.

error terms (σ). In our experiments, we systematically vary the spatial non-stationarity by setting the factor λ to values between 0 and 0.5 (see Section 2.1.1). Furthermore, we vary the signal-to-noise ratio by setting σ to values between 0 and 0.5, where $\sigma = 0.5$ corresponds to a low signal-to-noise ratio (i.e., strong noise).

3 Results and discussion

In the following, we first compare the performances of the models on our synthetic dataset, then derive recommendations for model selection, and finally validate these recommendations in experiments on five real-world datasets.

3.1 Results obtained from synthetic data

The model performances in terms of test-data RMSE are visualized in Figure 4, divided by data generating function (row), spatial non-stationarity (x-axis), and noise level (y-axis). Only the scenarios with 1000 samples and uniformly distributed noise ϵ are shown. As expected, the performance generally decreases with higher noise levels or higher degrees of spatial heterogeneity (see highest RMSE in the bottom right corner of each scenario depicted in Figure 4). For the linear DGP, one can clearly see the superiority of GWR in dealing with locally varying spatial data, as it is indeed very robust to the adjusted spatial heterogeneity. The linear models (GWR, OLS, and SAR) are also clearly better at dealing with noise in linear regression tasks, whereas non-linear regressors such as Random Forests may struggle from overfitting. However, the latter picture changes when considering a non-linear function. The non-linear models yet generally struggle more with spatial non-stationarity than their linear counterparts. Surprisingly, spatial RFs are consistently outperformed by other models for the linear case, probably due to overfitting local models on the limited number of samples. The figure further indicates that a spatial RF is also not the best model when it comes to non-linear scenarios, though better than GWR. In this case, the problem may be underfitting, given the lower number of samples that are fed into each local model.

3.1.1 Effect of the spatial heterogeneity of the errors

As explained in Section 2.1.2 we additionally simulate different distributions of the variance of the errors ϵ (see Figure 2). Figure 5 visualizes the RMSE for GWR and Regression Kriging by the noise level. The outcomes obtained for degrees of non-stationarity $\lambda \in \{0.3, 0.4, 0.5\}$

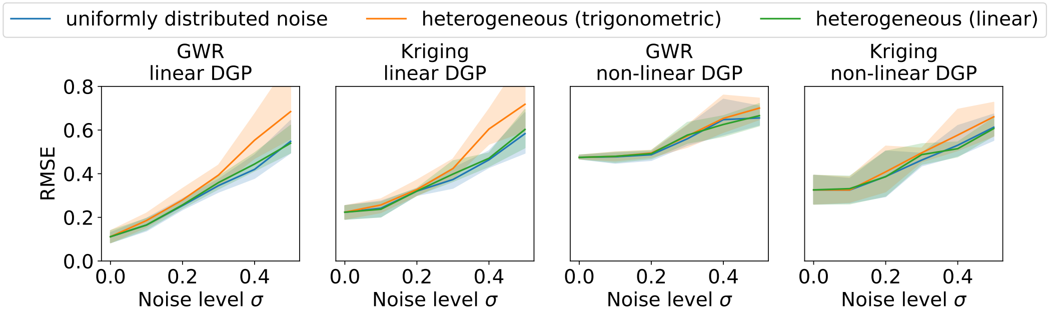


Figure 5 The average RMSEs with their 95% confidence intervals for varying noise levels (500 samples, averaged over scenarios with high degrees of non-stationarity). The RMSE is highest if the noise varies in the same fashion as the coefficients (heterogeneous – trigonometric) for the linear DGP. For the non-linear DGP the noise pattern has no significant influence.

are thereby averaged for obtaining an easier-to-interpret picture, so the blue lines (uniformly distributed noise) in Figure 5 correspond to the right part of the squares in Figure 4. In general, the type of distribution only has a minor effect compared to the average noise level, in particular for the non-linear GDP. However, at stronger noise levels, the scenario with trigonometrically varying noise is clearly the most difficult. Additionally, the variance of the RMSE increases in that case. Since the non-stationarity of the coefficients β is also modeled trigonmetrically, these findings indicate that the models particularly struggle to distinguish signal from noise if the variance of the errors is distributed similarly to the non-stationarity.

3.1.2 Comparing training and test errors

GWR and related local models are oftentimes only evaluated in terms of their fit to the input data, and not by means of inference on unseen data. Since GWR is based on linear models, the risk of overfitting is considered low, and evaluating the fit on test data is deemed unnecessary. Here, we make the case for evaluating models in terms of their predictive power, since even local linear models may overfit due to their higher number of parameters, and because local sample sizes are often small. To justify this argument empirically, we compare the RMSE on training and test data in our experiment. We find that Random Forest-based models (including Regression Kriging as we base it on RF) generally achieve very small training errors (RMSE < 0.01), which is expected since the individual decision trees overfit on the training data and only the boosting approach leads to good test performance. In Figure 6 we therefore only compare the results for SLX and GWR to showcase the danger of overfitting even linear models when they are local. Here, we consider $\lambda \in \{0, 0.1, 0.2\}$ as "weak non-stationarity" and $\lambda \in \{0.3, 0.4, 0.5\}$ as "strong non-stationarity", $\sigma \in \{0, 0.1, 0.2\}$ as "weak noise" and $\sigma \in \{0.3, 0.4, 0.5\}$ as "strong noise". The results are averaged over these scenarios for $n = 1000$ samples. Figure 6 shows that SLX as a global linear model hardly overfits on the data, whereas for GWR, which has considerably more parameters than global linear models, the training and test errors indeed diverge in some scenarios. For example, when there is strong non-stationarity but weak noise, the test error of GWR is 31% higher than its training error. This demonstrates the necessity to validate models on test data when employing them in predictive instead of purely analytical scenarios.

Additionally, overfitting may even lead to misinterpretations of analytical results of GWR, such as the visualization of the estimated coefficients on a map. The effect of overfitting on the spatial interpretation is application-dependent, but we exemplify the problem in

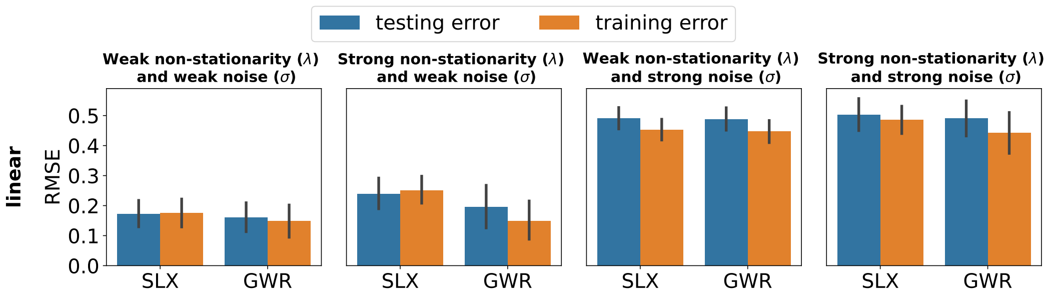


Figure 6 Comparing training and test errors in different scenarios. Even linear models such as GWR show overfitting behaviour, i.e., a lower test than train score, if there is either noise or non-stationarity in the data.

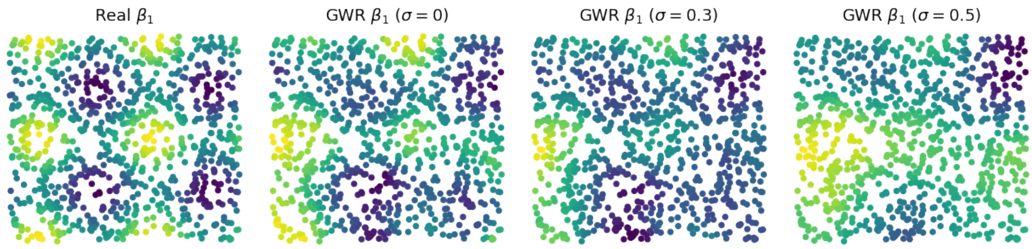


Figure 7 Comparing the GWR-estimated coefficients to the real coefficient for different signal-to-noise ratio. With noisy data, the spatial interpretation can be distorted.

Figure 7. The figure shows the true spatial variation of one coefficient β_1 in synthetic data ($n = 1000$) with moderate spatial heterogeneity ($\lambda = 0.3$), as well as the distribution of its estimate obtained with GWR. With decreasing signal-to-noise ratio, the spatial pattern of the estimated coefficient is perturbed. The pattern for $\sigma = 0.5$ indicates a single area with high β_1 on the left side of the region, in contrast to the true trigonometric pattern. This shows the potential for misinterpreting the results of a model with a bad fit to the data and calls for validation on test data before spatial analysis and interpretation. Of course, there is no unequivocal and generally agreed definition for when a model is overfitting, and overfitting may not be problematic as long as the test performance is sufficiently high. However, the *interpretation* of coefficients should be considered with caution in such case. For example, one could only analyze the coefficients of local models that were fit on a sufficient number of samples.

3.2 Proposed criteria for model selection

Our experiments on synthetic data allow to derive recommendations for choosing a model, dependent on the prediction task and on data availability. In general, the results in Figure 4 render linear models such as SLX most suitable for the linear DGP, with clear advantages of GWR in non-stationary scenarios. In contrast, Random Forest-based models are superior in the case of a non-linear DGP, while local RFs do not seem to provide many benefits. However, in real-world scenarios, the DGP is usually expected to be neither perfectly linear nor as complex as our non-linear scenario. It is therefore worthwhile to consider further factors such as the sample size. For this purpose, we analyze scenarios with strong non-stationarity ($\lambda \in \{0.3, 0.4, 0.5\}$) and weak noise ($\sigma \in \{0, 0.1, 0.2\}$) by sample size in Figure 8. Note that

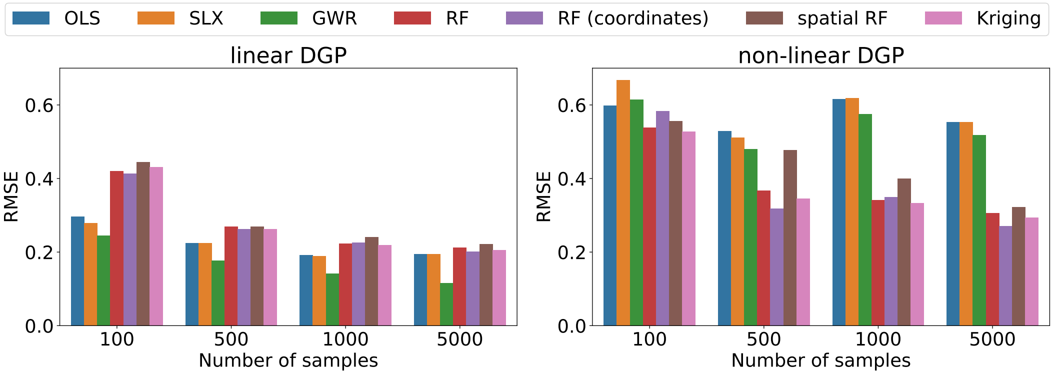


Figure 8 Comparing performance by the number of samples. The figure shows the average RMSE over all scenarios with strong non-stationarity ($\lambda \in \{0.3, 0.4, 0.5\}$) and weak noise ($\sigma \in \{0, 0.1, 0.2\}$).

the samples are constructed by *infill sampling* in a fixed spatial region, implying that a higher n leads to a higher sample *density*. For $n = 1000$ samples, the results correspond to the average over the top-right quadrant of each square in Figure 4.

As Figure 8 shows, RF models perform similarly well on linear data in scenarios with high sample density, whereas GWR is almost on-par for non-linear data when only 100 samples are provided. This observation leads us to derive the model selection tree presented in Figure 9: If there is clearly a high sample density over space, RFs should be used, whereas linear models are advisable in scenarios with very low sample density, or if the phenomenon is expected to show linear relations. Since the value of a "high" or "low" sampling density is application-dependent, this criterion must be decided on the basis of an analysis of the local number of samples, e.g., by the number of samples within the set range in a semivariogram. In scenarios with high non-stationarity, Kriging or spatial features in the RF are beneficial. Global RFs should be tested in any case, in order to validate the necessity of local models. It must be noted, however, that our analysis does not consider big data scenarios, where RFs may still perform well but would need to be replaced by more memory-efficient methods such as stochastic gradient descent.

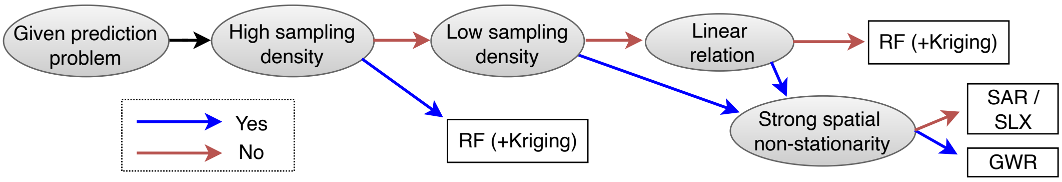


Figure 9 Proposed criteria for model selection. The model choices were derived from experiments with synthetic data of varying non-stationarity, sample size, and DGP (linear vs non-linear).

3.3 Results based on real-world data

We experiment with five benchmark datasets that have been used in previous work on spatial data analysis and prediction, e.g. [19, 22, 14, 1]. The following sub-section first introduces these datasets. Afterward, we discuss the results obtained.

3.3.1 Datasets

There are five real-world, publicly available datasets that we employ for validation:

- The **California housing dataset**[5] was generated from the 1990 California census. Our goal is to predict the median house price from the location and seven other variables, such as the size and number of rooms, age, income, and population size. The number of bedrooms is missing for 1% of the houses and we omitted those respective records.
- The **Atlantic mortality dataset**[6] captures county-level mortality rates from 2010–2012, from which we have extracted only one year's worth of data for our purposes. Rates of smoking and poverty, as well as PM25, SO2, and NO2 levels provided as annual means are utilized as covariates.
- We further use a dataset on **deforestation rates**[7] that was published by Santos et al. [27]. The dataset provides annual deforestation rates from 2000 to 2010 for 2418 grid cells (single values averaged over 10 years). The deforestation rate is to be predicted from 35 further variables about sociodemographics, spatial features, and economic information. The forestation rate is given as four *quantiles*, which is problematic for the framing as a *regression* problem. The results must therefore be taken with a grain of salt.
- The **Meuse river dataset**[8] is another standard dataset for experimental spatial analysis [25]. It is a rather small collection of soil measurements including copper, cadmium, and zinc. Usually, this dataset is used to predict zinc concentration from the other soil measurements as well as from further contextual information. For preprocessing, we omit the categorical "landuse" variable and two incomplete samples.
- Finally, a dataset on **plant richness** is included that was used for validating spatial random forests[9]. Plant species richness is given for 227 ecoregions in America, and there are 18 covariates with information on topography, land use, human population, and climate.

3.3.2 Results obtained from real-world data

To validate the model selection tree presented in Figure 9 on real-world data, we first compute an indicator for the degree of non-stationarity. The LOSH statistic [21] offers a way to estimate local heterogeneity in terms of a local, spatially-weighted variance estimator. When applying LOSH with a K-nearest-neighbor (KNN) weights matrix (here 20 neighbors), the global average of all LOSH values indicates the average heterogeneity with respect to the sample density. As shown in Table 1, we find LOSH values around 1.0 in the five real-world datasets, where the California housing and the Meuse datasets show lower local heterogeneity (øLOSH of 0.88 and 0.89) , and the plants data is subject to stronger local heterogeneity (øLOSH of 1.06). Table 1 further gives the number of samples and the number of covariates k as an indicator of the problem complexity. We then quantified the model performances in terms of RMSE, mean absolute error (MAE), and the R-squared score; however, all metrics yield the same ranking of methods, and we therefore only report the RMSE in Table 1.

[5] We use the public dataset available from Kaggle: `https://www.kaggle.com/datasets/camnugent/california-housing-prices?resource=download`.

[6] The data is available from `https://zia207.github.io/geospatial-r-github.io/geographically-wighted-random-forest.html`.

[7] Data downloaded from `https://github.com/FSantosCodes/GWRFC/tree/master/data`

[8] The data is included in the R package `sp`: `https://rsbivand.github.io/sp/reference/meuse.html`

[9] The data is available from `https://blasbenito.github.io/spatialRF/#data-requirements`.

Table 1 Model benchmarking on real-world data. We find that GWR performs better on the Atlantic dataset and the Meuse data, whereas non-linear models yield lower RMSE on datasets with higher sample density as expected (e.g. California housing).

Dataset	Samples	k	∅ LOSH	**RMSE**						
				OLS	SLX	GWR	RF	RF (coord.)	spatial RF	Kriging
Atlantic	666	7	1.00	8.65	8.64	**7.14**	7.54	7.34	8.18	7.43
California housing	20433	8	0.88	72244	63532	56156	61234	**48209**	67493	55173
Deforestation	2418	36	1.00	0.83	0.82	0.80	0.66	0.67	0.71	**0.66**
Meuse	153	11	0.89	51.65	54.80	**48.40**	68.13	68.13	88.70	64.16
Plants	227	18	1.06	2349	2334	2226	2216	2288	2507	**2120**

For validating the results, we compare to previous results reported for these datasets, and our scores improve over the ones reported in the data-accompanying tutorials[6,9], or achieve comparable results as related work [19].

We confirm previous results that spatial models achieve good results on these spatial datasets. However, RF-based methods perform better on several datasets, in particular, when the sample density is sufficiently high (e.g., California housing) or when the process analyzed is more complex (e.g., predicting quantiles in the deforestation dataset from 36 variables; or predicting plant richness from 18 covariates). A surprising result is the superiority of GWR above other model specifications for the Atlantic dataset (mortality rates) despite the intermediate LOSH value and sample size. This may be due to a rather linear dependency of Y on X, and is in line with previous findings [31]. Our results on real-world data, therefore, show the general applicability of our model selection criteria, but call for further efforts on quantifying spatial non-stationarity and problem complexity in spatial data.

4 Conclusions

While many promising regression methods were developed specifically for spatial data, there is a lack of analysis about the properties of data that render such models superior. We contribute to a better understanding of these conditions with an analysis systematically exploring the effects of non-stationarity, the signal-to-noise ratio, noise heterogeneity, the nature of the DGP (linear/non-linear), and sample size. Based on the experiments, we recommend using (local) linear models such as GWR for addressing problems encompassing a small sample size or strong non-stationarity. Further, we recommend using non-linear models such as Random Forests for prediction tasks involving larger spatial datasets, whereby locations should be fed into the model through additional spatial input features. RFs can further be combined with Kriging to better account for non-stationarity. While the type of data may give some indication of the non-stationarity and complexity, further work is necessary to assess spatial stationarity a priori. Promising avenues may be, for example, exploring spatial stationarity measures as proposed for time series [6], through better understanding localized (and varying) heterogeneity [30] or, alternatively, by controlling for complex forms of stationarity using Moran eigenvector filtering and its variants [29, 12]. We further argue that our results call for an increased significance of prediction for validating model performance. Even if a model is only used for analysis, the validity of the inferred coefficients should be evaluated via test data, since even linear local models are prone to overfitting in spatially structured noisy or non-stationary settings. At the same time, other factors that are not discussed in this work, such as model interpretability, may be important when it comes to model selection and may give preference to linear modeling even though non-linear models may be superior in terms of prediction. Future work could aim to combine the best of both worlds by improving the spatial interpretability of global models such as RFs.

Finally, our analysis is limited in scope regarding the considered properties and models. Follow-up work could put more focus on spatial autocorrelation and its interplay with non-stationarity, or explore other types of non-stationary non-linear relations. Another interesting path that some researchers have started venturing on is to integrate better modern machine learning models such as spatial neural networks with geospatial principles [13, 15]. We hope that our work inspires further efforts to properly benchmark new methods on both synthetic and real-world data, thereby improving our understanding of the use cases and advantages of spatially-explicit models.

References

1 Zia U Ahmed, Kang Sun, Michael Shelly, and Lina Mu. Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA. *Scientific Reports*, 11(1):1–15, 2021.

2 Colin M Beale, Jack J Lennon, Jon M Yearsley, Mark J Brewer, and David A Elston. Regression analysis of spatial data. *Ecology letters*, 13(2):246–264, 2010.

3 Chris Brunsdon, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443, 1998.

4 Alexis Comber, Christopher Brunsdon, Martin Charlton, Guanpeng Dong, Richard Harris, Binbin Lu, Yihe Lü, Daisuke Murakami, Tomoki Nakaya, Yunqiang Wang, et al. A route map for successful applications of geographically weighted regression. *Geographical Analysis*, 55(1):155–178, 2023.

5 Matthew J Cracknell and Anya M Reading. Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, 63:22–33, 2014.

6 Sourav Das and Guy P Nason. Measuring the degree of non-stationarity of a time series. *Stat*, 5(1):295–305, 2016.

7 Zhenhong Du, Zhongyi Wang, Sensen Wu, Feng Zhang, and Renyi Liu. Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. *International Journal of Geographical Information Science*, 34(7):1353–1377, 2020.

8 Andrew O Finley. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154, 2011.

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

10 A Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers*, 107(6):1247–1265, 2017.

11 Stefanos Georganos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff, and Stamatis Kalogirou. Geographical Random Forests: a spatial extension of the Random Forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2):121–136, 2021.

12 Daniel A Griffith and Yongwan Chun. Implementing Moran eigenvector spatial filtering for massively large georeferenced datasets. *International Journal of Geographical Information Science*, 33(9):1703–1717, 2019.

13 Julian Hagenauer and Marco Helbich. A geographically weighted artificial neural network. *International Journal of Geographical Information Science*, 36(2):215–235, 2022.

14 Tomislav Hengl, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler. Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

15 Konstantin Klemmer. *Improving neural networks for geospatial applications with geographic context embeddings*. PhD thesis, University of Warwick, Coventry, UK, 2022.

16 James LeSage. Spatial econometrics. In Charlie Karlsson, Martin Andersson, and Therese Norman, editors, *Handbook of research methods and applications in economic geography*, pages 23–40. Edward Elgar Publishing, Cheltenham, UK, 2015.

17 James P LeSage. A family of geographically weighted regression models. In Luc Anselin, Raymond J. G. M. Florax, and Sergio J. Rey, editors, *Advances in spatial econometrics*, pages 241–264. Springer, Berlin/Heidelberg, Germany, 2004.

18 Jin Li, Andrew D Heap, Anna Potter, and James J Daniell. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, 26(12):1647–1659, 2011.

19 Xiaojian Liu, Ourania Kounadi, and Raul Zurita-Milla. Incorporating spatial autocorrelation in machine learning models using spatial lag and eigenvector spatial filtering features. *ISPRS International Journal of Geo-Information*, 11(4):242, 2022.

20 Benjamin S Murphy. PyKrige: development of a Kriging toolkit for Python. In *American Geophysical Union Fall Meeting Abstracts*, volume 2014, pages H51K–0753, San Francisco, CA, USA, 2014.

21 J Keith Ord and Arthur Getis. Local spatial heteroscedasticity (LOSH). *The Annals of Regional Science*, 48:529–539, 2012.

22 R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

23 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

24 Martin Raubal. It's the spatial data science, stupid! In *Spatial Data Science Symposium "Setting the Spatial Data Science Agenda"*, Santa Barbara, CA, US, 2019. Center for Spatial Studies at the University of California.

25 MGJ Rikken and RPG Van Rijn. *Soil pollution with heavy metals: in inquiry into spatial variation, cost of mapping and the risk evaluation of Copper, Cadmium, Lead and Zinc in the floodplains of the Meuse West of Stein, The Netherlands: field study report*. University of Utrecht, 1993.

26 Sebastian Santibanez, Tobia Lakes, and Marius Kloft. Performance analysis of some machine learning algorithms for regression under varying spatial autocorrelation. In *Proceedings of the 18th AGILE International Conference on Geographic Information Science*, pages 9–12, Lisbon, Portugal, 2015.

27 Fabián Santos, Valerie Graw, and Santiago Bonilla. A geographically weighted Random Forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. *PloS one*, 14(12):e0226224, 2019.

28 Aleksandar Sekulić, Milan Kilibarda, Gerard B.M. Heuvelink, Mladen Nikolić, and Branislav Bajat. Random Forest spatial interpolation. *Remote Sensing*, 12(10):1687, 2020.

29 René Westerholt. Emphasising spatial structure in geosocial media data using spatial amplifier filtering. *Environment and Planning B: Urban Analytics and City Science*, 48(9):2842–2861, 2021.

30 René Westerholt, Bernd Resch, Franz-Benjamin Mocnik, and Dirk Hoffmeister. A statistical test on the local effects of spatially structured variance. *International Journal of Geographical Information Science*, 32(3):571–600, 2018.

31 Ryan Zhenqi Zhou, Yingjie Hu, Jill N Tirabassi, Yue Ma, and Zhen Xu. Deriving neighborhood-level diet and physical activity measurements from anonymized mobile phone location data for enhancing obesity estimation. *International Journal of Health Geographics*, 21(1):1–18, 2022.