MODELING RENTAL PRICE IN CALGARY UNDER SPATIAL BAYESIAN PARADIGMS

JIAJUN ZHANG
Directed Reading Program Report
McGill University

ABSTRACT

In this report, we model the monthly rental price in Calgary based on the data from 2024. We first fit a baseline linear regression model and accesses performances by computing Moran's I test and variograms to investigate potential spatial dependency among the residuals. To address the issue, our model involves a linear predictor term plus a isotropic spatial Gaussian processes with sill, nugget and range as parameters. We use a Bayesian framework where weakly-informative priors are selected for each parameter and use Markov Chain Monte Carlo (MCMC) methods to inference the parameters. We then examine the performance of the model and conclude the article by giving graphical illustrations of the spatial effect. Our model suggests that an additional bedroom is associated with an approximately 12.2% increase in rent, while holding other covariates and spatial effects fixed; An additional bathroom is associated with an approximately 22.1% increase in rent, while holding other covariates and spatial effects fixed. Finally, there is not strong posterior evidence that pet allowance has a clear effect on rent.

1 Introduction

Housing rental prices are determined by both dwelling-level characteristics and location-specific factors. The hedonic pricing framework treats housing as a differentiated good whose price reflects the implicit value of its attributes, such as bedrooms, bathrooms, and other unit features [Ros74]. However, rents are also strongly shaped by spatial context, including neighborhood amenities, accessibility, local demand, and nearby land use. As a result, nearby properties often exhibit similar prices even after controlling for observable structural characteristics. This motivates the use of spatial methods in housing price analysis, since ignoring spatial dependence may lead to incomplete inference and weaker prediction [BCH07, Osl10]. This report is an exercise of illustrating how to model monthly rental price while considering spatial effects. The report is organized as follows: Section 2 explains some basic concepts and the set up of our model; Section 3 provides the results and analysis of our model.

2 Method

2.1 Data

Our dataset is taken from Kaggle [Gav24] which contains 25000+ rental listings in Canada in June 2024. In this report, we focus on the rental data in the city of Calgary, Alberta. The dataset contains 12 covariates, namely: Address, Longitude, Latitude, Lease Term, Type of the unit (i.e., apartment, house, etc.), Number of Bedrooms (Beds), Number of Bathrooms (Baths), Square Feet of the unit, Furnishing (Yes or No), Smoking (Yes or No), Pets (Yes or No), Availability Date. In our model, the response variable y is the monthly rental price of the unit. According to exploratory data analysis using Bayesian Lasso, the covariates we consider are Number of bedrooms; Number of bathrooms; Square feet; Pets.

The prepossessing process of the dataset consists of decoding the covariates into numeric values, restricting the geographical location within Calgary and removing extreme outliers. In our model, beds, baths are decoded as non-negative integers, with the convention that a studio apartment will have 0 bedroom, and pets is decoded as a binary variable, (1 being "Yes", and 0 being "No"). After data prepossessing we have a total number of 7824 observations for Calgary.

2.2 Baseline Model

We first consider a baseline model being pure linear regression using the log price as outcome. The model takes the form $\log(y) = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon$, which is

$$\log(\text{price}) = \beta_1 + \beta_2 \cdot \text{Beds} + \beta_3 \cdot \text{Baths} + \beta_4 \cdot \text{Pets} + \varepsilon, \tag{1}$$

where $\varepsilon \sim N(0, \sigma^2)$, and β_1 is the intercept term. The residual of the data is given by the difference between observed value and fitted value under the model, namely:

$$e_i = \log(y_i) - \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

To evaluate our model, we compare the residual of different data points with their observed value. To avoid overfitting, we split the data into a training set (80% of data) and a testing set (20 % of data). We fit the baseline model then calculated the fitted value and residuals on the training set. We then test this model to calculate the predicted value and residuals on the testing set. The performance of the baseline model is recorded in Table 1 and Figure 1.

Variable	Estimate	Std. Error	p-value
Intercept	7.1499	0.0094	< 0.001
Beds	0.0698	0.0049	< 0.001
Baths	0.1976	0.0063	< 0.001
Pets allowed	0.1148	0.0067	< 0.001
$R^2 = 0.3822$, Adjusted $R^2 = 0.3819$, Residual SE $= 0.2628$			

Table 1: Estimated coefficients and significant figures for the baseline model.

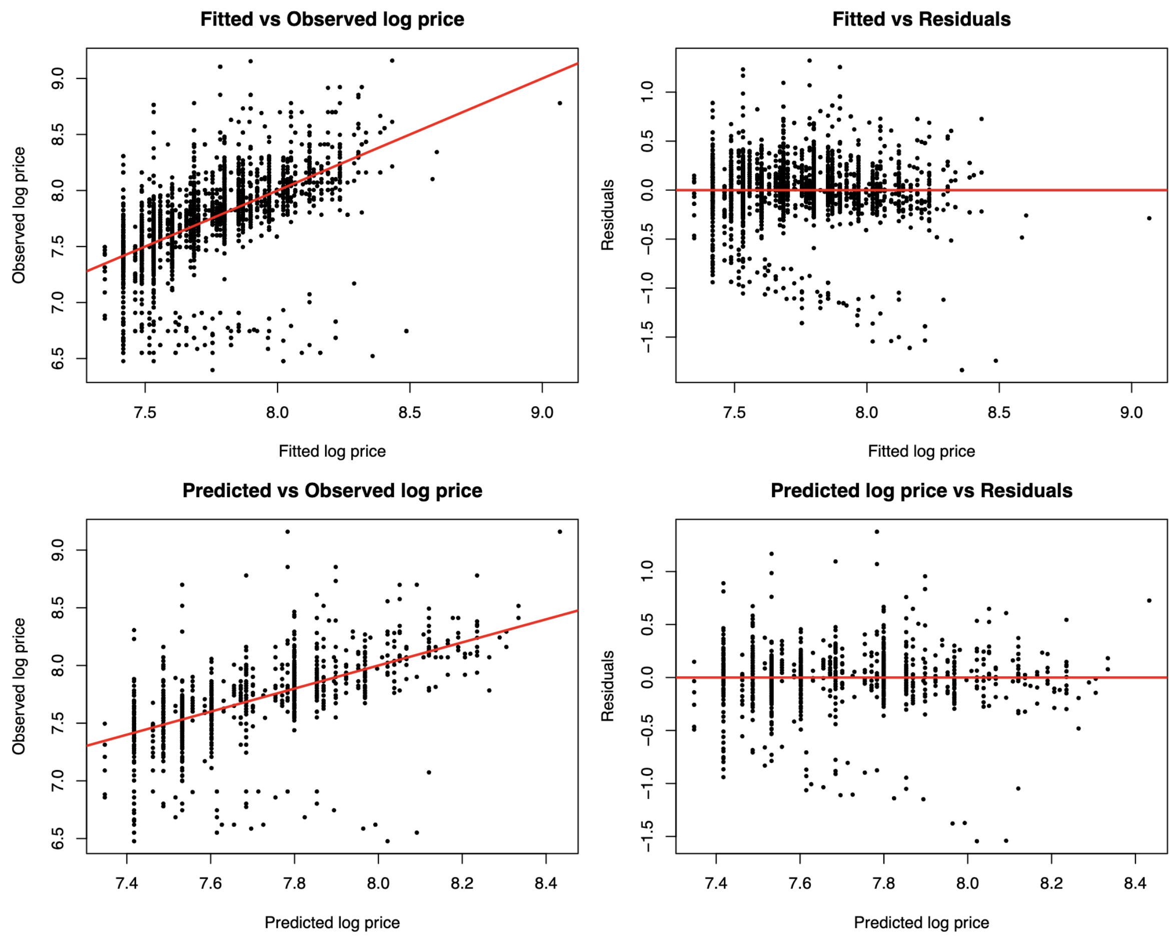

Figure 1: The 2×2 plots of the baseline model.

Table 1 shows that all three predictors are highly significant with low p values. While holding other variables fixed, an additional bedroom is associated with an estimated increase of 0.0698 in $\log(y)$, an additional bathroom with an increase of 0.1976, and allowing pets with an increase of 0.1148. Finally, the model explains about 38% of the variation in log rental price.

Figure 1 suggests bad performance of the baseline model, hence it is reasonable to seek for other potential factors that may impact $\log(y)$. Among them, we consider potential spatial dependency. To justify our hypothesis, we use Moran's I [Mor50] as a measure of spatial correlation on the residual side. The Moran's I statistic is defined by

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(e_i - \bar{e})(e_j - \bar{e})}{\sum_{i=1}^{n}(e_i - \bar{e})^2}. \tag{2}$$

In Equation (2), $\bar{e}$ is the mean of the observed residual;, e_i is the residual for observation indexed i; n is the total number of observations; w_{ij} are the entries of a spatial weight matrix $\boldsymbol{W}$ between observation indexed i and j. By convention $\boldsymbol{W}$ is symmetric and has 0 on its main diagonal. Because the observations are point-referenced rental listings rather than areal units, neighbors were defined using a k-nearest-neighbor graph. We used $k = 8$ to represent a local neighborhood around each listing while ensuring that every observation has the same number of neighbors. For observation indexed i, we choose its $k = 8$ nearest neighbors and assign equal weight $w_{ij} = 1/8$ for them. The weight for the rest of the pairs is set to 0.

To assess the significance of Moran's I, we perform the following hypothesis test: The null hypothesis H_0 is that there is no spatial autocorrelation. Under H_0, the Moran's I has an approximate normal distribution after standardization:

$$z = \frac{I_{\text{obs}} - \mathbb{E}[I]}{\text{Var}(I)} \sim N(0,1),$$

where $\mathbb{E}[I] = \frac{-1}{n-1}$, $\text{Var}(I) = \frac{NS_4 - S_3S_5}{(N-1)(N-2)(N-3)S_0^2} - (\mathbb{E}[I])^2$. $S_0, S_1, \cdots, S_5$ are all quantities defined through w_{ij} and e_i [CO81]. From our data, under the null hypothesis, the expected value $\mathbb{E}[I]$ is given by -0.0007255, and the variance is given by 5.6196×10^{-5}. Moran's I statistic for our dataset is 0.4297, with standard deviate 57.393 and p-value $< 2.2 \times 10^{-16}$. Hence, it is statistically significant to reject the null hypothesis, meaning the residuals are spatially correlated according to the test.

2.3 Varigoram and Covariogram

A variogram γ [Geo63] is a function measuring how residual changes based on the spatial distance. From now on, we use the following notation

$$\boldsymbol{Y}(\boldsymbol{s}) = \begin{pmatrix} Y(\boldsymbol{s}_1) & \cdots & Y(\boldsymbol{s}_n) \end{pmatrix}^\top$$

where $Y(\boldsymbol{s}_i)$ has realization y_i on spatial location $\boldsymbol{s}_i$ under a certain spatial process. The model we assume is

$$\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{3}$$

where $\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}$ is the fixed effect from the linear model introduced in (1). The second term $\boldsymbol{w}(\boldsymbol{s})$ is a zero-centered stationary isotropic Gaussian processes capturing the spatial dependency which cannot be explained by the linear predictor term. The third term $\boldsymbol{\epsilon}(\boldsymbol{s})$ is the uncorrelated independent and identically distributed (i.i.d) error called nugget, and $\varepsilon_i(\boldsymbol{s}_i) \overset{\text{i.i.d}}{\sim} N(\boldsymbol{0}, \tau^2)$ for all i. For any two spatial sites $\boldsymbol{s}, \boldsymbol{s}'$, define $\boldsymbol{h} = \boldsymbol{s}' - \boldsymbol{s}$, and we have $\mathbb{E}[w(\boldsymbol{s}+\boldsymbol{h}) - w(\boldsymbol{s})] = 0$. The variogram of the Gaussian process $w(\boldsymbol{s})$ with respect to $\boldsymbol{h}$ is given by

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\mathbb{E}[w(\boldsymbol{s}+\boldsymbol{h}) - w(\boldsymbol{s})]^2 = \frac{1}{2}\textbf{Var}(w(\boldsymbol{s}+\boldsymbol{h}) - w(\boldsymbol{s})).$$

The variogram $\gamma(\boldsymbol{h})$ increases within a certain range, and as distance increases, the spatial correlation decreases [Tob70]. Behaviorally, we would expect similar values of $\mathbb{E}[w(\boldsymbol{s})]$ and $\mathbb{E}[w(\boldsymbol{s}+\boldsymbol{h})]$ at short distance between two spatial sites. When $||\boldsymbol{h}||$ gets larger the similarity decreases resulting in a larger variogram. Finally when $||\boldsymbol{h}||$ gets large enough so that the spaial dependency is almost negligible, the variogram will stay stable at a certain threshold.

A theoretical variograram can be summarized using three parameters: *sill* σ^2, *nugget* τ^2 and *range* ϕ. Sill (σ^2) measures how much the variogram has increased after reaching stable. Nugget (τ^2) is the pure error term introduced in Equation (3) which is spatically uncorrelated. For example, it can be viewed as the price fluctuation for different units within the same apartment building. Range (ϕ) describes the spatial dependency strength, which can be viewed as the spatial distance when the variogram reaches sill. In our model, we compute the empirical variogram by

$$\widehat{\gamma}(d) = \frac{1}{2N(d)} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in N(d)} |Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)|^2, \tag{4}$$

where $N(d)$ is the set of pairs such that $||\boldsymbol{s}_i - \boldsymbol{s}_j|| = d$. Using Equation (4), we obtain the empirical variogram for our dataset:

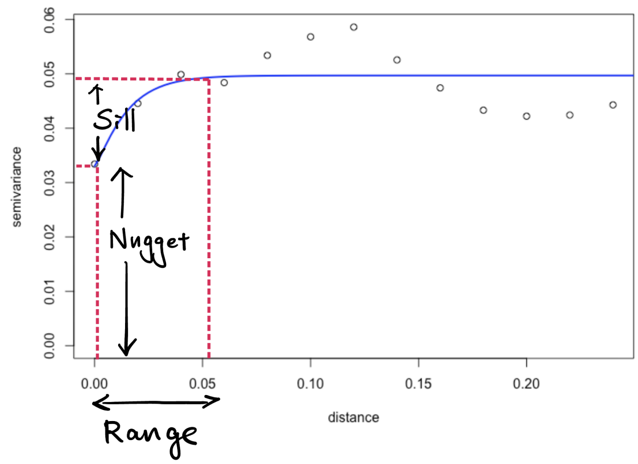

Figure 2: This figure shows the empirical semi-variogram ($\gamma(\boldsymbol{h})/2$) from our data. The blue curve is a smoothened curve using Matérn variance-covariance function.

In our analysis, we will assume our variogram is *isotrophic*, meaning $\gamma(\boldsymbol{h})$ depends only through $d = ||\boldsymbol{h}||$ and it is not direction wise. Isotrophic variograms are popular due to their simplicity and interpretability, and have a number of simple parametric forms as candidates for the variogram [BGC26]. The Matérn variogram [Mat86] is defined by

$$\gamma(d) = \begin{cases} \sigma^2\left[1 - \frac{(2\sqrt{\nu}d\phi)^\nu}{2^{\nu-1}\Gamma(\nu)}K_\nu(2\sqrt{\nu}d\phi)\right] + \tau^2 & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases},$$

where $d = ||\boldsymbol{h}||$, and K_ν is the modified Bessel function [Fra58] of order ν. ν controls the smoothness of the realized random field [BGC26]. In our model we fix $\nu = 1.5$. Apart from variogram, we are also interested in the covariance function $C(\boldsymbol{h})$ between any two spatial sites $\boldsymbol{s}$ and $\boldsymbol{s}' = \boldsymbol{s} + \boldsymbol{h}$. It can be shown that

$$C(\boldsymbol{h}) = \lim_{||\boldsymbol{u}|| \to \infty} \gamma(\boldsymbol{u}) - \gamma(\boldsymbol{h}),$$

which is monotoneically decreasing. Under isotrophy, the Matérn covariogram $C(d)$ is given by

$$C(d) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\phi d)^\nu K_\nu(\phi d) & \text{if } d > 0 \\ \sigma^2 + \tau^2 & \text{if } d = 0 \end{cases}, \quad \text{where } d = ||\boldsymbol{h}||. \tag{5}$$

Based on our explortary data analysis using Moran's I test and Matérn variogram for residuals in Sections 2.2 and 2.3, it is necessary to model spatial dependency.

2.4 Bayesian Inference

We use a Bayesian approach to estimate the parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi\right)^\top$. We assume $\boldsymbol{w}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s})$ defined through Equation (3) follows the spatial isotropy Gaussian processes with covariance function $\Sigma(\sigma^2, \tau^2, \phi)$.Under our assumption, we have

$$\boldsymbol{r}(\boldsymbol{s}) = \boldsymbol{w}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}) \sim MVN\left(\boldsymbol{0}, \Sigma(\sigma^2, \tau^2, \phi)\right).$$

The entries of the variance-covariance matrix are defined through the modified Matérn covariogram ($\nu = 1.5$) where

$$[\Sigma(\sigma^2, \tau^2, \phi)]_{ij} = \mathrm{Cov}(r(\boldsymbol{s}_i), r(\boldsymbol{s}_j)) = \begin{cases} \sigma^2 \left(1 + \frac{\sqrt{3}d}{\phi}\right) \exp\left(-\frac{\sqrt{3}d}{\phi}\right) & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases},$$

with $d = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$. Hence in the model $\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s})$, the data generating mechanism of $\boldsymbol{Y}(\boldsymbol{s})$ is specified using parameter vector $\boldsymbol{\theta}$ is

$$\boldsymbol{Y}(\boldsymbol{s})|\boldsymbol{\theta} \sim MVN\Big(\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}, \Sigma(\sigma^2, \tau^2, \phi)\Big).$$

Using conditional probability, the posterior distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{Y}(\boldsymbol{s})$ is

$$p_n(\boldsymbol{\theta}; \boldsymbol{Y}(\boldsymbol{s})) \propto f(\boldsymbol{Y}(\boldsymbol{s})|\boldsymbol{\theta}) \cdot p_0(\boldsymbol{\beta}) \cdot p_0(\sigma^2) \cdot p_0(\tau^2) \cdot p_0(\phi), \tag{6}$$

with some prior $p_0(\cdot)$ distribution on $\boldsymbol{\beta}, \sigma^2, \tau^2, \phi$, the conditional likelihood f of $\boldsymbol{Y}(\boldsymbol{s})$ given $\boldsymbol{\theta}$, and the posterior of $\boldsymbol{\theta}$. For the linear predictor, we choose a normal prior $N(0, 100)$ for $\beta_1, \beta_2, \beta_3, \beta_4$. As for the hyper parameter for the spatial Gaussian processes, despite σ^2 and ϕ cannot be identified, their product can be identified which is all one need for a consistent spatial prediction at a new spatial site [Zha04]. For computational and model simplicity, we still seek for independent inverse Gamma priors They are selected as follow.

$$\sigma^2 \sim \mathrm{InvGamma}(2, 0.03712); \quad \tau^2 \sim \mathrm{InvGamma}(2, 0.0092806); \quad \phi \sim \mathrm{InvGamma}(2, 0.129098),$$

Those are data-scaled priors. The scale for σ^2 and τ comes from the residual standard deviation of a preliminary model, while the scale for ϕ comes from the spatial scale of the observed locations.

Since we often want to make inferential statements about the parameters marginally, one can obtain the marginal posterior for each parameter from $p(\boldsymbol{\theta}; \boldsymbol{Y}(\boldsymbol{s}))$. For example, the marginal posterior for the linear predictor coefficients $\boldsymbol{\beta}$ arise from

$$p_n(\boldsymbol{\beta}|\boldsymbol{y}) = \iiint p_n(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi; \boldsymbol{y}) d\sigma^2 d\tau^2 d\phi. \tag{7}$$

However, Equation (7) has no closed form solution and can be high dimensional, hence a computational approach is needed. Therefore we resort to MCMC (Markov Chain Monte Carlo) methods. MCMC such as Gibbs sampling [GS90] and Metropolis-Hastings (MH) algorithm [Rob15], is an algorithmic approach that obtains samples from the posterior which is consistent and asymptotically accurate. After obtaining the posterior samples, one can use histogram based on it to illustrate the approximated posterior distribution up to arbitrarily accuracy. [GCS+25]. For example, the posterior mean is a minimizer under the L^2 loss:

$$\widehat{\boldsymbol{\theta}} = \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{\theta}^{(g)}.$$

We use Hamiltonian Monte Carlo (HMC) [Bet18] for posterior density estimation described in Equation (6). For efficient MCMC computation, we refer to the `rstan` software [Car].

3 Results and Discussion

3.1 MCMC diagnostics

Due to the computational complexity of the MCMC algorithm and time constraint, we are not able to implement the algorithm to the entire dataset. There are more advanced methods such as Fixed Rank Kriging (FRK) [CJ08], Nearest Neighbor Gaussian Process (NNGP) [DBFG16], and Local Approximate Gaussian Process (LaGP) [GA15] etc. For a review of more advanced computational methods for large spatial datasets, see Heaton et. al. [Hea19] for more detail. Given that it is a DRP project meant for undergraduate students and the short time constrain, we are not able to implement those advanced methods. Instead, we randomly choose 600 observations with no replicated spatial coordinates and split them into 80%/20% training / testing set. For the training set, the MCMC algorithm consists of 4 chains in total, with 500 warm up iterations and 1,000 sampling iterations. The MCMC trace plots and the posterior density plots are shown in Figure 3 and Figure 4. The posterior mean and 95% credible intervals for $\boldsymbol{\beta}$ and Matérn hyperparameters are reported in Table 2 below.

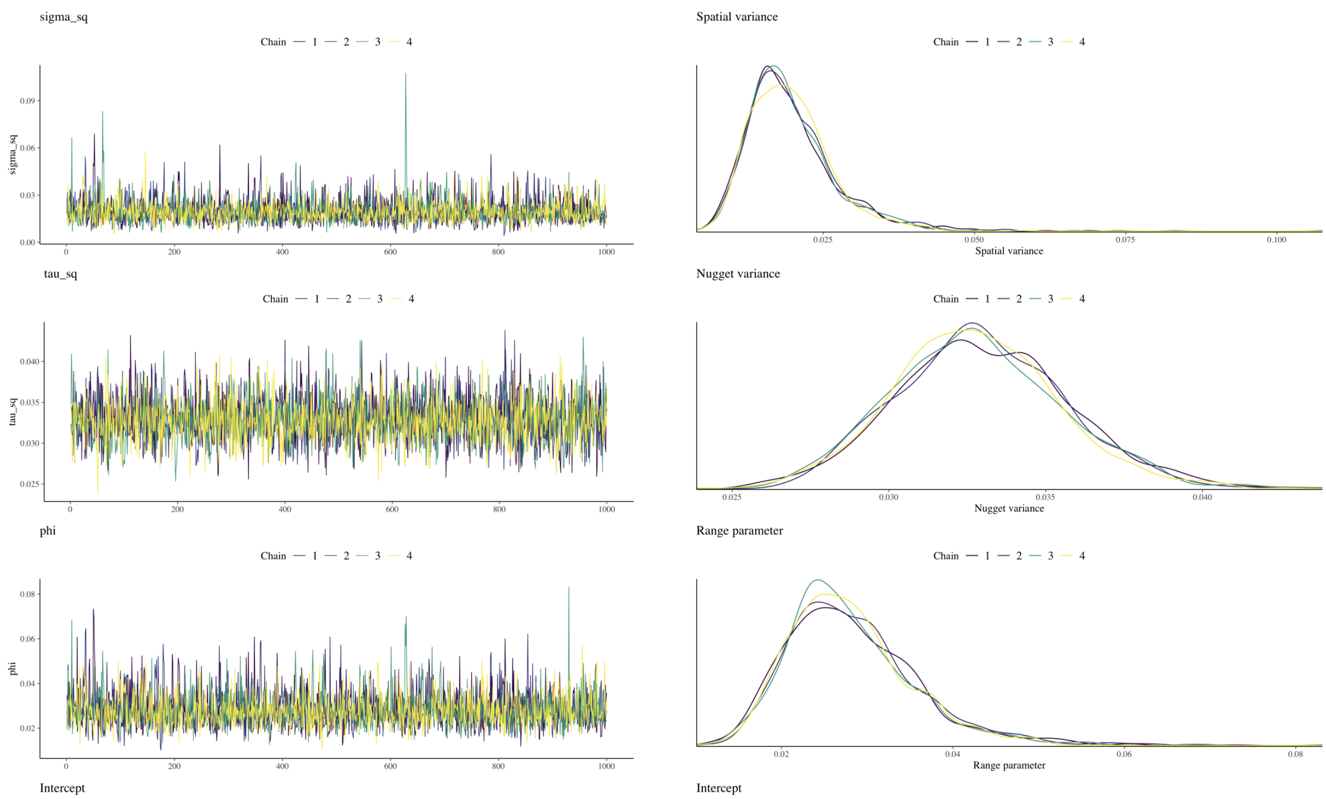

Figure 3: This figure shows the MCMC trace plots and the posterior densities for the Gaussian process parameters: Sill σ^2; Nugget τ^2 and Range ϕ.

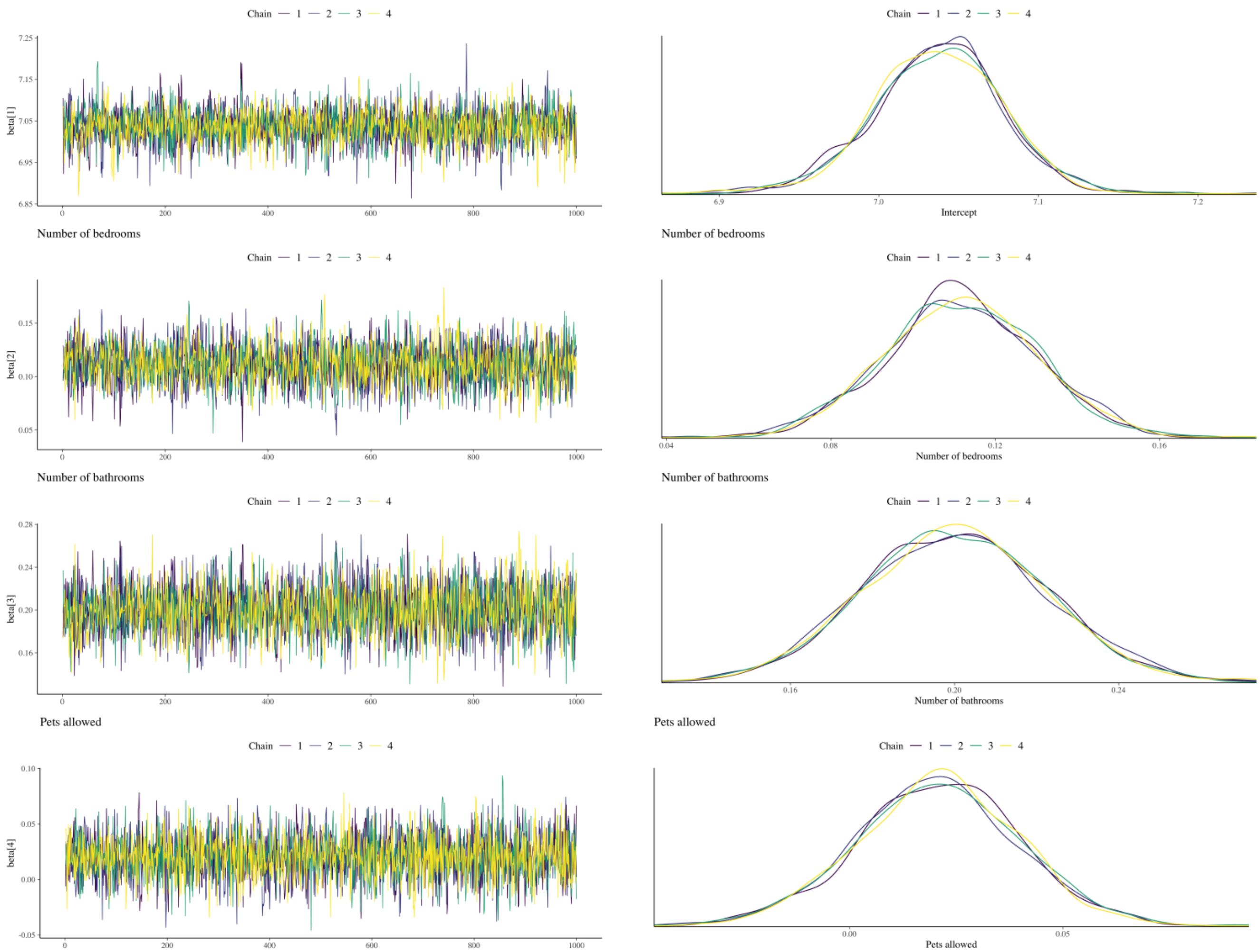

Figure 4: This figure shows the MCMC trace plots and the posterior densities for the linear predictor $\boldsymbol{\beta}$.

Parameter	Mean	Q2.5	Q97.5
Intercept (β_1)	7.0365	6.9538	7.1195
beds (β_2)	0.1149	0.0756	0.1469
baths (β_3)	0.2000	0.1573	0.2436
pets_allowed (β_4)	0.0206	-0.0169	0.0577
σ^2	0.0196	0.0093	0.0372
τ^2	0.0329	0.0279	0.0384
ϕ	0.0282	0.0165	0.0465

Table 2: Posterior means and 95% credible intervals for model parameters.

Table 2 reports the posterior summaries from the Bayesian spatial Gaussian process model fitted to log rental prices. The posterior mean of the intercept is 7.0365, with a 95% credible interval $(6.9538, 7.1195)$. The number of bedrooms has a positive posterior effect with posterior mean 0.1149 and 95% credible interval $(0.0756, 0.1469)$. In terms of the actual rent, while holding bathrooms, pet allowance, and the spatial effect fixed, our model suggests that an additional bedroom is associated with an approximately

12.2% increase in rent, with a 95% credible interval of approximately $(7.9\%, 15.8\%)$. Similarly, the number of bathrooms has a positive and relatively larger effect, with posterior mean 0.2000 and 95% credible interval $(0.1573, 0.2436)$. In terms of the actual rent, this corresponds to an estimated 22.1% increase in rent for one additional bathroom with a 95% credible interval of approximately $(17.0\%, 27.6\%)$, while holding the other covariates and the spatial effect fixed.

Both bedroom and bathroom effects are clearly positive because their credible intervals do not contain zero. On the other hand, the posterior mean of the pets coefficient is 0.0206, with 95% credible interval $(-0.0169, 0.0577)$. On the original rent scale, this corresponds to an estimated 2.1% increase in rent for listings that allow pets, with a 95% credible interval of approximately $(-1.7\%, 5.9\%)$, while holding the other covariates and the spatial effect fixed. Since this interval contains zero, we conclude that there is not strong posterior evidence that pet allowance has a clear effect on rent.

3.2 Spatial Prediction

After obtaining the estimates for $\boldsymbol{\theta}$, we move on to the prediction of the response $Y(\boldsymbol{s}_0)$ at a new value $\boldsymbol{x}_0$ with spatial coordinate $\boldsymbol{s}_0$. Let $\boldsymbol{X}$ denote the design matrix, then the joint distribution of $(\boldsymbol{y}, y_0)$ given the parameter vector $\boldsymbol{\theta}$ is multivariate normal:

$$\begin{pmatrix} \boldsymbol{y} \\ y_0 \end{pmatrix} \Bigg| \theta \sim MVN\left[\begin{pmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{x}_0^\top \boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{10} \\ \boldsymbol{\Sigma}_{01} & \boldsymbol{\Sigma}_{00} \end{pmatrix}\right],$$

where $\boldsymbol{\Sigma}_{11}$ is the covariance matrix among training locations, $\boldsymbol{\Sigma}_{10}$ is the covariance matrix between training locations and y_0, $\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10}^\top$ and $\boldsymbol{\Sigma}_{00} = \mathrm{Var}(y_0)$. Therefore, by the conditional distribution formula [And03] for the multivariate normal distribution,

$$y_0 \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{x}_0 \sim N(m_0, V_0),$$

where $m_0 = \boldsymbol{x}_0^\top \beta + \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, $V_0 = \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{10}$. Hence, the predictive distribution for $Y_0 = y_0$ is

$$p_n(y_0|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}_0) = \int f(y_0|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{x}_0) \cdot p_n(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) d\boldsymbol{\theta}. \tag{8}$$

In practice, MCMC methods may again be readily used to obtain estimates of Equation (8). Suppose we draw our posterior sample $\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(G)}$ from $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$, then the above integral may be computed as a Monte Carlo mixture of the form

$$p_n(y_0|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}_0) = \frac{1}{G}\sum_{g=1}^{G} p\Big(y_0|\boldsymbol{y}, \boldsymbol{\theta}^{(g)}, \boldsymbol{x}_0\Big). \tag{9}$$

We typically use composition sampling to draw one for one for each $\boldsymbol{\theta}^{(g)}$, that is, $y_0^{(g)} \sim p(y_0|\boldsymbol{y}, \boldsymbol{\theta}^{(g)}, \boldsymbol{x}_0)$. Then the collection $\{y_0^{(1)}, \cdots, y_0^{(G)}\}$ is a sample from the posterior predictive density, and so on can be fed into a histogram to obtain an approximate plot of the density and hence bypassing Equation (8). Figure 5 shows the histogram for the predicted log price for a randomly chosen observation from the testing set.

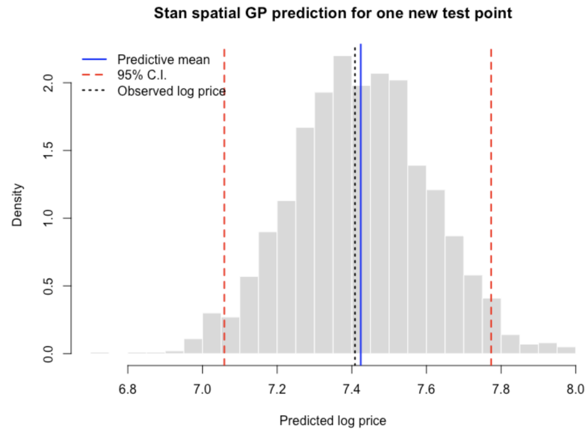

Figure 5: This figure shows histogram of the predicted log price. Using MCMC and the procedure described earlier, we obtain a sequence of posterior draws $y_0^{(1)}, \cdots, y_0^{(G)}$, and we report the predicted mean as the estimate for the log price for the observation.

For the prediction of the entire testing set, we should perform joint prediction via

$$p(\boldsymbol{y}_0|\boldsymbol{y},\boldsymbol{X},\boldsymbol{X}_0) = \int p(\boldsymbol{y}_0|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{X}_0)\cdot p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X})\boldsymbol{\theta} \approx \frac{1}{G}\sum_{g=1}^{G} p\Big(\boldsymbol{y}_0|\boldsymbol{y},\boldsymbol{\theta}^{(g)},\boldsymbol{X}_0\Big),$$

where $\boldsymbol{y}_0 = \begin{pmatrix} y_{01} & \cdots & y_{0M}\end{pmatrix}^\top$ is an unobserved vector with design matrix $\boldsymbol{X}_0$. As a result, we obtain a set of predicted log prices. Figure 6 shows the plot between observed log price and predicted log price with a 95% credible interval band, Figure 7 shows the plot between predicted log price and the residuals (in log-price scale).

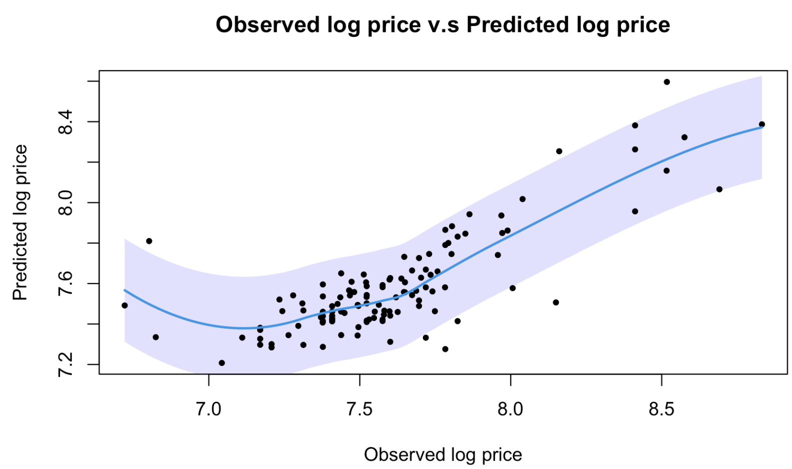

Figure 6: This figure shows the plot between observed log price and predicted log price on the testing set. A 95% credible interval is included in the plot.

Compare between Figure 1, Figure 6, and Figure 7, the plot under our new model shows tighter variation. This indicates that the spatial Gaussian process model captures the general trend in rental prices for unseen listings. The tail behavior is likely due to the lack of data on the tail. This suggests that although the model has good predictive ability, its power comes from sufficient count of data being available.

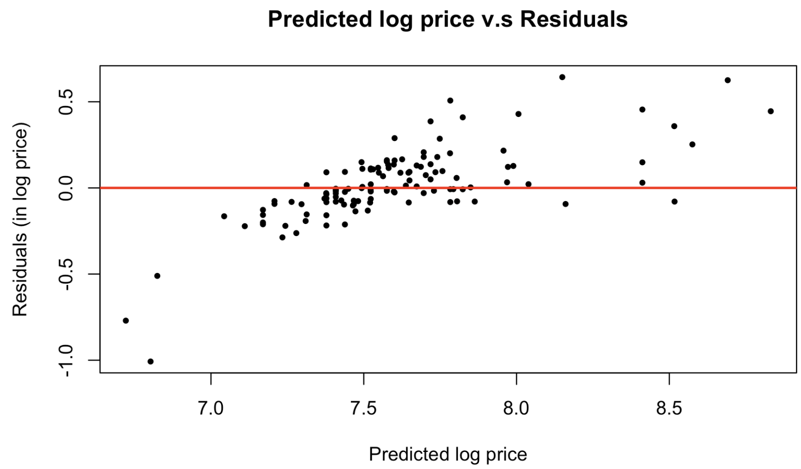

Figure 7: This figure shows the plot between predicted log price and residuals on the testing set.

The residual plot shows that most testing residuals are centered near zero, indicating reasonable predictive calibration for many listings. However, the residuals are not completely patternless. Some lower-priced listings are overpredicted, while several higher-priced listings have positive residuals, indicating underprediction. The residual spread also appears larger for higher predicted log prices. This suggests that the spatial GP model captures underlying spatial structure, but some systematic variation remains, possibly due to heteroskedasticity or unobserved listing characteristics such as building quality, exact location amenities, furnishing, or lease conditions.

3.3 Graphical Illustration

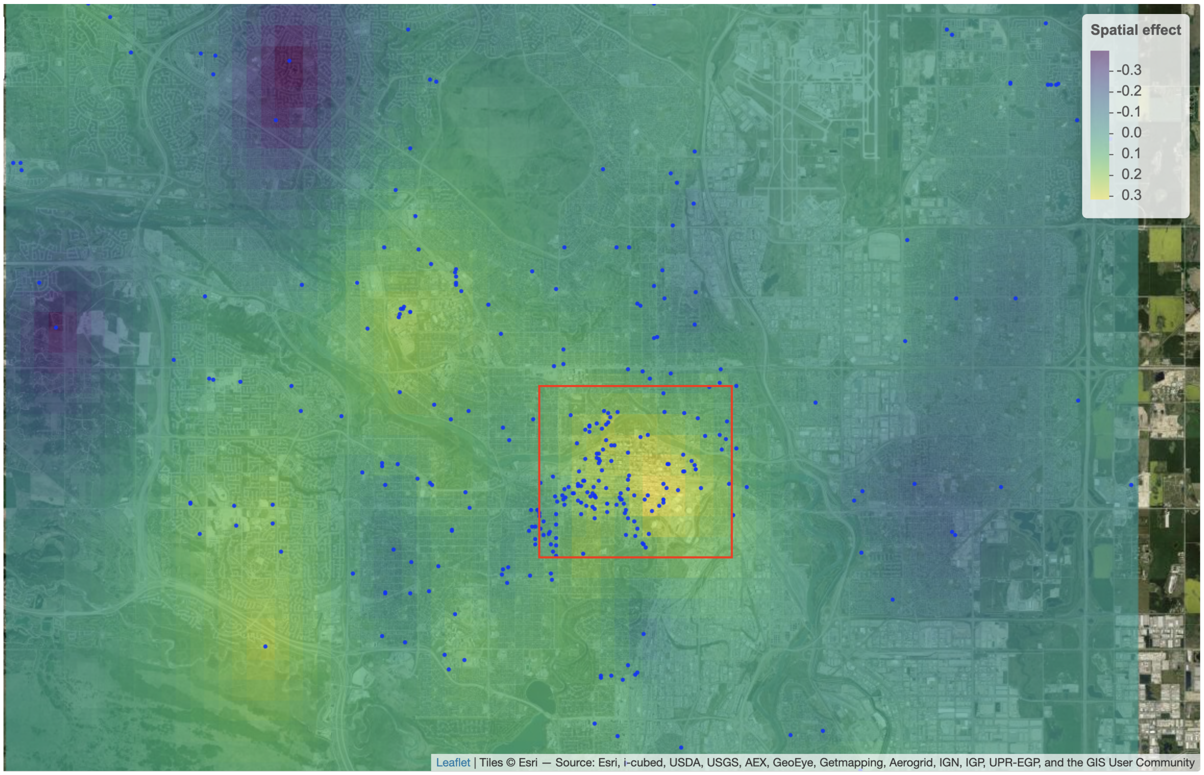

Figure 8: The spatial effect map based on our model.

The spatial effect map in Figure 8 provides a geographic interpretation of the residual spatial component $w(s)$, where the red rectangle outlines the downtown area. Positive values of $w(s)$ indicate areas where rental listings tend to be more expensive than predicted by the observed covariates, while negative values indicate areas where listings tend to be cheaper than predicted. In other words, the spatial effect captures location-specific variation not explained by covariates inside our model.

3.4 Conclusion and Discussion

In this report, we modeled the monthly rental price in Calgary using a linear predictor and a stationary isotropy Gaussian process. Our model suggested that the number of bedrooms and the number of bathrooms have positive effects, whereas an additional bedroom is associated with an approximately 12.2% increase in rent while holding other covariates and spatial effects fixed, and an additional bathroom is associated with an approximately 22.1% increase in rent while holding other covariates and spatial effects fixed.

We have also created a map that shows the spatial effects on monthly rental prices across Calgary. Several regions display positive spatial effects after controlling for the basic listing characteristics. This may be due to unobserved neighborhood-level factors such as proximity to transit, universities, parks, shopping, or newer residential developments. For example, Calgary rental submarkets such as Mission/-Beltline and Montgomery/Point McKay/University Heights have relatively high average rents in Canada Mortage and Housing Corporation (CMHC) rental data [MC], and amenity-rich areas such as Beltline and University District are characterized by walkability, services, green space, and access to urban amenities.

In contrast, regions with negative spatial effects correspond to areas where rents are lower than expected after accounting for the covariates. These negative residual spatial effects may reflect lower local demand, weaker access to amenities, older housing stock, or proximity to industrial and transportation corridors. Calgary contains major industrial planning areas in the Northeast, Nose Creek, Shepard, Southeast 68 Street, Southeast, and Stoney areas, and such surrounding land-use patterns may contribute to lower residential location premiums.

Overall, the spatial map supports the conclusion that location plays an important role in Calgary rental prices. The model captures not only the effects of unit-level features such as number of bedrooms and bathrooms, but also local spatial information that likely reflects neighborhood amenities, accessibility, development quality, and local market conditions.

4 Acknowledgment

I would like to express my sincere gratitude towards my mentor Peiyuan Huang for his supervision and guidance. I would also like to thank the DRP coordinators, the department of Mathematics and Statistics, and Professor Jessica Lin for organizing the DRP and create the opportunity for this experience.

References

[And03] Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3 edition, 2003.

[BCH07] Steven C. Bourassa, Eva Cantoni, and Martin Hoesli. Predicting house prices with spatial dependence: A comparison of alternative methods. *Journal of Real Estate Research*, 29(2):139–159, 2007.

[Bet18] Michael Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2018.

[BGC26] Sudipto Banerjee, Alen E. Gelfand, and Brandley P. Carlin. *Hierarchial Modeling and Analysis for Spatial Data*. CRC Press, 2026.

[Car] Bob Carpenter. *Stan: A Probabilistic Programming Language*. Journal of Statistical Software.

[CJ08] Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):209–226, 02 2008.

[CO81] A. D. Cliff and J. K. Ord. *Spatial Processes: Models and Applications*. 1981.

[DBFG16] Abhirup Datta, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand. On nearest-neighbor gaussian process models for massive spatial data. *WIREs Computational Statistics*, 8(5):162–171, 2016.

[Fra58] Bowman Frank. *Introduction to Bessel Functions*. 1958.

[GA15] Robert B. Gramacy and Daniel W. Apley. Local gaussian process approximation for large computer experiments. 2015.

[Gav24] Sergiy Gavrylov. *25000+ Canadian rental housing market June 2024*. Kaggle, https://www.kaggle.com/datasets/sergiygavrylov/25000-canadian-rental-housing-market-june-2024, 2024.

[GCS+25] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis (Third Edition)*. 2025.

[Geo63] Matheron Georges. Principles of geostatistics. 1963.

[GS90] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 1990.

[Hea19] Matthwe J. Heaton. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24:398–425, 2019.

[Mat86] Bertil Matern. *Spatial Variation*. 1986.

[MC] Canada Mortage and Housing Corporation. Calgary - average rent by bedroom type by neighborhood.

[Mor50] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[Osl10] Liv Osland. An application of spatial econometrics in relation to hedonic house price modeling. *Journal of Real Estate Research*, 32(3):289–320, 2010.

[Rob15] C.P. Robert. The metropolis-hastings algorithm. 2015.

[Ros74] Sherwin Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974.

[Tob70] W. R. Tobler. *A Computer Movie Simulating Urban Growth in the Detroit Region*. 1970.

[Zha04] Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99, 2004.