Modeling Rental Price in Calgary Under Bayesian Spatial Paradigms

Jiajun Zhang
Direct Reading Program

Department of Mathematics and Statistics, McGill University

2026.04.30

Introduction

In this work, we model the rental price y in the city of Calgary, based on the rental data from Kaggle.com, 2024.

- The **linear regression model** of the rental price y up to some transformation of g (if necessary) is given by

$$g(y) = \boldsymbol{x}^\top \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

  where $\boldsymbol{x}$ is the covariate vector, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2 \mathbb{I})$ is the i.i.d error.
- The original dataset contains 12 covariates. According to our exploratory data analysis, the covariates we consider are: β_1: The intercept; β_2: The number of bedrooms; β_3: The number of bathrooms; β_4 : Is pets allowed;

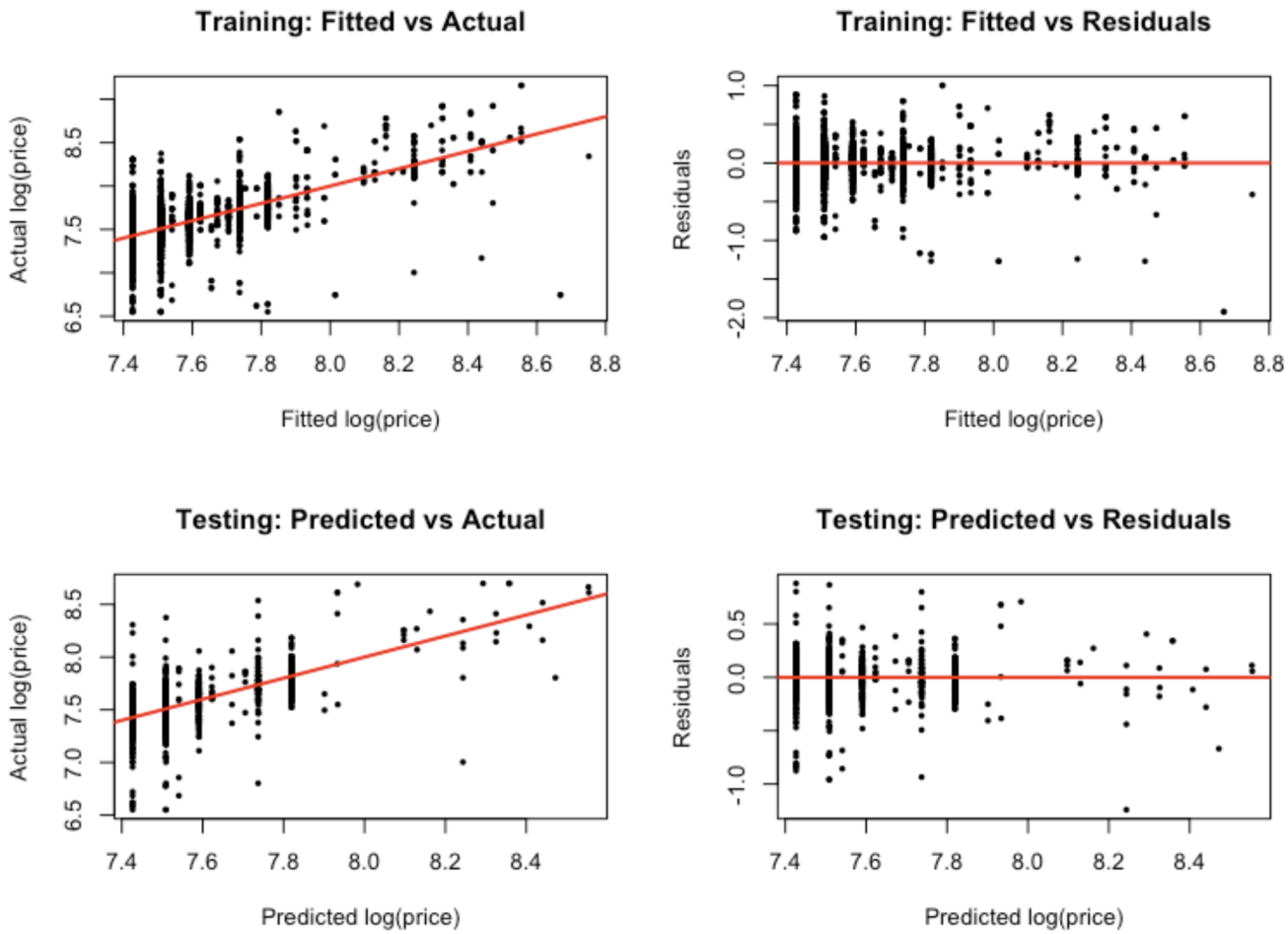

Figure: The performance of the baseline model on training and testing set.

Variogram

- A variogram γ is a diagnostic tool for detecting spatial dependence. For two locations $\boldsymbol{s}$ and $\boldsymbol{s}'$, it measures how different the residuals are as a function of their distance $d = ||\boldsymbol{s} - \boldsymbol{s}'||$.

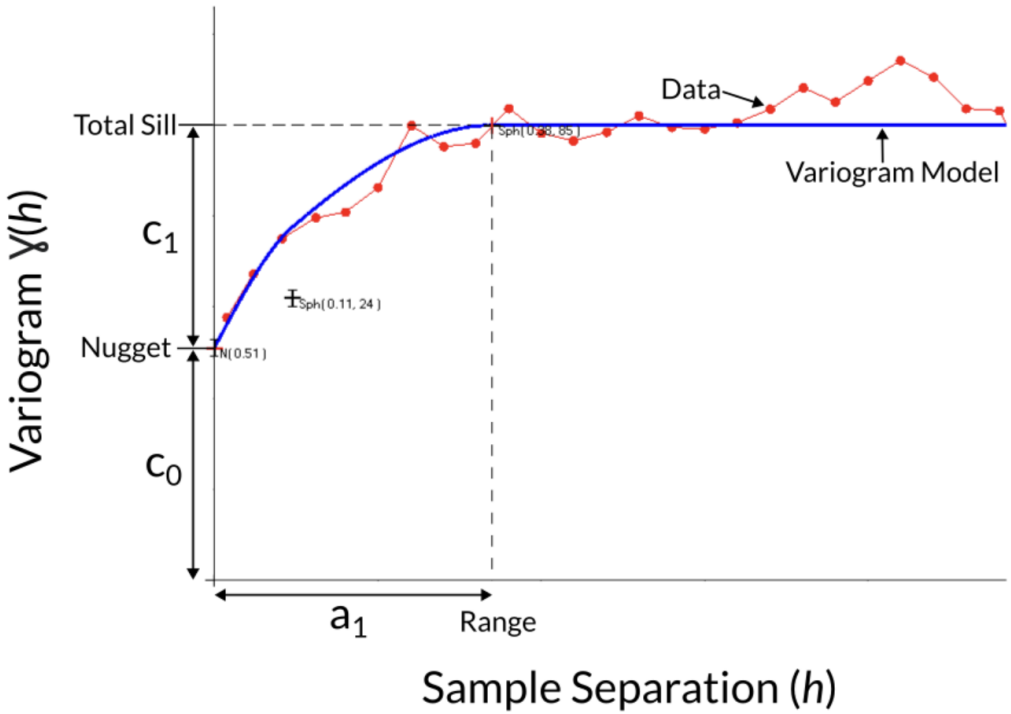

Figure: A theoretical variogram with three parameters: Nugget, Total sill, and range.

Empirical Variogram

In our model, the plot of the empirical variogram is given as below:

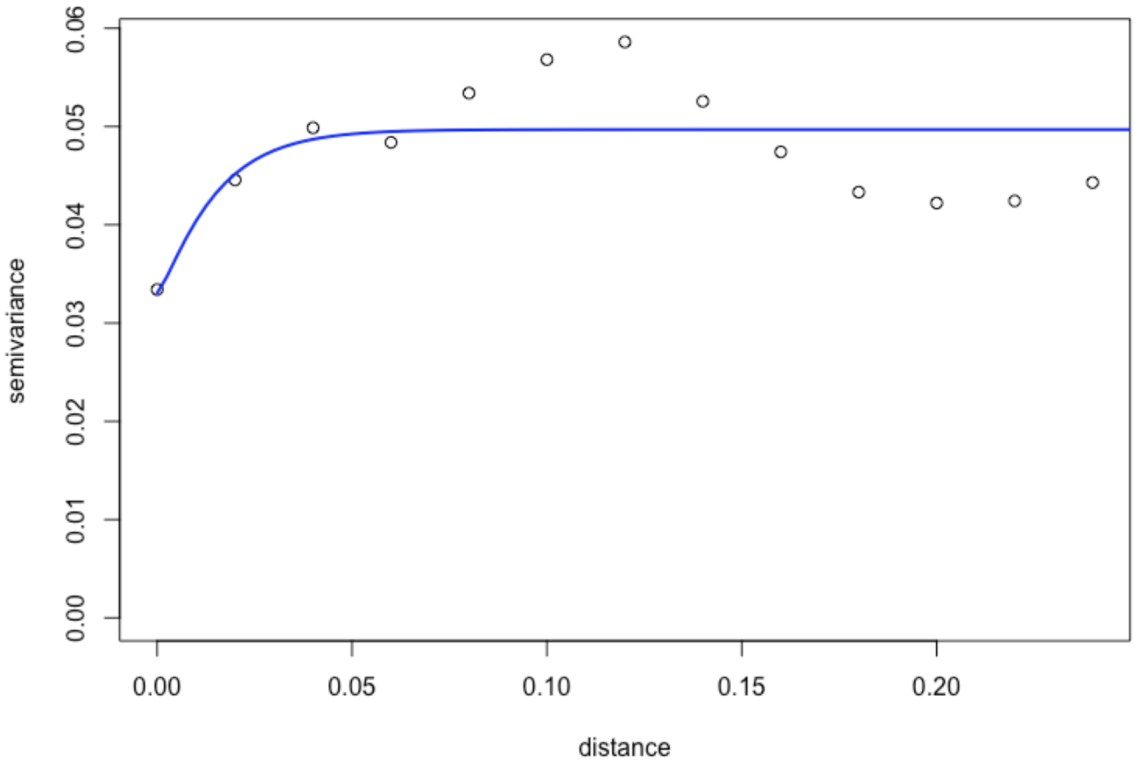

Figure: The empirical variogram of the linear regression model with smoothed curve.

Model Set-up

Instead, we consider the following model: For the observation $\boldsymbol{Y} = (Y(s_1), \cdots, Y(s_n))^\top$, we associate each y_i with its unique spatial coordinate s_i (longitude and latitude), denote by $Y(s_i)$, and define

$$\boldsymbol{Y}(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{r}(\boldsymbol{s}),$$

where

- $\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}$ is the **fixed effect** from predictor term;
- **The spatial dependency is modeled in $r(s)$**, which contains two terms:

$$\boldsymbol{r}(\boldsymbol{s}) = \boldsymbol{w}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}) :$$

  - $\boldsymbol{w}(\boldsymbol{s})$ is a zero-centered stationary Gaussian processes [Banerjee et al., 2026, Chapter 3] capturing **spatial error**, i.e., the spatial dependency that cannot be explained by the covariates $\boldsymbol{x}$.
  - $\boldsymbol{\epsilon}(\boldsymbol{s})$ is the **uncorrelated pure error** term, where $\boldsymbol{\epsilon}_i \overset{\text{i.i.d}}{\sim} N(0, \tau^2)$. The parameter τ^2 is called the **nugget**.

Spatial Gaussian Process

The (zero-centered) spatial Gaussian process $w(s)$ is an **infinite dimensional** stochastic process with the following properties:

- **Mean zero**: $\forall \boldsymbol{s} : \mathbb{E}[w(\boldsymbol{s})] = 0$;
- **Covariance**: The core idea of covariance function of $w(\boldsymbol{s})$ is that it is **monotonically decreasing**, modelling the idea that the further the distance, the less the spatial correlation. Using the Matern model [Matern, 1986], we have

$$\text{Cov}(w(\boldsymbol{s}), w(\boldsymbol{s}')) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\phi d)^\nu K_\nu(\phi d) & \text{if } d > 0 \\ \sigma^2 & \text{if } d = 0 \end{cases},$$

where $d = ||\boldsymbol{s} - \boldsymbol{s}'||$.

- K_ν is the modified Bessel function of order ν, where ν controls the smoothness of the realized random field [Banerjee et al., 2026]. In our model we fix $\nu = 1.5$;
- ϕ is the spatial decay parameter;
- σ^2 is the spatial variance parameter.

Model Set Up

- Combine with the nugget $\epsilon(\boldsymbol{s})$, and $\nu = 1.5$, the finite dimensional representation of $\boldsymbol{r}(\boldsymbol{s})$ follows a multivariate Gaussian:

$$\boldsymbol{r}(\boldsymbol{s}) \sim \mathcal{N}\Big(\mathbf{0}_n, \boldsymbol{\Sigma}_{n\times n}(\sigma^2, \tau^2, \phi)\Big);$$

- The entries of the variance-covariance matrix are defined using the Matern covariogram $[\boldsymbol{\Sigma}_{n\times n}(\sigma^2, \tau^2, \phi)]_{ij} = \text{Cov}(r(\boldsymbol{s}_i), r(\boldsymbol{s}_j))$, where

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

with $d = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$.

The Bayesian Approach

- Now, the parameter of interest is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi)^\top$. With the assumption we assumed, we have

$$\boldsymbol{y}|\boldsymbol{\theta} \sim \mathcal{N}\Big(\boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}(\sigma^2, \tau^2, \phi)\Big);$$

- By conditional probability, the posterior distribution $p(\boldsymbol{\theta}; y)$ can be written as

$$p(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{p(\boldsymbol{\theta} \cap \boldsymbol{y})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta});$$

- Since we already know the likelihood term of $\boldsymbol{y}|\boldsymbol{\theta}$, and we only need to impose a prior on $\boldsymbol{\theta}$.

Selecting the Priors

Under the Bayesian approach, our model is now

$$p(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi | \boldsymbol{y}) \propto p(\boldsymbol{y} | \boldsymbol{\beta}, \sigma^2, \tau^2, \phi) \cdot p(\boldsymbol{\beta} \cdot p(\sigma^2) \cdot p(\tau^2) \cdot p(\phi). \quad (2)$$

In our model, the priors are selected as follow:

- For the linear regression coefficient $\boldsymbol{\beta}$, we have $\beta_j \sim N(0, 100)$ for each $j = 1, 2, 3, 4$;
- σ^2, τ^2, ϕ are inverse Gamma priors given by
  - $\sigma^2 \sim \text{InvGamma}(2, 0.03712)$;
  - $\tau^2 \sim \text{InvGamma}(2, 0.0092806)$;
  - $\phi \sim \text{InvGamma}(2, 0.129098)$.
- These are weakly informative, data-scaled priors. The constants are chosen using the baseline residual variance and an initial empirical variogram estimate, so that the prior scale is consistent with the observed variation in the residuals [Gelman, 2006].

Learning the Parameters

The posterior distribution of parameters of interest in (2) have no closed form, hence we use a computational approach to approximate the posterior distribution.

- We use **Hamiltonian Monte Carlo** [Betancourt, 2018] (an advanced Markov chain Monte Carlo method) to draw samples from the posterior and simulate its distribution. The software we use is **Stan** [Carpenter, n.d.]. After running the program (2 chains, $1,500$ total iterations (including 500 burn-in) on a training set of size 480), we obtain a sequence of posterior draws:

$$\left\{ (\boldsymbol{\beta}_{(t)}, \sigma^2_{(t)}, \tau^2_{(t)}, \phi_{(t)}) \right\}_{t=1}^{M},$$

- Under the L^2 loss, we report the posterior mean from the draws:

$$\hat{\boldsymbol{\theta}} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{\theta}_{(i)}. \tag{3}$$

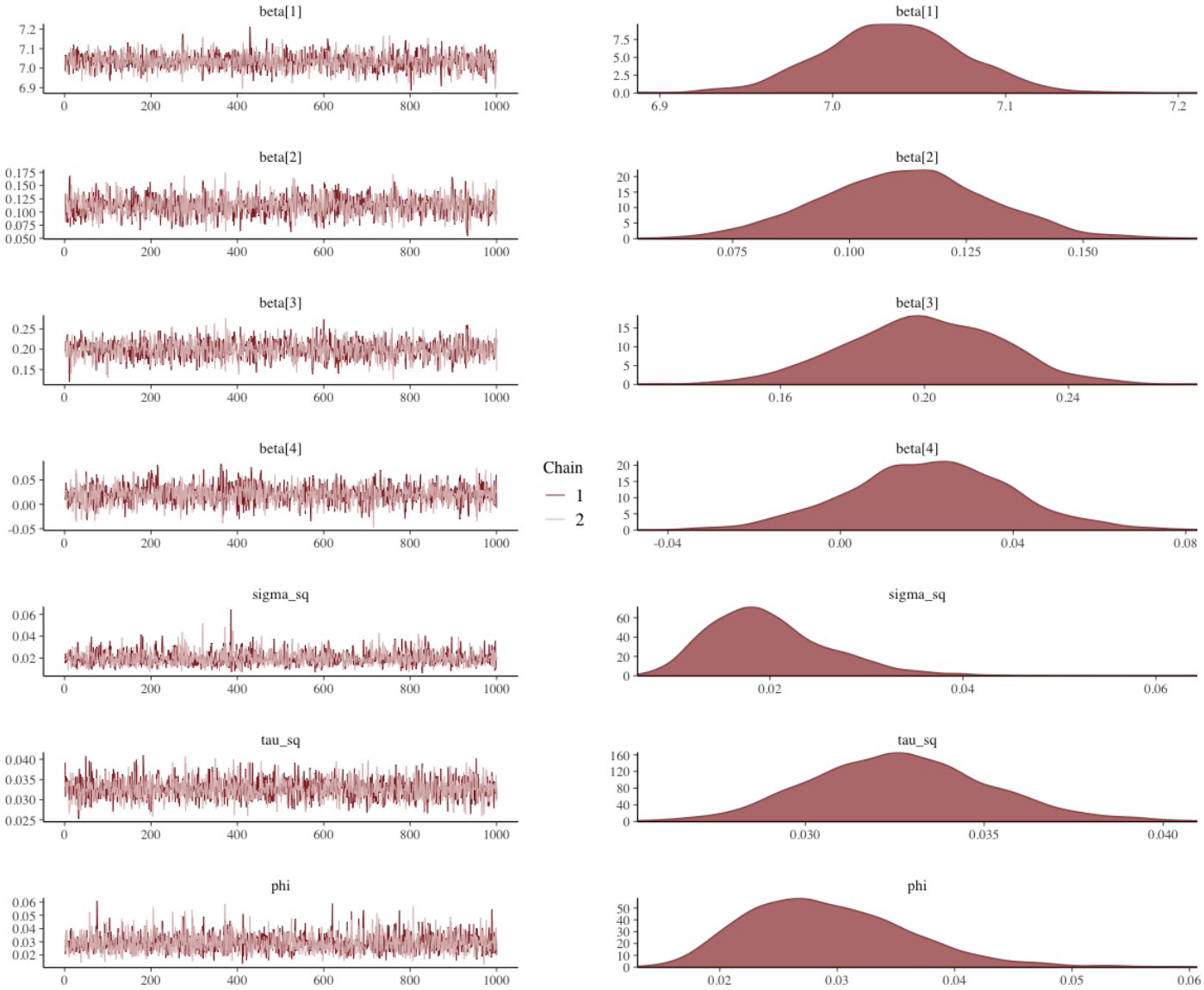

Figure: The MCMC trace plot

Spatial Prediction

After fitting the model to the training set, we now use this model to make spatial prediction on a new set of data.

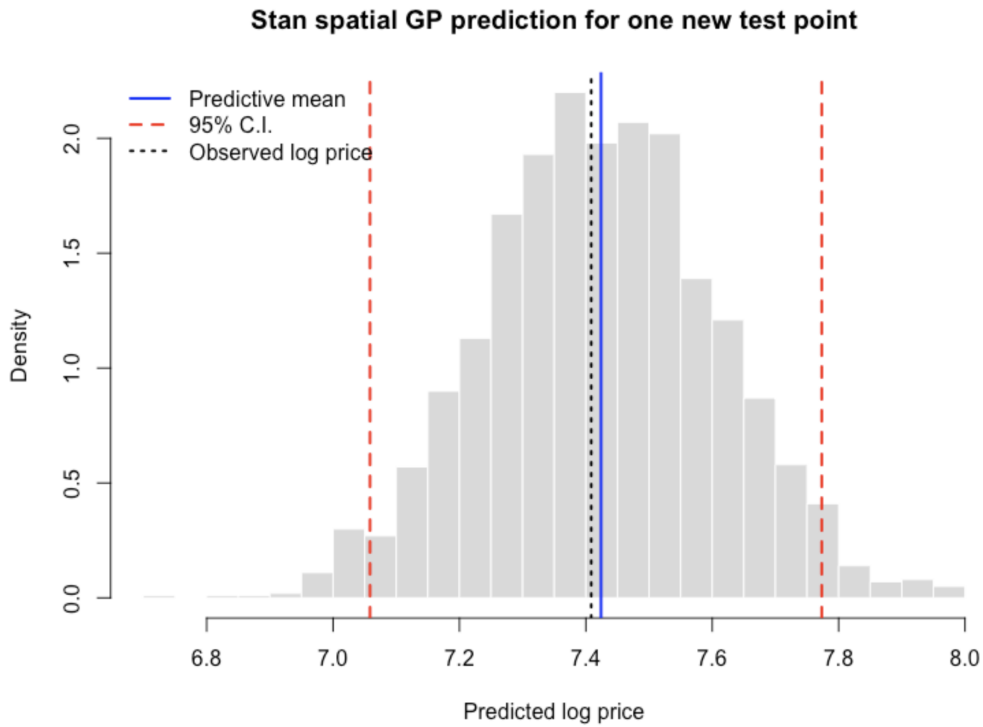

Figure: The histogram of posterior predictive log price at one new data point. We report the posterior mean by the same formula in (3).

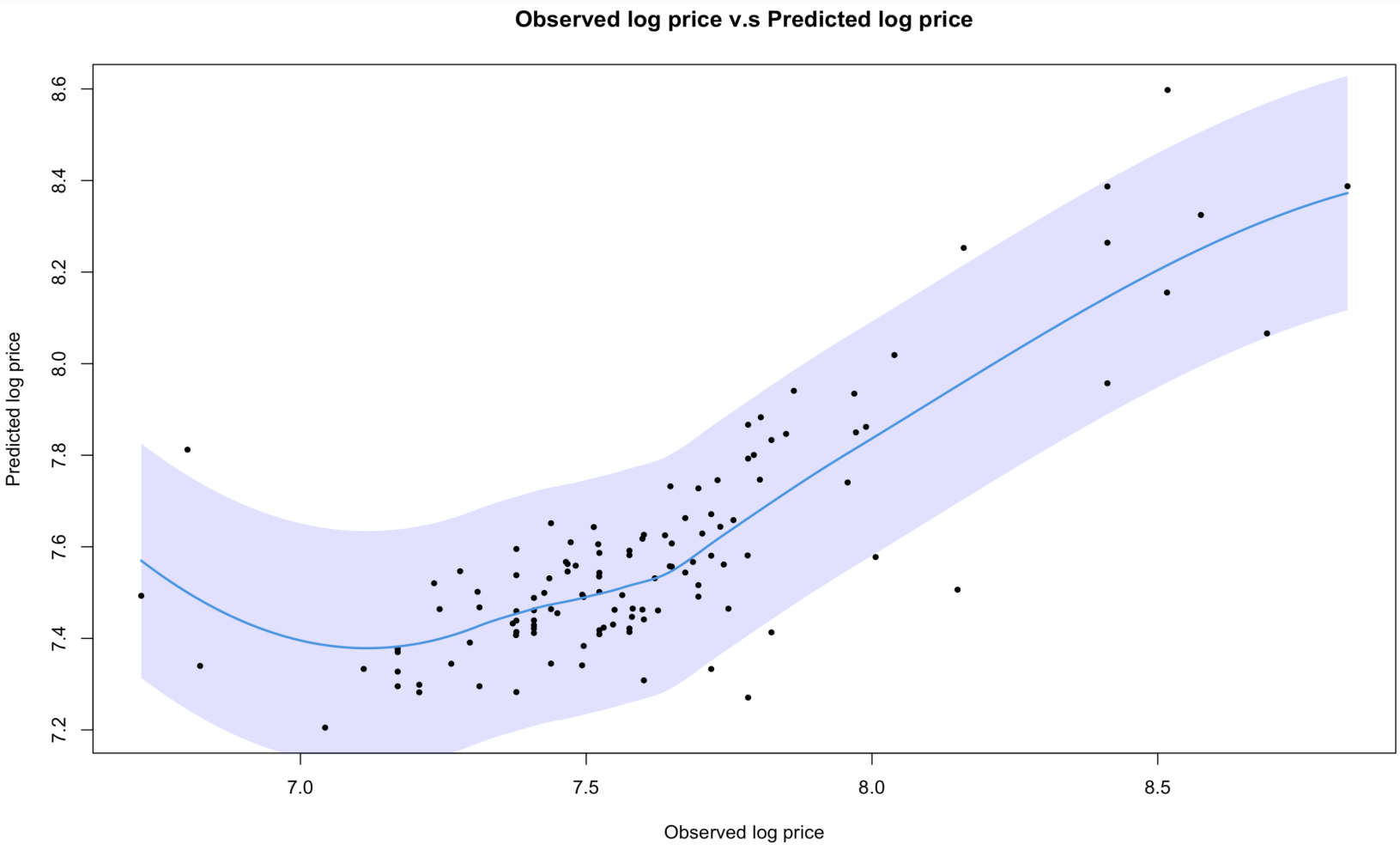

Figure: The observed log price v.s predicted log price on the testing set containing 120 data points. The $2.5\% - 97.5\%$ creditable interval is included in the plot.

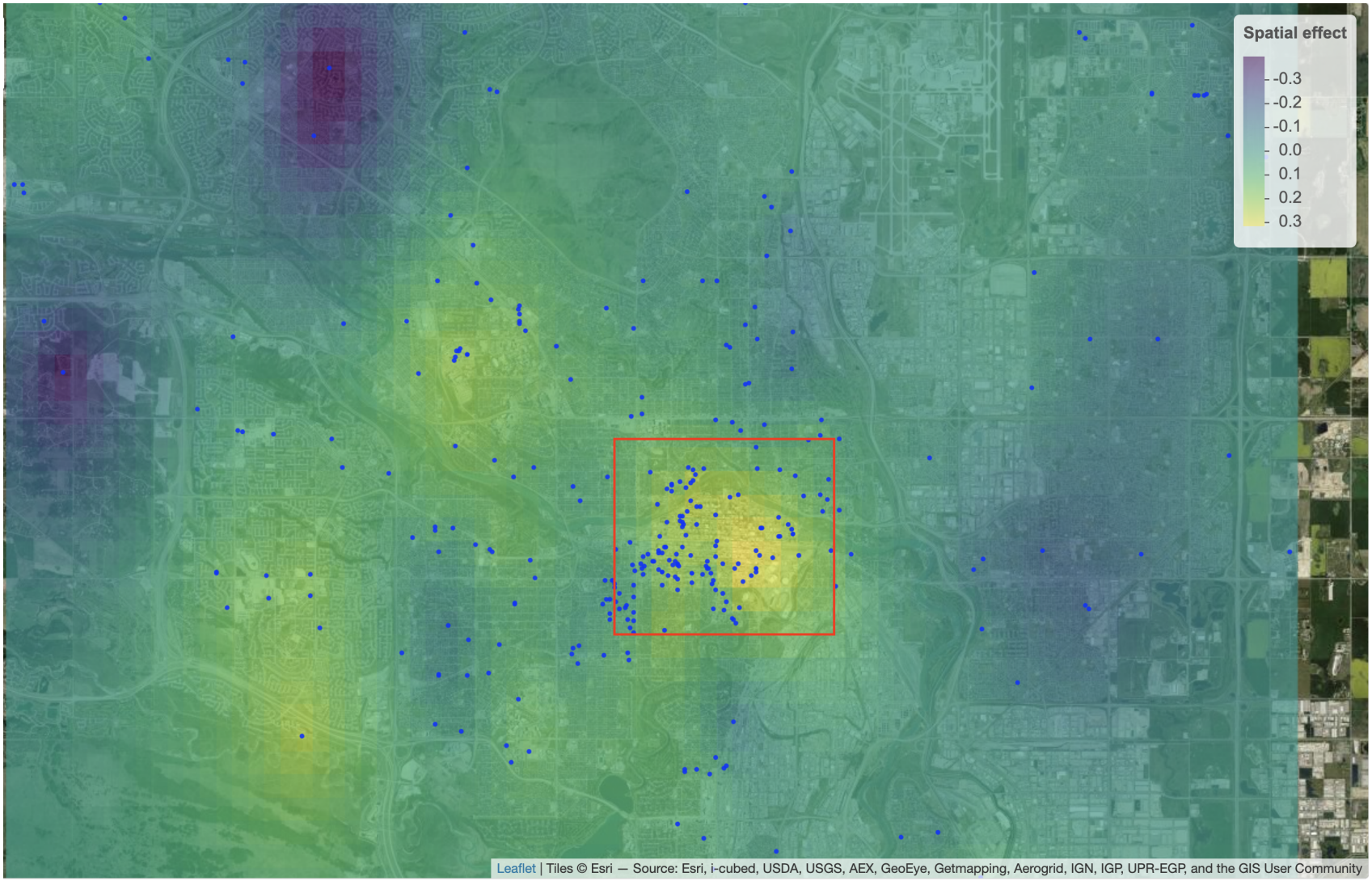

Figure: The spatial prediction plot based on posterior mean. The red rectangle indicates the downtown area.

Discussions

Q1: Why we use $\log(\text{price})$ instead?
A1: $\log(\text{price})$ makes the standard deviation lower. From our calculation, the log-transformed data has a SD around $50\$ - 120\$$; However without the log-transform the SD is around $175\$ - 300\$$. Also, using a log transformation is quite standard in retail research [Rosen, 1974].

Q2: How many data we used to fit our model?
A2: Due to huge computational complexity (as we need to invert the variance-covariance matrix after each iteration), we are only able to select 600 distinct spatial data, and use a $80\%/20\%$ split on the training and testing process. Fitting larger dataset requires some advanced methods, see Heaton, 2018 for more detail.

Next Steps

- Due to the computational complexity and the time constraint, we are only able to train our model based on 480 training data and 120 testing data. The next step will be using advanced methods (see Heaton, 2018) to fit the model on a larger data set.
- Another next step is to consider the influence of missing data. In our dataset, many observations have missing components and we simply discarded those data in the data cleaning process.

Banerjee, S., Gelfand, A. E., & Carlin, B. P. (2026). *Hierarchial modeling and analysis for spatial data*. CRC Press.

Betancourt, M. (2018). *A conceptual introduction to hamiltonian monte carlo*.

Carpenter, B. (n.d.). *Stan: A probabilistic programming language*. Journal of Statistical Software.

Gelman, A. (2006).Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Heaton, M. J. (2018).A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, *24*, 398–425.

Kaggle.com. (2024). *25000+ canadian rental housing market june 2024*.

Matern, B. (1986). *Spatial variation*.

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