TU Dortmund

Introductory Case Studies

Project 2: Discrete covariates

Lecturer:
Dr. rer. nat. Maximilian Wechsung

Author: Kartik Kamboj

Group number: 1

Group members: Bidhan Chandra Roy, Hemani Chand, Jatin Rattan, Kartik Kamboj

May 27, 2022

Contents

1 Introduction

Renting has become far more common in Germany than buying a home, owing to the lower cost of renting. A newly constructed large apartment with a kitchen and bathroom, as well as easy access to public transportation, is always favored at a fair price. The location of an apartment is also one of the deciding factors in the rent, and it can be classified as top, good, or average. An apartment with the top location is always favored, and it is also the most expensive, compared to one with a good or average location.

The goal of this study is to examine the data collected in 1999 about rent index data for 3082 apartments in Munich. The data comprises variables like size of the living area, construction year, quality of location, bathroom and kitchen, and central heating. All of these characteristics have an impact on an apartment's net rent, but in this project, we are simply looking at the quality of location. Furthermore, we are using statistical hypothesis testing methods like Kruskal-Wallis in the global test to get the significance of location quality on rent per square meter and the Mann-Whitney-Wilcoxon method to get the category of location quality, which are significantly different.

The dataset's source and quality are discussed in the second section. There is also a brief review of the data in terms of statistical tasks and methodologies in this section. The statistical methods utilized to conduct the hypothesis testing of the data are described in the third part. In the fourth section, these statistical procedures are implemented, and the results are understood and analyzed. A nonparametric test is conducted, comprising elements such as location quality and observations such as the apartment's rent per square meter and also different quality of location are evaluated pairwise. Finally, a summary of the findings of the dataset analysis is provided, as well as a quick explanation of the report's outlook.

2 Problem statement

2.1 Data set and data quality

The following dataset is gathered and provided by the lecturers of Introductory Case Studies from the website of the University of Goettingen (Göttingen, 2022). The Munich rent index dataset is the same dataset that is used in the book, Regression: Methods, Models and Applications (Fahrmeir et al., 2013, p. 6). The dataset includes net rent of the apartment, size of the living room, construction year, quality of bathroom, kitchen, location and central heating for Munich, collected in 1999. We have four categorical variables (bathroom, kitchen, quality of location and central heating), two continuous variables (net rent and living area) and one discrete variable (construction year). The bathroom and kitchen variables are categorized into two classes i.e., 0 as standard quality and 1 as premium quality, whereas quality of location is divided into three categories: 1 = average location, 2 = good location and 3 = top location. The central heating of the apartment is categorized as 1 and 0 i.e., yes and no, respectively. The net rent is the monthly rent of the whole apartment in Euros and is estimated to be one decimal place. The living area is the size of the apartment in square meters and the construction year varies from 1918 to 1997. Furthermore, the data quality in the study is good as there are no missing values in the essentials variables like net rent, living area and quality of location of the dataset.

2.2 Project Objectives

The goal of this project is to conduct a global test to see if the quality of the location has any significant impact on the rent per square meter. Before implementing the statistical methods, we need to perform some preliminary tests to check if the assumptions are met. Then, we conduct a nonparametric (rank-based) method, Kruskal-Wallis, to examine the impact of location quality on rent per square meter. The next aim is to examine the pairwise differences in rent per square meter for various location characteristics using the Mann-Whitney-Wilcoxon test. Then, to reduce the multiple testing problem, the Bonferroni correction is utilized to adjust the results. Finally, the data is analyzed and compared with the adjustment and also without it.

3 Statistical methods

This section explains the statistical methods that is used to analyze the data collection. All the following graphs were generated using the software R (R Core Team, 2020), version 4.0.5. The R packages used in the project for the graph plots are dplyr (Wickham et al., 2021), ggplot2 (Wickham, 2016) and gridExtra (Auguie, 2017).

3.1 Variance and Standard deviation

3.1.1 Variance

The variance measures how far a set of numbers deviates from their mean value. It is determined as the average of the sum of the squared deviations from the mean of a set of numbers and measured in squared units of the data. The arithmetic mean is denoted by $\bar{x}$, x_i is the i^{th} observation and n is the total number of observations if the data set contains the values x_1, x_2, x_3, ..., x_n. The variance (s^2) is determined as follows (Black, 2006, p. 59):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

3.1.2 Standard deviation

It also measures the dispersion of the data and is the square root of the variance. The standard deviation is measured in the same units as the data. If the data set contains the values x_1, x_2, x_3, ..., x_n, the arithmetic mean is denoted by $\bar{x}$, where x_i is the value of the i^{th} observation and n is the total number of observations. The following formula is used to determine standard deviation (s) (Black, 2006, p. 60):

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

3.2 Statistical Hypothesis testing

Data is at the heart of statistics. There is a massive amount of data, which can only be used for analysis or drawing conclusions. Hypothesis testing is used to determine such an interpretation and a conclusion. In general, hypothesis testing compares two mutually exclusive statements on a population of data using a sample of data and decides whether the population is true for one or the other (Wasserman, 2010, p. 149).

3.2.1 Null hypothesis and Alternate Hypothesis

A hypothesis test consists of two hypotheses: the null hypothesis (H_0) and the alternate hypothesis (H_1). The null hypothesis is the prevailing belief of the population. It states that there is no change in certain distribution functions of the population. An alternate hypothesis, on the other hand, is a claim that contradicts the null hypothesis and causes a difference in at least one distribution function. We do not reject the null hypothesis if we can show that the initial assumption is correct, and we only reject it if the initial assumption cannot be proven owing to a lack of evidence. Suppose we want to compare two distributions Θ_0 and Θ_1 of a sample space Θ, then null and alternate hypotheses are (Wasserman, 2010, p. 149-150):

$$H_0 : \Theta_0 = \Theta_1$$

$$H_1 : \Theta_0 \neq \Theta_1$$

3.2.2 Significance level

The significance level, denoted by α, is the chosen cutoff point that determines whether there is evidence to reject the null hypothesis. The significance level for a study is determined prior to data collection and is usually set at 5% (or 0.05) or lower, depending on the subject of study and the importance of the data (Wasserman, 2010, p. 155-156).

3.2.3 Test statistics

It is a numerical value generated from a sample of data using central tendency, variance, sample size and predictive factors under the null hypothesis. Depending on the null hypothesis's probability model, different hypothesis tests use different test statistics. For example, Chi-squared statistics is used in Kruskal-Wallis, F statistic is used in ANOVA, and the U-statistics is used in the Mann-Whitney-Wilcoxon test, which are addressed later in the report (Fahrmeir et al., 2013, p. 128).

3.2.4 critical value and p-value

The value chosen under the null hypothesis so that the chance of incorrectly rejecting the null hypothesis is less than the significance threshold (0.05) is known as the critical value. In general, we determine the critical value from the test statistics distribution table by using degrees of freedom (number of independent samples - 1) at a specific significance level. When the test statistical value surpasses the critical value, we usually reject the null hypothesis.

$$R = \{x : T(x) > c\}$$

Here, $T(x)$ is a test statistic function of an observation x, c is a critical value and R is the rejection area (Wasserman, 2010, p. 150).

The p-value is the smallest significance level at which we can reject the null hypothesis for the given data.

$$p - value = inf\,(\alpha : T(X^n) \in R_\alpha)$$

where $\alpha \in$ (0,1) is a size α test with rejection region R_α. T is a test statistics function on the sample of observations X^n. There is a statistically significant difference between the observed values when the p-value is less than the significance level. If the p-value is greater than the significance level, we can most likely fail to reject the null hypothesis and conclude that the observed values are not statistically significant (Wasserman, 2010, p. 156-159).

3.2.5 Type I and Type II error

Hypothesis tests are used by researchers and statisticians to obtain statistically relevant results. Even though hypothesis tests are trustworthy, they are susceptible to two sorts of errors. A type I error occurs when we reject a true null hypothesis due to a lack of evidence to support it, whereas a type II error occurs when we do not reject a false null hypothesis due to a lack of evidence supporting the alternate hypothesis. Type I errors are typically known as false positives, while type II errors are known as false negatives (Wasserman, 2010, p. 150).

	H_0 is True	H_1 is True
Reject H_0	Type I error	correct
Do not Reject H_0	correct	Type II error

3.3 QQ-plot

A Quantile-Quantile plot, commonly known as a QQ-plot, is a scatter plot that can be used to identify whether a set of data is likely to have come from a theoretical distribution like a Normal distribution. The data distribution is plotted against the expected normal distribution. A straight diagonal line or a reference line is formed, having the intercept and slope as the sample mean and standard deviation, respectively, of the data. The observed values are represented by points. If the data is normally distributed, the observed values follow the reference line exactly or are very close to it. Any variation from the line denotes a deviation from normality. The scatter plot around the straight line in the left skewed data is concave and the scatter plot around the straight line in the right skewed data is convex (Hay-Jahans, 2019, p. 146-153).

3.4 Parametric and Nonparametric Methods

Parametric and nonparametric methods are two types of statistical tests. Parametric tests can analyze interval or ratio data with some assumptions about the sample data drawn from the population, while nonparametric tests can analyze all types of data (nominal, ordinal, interval, or ratio) with fewer assumptions. The fundamental assumptions for parametric tests to follow are that sample data should be independent, normally distributed and have similar variance within the groups that are being compared (Wasserman, 2010, p. 119-135).

The nonparametric test requires all the observations of the samples to be independent. It is also known as a "rank-based test" because it puts the actual data into ranks (Wasserman, 2010, p. 303-319).

ANOVA is a parametric test and the Kruskal-Wallis and Mann-Whitney-Wilcoxon tests are both nonparametric tests, which are examined in the following subsections.

3.4.1 ANOVA

The analysis of variance, or ANOVA, determines whether there are statistically significant differences between two or more groups by comparing the means of the groups. It is a parametric test that adheres to all of its assumptions. A one-way ANOVA test is used when there is just one independent variable, while a two-way ANOVA test is used when there are two independent variables.

Suppose the following hypotheses, having $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_k$ as sample means, are being tested,

$H_0:$ Samples drawn from the population are equally distributed.

$H_1:$ At least one of the samples is different from others in distribution.

Then the F statistic, the test statistic used in ANOVA, determines whether or not the means of the various groups are significantly different and is computed as

$$F = MS_{between}/MS_{within}$$

where, $MS_{between}$ is the mean sum of squares for between-group variability

$$MS_{between} = \frac{\sum_{j=1}^{k}(x_j - \bar{x})^2}{(k-1)}$$

and MS_{within} is the mean sum of squares for within-group variability

$$MS_{within} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{(N-k)}$$

Here, k represents the number of groups, n_j represents the sample size of the j^{th} group, $\bar{x}_j$ represents the mean of the j^{th} group, $\bar{x}$ represents the overall mean of all the samples, x_{ij} represents the i^{th} element in the j^{th} group and N represents the total number of observations. $(k - 1)$ and $(N - k)$ represents degree of freedom for between-group variability and within-group variability, respectively.

If the observed F value is greater than the F-critical value (value obtained from the F distribution table) and the probability (p-value) of observing this F-statistics under the null hypothesis is sufficiently low (i.e., less than significance level, 0.05), then the null hypothesis is rejected otherwise not rejected (Black, 2006, p. 406-412).

3.4.2 Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric alternative to one-way ANOVA test when the parametric assumptions are not met except the assumption of independent sample data has to satisfy. This test is based on ranking data i.e., it takes the actual data, puts them into ranks and then calculates the significant difference between the groups. The Kruskal-Wallis test statistics converges into asymptotically chi-squared (χ^2) distribution.

Let there be J samples drawn from a population and the hypotheses to be tested at significance level (α =0.05) are as follows

H_0 : All independent samples have the same distribution.

H_1 : At least one of the independent samples differs in distribution.

The Kruskal-Wallis test statistics is given by

$$\chi^2 = \frac{12}{n(n+1)} \left(\sum_{j=1}^{J} \frac{R_j^2}{n_j} \right) - 3(n+1)$$

where n is the sum of sample sizes for all samples, J is the number of samples, R_j is the sum of ranks in the jth sample and n_j is the size of the jth sample The χ^2 value is compared with the critical value ($\chi^2_{\alpha,df}$) of the chi-square distribution at $\alpha = 0.05$ under the null hypothesis and if the value of the test statistic is greater than the $\chi^2_{\alpha,df}$ then the null hypothesis is rejected. Simultaneously if the probability (p-value) of observing the chi-square distribution under the null hypothesis is less than the significance level then the null hypothesis is rejected (Black, 2006, p. 694-696).

3.4.3 Mann-Whitney-Wilcoxon test

The Mann-Whitney U test, also known as the Wilcoxon rank sum test, is a nonparametric hypothesis test that confirms pairwise differences in group levels between one or two samples. It is necessary to satisfy the assumption of two independent groups having at least ordinally scaled properties.

It combines the two distinct groups into one, keeping the original group's identity, and ranks them according to their value. The ranks are assigned from smallest to largest value, with the smallest value receiving first rank and the largest value receiving last rank. The ranks of each group are then added together, and the accompanying U-statistics are utilized in determing the p-value.

Suppose following hypothesis is being tested,

H_0 = All the independent groups are have same distribution.

H_1 = At least one of the independent groups differ in distribution.

The α is at 0.05. The U statistics, U_1, for rank sum (W_1) of group 1 and, U_2, for rank sum (W_2) of group 2 are

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \quad and \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2$$

Now, for example, samples are small in size (less than 10) then U, minimum of (U_1,U_2), is compared with the critical value at significance level from the U-distribution table. If the test statistic is greater than the critical value then the null hypothesis is rejected otherwise not rejected. But for large sample sizes (more than 10), test statistics converges to approximately normally distributed and critical value is calculated using z-statistics under the null hypothesis.

$$\mu_U = \frac{n_1.n_2}{2}, \sigma_U = \sqrt{\frac{n_1.n_2(n_1+n_2+1)}{12}}, z = \frac{U-\mu_U}{\sigma_U}$$

where, μ_U is the expected value of U and σ_U is the standard error of U. If the z-statistical value is greater than the z-critical value (from the z-distribution table) and p value less than α, the null hypothesis is rejected (Black, 2006, p. 678-685).

3.5 Bonferroni correction

The likelihood of making a Type I mistake increases when two or more statistical analyses are performed on the same sample of data. As a result, the family-wise error rate (FWER) increases. The probability of making at least one Type I error while doing multiple hypothesis tests on the same data is known as the family-wise error rate (Wasserman, 2010, p. 165-166).

$$\alpha_{FW} = 1-(1-\alpha)^k$$

where, k is the number of comparisons performed, α is the significance level and α_{FW} is the family wise error rate.

To account for this error, Bonferroni correction is used. The Bonferroni corrected/adjusted value can be calculated by dividing the significance level by the number of statistical analyses performed. This adjusted value is then compared with the p-values.

$$\alpha_{adjusted} = \frac{\alpha}{k}$$

Another way is that the bonferroni adjusted p-value can be calculated by multiplying observed p-values with the number of statistical analyses and then compare with the significance level. Let there be k observed p-values for each analysis then any observed p-value less than the corrected p-value is declared to be statistically significant.

$$p-value_{adjusted} = p-value_{observed} * k.$$

In both the scenarios, there can be a situation where we can get significant solutions without applying the correction but after applying the correction there are no significant solutions due to decrease in adjusted significant value or increase in the adjusted p-value (Wasserman, 2010, p. 165-167).

4 Statistical analysis

All of the previously stated statistical approaches are utilized to perform statistical hypothesis tests on the data set in the following section. The Kruskal-Wallis H test is employed in the first task, while the Mann-Whitney-Wilcoxon test is utilized in the second. The data comprises of information on net rent of an apartment, which includes the entire size of the apartment, but both tasks require a comparison with rent per square meter, which is not included in the dataset. So, a new element is caculated first: rent per square meter, which is basically net rent divided by the apartment's living area.

Before applying the test to both tasks, three fundamental parametric assumptions are verified:

Assumptions of independence: When one observation does not impact or affect other observations, it is said to be independent. We can presume that rent in a top location has no bearing on rent in an average or good location, and vice versa, because we have no repeated data. Good, top, and average location appear to be more or less independent due to the nature of the dataset.

Assumptions of Normality: To fulfill the normality assumption, a set of observations follows a straight line in the QQ-plot. In Figure 1 in the Appendix on page 15, we have plotted all the categories of location quality in three QQ-plots. We can see from it that average and good locations follow approximately a normal distribution as the observations are less deviated from the reference line. The top location QQ-plot shows a few observations which are deviated from the reference line, but as the dataset has a large number of observations, we can ignore these observations and consider the plot to be approximately normal distribution. Hence, the normality assumption is fulfilled.

Assumption of homogeneity of variance: Firstly from the Figure 2 in the Appendix on page 16, we can see that the median values for all the three location qualities differ with each other and also the length (IQR) of each box differs. To further confirm the observation, Table 1 depicts the variances of each location quality. The variance of top, good and average location is approximately 29, 26 and 20 respectively. To fulfill homogeneity of variance assumption the variance has to be roughly equal but in this data we cannot assume equal variances in the location quality as they differ with some amount. Hence, the homogeneity of variance assumption is not fulfilled.

	Location	Variance
1	Average (1)	19.7928
2	Good (2)	25.7956
3	Top (3)	28.9518

Table 1: Variance of Quality of location

4.1 Global test using Kruskal-Wallis

All of the assumptions must be met for a parametric test, however we cannot proceed with the one-way ANOVA test due to a breach of homogeneity of variance. Since non-parametric tests do not require any assumptions other than data independence, we are employing the Kruskal Wallis test to investigate the effect of location quality on rent per square meter. The significance level for the Kruskal-Wallis test is set to 0.05.

The hypotheses we are putting to the test are

H_0 : The independent samples are from the same distribution.

H_1 : At least one of the independent samples is from another distribution.

Through the Kruskal-Wallis test, we acquire a test statistical value of 23.451 as given in Table 2. The degree of freedom (k-1) is 2, with k equaling 3, the number of groups (top, good and average). As a consequence of the test statistical value and degree of freedom, we acquire a very small p-value, less than the significance level (α) of 0.05. Hence, we can reject the null hypothesis and consider that at least one of the location qualities differs in distribution with others and has a significant difference in the rent per square meter variable based on these facts.

	Data	test statistics (χ^2)	degree of freedom	p-value
	rent per sqmtr by quality.of.location	23.451	2	8.087e-06

Table 2: Summary of Kruskal-Wallis test

4.2 Pairwise comparison using Mann-Whitney-Wilcoxon Test

The preceding section says that there is at least one significant quality of location, and this section examines which category of quality of location is significant in terms of rent per square meter. For testing, the significance threshold (α) is set to 0.05.

The hypotheses test we are doing right now are

H_0 : All the location qualities belong to the same distribution.

H_1 : At least one of the location quality comes from different distributions.

Table 3 illustrates the results of the pairwise Wilcoxon test difference between top, good, and average locations with respect to rent per square meter. The test yields a p-value with no adjustments or corrections, as well as an adjusted p-value with the Bonferroni correction. All of the p-values for the pairwise differences in location quality are below the significance level (0.05). As a result, we reject the null hypothesis and conclude that their distributions are all distinct.

Due to three location qualities, the comparison is done three times. Hence, we apply the Bonferroni adjustment since there is a potential for making a Type I error in this test due to multiple testing. This includes multiplying the p-value by the number of comparisons, which in this case is 3. Despite the fact that the adjusted p-value has increased, it still remains below 0.05. The null hypothesis is therefore rejected. Overall, we may deduce that each location characteristic is distributed differently and has a significant difference in rent per square meter.

	Quality of location pair	p-value	adjusted p-value
1	Good-Average	0.0021	0.0062
2	Top-Average	3.6e-05	0.0001
3	Top-Good	0.0025	0.0074

Table 3: Pairwise comparisons with and without Bonferroni adjustment

5 Summary

For this study, we investigated the Munich city rent index dataset gathered in 1999. The Introductory Case Studies lecturers compiled the dataset from the University of Göttingen's website and provided to us. It contains information about 3082 apartment with the net rent and six other characteristics that influence apartment rent. The purpose of this report is to conduct hypothesis testing and determine whether the quality of the location element has a substantial impact on the rent per square meter of an apartment.

The hypothesis test was performed using the Kruskal-Wallis test (rank-based test) because the homogeneity of variance assumptions were violated. Furthermore, the p-value derived from the Kruskal–Wallis test indicates that there is a significant difference in at least one of the distributions of top, good, and average location attributes.

The Mann-Whitney-Wilcoxon test was used to determine which characteristics of a location have a substantial impact on rent per square meter. The test evaluated location quality in pairs, and the results revealed that all types of location quality are considerably different in distributions and even have a pairwise difference in rent per square meter. While doing the pairwise test three times, the occurrence of the family-wise error rate increased. Hence, Bonferroni correction was used to reduce such errors, with some adjustments in the observed p-values of the test. The adjusted p-values obtained from Bonferroni adjustment are still statistically significant, so the null hypothesis is again rejected. All the three independent location qualities have different distributions and therefore have a significant difference in the rent per square meter.

In general, we can extract more records and expand the dataset to include new criteria like laundry facilities, number of bedrooms, and internet facilities to improve the analysis and acquire more reliable and accurate results. A huge amount of data can aid in the support of a hypothesis and the reduction of hypothesis errors.

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL `https://CRAN.R-project.org/package=gridExtra`. R package version 2.3.

Ken Black. *Business statistics for contemporary decision making.* John Wiley & Sons, Inc., 2006. ISBN 978-0470-40901-5.

Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression: Models, Methods und Applications.* Springer New York, 2013. ISBN 978-3-642-34332-2.

Georg-August-Universität Göttingen. *Georg-August-Universität Göttingen*, 2022. URL `https://www.uni-goettingen.de/de/551218.html`. (visited on 22nd May 2022).

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics.* CRC Press, Taylor & Francis Group, Boca Raton, 01 2019. ISBN 9780429448294.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer, 2010. ISBN 9781441923226 1441923225.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.4.

Appendix

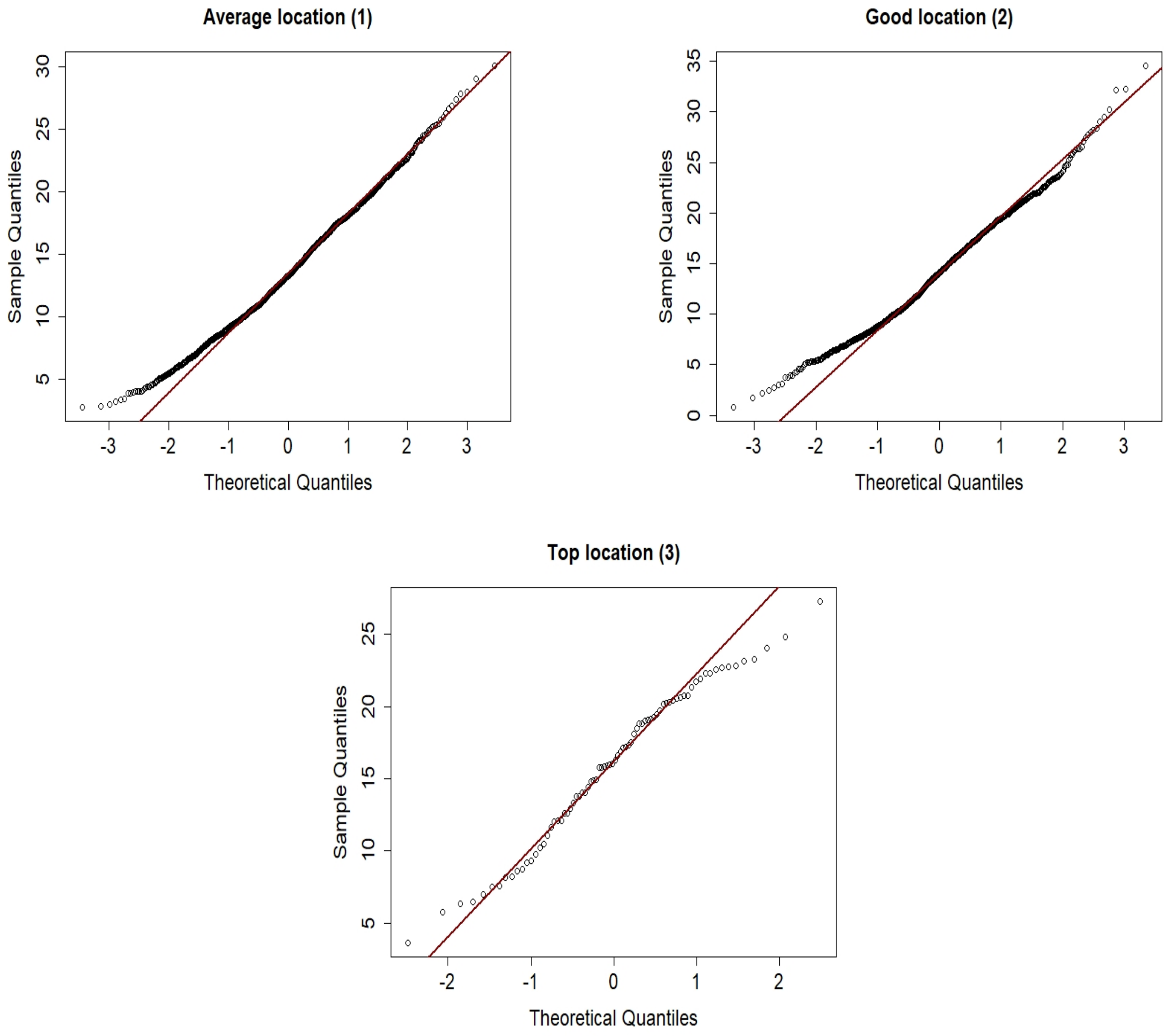

Figure 1: Normal QQ-plots for three location qualities

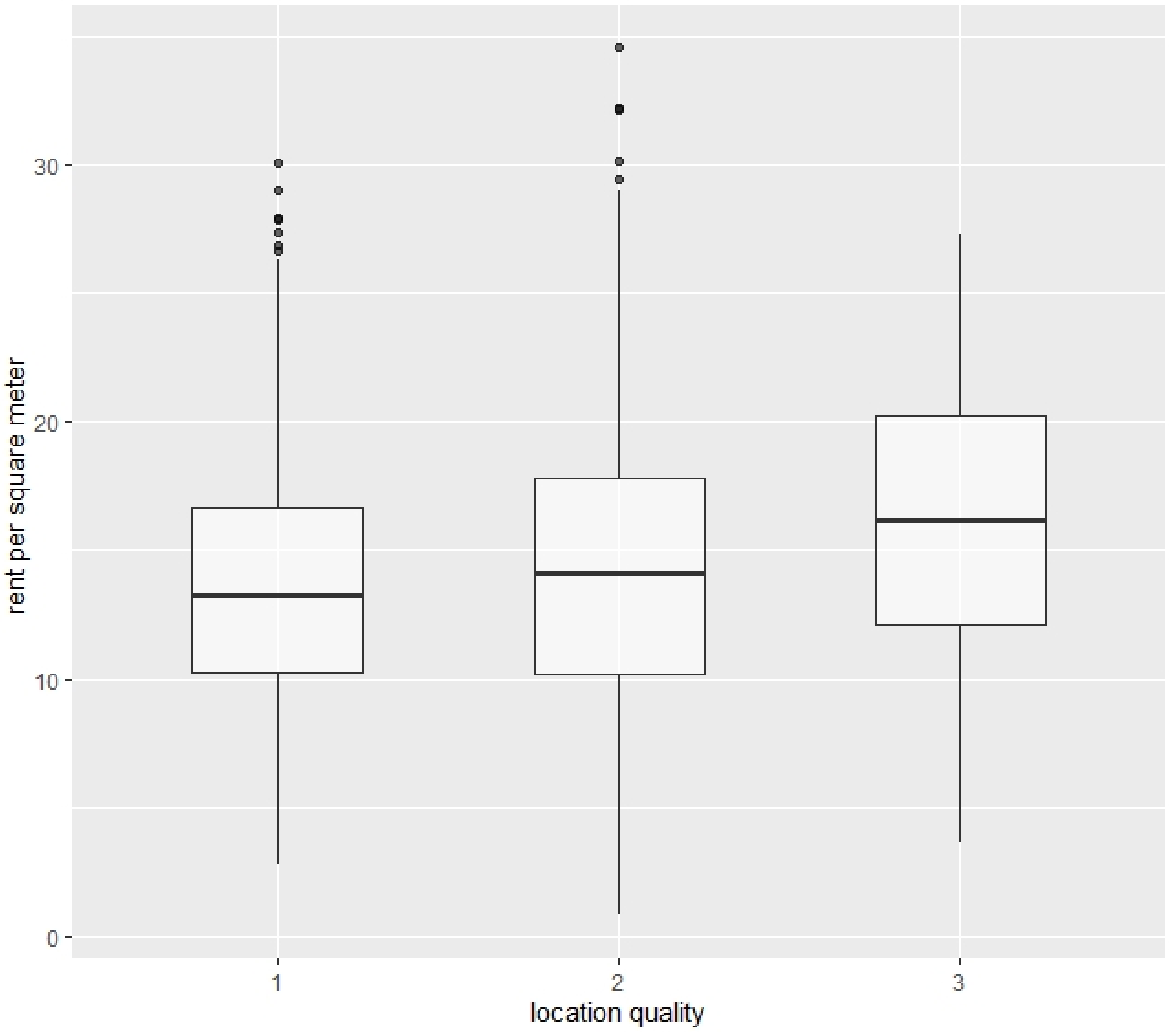

Figure 2: Boxplot for three location qualities