Aalto-yliopisto
School of Science
Department of Mathematics and Systems Analysis

On residential rents in the Helsinki metropolitan area

Bachelor's thesis
18.1.2017

Max Merikoski

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of bachelor's thesis

Author Max Merikoski		
Title of thesis On residential rents in the Helsinki metropolitan area		
Degree programme Engineering physics and mathematics		
Major Systems analysis		**Code of major** SCI3029
Supervisor Pauliina Ilmonen		
Thesis advisor(s) Pauliina Ilmonen		
Date 18.01.2017	**Number of pages** 22	**Language** English

Abstract

In this study, apartment rents in four different areas of the Helsinki metropolitan area were compared. The study was focused on single and two bedroom apartments sized between 20-60 m^2. One of the goals of the study was to find appropriate statistical methods for the comparison. The areas' price and size distributions were first analyzed visually as well as using descriptive statistics. The selection of the methods was based on this analysis.

The compared areas were Kamppi, Kallio, Matinkylä and Myyrmäki. These areas were chosen because they had sufficient sample sizes and were considered to represent areas with clearly distinct profiles.

The methods used consisted of a linear least absolute deviations regression and nonparametric statistical tests. The comparison of the price levels of the areas was done using the residuals of the regression line. The regression was used to model the dependency between the apartments floor area and its price per square meter.

Clearly the most expensive area was Kamppi, and Kallio after that. The prices between Matinkylä and Myyrmäki were confirmed to be very similar.

One of the important problems in the thesis was to eliminate the effect of apartment size in the price comparison between the areas. This was done using robust linear regression and allowed a better accuracy.

Keywords residential, apartment rents, Helsinki, linear regression, nonparametric tests

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Tekniikan kandidaatintyön tiivistelmä

Tekijä Max Merikoski

Työn nimi On residential rents in the Helsinki metropolitan area

Koulutusohjelma Teknillisen fysiikan ja matematiikan koulutusohjelma

Pääaine Systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja Pauliina Ilmonen

Työn ohjaaja(t) Pauliina Ilmonen

Päivämäärä 18.01.2017 **Sivumäärä** 22 **Kieli** Englanti

Tiivistelmä

Tässä työssä vertailtiin neljän eri pääkaupunkiseudulla sijaitsevan asuinalueen vuokrahintoja. Tarkastelun kohteena olivat 20-60 m^2 yksiöt ja kaksiot. Yhtenä työn tavoitteena oli sopivien menetelmien löytäminen vuokratasojen vertailua varten. Alueiden hinta- ja asuntojakaumia tutkittiin ensin visuaalisesti sekä tilastollisia tunnuslukuja hyödyntäen. Menetelmien valinta ja toteutus perustui tähän analyysiin.

Vertaillut alueet olivat Kamppi, Kallio, Matinkylä ja Myyrmäki. Alkuperäinen aineisto rajattiin kyseisiin alueisiin johtuen niiden erilaisista hintaprofiileista ja vuokrailmoitusten määristä.

Menetelminä työssä käytettiin robustia lineaarista regressiota ja epäparametrisia tilastollisia testejä. Alueiden hintojen vertailu toteutettiin alueiden neliöhintojen ja asunnon koon välistä riippuvuutta selittävän regressiosuoran residuaalien avulla.

Asuinalueista kalleimpia olivat Kamppi ja Kallio. Matinkylän ja Myyrmäen vuokratasot vahvistuivat hyvin samankaltaisiksi.

Tärkeä työssä ratkaistu ongelma oli asuntojen pinta-alan vaikutuksen poistaminen hintojen vertailussa käyttäen lineaarista regressiota. Tämä salli alueiden hintojen tarkan vertailun.

Avainsanat asuntojen vuokrat, pääkaupunkiseutu, lineaarinen regression, epäparametriset testit

Contents

1 Introduction

The purpose of this thesis is to compare the rental prices of residential apartments in four different urban areas of the Helsinki metropolitan area (HMA). These are Kamppi, Kallio, Matinkylä and Myyrmäki. Their approximate locations can be seen in Figure 1.

The main goal in the thesis was to create a robust method of comparing price data between different areas using statistical analysis. Simple and common location statistics like the median or average can give an incomplete and unreliable measure of prices. For example, when comparing data sets that include apartments of different sizes as in this thesis.

The effect of the apartment size was considered in the analysis with a linear regression after which statistical tests were used to analyze the differences in the residuals of prices in each area.

The original data including apartment locations, sizes and requested rents was extracted from a popular online aparment rental advertisement site, oikotie.fi. The four areas are all urban subdivisions of their respective cities and were chosen based on their expected price level and reputation, and sample size.

The analysis is focused on 1 and 2 room apartments sized between 20m^2 to 60m^2. Small one and two bedroom apartments were seen as the most interesting and relevant subgroup to focus the study on, due to their high demand in the HMA [16], and the possible effect of this high demand in prices.

2 Other studies

The property market in general has been a topic of research in many fields. Perhaps the most popular subject of all have been residential properties because almost all of us live in them. The motivation to predict changes in demand and supply of housing is usually financial. Ultimately the goal is to predict the price fluctuations in rents or prices of entire properties in order to make better decisions in economics and capital investments.

Hedonic regression is one of the popular methods for estimating property values. It uses regression analysis to create hedonic models that can be used to calculate the price of a building by combining different characteristics such as number of rooms, whether the apartments have balconies or what kind of a view they have, the type of flooring, etc. to calculate an estimate for the overall value. One

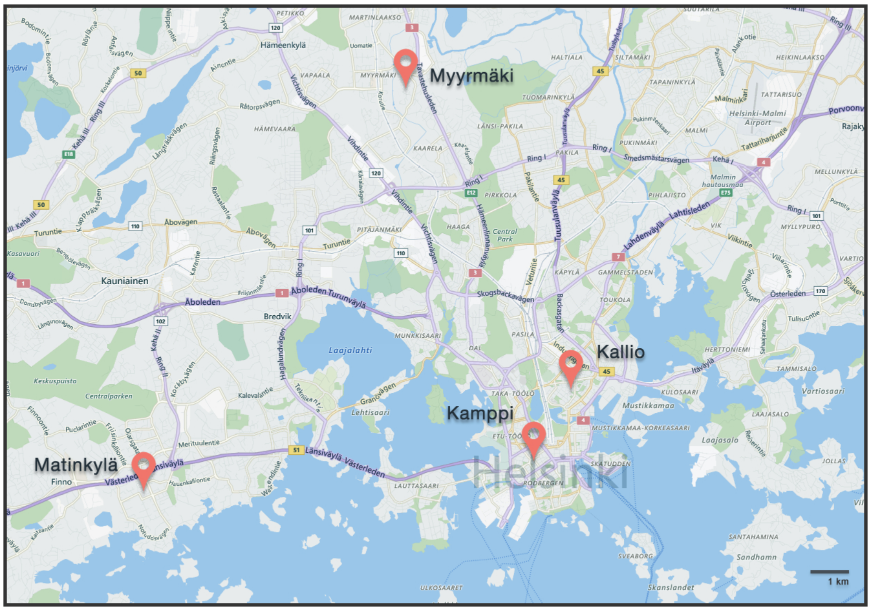

Figure 1: Map showing the locations of the examined areas. Kamppi is located right next to the central business district of Helsinki. (Bing Maps 2016)

summary of these methods is given by Matt Monson, 2009 [10]. One issue with spatial hedonic price models can be that they are based on data collected through time, but the effect of time on the data is unknown [3].

As an example of a simpler approach to the effect of public transportation and apartment rents is the article titled "Mass transportation, apartment rent, and property values" by Sirmans et al. [2]: "Our empirical results show that distance from a metro station has an adverse effect on apartment rent, i.e., each one-tenth mile increase in distance from the station results in a decrease in rent per apartment unit of about 2.50%." A similar study, but instead using spatial hedonic pricing models, was done comparing three central municipalities in Lisbon, Portugal [9]. Martinez et al. found results that suggest that the proximity to one or two metro lines leads to significant property value changes.

Predicting demand is something that has also been studied using different methods. Linear regression analysis and genetic algorithms were compared in a study that found GA-LRA models to be the most effective out of those tested in providing the most accurate forecasts over long time horizons [17]. These models combine regression analysis and genetic algorithms.

3 Overview of the property market and the areas

3.1 The residential property market

The residential portfolio transaction volume in the property investment market has grown more than 10 times in the last 5 years, which tells how popular this segment of real estate currently is. Rents have been growing everywhere in the HMA since at least the 1990's, although, the KTI Residential rent index shows a negative trend since 2010 in the annual change which implies that growth has slowed down [7]. The KTI Residential rent index shows the annual percentage change in rents in all major cities of Finland.

Small apartments were chosen as the focus of this study because of their high demand. The average household size has declined steadily in Finland for a long time [18] and in fact , the proportion of people living in single person households grew the most with people over 35 years old in 2015 [19], which highlights that this increased trend in smaller households is not only happening with young people. Instead the people creating demand for small apartments are a heterogenous group not characterized by age or lifestyle, but more by their preference to urban environments [1].

3.2 The areas

A short summary of all the areas are described in this section.

One of the main differences in the four areas is that in Espoo and Vantaa, which are less densely populated than especially the more central parts of Helsinki, subareas like Matinkylä or Myyrmäki tend to become more focused on creating and maintaining a network of services intended to serve the needs of the subarea's local residents. [20]

This can be thought of as one of the main differences in living in a city environment compared to a suburban environment. Even though Matinkylä and Myyrmäki are both important urban centres in their respective cities, they represent a more self-serving area type than Kamppi or Kallio, which are both more dense and urban city environments.

3.2.1 Kamppi

The area of Kamppi is located right next to the central business district of Helsinki and is home to the 2006 opened shopping centre bearing the subdivision's name "Kamppi". It is characterised by dense construction and heavy traffic. Kamppi can be considered as a middle ground between the most central part of Helsinki and the areas surrounding it [14].

The services provided by the shopping centre, including the public transportation connections have had a major impact on the area. Due to the large amounts of business and retail in Kamppi, and its great traffic connections, it has become one of the most expensive residential areas of Helsinki [14] [4]. This is also reflected in the median salaries of the local population.

3.2.2 Kallio

Kallio is located only about a kilometre from the city centre, on the eastern side of the Helsinki peninsula. It has excellent public transportation connections to the rest of the HMA and to the city centre.

The surrounding districts of Sörnäinen and Alppiharju form together with Kallio an area that is home to about 40,000 people. These three districts also have the lowest average household size in all of Helsinki while simultaneously having the lowest rooms per people ratio [4]. This means that Kallio is an area very much defined by a large amount of people living alone in small apartments.

The population of Kallio stayed quite unchanged in the years before 2010, but has clearly been growing since that time, and is forecasted to keep doing so [4].

3.2.3 Matinkylä

The most central point of Matinkylä is the Iso Omena shopping centre which opened in 2001. Since then, the shopping centre has become the defining land mark for the centre of Matinkylä. Also, an entirely new section of Iso Omena was opened in 2016 and the upcoming western metro extension will also further cement the importance of the shopping centre. It is also notable that the shopping centre is located next to the Länsiväylä, a busy highway connecting Espoo to the heart of Helsinki.

Matinkylä as a socioeconomical area is defined largely by the contrast between two extremes: the exceptionally highly educated, high income population of

Nuottaniemi and the population of Tiistilä and Matinmetsä which are described by a lower education level and a higher unemplyoment level [6]. The are surrounding Iso Omena can be thought of as a middle ground of these two subareas.

Most of the residential properties near Iso Omena have been built during the 1990's and 2000's, and are quite modern. The older buildings in Matinkylä, mostly from the 1970's to 1980's are located in the Tiistilä and Matinmetsä areas. The seaside areas of Matinkylä, such as Nuottaniemi, are also mostly built during the 2000's [8].

3.2.4 Myyrmäki

Myyrmäki is one of the most important urban areas in the city of Vantaa with about 55,000 residents [22]. It is also one of the highest employing areas of the city. Myyrmäki is mostly focused on servicing the needs of its residents, as opposed to business to business activity [5]. The busiest area of Myyrmäki is the Myyrmanni shopping centre which is located next to the suburb's most important transportation element, the Myyrmäki railway station.

In the way business and retail functions in Myyrmäki, it is quite similar to Matinkylä, although when it comes to transportation and traffic, Myyrmäki is not located next to an important road or highway, nor is it part of the subway system, but it is conveniently connected to the rest of the HMA with the Myyrmäki railway station.

4 Methods

To compare the price levels of the different areas, a least absolute deviation regression model is applied to the data shown in Figure 2. The regression is used to remove the effect of apartment size on price distributions so that the residuals from each area can be used to compare price levels. Without doing this, the different amounts of smaller or larger apartments would skew any measures of an area's price. In order to calculate the regression, it is assumed that each area has a similar dependency between apartment size and price per square meter.

The residuals of each area's price data are then calculated and can be compared to each other with measures like the sample median to find obvious differences in prices. Finally, the Wilcoxon rank sum test and the Kruskal-Wallis test are used to test for differences in the medians of the areas and to provide a statistical significance to the results.

4.1 Descriptive statistics

The various data analysed in this thesis are described using common measures in descriptive statistics. The following measures are defined here: mean, median, variance, standard deviation, median absolute deviation, skewness and kurtosis.

4.1.1 Definitions

Throughout this section we assume that $(x_1, ..., x_n)$ are independent and identically distributed observations of a random variable X. In the following definitions, we assume that all the expected values $E(\cdot)$ exist as finite quantities.

Definition 4.1. Mean

The mean value of X is defined as

$$\mu = E(X)$$

Mean is usually estimated with the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Definition 4.2. Median

Let $y_1 \leq y_2 \leq ... \leq y_n$ be ordered values of the observations $(x_1, ..., x_n)$.

Then the sample median of $(x_1, ..., x_n)$ is the middle value of y.

It follows that if n is even, then the sample median $\hat{m}$ can be calculated by

$$\hat{m} = \frac{(y_{n/2} + y_{n/2+1})}{2}$$

If n is odd, the sample median $\hat{m}$ can be calculated by

$$\hat{m} = y_{(n+1)/2}$$

The population median m_x is defined by

$$P(x < m_x) \leq \frac{1}{2} \quad \text{and}$$
$$P(x \leq m_x) \geq \frac{1}{2}$$

Definition 4.3. Variance

The variance of X is defined as

$$\sigma^2 = E[(X - \mu)^2]$$

where the standard deviation of X is σ.

Variance is usually estimated with the sample variance s_x^2 defined as

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

Definition 4.4. Median absolute deviation

Sample median absolute deviation (MAD) is defined as the sample median of the absolute deviations of the observations from its median. It can be calculated as

$$\text{MAD} = Median\{|x_1 - \hat{m}|, ..., |x_n - \hat{m}|\}$$

Population median absolute deviation MAD_x is defined as

$$P(|x - m_x| < \text{MAD}_x) \leq \frac{1}{2} \quad \text{and}$$
$$P(|x - m_x| \leq \text{MAD}_x) \geq \frac{1}{2}$$

Definition 4.5. Skewness

The skewness γ_1 of a distribution is defined as

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

where μ_j is the jth central moment i.e.

$$\mu_j = E\left[\left(\frac{x-\mu}{\sigma}\right)^j\right] \tag{1}$$

Skewness is usually estimated with sample skewness g_1 defined as

$$g_1 = \frac{k_3}{k_2^{3/2}}$$

where k_j is the jth central sample moment, i.e.

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

Definition 4.6. Kurtosis

Excess kurtosis, usually denoted by γ_2 is defined by

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$$

where μ_j is as in Equation 1 in Definition 4.5.

Excess kurtosis can be estimated with the sample excess kurtosis g_2 defined as

$$g_2 = \frac{k_4}{k_2^2}$$

where k_j is as in Equation 2 in Definition 4.5.

4.2 Least absolute deviation regression

The least absolute deviations (LAD) method of finding a linear function to describe the dependency of the prices per square meter to the floor area of the apartments is used in the regression part of this study. The LAD method minimizes the sum of absolute values of errors. The method is also known as the L1 method, since the the minimized distances are of the L1 norm.

The LAD method is not as sensitive to outliers in the data compared to the more widely used least squares (LS) method for example, since large deviations are not weighed as much as with the quadratic term that is used in the LS method. Therefore, it produces more robust estimates.

The drawback of the LAD method is that it is more inconvenient to calculate than the LS method, since it does not have a closed-form solution. The solution method used in this thesis is represented here.

Definition 4.7. LAD regression

Given data (x_i, y_i), the linear regression is fitted by solving the following problem.

$$\arg\min_{a,b} \quad \sum_{i=1}^{n} |y_i - (ax_i + b)| \tag{3}$$

R function

The R package [12] and its function **’lad’** is used to find the parameters a and b in

$$f(x) = ax + b \tag{4}$$

The function formulates the problem as a linear program and uses the Barrodale and Roberts algorithm which is a Simplex based method.

4.3 Kruskal-Wallis test

The Kruskal-Wallis test is a non-parametric equivalent of the one-way analysis of variance (ANOVA). It is used to test the null hypothesis that k different different random samples come from the same distribution [15]. The test is usually interpreted as a test of similarity of medians, since it is especially sensitive to differences in median.

The test is chosen because it does not make assumptions about the distributions of the observations. In this case the null hypothesis is interpreted as indicating that the median of the tested residuals are the same for all 4 areas. The alternative hypothesis is then that at least one population median of the observed residuals is different from at least one other sample.

Definition 4.8. Rank

Let $X_1 \leq X_2 \leq ... \leq X_N$ be real-valued observations from a continuous distribution positioned in increasing order.

The rank R_{Nj} of X_j is its position number in the ordered set $X_1, ..., X_N$. The rank vector R_N is the set of these ranks.

If $X_1, ..., X_N$ are all different, then R_{Ni} is defined by the equation

$$X_i = X_{N(R_{Ni})} \tag{5}$$

Definition 4.9. Kruskal-Wallis test

Suppose the problem is to test the hypothesis that k independent random samples $(X_1, ..., X_{N1}, X_{N1+1}, ..., X_{N2}, ..., X_{Nk-1+1}, ..., X_{Nk})$ are identical in distribution. Let us also assume that the k samples are continuous and identically distributed, but may have different medians m_j.

Let $N = N_k$ be the total number of observations, and let R_n be the rank vector of the pooled sample $(X_1, ..., X_N)$, as defined in Definition 4.8. Let n_j equal to the number of observations in the jth sample.

Null hypothesis $H_0 : m_1 = m_2 = ... = m_k$.

Alternative hypothesis H_1 : At least two of the medians m_j differ from each other.

The Kruskal-Wallis test statistic is usually written in the form [21]

$$\frac{12}{N(N-1)} \sum_{j=1}^{k} n_j \left(\bar{R}_{j.} - \frac{N+1}{2} \right)^2, \quad \text{where} \tag{6}$$

$$\bar{R}_{j.} = \frac{\sum_{i=N_{j-1}+1}^{N_j} R_{Ni}}{n_j} \tag{7}$$

This test statistic measures the distance of the average scores of the k samples to the average score $(N+1)/2$ of the pooled sample.

R function

The R function **kruskal.test** [11] is used to calculate the test statistic and p-value.

4.4 Wilcoxon rank sum test

Also known as the Mann-Whitney U-test, the Wilcoxon rank sum test is used to test the hypothesis that the distributions of two separate samples are identical. The test was chosen because as a non-parametric test, it doesn't require assumptions about the probability distributions of the test samples. The Wilcoxon rank sum test is known to especially detect differences in medians.

In this thesis, the test is used to test the similarity of the prices per square meter in different areas. The alternative hypothesis of the test is that one of the two areas being tested has a higher price. All four areas' price residuals from the linear regression are compared resulting in six two-sample tests.

Definition 4.10. Wilcoxon rank sum test

Let there be two independent random samples, $X = (x_1, ..., x_m)$ and $Y = (y_1, ..., y_n)$. Assume that they come from continuous distributions. Set $N = m+n$ and let R_N be the rank vector of the pooled sample $(X_1, ..., X_m, Y_1, ..., Y_n)$.

Null hypothesis: $H_0 : X$ and Y have identical distributions.

Alternative hypothesis H_1 : The distribution of Y is stochastically larger than that of X.

The Wilcoxon statistic is defined as [21]

$$W = \sum_{i=m+1}^{N} R_{Ni} \tag{8}$$

For large values of the Wilcoxon statistic, the null hypothesis is rejected.

R function

The R function **wilcox.test** [13] was used to calculate the wilcoxon test statistic and p-value. The non-paired two-sample version of the test is used.

5 Descriptive analysis of the data

To describe the data in each four areas, the prices per square meter and apartment sizes in square meters have been plotted in Figures 3 and 4. Descriptive statistics

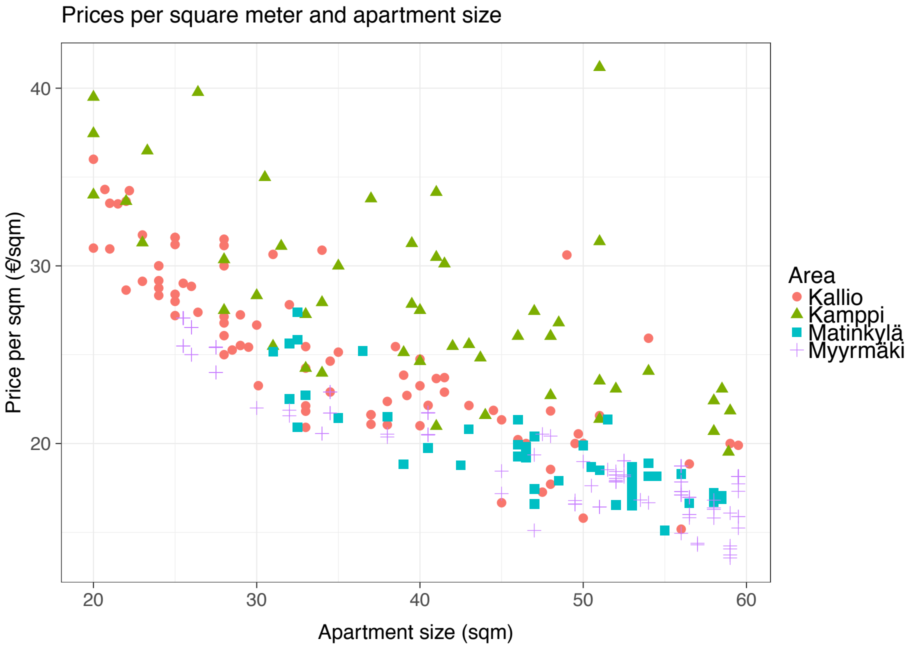

Figure 2: The rent per square meter plotted with apartment size.

have also been calculated and they can be seen in Tables 1 and 2. A scatter plot of the relationship between apartment price and size is shown in Figure 2.

A correlation between apartment size and price can be seen in Figure 2. Larger apartments are being rented for a lower price per square meter than smaller apartments. Some clear outlying data points can also be seen. All areas show a negative linear trend in price per square meter as apartment size increases.

To use a method that is less sensitive to the outliers, the least absolute deviations method for finding a linear approximation function for the data in Figure 2 is used. The estimated regression line is given in Figure 5.

5.1 Distributions of size and price in the areas

In this section Figures 3 and 4 and Tables 1 and 2 will be used to describe the distributions of the two variables, price per square meter and apartment size.

Looking at Figure 3 it can be seen that the distributions of apartment sizes are inconsistent with each other.

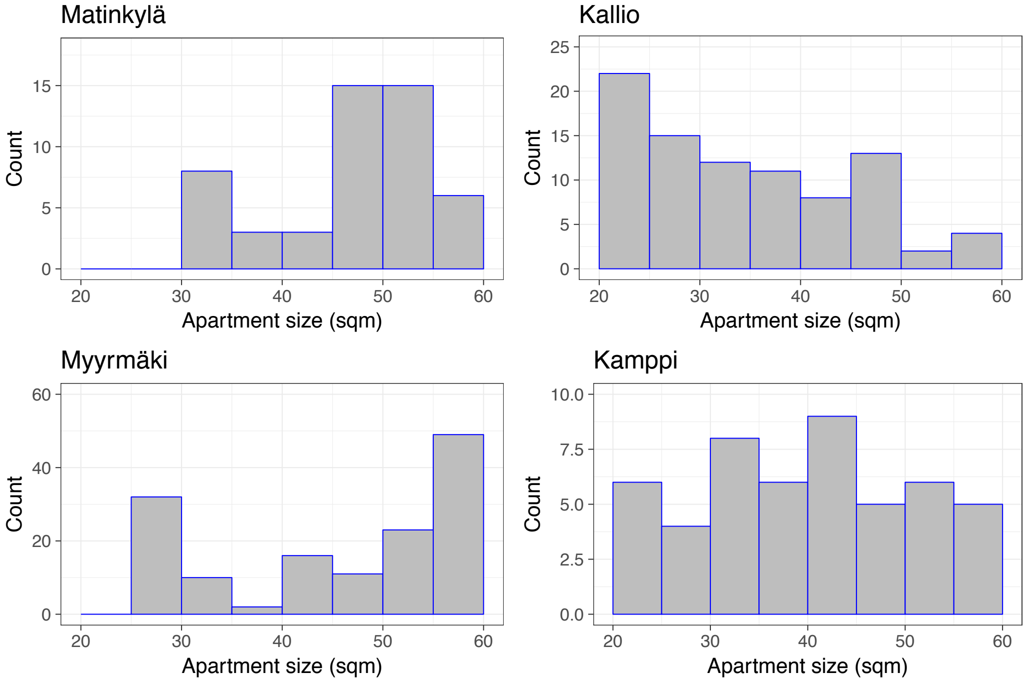

Figure 3: Apartment size histograms of all four areas.

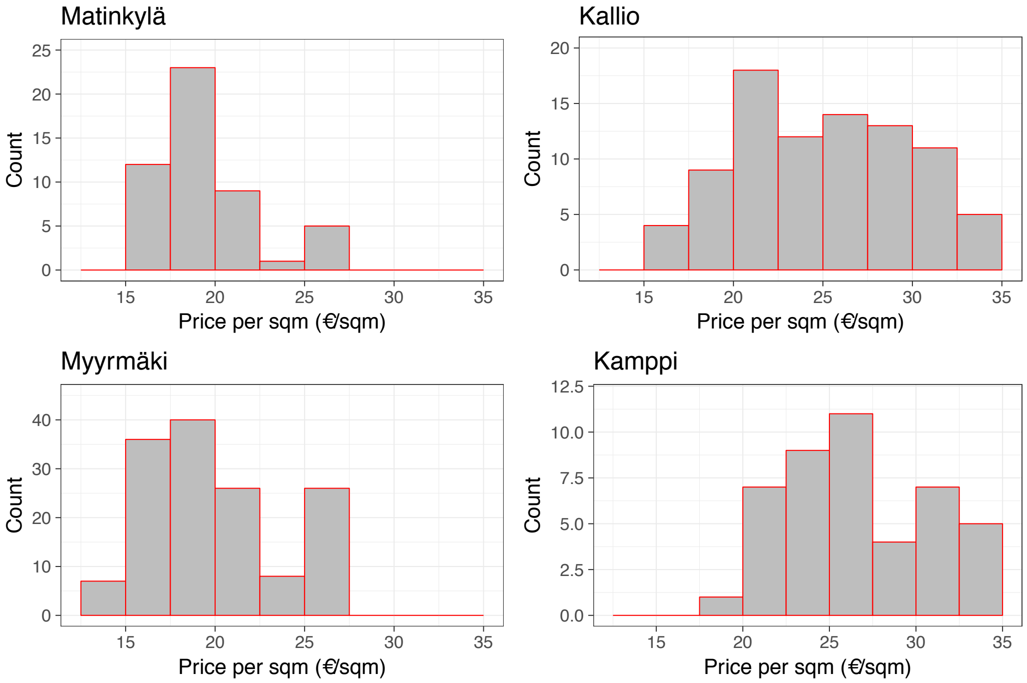

Figure 4: Histograms of prices per square meter in all four areas.

Table 1: Descriptive statistics of the apartment size data in each area.

Size	Kallio	Kamppi	Matinkylä	Myyrmäki
Mean	35.1	40.0	46.9	45.4
Median	33.0	41.0	47.0	50.5
MAD	8.00	8.00	6.00	9.00
Variance	113.2	124.9	66.3	154.8
Std. Dev	10.6	11.2	8.1	12.4
Skewness	0.5	-0.1	-0.6	-0.5
Kurtosis	-0.9	-0.8	-0.7	-1.3

Table 2: Descriptive statistics of the prices per square meter data in each area.

Sqm Price	Kallio	Kamppi	Matinkylä	Myyrmäki
Mean	25.3	27.9	19.6	19.9
Median	25.1	27.3	19.0	18.4
MAD	3.81	3.84	1.45	2.07
Variance	23.5	29.7	7.3	13.8
Std. Dev	4.8	5.5	2.7	3.7
Skewness	0.1	0.7	1.1	0.6
Kurtosis	-0.8	-0.2	1.1	-0.8

5.1.1 Kamppi

The apartment sizes in Kamppi are quite varied, and the distribution is relatively flat. This is indicated especially by a kurtosis of -0.8. A median size of 41 m^2 is in the middle of our range of 20 m^2 to 60 m^2.

Kamppi also has the most expensive apartments of all areas. The median price per square meter in Kamppi is 27.3 €/m^2. Of all the areas, it also has the highest variance in price, shown by a MAD of 3.84 €/m^2 and a standard deviation of 5.5 €/m^2.

5.1.2 Kallio

Apartments of all sizes are represented in Kallio, however the bulk of apartments are smaller, with a median size of 33 m^2. It can be said that the data is skewed to the right , which is also indicated by a skewness of 0.5. The kurtosis of apartment prices and sizes is also something to note in Kallio, with respective values of -0.8 and -0.9.

Table 3: Comparing the amount of unique addresses per area to all apartment adverts in the data set.

Adverts per area	Kamppi	Kallio	Matinkylä	Myyrmäki
Unique Addresses	41	73	26	30
Total amt of apartments	49	87	50	143

A median price of 25.1 €/m^2 puts Kallio higher in price than Matinkylä or Myyrmäki, but not quite as high as Kamppi. From Table 3, Kallio also has a relatively large amount of data with a high proportion of unique adverts. All in all many apartment prices and sizes are represented in Kallio.

5.1.3 Matinkylä

Most apartments in Matinkylä are in the region of 20 €/m^2. The median price is 19 €/m^2. Matinkylä's prices have the lowest variance out of all four areas which is shown by a MAD of 1.45 €/m^2 and a standard deviation of 2.7 €/m^2. In apartment sizes, Matinkylä is quite similar to Myyrmäki in that neither have many apartments below 35 m^2. Matinkylä also has the lowest amount of uniqe adverts, with a ratio of 26 out of 50 adverts being in unique addresses as shown in Table 3.

5.1.4 Myyrmäki

Myyrmäki has a distribution in apartment sizes with the biggest apartments taking up a major portion of the sample. The median apartment size in Myyrmäki is 50.5 m^2 and the median price is 18.4 €/m^2.

The distribution of apartment price and size in Myyrmäki are heavily affected by a large portion of the sample data consisting of only 2 new apartment buildings' rental adverts. Myyrmäki still has some variance in prices with a standard deviation of 3.7 €/m^2 and a MAD of 2.07 €/m^2.

6 Results

6.1 LAD Regression

Figure 5 shows the regression line that was fitted using the methods described in Section 4.2. The following equation describes the linear regression.

$$f(x) = -0.3263x + 35.3791, \quad \text{where} \tag{9}$$

x is the size of the apartment in square meters.

6.2 Price differences

The prices per square meter from Kamppi, Kallio, Matinkylä and Myyrmäki were used to model the relationship between apartment size and price using a least absolute deviations linear regression, shown in Figure 5. The residuals in Figure 6 were then compared using the Wilcoxon rank sum test and Kruskal-Wallis test. The results of these tests can be seen in Tables 5 and 6. The medians of the residuals are shown in Table 4.

The Kruskal-Wallis test's small p-value of $8.7 \cdot 10^{-26}$ indicates that the null hypothesis of identical medians in the price residuals can be rejected. A more detailed comparison of the prices' medians is done with the Wilcoxon rank sum test, which is done for all six pairs of areas. All but one test pair of areas resulted in a very small p-value, further explaining the result of the Kruskal-Wallis test. The Wilcoxon test resulted in a p-value of 0.375 for the Myyrmäki-Matinkylä pair. Thus, the null hypothesis of identical distributions is not rejected even with high confidence levels.

The sample medians of the residuals in Table 4 shows that the highest prices are found in Kamppi, after that Kallio is the second most expensive area while Matinkylä and Myyrmäki have the lowest rents, sharing a similar price distribution. This argument is also supported by the descriptive statistics in Table 2 in Section 5.1, which shows the similarities of the areas prices mean and median values.

The high prices of Kamppi are explained by its location. It is next to the central business district (CBD) of Helsinki, and depending on the definition, it can be partly included in the CBD. Apartments of all types are high in demand in this area. Kallio is also much closer to the centre of Helsinki than Matinkylä or Myyrmäki, which most likely explains a large part of the price difference it has to the latter two.

Table 4: Sample medians of the price per square meter residuals.

Area	Median
Kallio	1.01
Kamppi	5.17
Matinkylä	-0.34
Myyrmäki	-0.44

Table 5: Wilcoxon rank sum test values for the price per square meter residuals.

Test pair	W	p-value
Myyrmäki, Matinkylä	3273	0.375
Myyrmäki, Kamppi	306	$1.6 \cdot 10^{-21}$
Myyrmäki, Kallio	3293.5	$1.1 \cdot 10^{-9}$
Matinkylä, Kamppi	140	$3.2 \cdot 10^{-14}$
Matinkylä, Kallio	1186	$9.9 \cdot 10^{-6}$
Kamppi, Kallio	3485	$8.6 \cdot 10^{-10}$

The similarity of the price distributions of Matinkylä and Myyrmäki is interesting, because both areas could be historically said to have had the reputation of being for lower income residents. This is less of a factor nowadays, and especially Matinkylä is a more demographically diverse area nowadays in this sense. However, with the construction of the Länsimetro (western metro extension), the expansion of the shopping centre Iso Omena, and the addition of many new apartment buildings in the last years, Matinkylä is seen as a quite modern and central part of Espoo nowadays. Even before that, with Iso Omena being situated next to the Länsiväylä motorway, Matinkylä has been one of the many important urban centres of Espoo. Myyrmäki is similar to Matinkylä since it also has a big central shopping centre, Myyrmanni, that attracts visitors from outside of Myyrmäki as well as being home to an important link in the public transportation system by having a railway station that connects it to the rest of the HMA and being an important urban centre of the city of Vantaa.

Table 6: Kruskal-Wallis test values for the price per square meter residuals.

Test value	Chi-squared	d.f.	p-value
All areas	119.76	3	$8.7 \cdot 10^{-26}$

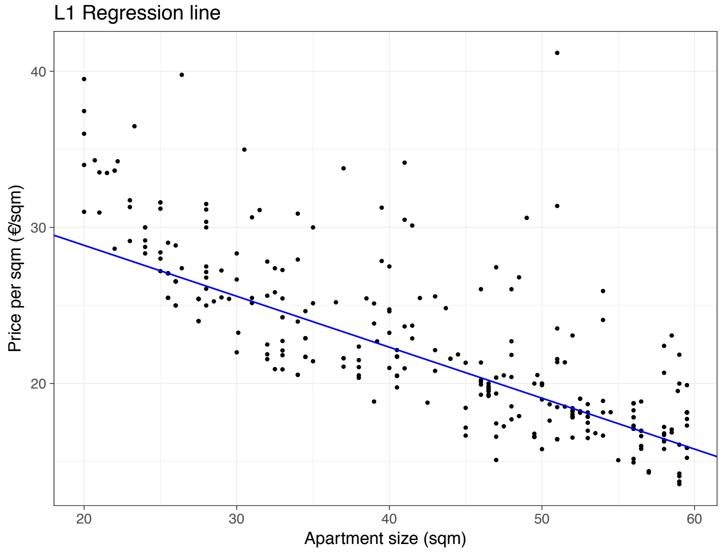

Figure 5: Prices per square meter as a function of apartment size, with the LAD-model.

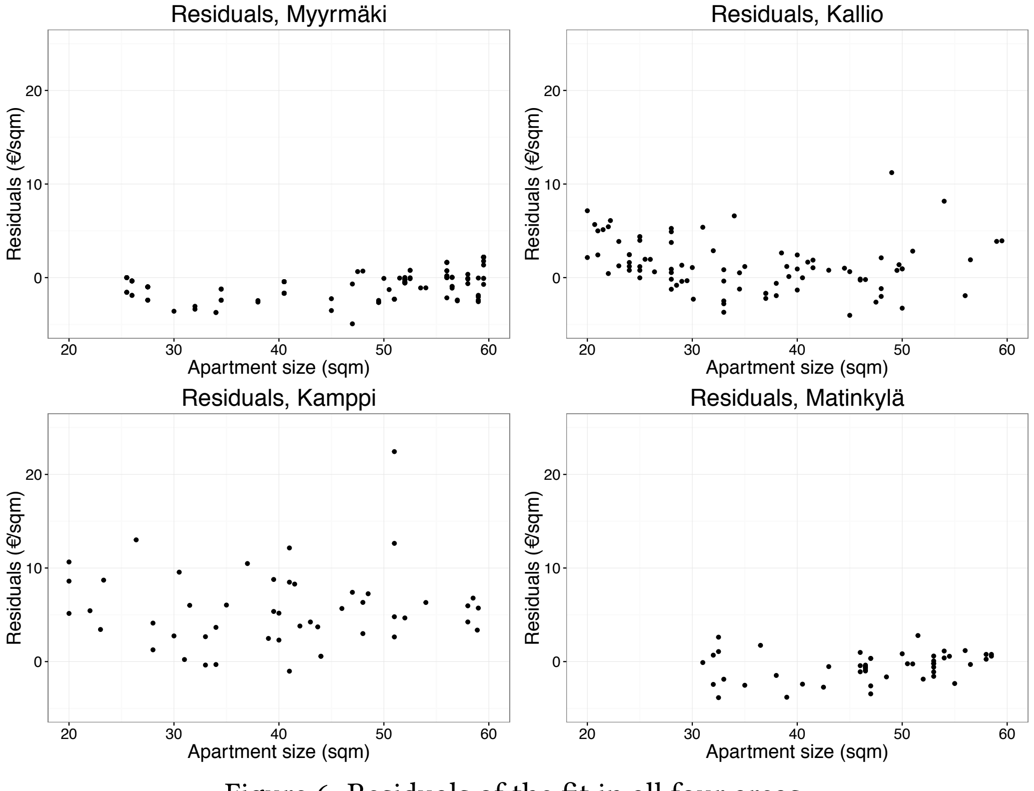

Figure 6: Residuals of the fit in all four areas.

7 Conclusions

The most interesting result was that the prices of Matinkylä and Myyrmäki were found to be so similar. It is contrary to the expectation of the areas' current images. Kamppi is found to be the most expensive area as expected. Kallio was also expected to be less expensive than Kamppi, but more expensive than the other two areas.

With these results, the method of removing the size factor from price data of apartments with an LAD regression is found to be an effective way to simplify the problem of comparing apartments of different sizes across different areas. One could take a larger sample of apartments and use the methods described in this thesis to conduct a more thorough examination of rent differences in the HMA. The benefit of statistical testing of the residuals is that the results are then confirmed with a statistical significance and the results can be interpreted at a population level, in the sense that if more data is collected, the results can be expected to stay the same.

Statistics such as the mean or median are usually used to describe rent levels of areas without taking into account the effect of apartment sizes. The distribution in apartment sizes can have a significant effect on any location measures such as a median or average taken from a large sample of apartments without taking the aforementioned effect into account.

It should be noted that the effect of apartment quality and age was not included in any way in the analysis. In addition, the data used in this study only represents a snapshot in time, and the effect of time in the results is unknown, although for the findings of this thesis, most likely very small.

References

[1] Katri Backman. Yksinasuvien helsinkiläisten asumispreferenssit. Master's thesis, Helsingin Yliopisto, 2015.

[2] John D. Benjamin and G. Stacy Sirmans. Mass transportation, apartment rent and property values. *Journal of Real Estate Research*, 12(1):1–8, 1996.

[3] Nicolas Devaux and Jean Dubé. About the influence of time on spatial dependence: A meta-analysis using real estate hedonic pricing models. *Journal of Real Estate Literature*, 24(1), 2016.

[4] Helsingin Kaupungin Tietokeskus. Helsinki by district. www.hel.fi/tietokeskus, 2015.

[5] Helsingin kaupunkisuunnitteluvirasto. Pääkaupunkiseudun työpaikkakeskittymät - klustereitako? 2013.

[6] Monika Kral-Leszczynska. Asuinalueiden sosiaalinen eriytyminen - tapaustutkimus matinkylästä. Master's thesis, Helsingin Yliopisto, 2012.

[7] KTI Finland. Kti market review, autumn 2016.

[8] Hannu Kytö and Monika Kral-Leszczynska. Asuinalueiden elinkaarikestävyys pääkaupunkiseudulla. *Kuluttajatutkimuskeskus. Tutkimuksia ja selvityksiä 2/2014*, 2014.

[9] L. Martinez and José Viegas. Effects of transportation accessibility on residential property values. *Transportation Research Record: Journal of the Transportation Research Board*, 2115, 2014.

[10] Matt Monson. Valuation using hedonic pricing models. *Cornell Real Estate Review*, 7, 2009.

[11] R Documentation. Kruskal-wallis rank sum test, 2016. Accessed: 2016-11-02.

[12] R Documentation. L1pack: Routines for L1 Estimation, 2016. Accessed: 2016-10-15.

[13] R Documentation. Wilcoxon rank sum and signed rank tests, 2016. Accessed: 2016-11-02.

[14] Katja Raunio. Kampin keskus: kasvavan kaupungin välitilassa. Master's thesis, Helsingin Yliopisto, 2015.

[15] Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists (Fifth Edition).* Academic Press, 2014.

[16] Suomen Hypoteekkiyhdistys. Hypon asuntomarkkinakatsaus q4/2016. 11 2016.

[17] S. Thomas, Martin, Skitmore, Keung, Fai, and Wong. Using genetic algorithms and linear regression analysis for private housing demand forecast. *Building and Environment*, 43(6), 2008.

[18] Tilastokeskus. Liitetaulukko 1. asuntokunnat koon mukaan ja asuntokuntien keskikoko 1960–2015. http://tilastokeskus.fi/til/asas/2015/asas_2015_2016-05-24_tau_001_fi.html, 2016.

[19] Tilastokeskus. Yksinasuvien määrä kasvoi eniten vanhemmissa ikäryhmissä 2015. http://tilastokeskus.fi/til/asas/2015/asas_2015_2016-05-24_tie_001_fi.html, 2016.

[20] Topi Tjukanov. Elinkeinojen sijoittuminen muuttuvassa kaupunkirakenteessa - kehäradan vaikutukset vantaalla. Master's thesis, Helsingin Yliopisto, 2014.

[21] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2007. 8th printing.

[22] Vantaan Kaupunki, tietopalveluyksikkö. Vantaan väestö 2014/2015. Vantaan Kaupunki, Tietopalvelu B6: 2015.

A Appendix 1: Yhteenveto

Yhteenveto

Tämän kandidaatintyön tarkoituksena oli vertailla asuntojen vuokrahintoja neljän pääkaupunkiseudun alueen välillä hyödyntäen tilastollista analyysiä. Analyysiin valitut alueet olivat Helsingistä, Espoosta ja Vantaalta: Kamppi, Kallio, Matinkylä ja Myyrmäki.

Pääongelma tutkimuksessa oli valita hintojen vertailuun sopivat tilastolliset menetelmät ja sitä kautta luoda robusti menetelmä hintojen vertailuun eri kokoisia asuntoja sisältävien alueiden välillä. Yksinkertaiset ja yleiset sijaintiluvut, kuten otoskeskiarvo -ja mediaani antavat usein epätarkan kuvan, kun tutkitaan ja vertaillaan erilaisten asuntojakaumien sisältävien alueiden hintoja.

Menetelmät, joita tässä työssä käytettiin, olivat lineaarinen regressio käyttäen ”Least absolute deviations” -menetelmää ja Wilcoxonin järjestyssummatesti sekä Kruskal-Wallisin testit alueiden hintojen vertailuun.

Alkuperäinen aineisto hankittiin oikotie.fi:stä, joka on suosittu internet-sivusto asuntojen myynnin ja vuokrauksen ilmoittamiseen. Aineisto sisältää tiedon asunnon pyydetystä vuokrahinnasta, pinta-alasta ja kaupunginosasta. Kyseiset neljä aluetta valittiin tutkimuksen kohteiksi johtuen niiden odotetusta hintatasosta, maineesta, asuntoilmoitusten määrästä ja asuinalueiden erilaisesta hintaprofiilista.

Analyysissä keskityttiin yhden ja kahden makuuhuoneen asuntoihin kooltaan 20 – 60 m^2. Rajaus tehtiin yksiöihin ja kaksioihin johtuen niiden suuresta kysynnästä pääkaupunkiseudulla.

Alueiden asuntojen kokojen ja hintojen jakaumia tutkittiin ensin kuvailevasti käyttäen visuaalisia menetelmiä ja tilastollisia tunnuslukuja.

Huomattiin, että Kampissa on korkein mediaanihinta sekä suurin vaihtelu neliöhinnoissa. Kampissa oli myös muihin alueisiin verrattuna hyvin paljon eri kokoisia asuntoja. Sen sijaan esimerkiksi Myyrmäen aineistossa oli suhteellisesti paljon enemmän yli 40 m^2 kokoisia asuntoja kuin sitä pienempiä. Matinkylän asuntojakauma muistutti tässä mielessä Myyrmäkeä, sillä Matinkylässäkin suurin osa asunnoista oli yli 40 neliöisiä. Kallion tapauksessa jakauma oli toiseen suuntaan vino, ja siellä suurin osa asunnoista oli pienempiä yksiöitä.

Mediaanihinnoiltaan Kampin jälkeen suuruusjärjestyksessä tulivat Kallio, Matinkylä ja Myyrmäki. Keskenään samankaltaisempia hinnat ja asuntojen jakaumat olivat Kampin ja Kallion välillä sekä Matinkylän ja Myyrmäen välillä.

Alueiden asuntojen neliöhinnat ja pinta-alat piirrettiin kuvaajaan, jolloin niiden välillä huomatiin riippuvuus. Kaikilla alueilla asuntojen neliöhinnat laskivat, kun pinta-ala kasvoi. Todettiin, että alueiden hintatasoja voitaisiin vertailla ensin mallintaen riippuvuus käyttäen regressioanalyysia, ja sen jälkeen vertailemalla kunkin alueen hintaresiduaaleja tästä regressiosta. Visuaaliseen tarkkailuun perustuen todettiin, että riippuvuutta voidaan mallintaa lineaarisella regressiolla.

Lineaarinen regressio sovitettiin koko aineistoon, jolloin oletettiin, että kaikilla alueilla pinta-alan ja neliöhinnan välinen riippuvuus on samanlainen. Riippuvuuden mallintamisen hyötynä oli se, että alueet sisälsivät eri kokoisia asuntoja eriävissä määrin, jolloin suora hintojen vertailu alueiden välillä olisi vääristynyt huomattavasti riippuen asuntojen pinta-alojen jakaumasta.

Lineaarisen regression sovitusmenetelmäksi valittiin nk. ”Least absolute deviations” -menetelmä, joka on yleisesti käytettyyn pienimmän neliösumman menetelmään verrattuna robustimpi poikkeavien havaintojen suhteen. Least absolute deviations -menetelmä perustuu siihen, että lineaarisen mallin parametrit valitaan minimoimalla residuaalien itseisarvojen summa.

Tilastollinen testaaminen ja alueiden hintavertailu tehtiin lineaarisen mallin residuaaleilla. Residuaalien otosmediaaneja vertailemalla alueiden keskinäinen hintajärjestys säilyi samana kuin aikaisemmin. Kamppi erottautui nyt muista alueista aikaisempaa enemmän, kun taas muut kolme aluetta osoittautuivat hinnoiltaan keskenään samankaltaisemmiksi kuin alkuperäisen aineiston tunnuslukujen analyysin perusteella vaikutti. Kampin hintaresiduaalien otosmediaani oli 5,17 €/m^2 , Kallion 1,01 €/m^2, Matinkylän -0,34 €/m^2 ja Myyrmäen 0,44 €/m^2.

Alueiden residuaaleja testattiin kahdella erillisellä testillä, jotka olivat Kruskal-Wallisin testi ja Wilcoxonin järjestyssummatesti. Molemmat testit testaavat jakaumien identtisyyttä, ja koska ne ovat herkkiä mediaanien eroille, testit voidaan tulkita testeinä mediaanien identtisyydelle. Molemmat testit ovat epäparametrisia ja perustuvat järjestyslukuihin, joten ne eivät vaadi oletuksia residuaalien jakaumista.

Kruskal-Wallisin testiä käytettiin ensin testaamaan kaikkien alueiden hintaresiduaaleja yhdessä. Tämä testi tuotti hyvin pienen p-arvon, ja nollahypoteesi residuaalien mediaanien identtisyydestä hylättiin. Tulos ei ollut yllättävä, sillä jo aikaisemmat tunnuslukuvertailut viittasivat siihen, että alueiden välillä on huomattavia hintaeroja.

Wilcoxonin järjestyssummatestin tulokset selittävät Kruskal-Wallisin testin tuloksen hyvin, sillä kaikista paitsi Myyrmäen ja Matinkylän alueen välisistä testeistä saadut p-arvot viittasivat selvästi nollahypoteesien hylkäämiseen. Matinkylän ja Myyrmäen residuaalien välinen Wilcoxonin testi tuotti p-arvon 0,375. Tämän perusteella voidaan todeta, että nollahypoteesia mediaanien identtisyydestä ei tarvitse hylätä millään tavanomaisella luottamustasolla.

Tilastollisen testaamisen etu työn tuloksia käsitellessä on se, että ne voidaan yleistää populaatiotasolla. Tämä tarkoittaa, että mikäli Matinkylästä ja Myyrmäestä etsittäisiin lisää asuntoja vertailua varten, hintojen voitaisiin odottaa olevan edelleen identtiset.

Työ toteutettiin aineistopohjaisesti siten, että alkuperäisen kysymyksen asettelun jälkeen analyysin rajaus sekä menetelmien valinta ja perustelu tehtiin tutkimalla ensin aineistoa hyödyntäen kuvaajia ja tunnuslukuja. Työn tärkein vaihe oli lineaarisen regression tarpeen tunnistaminen ja hyödyntäminen asunnon pinta-alan ja neliöhinnan välisen riippuvuuden mallintamiseen. Tällä tavalla pystyttiin huomioimaan alueiden asuntojen kokojakaumat siten, että ne eivät vaikuttaneet hintojen vertailuun. Least absolute deviations –menetelmän valinta perusteltiin sen robustisuudella poikkeavia havaintoja kohtaan. Regressioanalyysin käyttäminen johti myös siihen, että hintatasojen vertailu toteutettiin residuaaleja vertailemalla.

Työn pohjimmaiseen kysymykseen alueiden hintaeroista pystyttiin vastaamaan selvästi. Myös tavoitteena ollut alueiden hintojen vertailuun sopivien tilastollisten menetelmien valinta ja soveltaminen onnistui. Odotusten mukaisesti, Kamppi osoittautui kalleimmaksi ja Kallio toisiksi kalleimmaksi alueeksi, mutta Matinkylän ja Myyrmäen hintojen samankaltaisuutta olisi tuskin havaittu ilman työssä käytettyjen menetelmien takaamaa tarkkuutta.