Spatial Statistics Project 1

Wanru Cheng & Kuan Teh Wan

November 2024

1 Introduction

This assignment focuses on the reconstruction of spatial patterns in housing prices within the city of Utrecht, Netherlands. The primary goal is to explore and compare two classic methods: **Ordinary least squares (OLS) regression** and **Universal Kriging**.

OLS regression is a statistical method used to model the relationship between a dependent variable (housing price) and independent variables (covariates like lot area, number of balconies, and location). By fitting a linear regression model, we can estimate the impact of these covariates on housing prices.

Universal Kriging is a geostatistical technique that leverages spatial autocorrelation to predict values at unobserved locations. It considers the spatial structure of the data and estimates the uncertainty associated with the predictions.

By applying these two methods, we aim to:

- **Reconstruct the spatial component of housing prices:** Map the spatial distribution of housing prices across Utrecht.
- **Identify key factors influencing housing prices:** Determine the relative importance of covariates like lot area, number of balconies, and location in explaining price variations.
- **Compare the performance of OLS and Kriging:** Evaluate the strengths and weaknesses of each method in capturing the spatial patterns of housing prices.

Through this analysis, we can gain valuable insights into the factors driving housing prices in Utrecht.

2 OLS Regression

First, we have to determine which covariates would be suitable for the Regression model to reconstruct the housing prices. To do that, we start by inspecting and comparing the plots of observations based on different covariates, trying to identify any relationship between them. The plot is shown below:

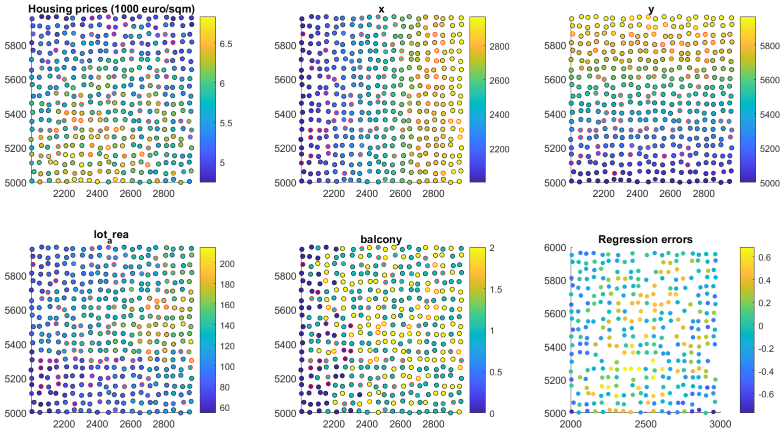


Figure 1: Observations based on different covariates

Then, we found that there's a positive correlation between *lot_area* and the housing price. Since observations where have larger lots also tend to have higher housing prices according to the plot. Secondly, it also matches our intuition that, larger lot offers more space, privacy and also potential expansion, which is valuable to buyers so it makes sense to have higher prices. With these two reasons, we conclude that *lot_area* will be important for modeling prices.

Next, inspecting the *balcony* plot, we can definitely see that there are some spatial variation. However, there's no clear positive correlation like we saw with *lot_area*. Moreover, the number of balconies of a property could only be a local preference or depend entirely on climate, which doesn't really connect directly to the price. Thus, we decided to discard this covariate for the Regression model.

With that being said, we have OLS Regression model:

$$price = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 lot\_area$$

As well as the estimates of the beta coefficients and confidence intervals:

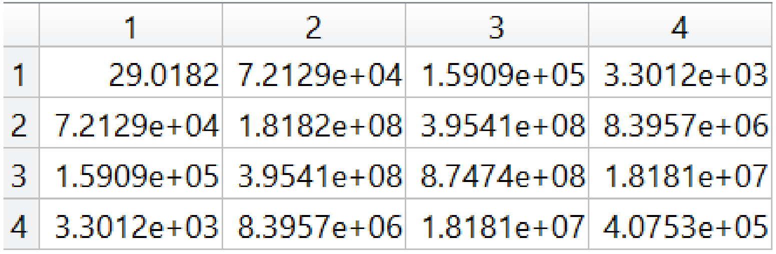
	Estimates	C.I
β_0	13.8572	(13.0703, 14.6441)
β_1	-0.00004	(-0.0008, -0.0003)
β_2	-0.0014	(-0.0015, -0.0012)
β_3	0.0028	(0.0014, 0.0043)

Table 1: OLS Regression Model beta estimates and confidence interval

with the estimated error variance

$$\sigma_\epsilon^2 = 0.0910$$

and the uncertainty in β as shown:



	1	2	3	4
1	29.0182	7.2129e+04	1.5909e+05	3.3012e+03
2	7.2129e+04	1.8182e+08	3.9541e+08	8.3957e+06
3	1.5909e+05	3.9541e+08	8.7474e+08	1.8181e+07
4	3.3012e+03	8.3957e+06	1.8181e+07	4.0753e+05

Based on this model, the result of the predictions against the validation data is presented below. We can see that the linear relationship between the predicted and the validation data is almost at 45 degree which indicates good prediction.

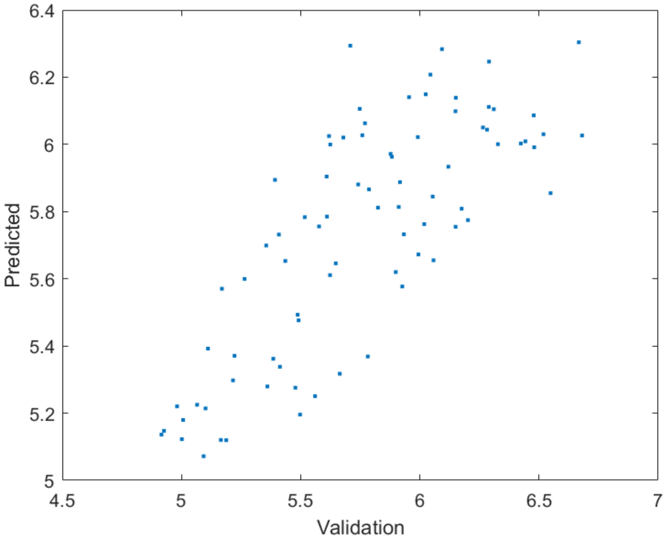


Figure 2: Result of predictions against validation data

Then, here we take a look at how the predictions spatial effect compare to the validation points with correct values.

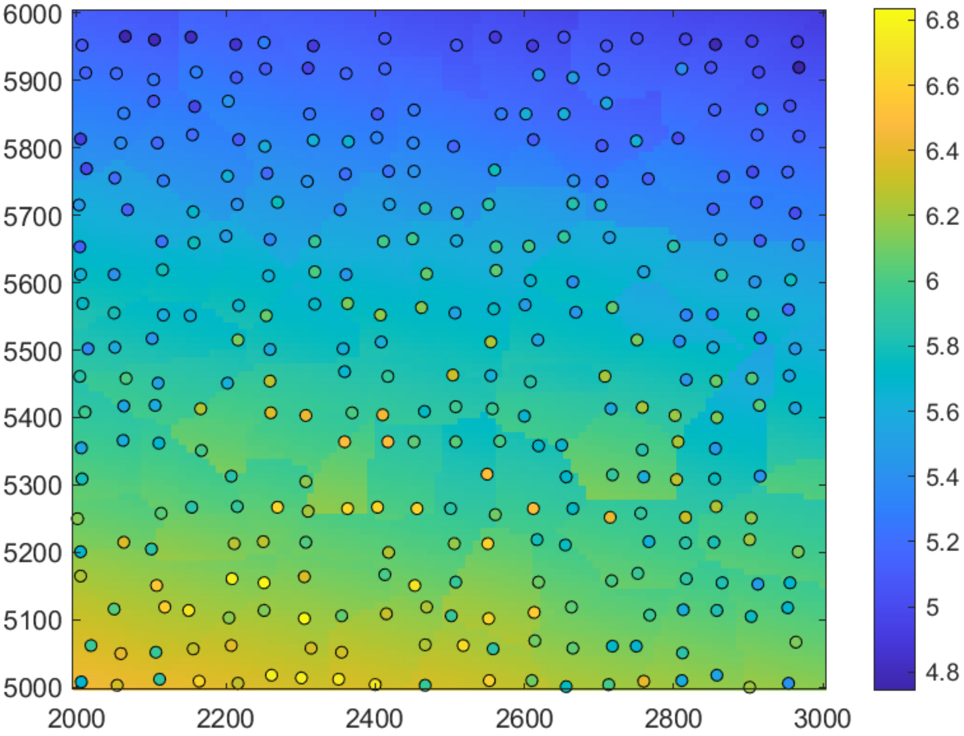
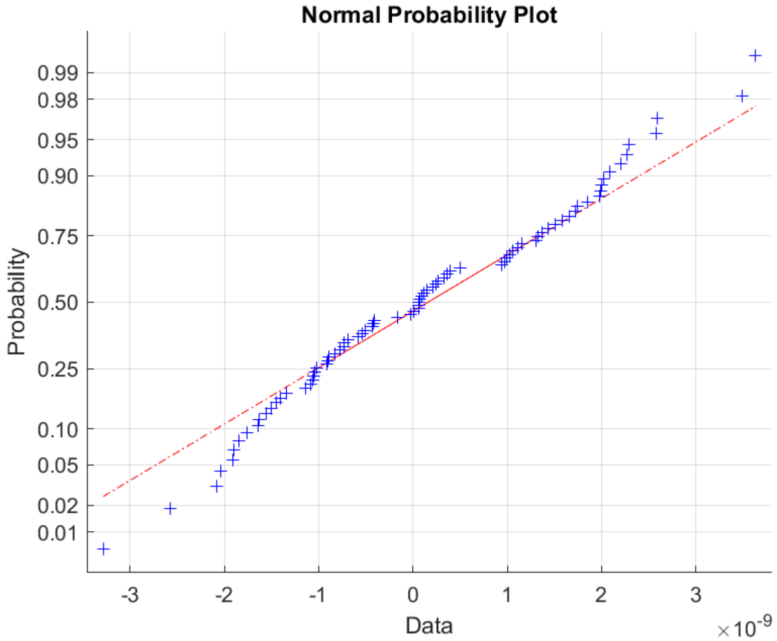


Figure 3: Predictions spatial effect/Validation data

As we can see from the plot, it does capture the pattern to some degree of the validation data. However, it still lacks the ability to capture the spatial effect of the validation data in the middle.

Lastly, the OLS regression model comes with a MSE of 0.0810 and with the standardized residual shown as below, which follows a normal distribution.



3 Universal Kriging

In this section, we first need to analyzed the spatial dependence in the residuals of the OLS model we obtained in the last section. As we can recall, we eliminated balcony as a covariate and obtained the residuals from

$$resi\_OLS = Y - X\_OLS * beta$$

.

Specifically,

$$X\_OLS \;=\; X\_OLS = [ones(size(X, 1), 1), x, y, lot\_area]$$

. Then, we used non-parametric approach to estimate the covariance, and the bootstrap approach with 100 permutations to obtain the result. It is presented in figure 4.The observed binned covariance (blue bar) lies outside the 95% confidence interval at distance 0200,400700, which suggests that there is significant spatial dependence at those distance ranges. On the other hand, the spatial dependence is not significant.

Next, we estimated all 5 covariance functions using covest_ml.m, the parameter lists are in table 2

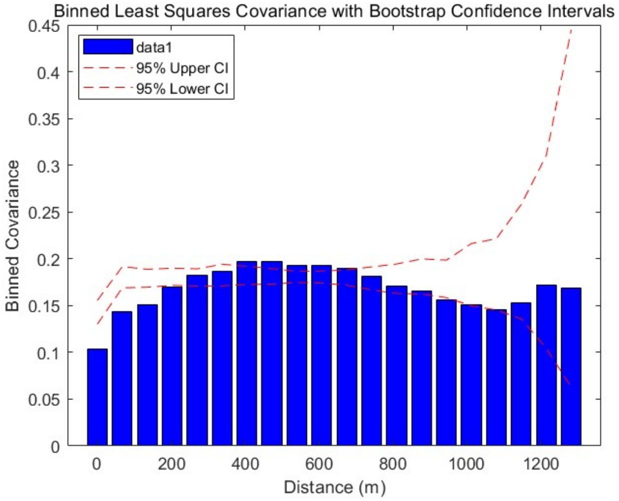


Figure 4: Binned Least Square covariance with confidence intervals

Covariance functions	Params
Matern	sigma2=8.1917e-27,kappa=9.5861e-04,nu=0.9992

Table 2: Parameter list for 5 covariance functions

Appendix A: Theory

Q1:

Covariance matrix C for the three points:

$$C = \sigma^2 * \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

The diagonal elements are r(0) = σ^2, representing the variance of each point. The off-diagonal elements are r(h) = $\rho\sigma^2$, representing the covariance between points separated by distance h.

Q2:

Using the property of a covariance matrix, it must be positive semi-definite. This means that all its eigenvalues must be non-negative.

Then the eigenvalues are the solutions to the characteristic equation:

$$det(C - \lambda I) = det(\sigma^2 * \begin{bmatrix} 1-\sigma & \rho & \rho \\ \rho & 1-\sigma & \rho \\ \rho & \rho & 1-\sigma \end{bmatrix}) = 0$$

Solving this we have the three eigenvalues:

$$\lambda_1 = \sigma^2(1 + 2\rho)$$

$$\lambda_2 = \lambda_3 = \sigma^2(1 - \rho)$$

Thus, for the matrix to be positive semi-definite, all eigenvalues must be non-negative:

$$\lambda_1 \geq 0$$

$$\lambda_2 = \lambda_3 \geq 0$$

And then we have

$$1 + 2\rho \geq 0$$

$$1 - \rho \geq 0$$

And by solving the equalities we have the limit on ρ:

$$-\frac{1}{2} \leq \rho \leq 1$$

Q3:

Similarly, the Covariance matrix for the 4 points in a regular tetrahedron is:

$$C = \sigma^2 * \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

As before the covariance matrix must be positive semi-definite. Which again, means all eigenvalues must be non negative.

Solving the characteristic equation $det(C - \lambda I)$ will give us the eigenvalues. However, this is more complex than the 2D 3D case.

Thus, we'll be using the properties of symmetric matrices to simplify the process. Using "Gershgorin Circle Theorem", it states that each eigenvalue of a matrix lies within at least one of the Gershgorin disks. For our covariance matrix, the Gershgorin disks are centered at σ^2 with radius $3\sigma^2|\rho|$. To ensure all eigenvalues are non-negative, we need the disks to be entirely within the positive real axis. This condition gives us:

$$3\sigma^2|\rho| \leq \sigma^2$$

Therefore, the limits on ρ:

$$-\frac{1}{3} \leq \rho \leq \frac{1}{3}$$

Q4:

In higher dimensions, we can extend the concept of a regular tetrahedron to a regular simplex. For a regular simplex with n+1 vertices in n dimensions, the covariance matrix will be an (n+1)x(n+1) matrix with diagonal elements σ^2 and off-diagonal elements $\rho\sigma^2$.

Again, applying the Gershgorin Circle Theorem, we then have the limits:

$$-\frac{1}{n} \leq \rho \leq \frac{1}{n}$$