Universitat Pompeu Fabra Barcelona
Faculty of Economics and Business

Academic year 2020-2021

Final Year Project

The Rental Housing Market in Barcelona
A Nonparametric Analysis
Corrected & Printable Version

Cunquero Orts, Josep (Degree in Economics)

Tutors

Free Modality Tutor: Lugosi, Gabor

UPF Tutor: Gundin Castro, María

Code: EME12

Abstract

The quality of geolocation services and the amount of publicly available data on the internet have increased substantially in the last decade. This paper studies the Barcelonese rental real estate market by exploiting both geolocation services and user-generated public data. This allows for the creation of a dataset which is later analysed using kernel smoothing methods. The key findings are that (i) rent prices are approximately log-normally distributed, that (ii) rent prices are lower in areas further away from the city centre, and that (iii) the previous point has to be complemented with a geospatial analysis that accounts for spatial heterogeneity.

Keywords— econometrics, kernel smoothing, user data, housing market

Acknowledgements

I want to personally thank my tutors, Gabor Lugosi and María Gundin, for their time and their patience. I would also like to thank Prof. Dr. Cathrine Aeckerle-Willems for introducing me to the world of nonparametric econometrics; Dr. Federic Udina for aiding me in the practical application of kernel methods; and Dr. Ingo Steinke for counselling me in the relative to hypothesis testing.

This paper stands upon the shoulders of giants. In particular, modern scientific research is greatly indebted with the creators, publishers and maintainers of open-source software. They are the least recognised link in the scientific production chain, partially due to their anonymous and altruistic character, but also partly due to poor citation practices. I want to hereby explicitly thank the anonymous open-source software contributor.

Finally, I want to thank the reader, who takes her time to read this paper.

Contents

Acronyms

ANOVA	Analysis of Variance
API	Application Programming Interface
DGP	Data Generating Process
HTML	Hyper Text Markup Language
LOOCV	Leave-One-Out Cross Validation
MISE	Integrated Mean Squared Error
OHRW	Online Housing Rental Website
PDF	Probability Density Function
PEUAT	Pla Especial Urbanístic d'Allotjaments Turístics (Special Urbanistic Plan of Tourist Accommodations)
URL	Uniform Resource Locator

Glossary

$\vec{b}$	A vector
$\boldsymbol{B}$	A matrix
$\boldsymbol{B}^T$	The transpose of matrix $\boldsymbol{B}$
$\boldsymbol{B}^{-1}$	The inverse of matrix $\boldsymbol{B}$
$\triangleq$	Equals by definition
$\|\bullet\|$	Euclidean norm
$\ln$	Natural logarithm
$\neg p$	Not p, i.e., proposition p is false
$\mathbb{D}$	The domain of a function
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers

1 Introduction

> "*We do not want a Spain of proletariat, but a Spain of [home] owners.*"
> Jose Luís de Arrese, Spanish Minister of Housing 1957-1960 [De Arrese, 1959]

Despite being the country with the second highest rate of home ownership in Europe [Sweeney, 2021], the Spanish share of population that owns its own accommodation has been consistently falling since 2011 [Romero, 2020], and currently lies at the minimum since the series started in 2004 [INE, 2019]. This tendency has accentuated the social and economic weight of alternative tenancy modalities, in particular rental tenancy. The rental housing market has an especially important role in the social fabric of urban areas, as it offers housing opportunities to specific social groups—for instance migrants, young workers or new marriages—which due to precarious economic situation or uncertainty, cannot, or are not willing to, face the costs of ownership [Hu, He, Han, Xiao, Su, Weng, and Cai, 2019]. In the city of Barcelona, 30% of all accommodations were in a rental agreement in 2016, a number that escalates to even 55% in certain neighbourhoods [Gabinet Tècnic de Programació, 2016]. These numbers should convince the reader that a study of the housing rental market in Barcelona is key to understand, among other things, how accessible housing is for a substantial share of the population.

The first objective of this study is to estimate the density distribution of the rental prices in the city of Barcelona, both in rent prices and in rent per m^2. This estimation is conducted by means of kernel smoothing. The second objective is to study the geospatial distribution of rental prices per m^2. This second goal is less theoretically bounded and allows for different approaches. Two analysis will be conducted: the distribution of rent per m^2 with respect to peripherality, and with respect to the latitude-longitude space; both of them by means of kernel smoothing as well.

The data on which the analysis is conducted is extracted from various OHRWs via web scraping. In contraposition to official census records, this methodology provides access to large quantities of real user-generated data, which is rich in detail and up-to-date. Indeed, the advertisers have an incentive to provide as much information as possible to attract potential tenants, which results in very complete dataset. On the other hand, the resulting datasets are subject to a series of potential statistical obstacles, which among others are (i) sampling biases, (ii) user-generated inaccuracies and errors, and (iii) the presence of duplicates. Because of this, the datasets are thoroughly cleaned before any analysis. But even after the clean-up of the data, the character of this study is purely descriptive, and a causal analysis is not pursued. Causal inference on rent prices would require a different set of statistical tools and a more specifically planed dataset to control for external variables, such as attractions near an accommodation.

The results show that rents, in monthly prices and in m^2 prices, follow strongly skewed distributions with thick right tails, that behave approximately normally after taking logarithms. Moreover, the study also reveals that rents fall with distance with respect to the city centre, but that this generalisation has to go hand in hand with a less simplifying description — for instance, with an analysis on how rent prices are distributed in the geographic space, which is also conducted in this project.

This paper is part of an ever growing current of literature that uses data from web portals to draw conclusions about housing markets. Similar papers are [Rae, 2015] for the UK , [Boeing and Waddell, 2016] for the US, [Schernthanner, Steppan, Kuntzsch, Borg, and Asche, 2017] for Berlin, [Chapelle and Eyméoud, 2018] for France, [Clark and Lomax, 2018] for England, [Loberto, Luciani, and Panagallo, 2018] for the Italian sale market, [Hu, He, Han, Xiao, Su, Weng, and Cai, 2019] for Shenzhen, [Equip Observatori Metropolità de l'Habitatge de Barcelona, 2020] for Barcelona and [Garcia-López, Jofre-Monseny, Martínez-Mazza, and Segú, 2020] for the rental tourist accommodations in Barcelona. This new, progressively more established research field is based on the consensus that the results are valid as long as (i) a carefully laid protocol to weed out duplicates and errors is implemented and (ii) the results are validated by external sources.

The rest of the paper proceeds as follows. In Section 2 the data, as well as its gathering and cleaning processes are described. Section 3 presents the statistical methods used to analyse the data. In section 4 the results are presented and discussed. Section 5 concludes.

2 Data

> "*And most studies have relied on self-reports from participants, which suffer from cognitive biases [...]. For the first time, we can begin to observe the real-time interactions of people [...].*"
> Duncant J. Watts [Watts, 2007]

The data for this project was extracted from OHRWs via web scraping. Their selection was based essentially on two criteria: popularity (i.e., sheer size) and availability of data; being the latter the most crucial one. Since the ethical and legal implications of web scraping are still not well defined, I limited myself to those portals that did not explicitly ban the automated extraction of public information from their websites. Moreover, some OHRWs do not publish certain pieces of information about advertisements. Those that omitted crucial information, such as the last update or the address, were excluded. Four OHRWs, namely Fotocasa, Habitaclia, Idealista and Pisos, were finally selected and scraped. The scraping, a time-consuming process, was carried out the 3rd and 4th of May, 2021.

The average advertisement consists of a few pictures of the accommodation; the price, surface and other qualities; a user-typed description and a physical address. The large volume of advertisements means that manually checking each one of them is not feasible, and that the process has to be automated. The scraping process heavily depends on the display structure of each OHRW, and in consequence the process had peculiarities for each one of the four portals analysed. This means a different Python program had to be written for each of the 4 OHRWs. To access the code, see Appendix E. The general procedure, however, is as follows. First it is imperative list all URLs for all the individual advertisements to be analysed. Once this is done, for each advertisement a request is sent to the OHRW to access the page corresponding to the URL. Once the access is granted, the content of the page, in HTML format, is downloaded. The HTML file contains, in a raw state, all the information of a webpage, as well as information about its display. After storing the file, the HTML is interpreted. A script written beforehand so that it is tailored to the structure of the portal then grabs specific pieces of information that correspond to the qualities of the accommodation. Because all the advertisements of a particular OHRW have the same display structure, the qualities of the accommodation are always stored in the same containers. These regularities are key to be able to automatise the procedure. This process is run for every URL harvested and the data is then stored in an Excel file. The detailed procedure is described in Appendix A.

One of the goals of the paper is to assess whether rent prices are higher in the centre of the city or in the periphery. In order to do so, it is imperative to first define what geographic point is the centre of Barcelona. Plaça Sant Jaume is taken to be the city centre, so that the city divisions enforced in this paper are harmonic with the PEUAT, a city planning project established by the City Council of Barcelona that divides the city into zones depending on the intensity of the prevalence

of tourist accommodations, being zone 1 the one with highest pressure, and 3 & 4 the ones with the lowest [Blanco-Romero, Blázquez-Salom, and Cànoves, 2018]. Accordingly, if we draw two concentric circumferences on the map of radii 1.4 km and 3 km centred on Plaça Sant Jaume, the area in the smallest circumference (the centre), the area in the biggest circumference and not in the smallest circumference (the middle ring), and the area not included in any circumference (the periphery) *loosely* correspond to the PEUAT zones 1, 2, and 3 & 4, respectively (see Appendix C for an illustration). Although the high tourist pressure – high housing rent prices relationship is not necessarily a one-to-one relationship, the PEUAT zones provide a useful benchmark to define different areas of the city with varying degrees of rental demand, which later will be analysed differently. The second step is to geocode each observation; that is, to assign each observation a pair of accurate latitude-longitude coordinates. This is achieved using the geocoding API from Google Maps. This service from Google geocodes each accommodation given a locator text string[1]. Once all the observations are geocoded, calculating their distance with respect to the city centre is but a mathematical question. The geocoding process is automated using Python.

2.1 Cleaning of the Data

User-generated data is prone to duplicates, errors and incomplete observations. Therefore, an in-depth cleaning of the data is required. The guiding principle has been to avoid setting artificial restricting upper caps on variables, to avoid falling into reductionism. Geospatial data is characterised by containing extreme observations, and omitting them would severely harm the representativeness of the resulting dataset. The only restrictions that are applied are those that, if violated, imply that crucial information is omitted, imply the observation is out of the scope of the study, or *necessarily* imply an error. Therefore, those observations

- that do not provide a value for the rent price, the surface or the last update;
- the last update of which was more than a month ago (not representative of the current market anymore);
- that have a surface of under 10 m^2 (not an accommodation);
- that have a rent per m^2 of under 5 Euros;
- the rent price of which is over 50.000 Euros (user should have selected "sale", instead of "rent");
- that are more than 9 km away from the city centre (hence not in Barcelona);
- that, if applicable, have more than 30 bedrooms (probably a whole building or mistake);
- that, if applicable, have more than 20 bathrooms; or
- that are duplicates (see Appendix B for duplicate detection)

are dropped out of their dataset. See Appendix D for a complete description of the final datasets.

[1] In most cases in this study, an address. If missing, the neighbourhood.

2.2 Four Samples — One Population?

Throughout the paper the four OHRWs are analysed separately. This is done because of a variety of reasons. First, there is no a priori reason to believe that the observations from the different portals come from the same population or DGP [Equip Observatori Metropolità de l'Habitatge de Barcelona, 2020]. Second, each dataset has a different set of variables and peculiarities, such as, for instance, different property subtype classifications. Lastly, because there is no reason to believe that a single accommodation cannot be advertised in different OHRWs. Because of the different structure of each dataset, duplicate detection in a pooled dataset would be far too complicated.

Whether or not the different datasets come from the same DGP is actually a testable hypothesis. In order to test whether two samples come from the same distribution the Kolmogorov-Smirnov test can be used. The version of the Kolmogorov-Smirnov used here tests

- H_0: The samples are drawn from the same distribution,
- H_1: $\neg H_0$.

This is tested by means of the Kolmogorov-Smirnov statistic, defined as

$$D_{n_1,n_2} = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)|, \tag{2.1}$$

where sup is the supremum, and F_{1,n_1}, F_{2,n_2} are the empirical cumulative distribution functions of the two samples to be tested. This specification is tested on the variable rent per m^2, since it is arguably the variable of highest relevance in the paper. To run the tests, a normally distributed mean zero random component with minuscule variance is introduced to eliminate ties.

Table 2.1: p-values for the Kolmogorov-Smirnov test for each pair of samples

	Fotocasa	Habitaclia	Idealista	Pisos
Fotocasa		6.333×10^{-5}	0.427	0.060
Habitaclia			8.947×10^{-9}	4.785×10^{-7}
Idealista				0.675
Pisos				

Some conclusions can be extracted from the results on Table 2.1. First, that it was justified to doubt that the observations from the four OHRWs come from the same distribution, and that is justified to conduct a discriminated analysis. Second, that the Idealista-Fotocasa and the Idealista-Pisos pairs seem to originate, respectively, from the same population for each pair and, at a 5% confidence level, so does the Fotocasa-Pisos pair. Third, that the Habitaclia dataset comes from a radically different DGT from the others.

It is not clear why the Habitaclia dataset stands out so clearly. Here I provide two possible explanations. The first one is that perhaps the structure of the portal itself or the requirements to post an advertisement somehow self-select the type of advertiser or make the OHRW more prone to be filled with duplicates or statistical anomalies that might have gone unnoticed. The second hypothesis is that a single luxury real estate agency might post the advertisements exclusively in Habitaclia, thus unbalancing the sample.

2.3 Categorical Variables

Even though the paper mainly focuses on continuous variables, the categorical variables should be shortly discussed. As expected in a densely populated city like Barcelona, between 70% and 90% of the accommodations, depending on the OHRW, are labelled as flats. A further 8% to 20% corresponds to apartments, and between a 1% and as high as a 4% correspond to studios. Under 1% of all advertisements correspond to houses or chalets. Finally, a myriad of other categories fill the remaining shares. It is also worthwhile mentioning that although OHRWs give individuals the power to take the renting of their accommodation in their own hands, between 88% and 99% of all advertisements are labelled as published by real estate agencies or professionals. This could imply that in fact, the price-setting decisions are taken by a smaller number of parties than what it could appear to be at first glance.

2.4 Validity of the Data

A final step is required to generalise the results of the study to the whole real estate rental market of Barcelona, and that is to check if the data accurately corresponds with official reports and census. There are several possible approaches. [Chapelle and Eyméoud, 2018], for instance, calculate the share of observations for each geographic subdivision with respect to the total, and compare these shares with the ones in the governmental registries, to ascertain whether there is infra- or over-representation in each geographic subdivision. Here a simpler comparison is conducted, namely comparing the sample average rent per m^2 with the official governmental statistics elaborated by [INCASÒL, 2021].

A one-way ANOVA test rejects at every usual confidence level the hypothesis that the four samples have simultaneously an average rent per m^2 equal to the highest quarterly mean rent per m^2 in 2020, namely 14.3 Euros per m^2. As a matter of fact, each individual one-sample t-test also unmistakeably reject the null hypothesis that the average rent per m^2 is equal to 14.3. A closer look reveals that both the average surface, and especially the average rent price, are higher for every sample than in governmental statistics.

There are several explanations for this lack of coherence. One factor can be that higher quality, higher price or larger size accommodations are over-represented in OHRWs [Thomschke, 2015]. Another factor could be that landlords may advertise a given rent, but during the bargaining process with the potential tenant, the final agreed rent takes a lower value. The latter explanation would contradict with several studies. According to [Saiz, 2010], in urban areas under geographic constraints, such as Barcelona[2], the housing supply is especially inelastic, hence tenants have little to no bargaining power. The authors of [Chapelle and Eyméoud, 2018] furthermore argue that advertised rents converge with real rents because (i) the availability of public information on rents generates strong competition forces among landlords, and (ii) tenants are the most impatient party and therefore tend to adopt price-accepting roles. Yet in the end the last explanation seems to hold true. In the official report [Equip Observatori Metropolità de l'Habitatge de Barcelona, 2020], the authors estimate the "supply" and the "demand" for two OHRWs for the city of Barcelona. They estimate the supply price by averaging the rent prices advertised by landlords, and the demand price by averaging the requests

[2] Barcelona is delimited by the Besòs river to the North, the Llobregat river to the South, the Mediterranean to the East and Collserola to the West.

that potential tenants send to the advertisers. The average supply rent they obtain is 1267 Euros per month, quite similar to the averages found in this study (see Appendix D). The average demand rent they obtain, however, is 857 Euros per month. This little correspondence between supply and demand could mean that only lower-than-average priced accommodations are rented, or that tenants do have a substantial bargaining power and are able to drive rents down. In any case, the lack of correspondence can very well explain the difference between advertised rents and contract rents. The matter is nevertheless far from concluded and the arguments sketched above are but hypotheses.

The conclusion is that the results of this paper clearly *cannot* at any rate be generalised to the whole market. From this point onwards, any references to "rent prices" or "rent per m^2", referring to the variables of the dataset, allude, if anything, to the observed advertised prices, or alternatively, to the supply side of the market.

3 Statistical Methodology

> "*Everything is related to everything else, but near things are more related than distant things.*"
> Tobler's First Law of Geography [Tobler, 1970]

Essentially only two kinds of statistical analysis are conducted: (i) density estimation, and (ii) regression analysis; both of them using kernel smoothing methods. A kernel is a mathematical function $K : \mathbb{R} \to \mathbb{R}$ fulfilling that $\int K(u)du = 1$. In particular, throughout the paper only the Gaussian kernel is used, defined as

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \tag{3.1}$$

3.1 Density Estimation

Throughout the paper it is often attempted to estimate the density function $f(x)$ of a random i.i.d. sample $x_1, \ldots, x_n$, $n \in \mathbb{N}$ in $\mathbb{R}$, by means of an estimator $\hat{f}(x)$. I take the Parzen–Rosenblatt estimator in the spirit of Tsybakov [Tsybakov, 2009]. For a point $x_0 \in \mathbb{R}$, $\hat{f}(x_0)$ is defined as

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\|x_0 - x_i\|}{h} \right), \tag{3.2}$$

where $h > 0$ is the user-defined bandwidth. In practice $\hat{f}(\bullet)$ is evaluated at a grid of values in a meaningful interval.

The optimal choice of the bandwidth is of extreme importance to obtain a curve that is properly tailored to the data. In the context of density estimation, in this paper the plug-in method proposed by [Cao, Cuevas, and González Manteiga, 1994] is used. Using this method, the bandwidth is chosen so that it minimises the MISE, defined as

$$\text{MISE} \triangleq E_f \left[\int \left(\hat{f}_n(x) - f(x) \right)^2 dx \right]. \tag{3.3}$$

It can be shown that the bandwidth that minimises Expression 3.3 has asymptotically the form

$$h_0 = d_k^{-2/5} c_k^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5} n^{-1/5}, \tag{3.4}$$

where

$$d_k = \frac{1}{2}\int u^2 K(u)du, \qquad c_k = \int K^2(u)du \tag{3.5}$$

are constants. Since the term $\int f''(x)^2 dx$ is unknown, Cao et al propose substituting it by

$$\hat{S}(p) = n^{-2}p^{-5}\sum_{i,j} K^{iv}\left(\frac{\|x_i - x_j\|}{p}\right), \tag{3.6}$$

where p is an auxiliary window defined as

$$p = \left(\frac{2K^{iv}(0)}{d_k}\right)^{1/7} \hat{T}^{-1/7} n^{-1/7}, \tag{3.7}$$

where $\hat{T}$ is an estimator of $T = \int f'''(x)^2 dx$, estimated by assuming normality of distribution. The plug-in method consists simply in estimating h_0 in Expression 3.4 by means of "plugging-in" Expression 3.6 instead of the unknown term. The result is the optimal plug-in bandwidth h_{PI}. Unless explicitly stated, in all instances of density estimation the plug-in approach exposed above is used to choose the bandwidth.

3.2 Regression

In some occasions in the paper it will be interesting to study how a dependent variable y behaves in the space of a set of independent variables $x_1, \ldots, x_d$, $d \in \mathbb{N}$. In this case, I use the local radial linear approach as described by [Hastie, Tibshirani, and Friedman, 2017]. Assuming there exists a continuous function $g : \mathbb{R}^d \to \mathbb{R}$ such that $y = g(x_1, \ldots, x_d) + \varepsilon$, with

$$E[\varepsilon] = 0, \qquad Var(\varepsilon) < \infty, \tag{3.8}$$

the function g at a point $x_0 \in \mathbb{R}^d$ given a sample of size $n \in \mathbb{N}$ can be estimated by

$$\hat{g}(x_0) = \overrightarrow{b(x_0)}^T \left(\boldsymbol{B}^T \boldsymbol{W}(x_0)\boldsymbol{B}\right)^{-1} \boldsymbol{B}^T \boldsymbol{W}(x_0)\vec{y} \tag{3.9}$$

$$= \sum_{i=1}^{n} l_i(x_0) y_i, \tag{3.10}$$

where

- $\overrightarrow{b(x_0)} = (1, x_{0,1}, \ldots, x_{0,d})^T$ is a $((1+d)\times 1)$ vector,
- $\boldsymbol{B}$ is a $(n \times (1+d))$ matrix where the i-th row is $\overrightarrow{b(x_i)}^T$,
- $\boldsymbol{W}(x_0)$ is a diagonal $(n \times n)$ matrix where $\boldsymbol{W}_{i,i}(x_0) = K\left(\frac{\|x_0 - x_i\|}{h}\right)$,
- $\vec{y}$ is the $(n \times 1)$ vector of values of the dependent variable, and
- $l_i(x_0)$ is the weight given by the procedure to the i-th observation.

The estimation of $\hat{g}(\bullet)$ is dependent on the invertibility of the square matrix $\boldsymbol{B}^T\boldsymbol{W}(x_0)\boldsymbol{B}$, which is only violated due to computational constraints at estimation at regions with acute data sparsity.

Again, in practice the function $\hat{g}(\bullet)$ is evaluated for a grid values. As it is the case with density estimation, the choice of the bandwidth is quite transcendental. Here I take the approach proposed by [Li and Racine, 2007]. Let the LOOCV error be defined as

$$\text{CV} \triangleq \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}_{-i}(x_i))^2, \tag{3.11}$$

where $\hat{g}_{-i}(x_i)$ is the evaluation of Function 3.9 at x_i, but leaving x_i out of the calculation. The bandwidth is chosen such that it minimises Expression 3.11:

$$h_{CV} \triangleq \underset{h>0}{\arg\min}\, CV(h). \tag{3.12}$$

Unless explicitly stated, the LOOCV method exposed above is used in all instances of nonparametric regression. Because of the large computational cost of estimating the value of the CV function, it is only calculated for a smaller random sample of observations. Finally, because the dependent variable often contains several repeated values, which is something intrinsic of the nature of the data and interferes with the bandwidth selection process, the LOOCV method is run on a variation of the dependent variable vector, where a random normally distributed component with mean zero and minuscule variation is introduced.

3.3 Remarks

A nonparametric approach is taken in this paper because kernel smoothing works remarkably well when there is uncertainty about the particular specification of the function to be estimated. It allows for a *atheoretic* approach to the subject of study, without having to adopt any theoretical framework or a priori assumptions, besides differentiability conditions. Moreover, kernel smoothing is particularly well suited for geospatial data. The property of the Gaussian kernel that

$$\frac{\partial K\left(\frac{\|x_0 - x_1\|}{h}\right)}{\partial\, \mathrm{d}(x_0, x_1)} < 0, \tag{3.13}$$

where $\mathrm{d}(\bullet, \bullet)$ is a distance function, perfectly harmonises with Tobler's First Law of Geography [Tobler, 1970]. Consequently, the estimation at a point is mainly dependent on its immediate environment, almost irrespectively of the behaviour of the function at regions further away, where the data might be of a radically different quality.

All the statistical procedures explained in this section are applied by means of a set of functions I have defined myself on the programming language R. See Appendix E to access the code.

4 Results

> "*No average location exists on the Earth's surface.*"
> Bin Jiang [Jiang, 2015]

This section presents and discusses the results of the study and is structured as follows. The first subsection deals with the distribution of the key variables of the study, mainly through, though not limited to, the estimation of their density functions. The second part discusses how the distributions change with respect to peripherality, but since the complexity and richness of geospatial data can hardly be captured by a one-variable analysis, the third subsection shows how the key variables behave in the latitude-longitude space.

4.1 Density Distributions

A distribution function maps from the sample space to the *relative* likelihood of an event (e.g., a variable taking a certain value) happening. It is useful to judge which ranges of values are more common than others. Figure 4.1 displays the distribution functions of arguably the two most important variables of the study, the rent price and the rent price per m^2. Both variables are very similarly distributed. They take unimodal distributions, i.e., they are characterised by a single peak, which concentrates in its immediate neighbourhood the vast majority of observations. The PDFs are asymmetrical, strongly positively skewed, with a long right tail that extends far beyond the plotted interval[1]. The variable reaches very quickly the peak and then slowly drops. Their distribution is characterised by a relatively large variance, stretched by some extreme values; and by a slightly uninformative mean value, which takes a larger value than the median. Like many economic variables, the two variables seem to behave log-normally, as shown by the fact that the natural logarithm of the variables has an approximately normal PDF (see Figure 4.2).

The shape of the PDF has social distributive implications. The rapid, almost vertical growth of the curves right before the peak means that there are very few surprisingly cheap flats, in contrast with a relatively wide selection of incredibly expensive accommodations. Potential tenants in the search for an inexpensive dwelling might experience that there is a tangible "floor" on rental prices, around 600 Euros per month or, alternatively, 10 Euros per m^2. Accommodations below those lower bounds are extremely rare. If actual rental prices behave similarly than advertised rental prices, then a policy implication is that price ceilings would hardly increase the supply of inexpensive accommodations. Instead, the upper cap would simply sever the tail and reduce the spread of the distribution[2].

[1] The plots sacrifice showing the total length of the tail in exchange for a better definition of the peak
[2] Unless the rent ceiling would be set to take a radically low value, of course.

Examination of the plots shows that there are no big differences among OHRWs. If anything, the sample from Habitaclia seems to be even more skewed, with a higher mean, and a tail that is thicker and that fades away slower than for the remaining three samples. As for the other samples, the curves for Fotocasa and Idealista seem to be especially similar. These findings corroborate the results of the testing in Section 2.2.

In this case, estimating the density distribution with a smaller than optimal bandwidth (Plots 4.1c and 4.1d) reveals interesting patterns. A close inspection of Plot 4.1c shows that there exist regular peaks around round numbers. In particular, there is a remarkable drop after the value 1000, and strong peaks at the values 900, 1100, 1500, and at every following multiple of 500. This clustering of observations suggest that price-setters are prone to set round prices[3] if the "expected" price is close to a round number. Intuitively, round numbers are more attractive in terms of marketing strategies. This also implies that price-setters take the monthly rent to be the important variable, and let the rent per m^2 be determined by the division by the surface, and not the other way around. This behaviour does not seem to be universal, however, given the sudden peak around the rent per m^2 value of 20 in Plot 4.1d . There appears to exist a minority of landlords who set the rent per m^2 and let the monthly rent be determined by the multiplication by the surface. This predisposition to round prices seems to hold across the different OHRWs.

Figure 4.2 displays the PDFs of the natural logarithm of the key variables. As already hinted, their approximately normal shape indicates that the original variables might be accurately modelled by a log-normal distribution. The similarity is not perfect, however. Even after the logarithmic transformation, the distributions still have a slight positive skew. Also note that the marked spike in Plot 4.2c at exactly the value 7 is accidental. The drop just before corresponds to the drop in Plot 4.1c right after the value 1000, and the spike is explained by the fact that $e^7 \approx 1100$.

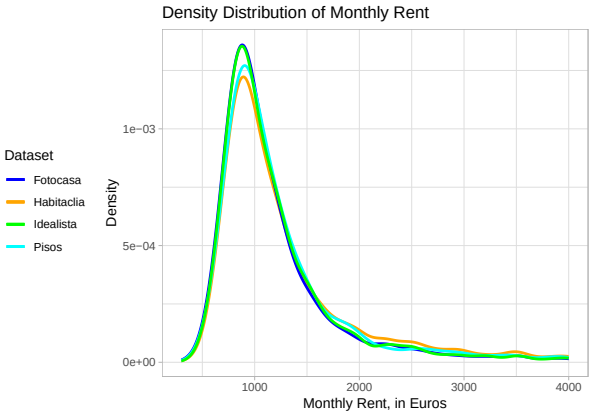
The figures and explanation until here provide a detailed description of the variables monthly rent and rent per m^2, but fail to convey a sense of the length of the tail and the presence of extreme values. I attempt to close this gap by means of Figure 4.3. Plot 4.3a shows that a small proportion of observations take extremely high values, roughly between 2 and as far as 10 times higher than the median values. The curve in Plot 4.3a immediately drops and reaches an ordinary value, which then very gradually decreases. The plot clearly differentiates between two regions: the vertical fall at the beginning, corresponding to the surprisingly expensive (per m^2) accommodations, and the flat region of the plot, corresponding to the body of the distribution. It is also worthwhile mentioning that, sample sizes notwithstanding, the Habitaclia sample stands out as the sample with more outliers: around 1% of all accommodations have a rent per m^2 over 50, while this proportion is around or under 0.5% for the rest of distributions. It is also the case that the curve described by the Habitaclia sample sinks slower, even correcting the sample size differences, than for the other datasets.

This concludes the analysis of the PDFs of the key variables of the study. The following section deals with how the aforementioned distributions change with respect to the distance to city centre.

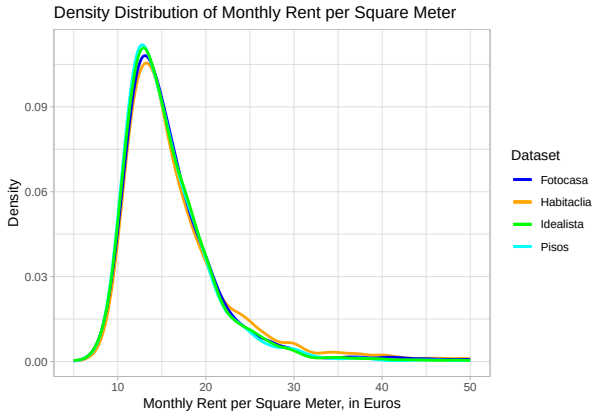
[3] Whether landlords tend to round more upwards than downwards is difficult to assess.

Figure 4.1: Densities of key variables.

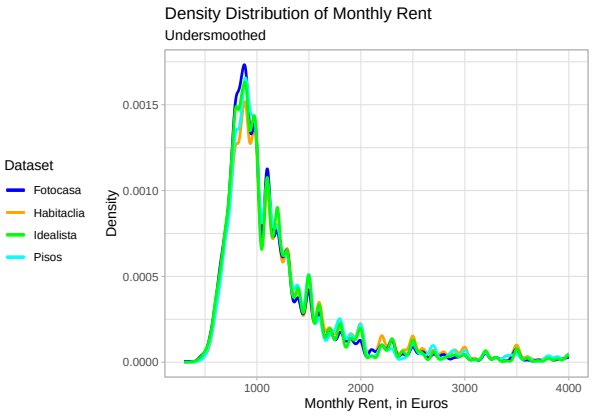
(a) PDF of rent on $\mathbb{D} = [300, 4000]$.



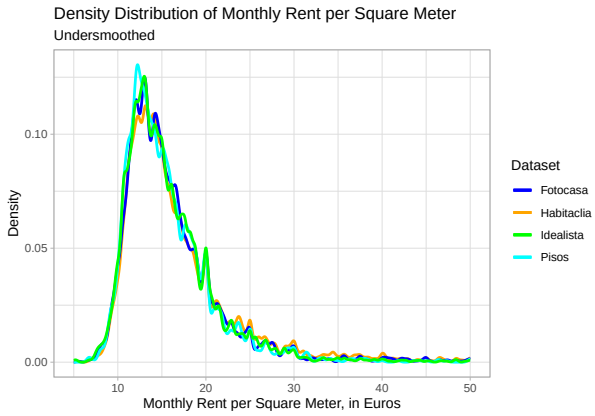
(b) PDF of rent per m^2 on $\mathbb{D} = [5, 50]$.



(c) PDF of rent on $\mathbb{D} = [300, 4000]$.



(d) PDF of rent per m^2 on $\mathbb{D} = [5, 50]$.



4.2 The Role of Peripherality

The first approach at studying the geospatial distribution of rent prices involves exploiting the classification presented in Section 2. Accordingly, we can divide all observations in three groups:

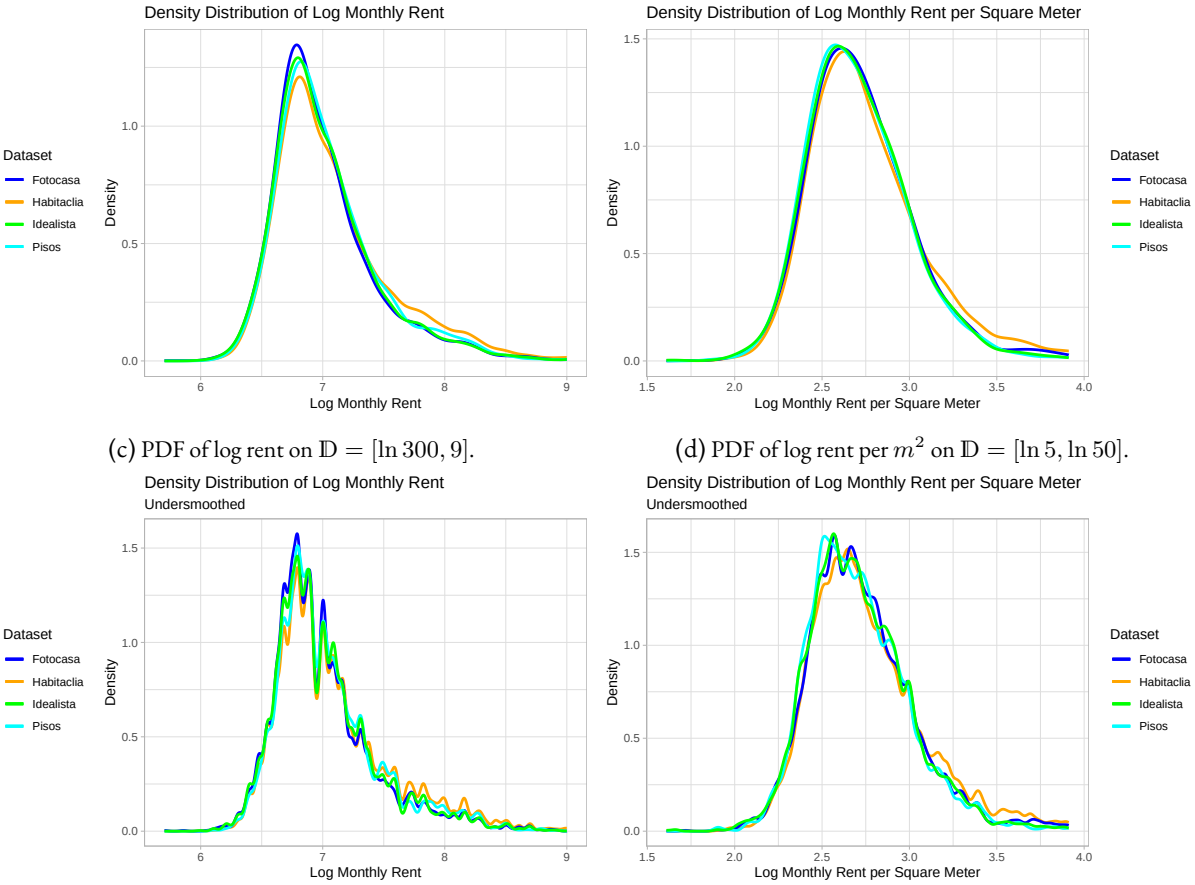
- Centre: Accommodations with a distance with respect to city centre under 1.4 km.
- Middle Ring: Accommodations with a distance with respect to city centre between 1.4 and 3 km.
- Periphery: Accommodations with a distance with respect to city centre over 3 km.

Figure 4.4 displays the PDFs of the natural logarithm of rent per m^2 by Distance Group for all four OHRWs. At first glance there appears to be a clear pattern. The further away from the city centre, the farther to the left that the distribution tilts. Indeed, it holds for all OHRWs that the mean monotonically decreases as we switch from the more centred group to the more peripheral group. In some OHRWs, the distribution for the Centre group is even negatively skewed.

Figure 4.2: Densities of log key variables.

(a) PDF of log rent on $\mathbb{D} = [\ln 300, 9]$.

(b) PDF of log rent per m^2 on $\mathbb{D} = [\ln 5, \ln 50]$.



(c) PDF of log rent on $\mathbb{D} = [\ln 300, 9]$.

(d) PDF of log rent per m^2 on $\mathbb{D} = [\ln 5, \ln 50]$.

The same simple pattern does not arise with the variances of the distributions. One would expect that the accommodations in the city centre would be more homogeneous in prices, and see a progressive increase in diversity accommodation typologies, and hence prices, as we move further away from the city centre. As a matter of fact, this is what happens in the Habitaclia sample, and less acutely in the Idealista sample. But contrary to the intuition, in the Fotocasa sample the maximum standard deviation is achieved in the middle ring group, and in the Pisos dataset, this precise group is the one with the lowest standard deviation.

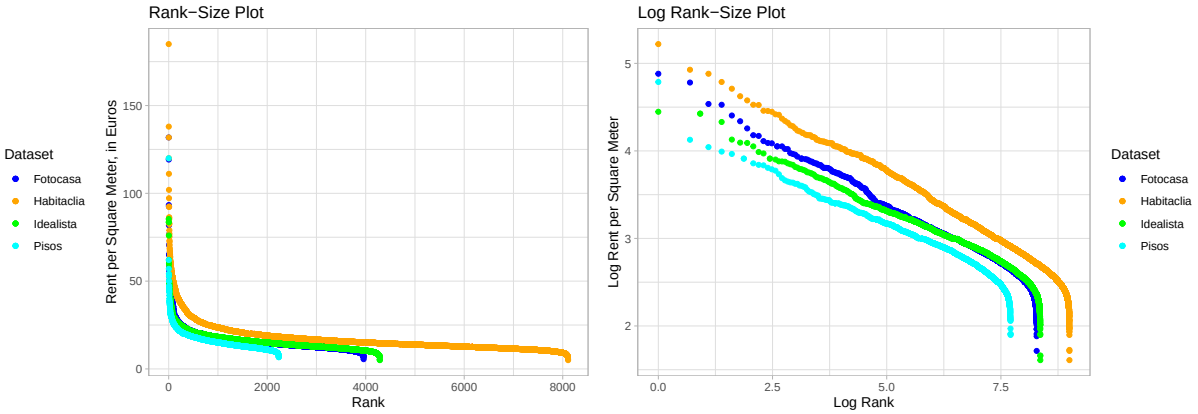
The division of the samples in three groups is an illustrating introduction to understanding the distribution of rent prices across the Barcelona. It already hints at the fact that the relation between peripherality and rent prices is strongly nonlinear. The next step is to run a nonparametric regression of rent per m^2 on distance to city centre. The results are displayed in Figure 4.5.

The regression in Figure 4.5 clearly shows that the overall tendency is that prices are lower the further away from the city centre. But this tendency is far from linear, monotonic or universal. The curves for the different OHRWs display constant ups and downs that could very well correspond to clusters of accommodations with similar characteristics, or, in other words, neighbourhoods. In particular, the "hill" at around 3 km could correspond to the Gràcia and Front Marítim del Poblenou

Figure 4.3: Rank analysis assigns each observation a rank corresponding to its relative position inside the sample. The observation with the highest value gets rank one, the observation with second highest value gets rank two, and so on.

(a) A rank plot plots the rank in the x-axis and the value of the observation in the y-axis. It is useful to display variables that are characterised by a few extreme values and a majority of low values.

(b) The approximately linear form of the curve might indicate that the distribution of rent prices per m^2 could comply with Zipf's law [Kyriakidou, Michalakelis, and Varoutas, 2011]. Zipf's law, in its simplest form, dictates that the second rank observation takes half of the value of the observation with first rank, the third rank observation takes a third of the value of the observation with first rank, and so on.



neighbourhoods; and the spikes between 5 and 7 km from the city centre could correspond to the Upper Diagonal districts. The results for distances above 6, specially the sudden increase at the end of the domain, should be taken with a grain of salt, however. The reason is that this is a region with a substantial data scarcity. The divergence between the different OHRWs indicates that at the boundary of the domain, the regression is mainly dependent on a few points, which could correspond to poor and isolated neighbourhoods (Idealista) or mountain chalets (Pisos). Furthermore, the slight increase of the curves at the lower bound of the domain (between 0 and 0.5 km with respect to the city centre) indicate that the city centre is not as homogeneous as one might have thought, and that the most inner nucleus of the city actually has lower rent prices than the rest of what we call the historic centre[4]. A potential reason why is that buildings in this area are old and cramped, needing renovations or lacking modern characteristics such as efficient heating or elevators.

It is also remarkable that advertisements in the portal Habitaclia have consistently higher prices per m^2 in the range $[0, 4]$ than the rest of OHRWs, despite the fact that all of them seem to follow approximately the same tendency and the same patterns of ups and downs. The interval $[0, 4]$ is also the most densely populated range which, together with the fact that the density distributions of distance to city centre for the different OHRWs are almost identical, could imply that for accommodations with similar characteristics, prices are higher if advertised in Habitaclia.

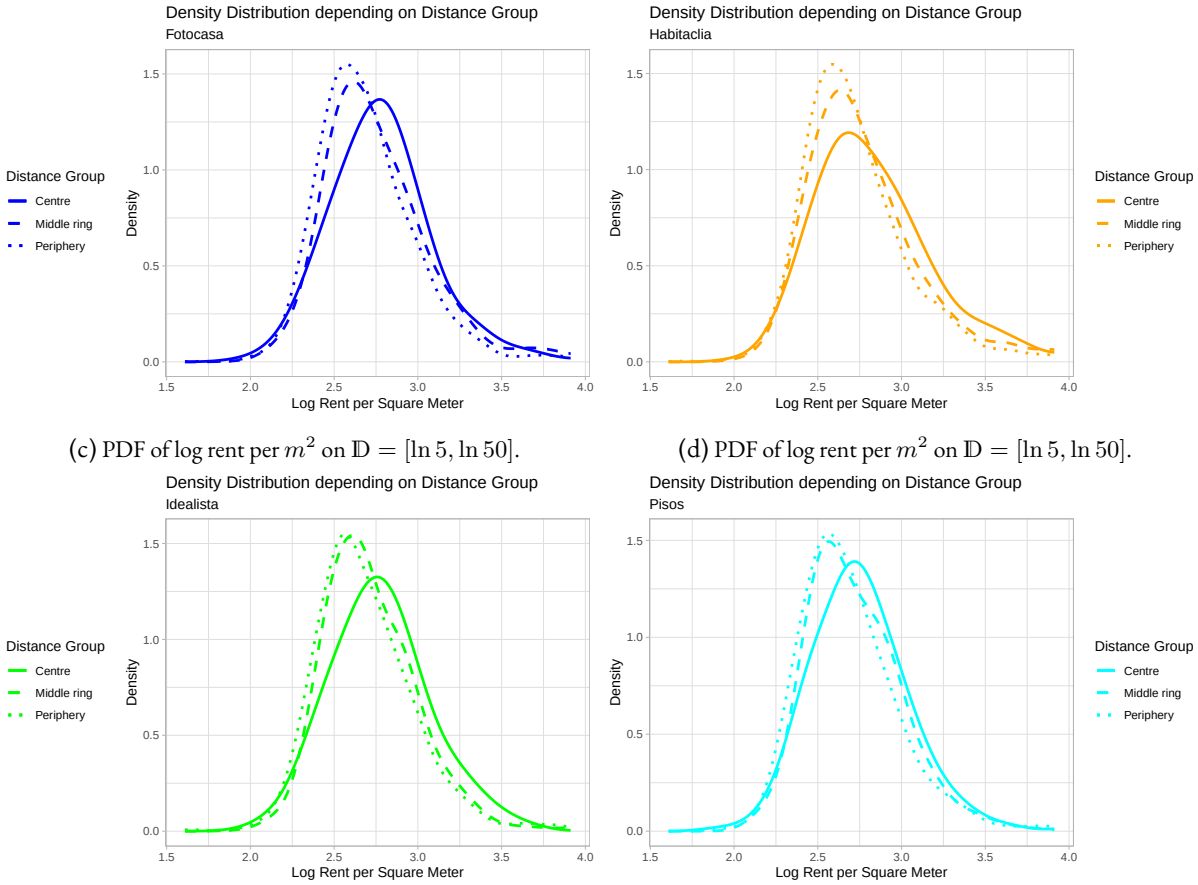
This section should not be concluded without mentioning that the results are sensitive to the choice of what the city centre is. This is the reason why I have tried not to define what the centre is in an arbitrary way, and why I have based the decision on the City Council's planing model, but the

[4] That is, the area that used to be inside the city walls.

Figure 4.4: Densities depending on Distance Group.

(a) PDF of log rent per m^2 on $\mathbb{D} = [\ln 5, \ln 50]$.

(b) PDF of log rent per m^2 on $\mathbb{D} = [\ln 5, \ln 50]$.



(c) PDF of log rent per m^2 on $\mathbb{D} = [\ln 5, \ln 50]$.

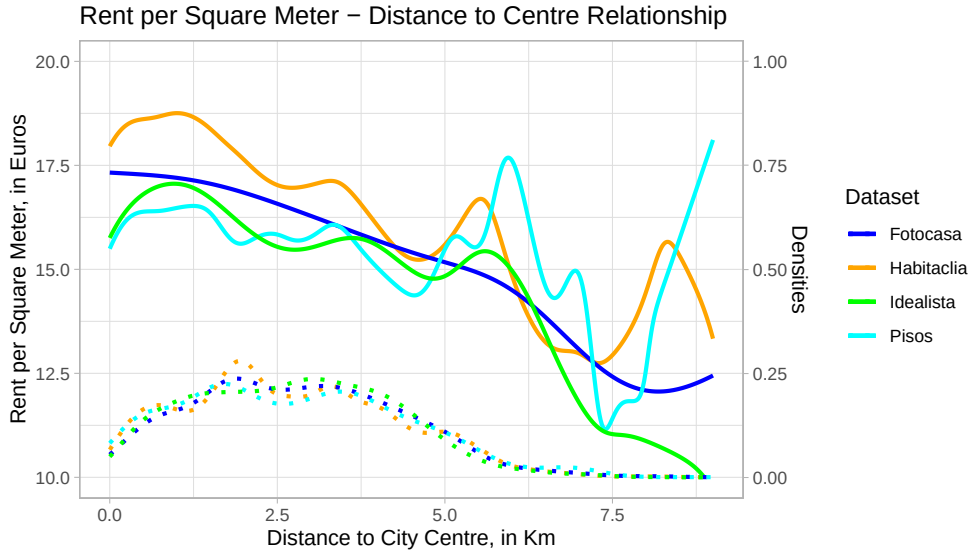
(d) PDF of log rent per m^2 on $\mathbb{D} = [\ln 5, \ln 50]$.

choice is still debatable. Other potential candidates were the Plaça Catalunya or the Gran Teatre del Liceu.

4.3 Geospatial Distribution

If the paper ended here, then the negative relation between rent prices and peripherality found in Section 4.2 could be seen as an overgeneralisation, an attempt to oversimplify the heterogeneous distribution of rents, which naturally is a much more complex issue than what a one-variable regression can properly explain. I am of the opinion that its validity as simplification is only justified after understanding how rent prices are distributed geographically. This is what attempt to display in Figure 4.6. These four plots summarise very well the whole paper. Firstly, because they show that in some very restricted parts of the city rent prices are four- and even fivefold what rent prices for the rest of the city are. In fact, this is the reason why the colour scale of the legend needs to be nonlinear; otherwise, the differences in rents would pale in comparison to the contrast between the whole city and those particularly expensive areas. Secondly, because they illustrate how rent prices progressively fall,

Figure 4.5: Kernel Regression on $\mathbb{D} = [0, 9]$. Independent variable: Distance to city centre. Dependent Variable: rent per m^2. The density distribution of distance to city centre is added (the dotted lines) for reference.



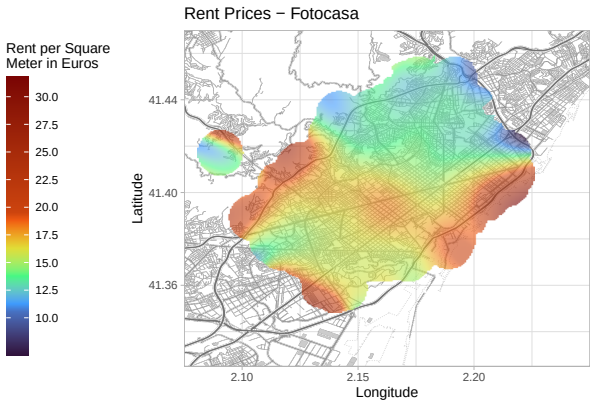
on average, as we increase the distance with respect to the city centre. This last statement could confuse the reader, who could think that the beachside and mountainside high-priced areas, in dark red, should have an upwards effect in the middle to last regions of Plot 4.5. But in fact, these parts of the city contain relatively few observations, and hence have little impact. This does not necessarily come from an under-representation in the dataset. According to [Ajuntament de Barcelona, 2020], these two districts have a very low share of rented accommodations with respect to total accommodations. This lack of supply also partly explains the high supply rent prices per m^2.

The four OHRWs convey a very similar image. Rents are higher than average in Ciutat Vella; the Eixample and particularly in the Antiga Esquerra de l'Eixample; Vila de Gràcia and the Upper Diagonal neighbourhoods. Rent prices are even remarkably higher for the areas closest to the beaches, that is, the Barceloneta and Front Marítim de Sant Martí; and also for the neighbourhoods on the feet of the Collserola mountain range. Rent prices per m^2 are lower than average to the North, next to the Besòs river, corresponding to the Nou Barris and Sant Andreu districts; and also to South, next to the limit with l'Hospitalet, corresponding to the Les Corts and Sants-Montjuïch districts. Occasionally there are some disagreements between samples. For instance, according to the Idealista sample, the northern limit of the Horta-Guinardó neighbourhood is very expensive; in Plot 4.6a, the industrial southernmost part of the city is portrayed as an area of high rents; and similarly happens in Plot 4.6d in the boundary with Badalona. Because these disagreements occur only in one sample at a time and all other three plots display the opposite, these divergences should be attributed to outliers which, because of the relative sparsity of data in the boundaries of the city (see Plot 4.5), have a lot of weight and substantially modify the estimation.

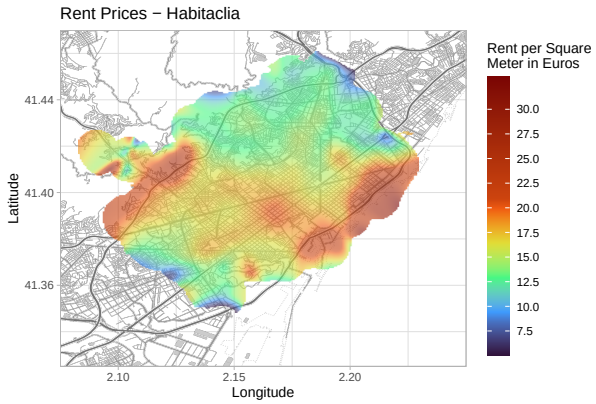
The results with respect to geographic distribution are very similar, even though in a higher degree of detail, to those presented by the City Council [Ajuntament de Barcelona, 2020; Garcia, Marfà, Camprodon, Logan, and Funollet, 2021]. Still, rent prices per m^2 are higher in the datasets of this

Figure 4.6: Kernel Regression. Independent variables: latitude and longitude. Dependent Variable: rent per m^2. In practice, the estimation at each point $x_0 \in \mathbb{R}^2$ is almost equivalent to a weighted average of the observations nearby based on their distance with respect to x_0, except that the local linear estimation has better properties at the boundaries. The results are only plotted at domain points such that their neighbourhood of radius 0.1 in the coordinate space contains at least 3 observations, to avoid extrapolating to data sparse regions. Without any loss of information, points with a value over 55 or under 5 Euros per m^2 are not plotted. Beware that each plot has a different legend and that these are nonlinear in colour. Map tiles by Stamen Design, under CC BY 3.0. Geographic data by OpenStreetMap, under ODbL.

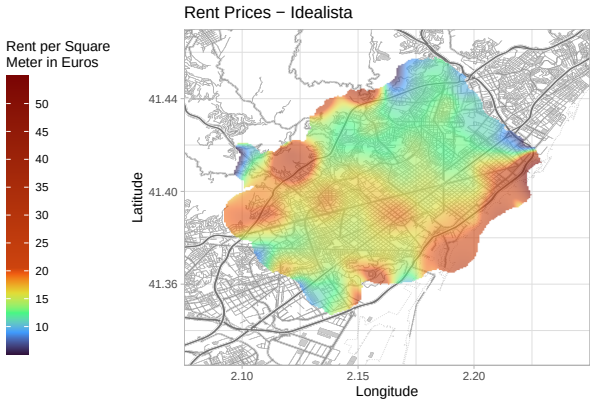
(a) Kernel regression on $\mathbb{D} = [2.075, 2.25] \times [41.325, 41.47]$.



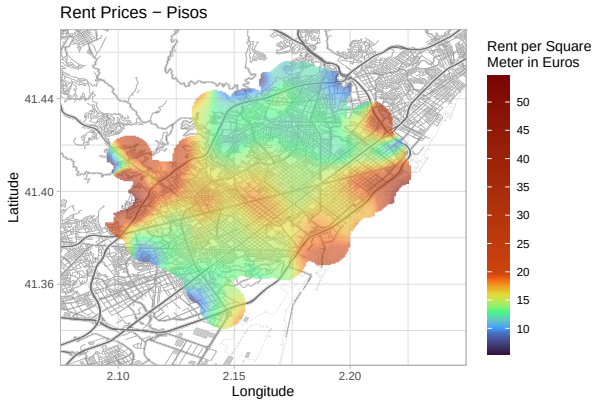
(b) Kernel regression on $\mathbb{D} = [2.075, 2.25] \times [41.325, 41.47]$.



(c) Kernel regression on $\mathbb{D} = [2.075, 2.25] \times [41.325, 41.47]$.



(d) Kernel regression on $\mathbb{D} = [2.075, 2.25] \times [41.325, 41.47]$.



project than in the studies from the City Council. Here there are also differences between OHRWs. Some of them, namely Idealista and Pisos, seem to have more concentrated very high values (see the legends of Plots 4.6c and 4.6d). This could be caused by a smaller bandwidth for the regression for those datasets, which does not "dissolve" large values as much as a wider bandwidth would. Since the results are quite sensitive to the bandwidth selection, it is justified to choose the kernel regression bandwidth by the optimality criterion described in Section 3.2.

Finally, the small differences between OHRWs in Figure 4.6, and specially the small differences with respect to the Idealista sample plot mean that the quality of the geocoding is remarkably good[5]. This is the only kind of validation of its quality pursued in this paper. On the other hand, there are only two typologies of geocoding mistakes, and neither of them is easily surmountable. The first typology corresponds to the geocoding service not properly understanding the address. Mistakes of this kind are hardly susceptible of being mended in datasets of large size. The second error type occurs if the address the landlord gives is already erroneous or incomplete, which is unverifiable.

[5] Idealista is the only OHRW that provided the geocoding directly. Idealista uses Google services to geocode the advertisements as well, but when creating an advertisement, after writing the address, landlords are shown the location of the address on a map and need to confirm its accuracy. This means that its likeliness to contain geocoding errors is reduced.

5 Conclusions

> "*[...] market segmentation suggests that selection biases are pervasive and generalizable conclusions can only be held through the fusion of data from various sources.*"
> Bin Jiang and Jean Claude Thill [Jiang and Thill, 2015]

This paper has shown that rent, both in monthly and in m^2 prices can be thought as being log-normally distributed, with a remarkably high variance which results from the notorious presence of extreme values. Methodologically, when dealing with such variables, techniques like medians and rank plots are better suited than traditional ones (e.g., means, histograms). It has also shown that there exists a negative relationship between rent prices and distance with respect to the centre of the city, which is further legitimised by the geospatial visualisation conducted in Section 4.3. The geographic distribution of rental prices found in this paper is coherent with results found by other studies and reports.

Additionally, the results of the paper show that there are there are nonnegligible differences between samples due to unknown reasons, but potentially owning to (self-)selection biases. After all, OHRWs are not random sample experiments in lab-like conditions, but rather databases created by private firms that might have an interest in focusing towards a specific subset profile of accommodations. Web scraping practitioners should beforehand consider what the marketing strategies of such websites are, and what implications could that have for the quality of the data.

The results of the paper are somewhat bounded. In what follows, I would like to point out the limitations, hint how they could be surpassed and indicate where further research might be done.

Even though the project is quite internally valid, there are some aspects that must be taken into account in further research, in particular regarding data cleaning. A protocol for duplicate detection has to be flexible enough as to be able to neither delete legitimate non-duplicate observations nor to allow duplicates to remain in the dataset. The approach taken here attempted to strike a balance between these two extremes by writing an algorithm that, among observations that share the same characteristics, judges whether they are duplicates or not based on the string similarity of their descriptions. Alternatives would be, as in [Chapelle and Eyméoud, 2018], to simply drop repeated values, or, as in [Loberto, Luciani, and Panagallo, 2018], use machine learning techniques to let the AI itself decide what a duplicate is. Furthermore, there remains the concern that some advertisements might only be "attempts" of landlords who are just getting familiarised with the OHRW, or that simply do not correspond to any real accommodation at all. This is a trickier question to address; the optimal approach here, and for many of the data quality problems that data practitioners often face, is to engage with the DGP itself. In this case, this translates to working hand in hand with the OHRW. An example of institutions working together with firms to obtain quality data is the collaboration agreement

between OHRWs and the Barcelona City Council that gave as a result the report [Equip Observatori Metropolità de l'Habitatge de Barcelona, 2020]. The authors of the report had access to the number of views each advertisement had, and decided to drop out of the dataset those observations that had less than 2 visualisations, lest they be, as mentioned above, mere testing attempts.

On the other hand, the quality of the geocoding seems to be satisfactory, and arguably there is no geocoding service provider that can surpass the quality of Google services. Open-source providers, in this project, failed to provide an acceptable level of geolocation accuracy, but could be helpful if the quantity of data to be geolocated is so large that it exceeds the free quota of Google services.

The external validity is severely jeopardised by the lack of correspondence with census data, in terms exposed in Section 2.4. The cause for this lack of correspondence is unknown, and further research might be needed to understand it. I can but hypothesise that it is both due to an over-representation of high-quality accommodations and due to the bargaining process setting a lower rent than what was advertised. Even though the results might accurately show the behaviour of the rent housing supply in Barcelona, this is not enough to generalise the results to the whole market. Here the potential ways to bridge the difference are diverse. A traditional approach might be to disregard the rent price advertised by the landlord and to construct an hedonic price based on the qualities of the accommodation, its location and the attractions available in the vicinity. The hedonic prices might be constructed using a traditional regression or by using new approaches, such as machine learning algorithms. Another solution that again involves some sort of institutional cooperation is to use the Rent Price Index that the Catalan Housing Agency elaborates. Lastly, if somehow OHRWs monitor the rent price that both parties agree on after the negotiation, this could be a more reliable source of data, although it is likely to raise ethical and legal concerns.

Methodologically, there are some aspects that can be improved in further research. Firstly, non-parametric estimation does not mean nontestable estimation. Kernel regression results can be accompanied by confidence interval bands. These were not included in this paper because (i) they involve too complex and computationally taxing calculations, and because (ii) they would make the data visualisation too burdensome. But this does not mean that in smaller samples they should not be used to convey a sense of how robust the results are. Also, I have already mentioned that price-setters seem to favour round rent prices. This poses some problems for bandwidth selection in the context of density estimation (for instance, the Cross-Validation Least-Squares method returned absurdly small bandwidths, something notorious for samples with spikes [Devroye, 1989]), that suggest that other statistical methods, such as kernel density estimation models specifically tailored to heaped data might yield better results.

Finally, it is possible that Zipf's law applies to the distribution of rent per m^2. My own testing shows that the data is very likely to complain with the law[1], and this could be yet another field of nature where Zipf's law applies. Further research would be needed.

[1] The test used to assess whether a sample is Zipf-distributed shows that it indeed does, but since this is not the objective of this paper, I have omitted them.

Bibliography

Ajuntament de Barcelona. (2020). Mercat de lloguer per barris. Retrieved May 28, 2021, from https://ajuntament.barcelona.cat/barcelonaeconomia/ca/mercat-immobiliari/mercat-de-lloguer/mercat-de-lloguer-barris-0

Appelhans, T., Detsch, F., Reudenbach, C., & Woellauer, S. (2020). *Mapview: interactive viewing of spatial data in r* [R package version 2.9.0]. https://CRAN.R-project.org/package=mapview

Blanco-Romero, A., Blázquez-Salom, M., & Cànoves, G. (2018). Barcelona, housing rent bubble in a tourist city. Social responses and local policies. *Sustainability (Switzerland)*, *10*(6), 1–18. https://doi.org/10.3390/su10062043

Boeing, G., & Waddell, P. (2016). New Insights into Rental Housing Markets Across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *SSRN Electronic Journal*, (December). https://doi.org/10.2139/ssrn.2781297

Cao, R., Cuevas, A., & González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, *17*(2), 153–176. https://doi.org/10.1016/0167-9473(92)00066-Z

Chandra, R. V., & Varanasi, B. S. (2015). *Python requests essentials*. Packt Publishing Ltd.

Chapelle, G., & Eyméoud, J.-B. (2018). *Can Big Data increase our knowledge of local rental markets? Estimating the cost of density with rents* (tech. rep.). Minsitère de l'Économie, des Finances et de la Relance. Paris. https://www.tresor.economie.gouv.fr/Articles/283bb902-a878-44a4-86fa-991a3240e07f/files/284f9f1d-6138-4297-9f4a-a378583ba174

Clark, S. D., & Lomax, N. (2018). A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques. *Journal of Big Data*, *5*(1). https://doi.org/10.1186/s40537-018-0154-3

De Arrese, J. L. (1959). No queremos una España de proletarios, sino de propietarios, 41–42. https://www.abc.es/archivo/periodicos/abc-madrid-19590502-41.html

Devroye, L. (1989). On the non-consistency of the L^2-Cross-Validated Kernel Density Estimate. *Statistics & Probability Letters*, *8*(5), 425–433.

Equip Observatori Metropolità de l'Habitatge de Barcelona. (2020). *L'oferta i la demanda d'habitatges de lloguer. Dades mensuals de portals immobiliaris.* (tech. rep.). Ajuntament de Barcelona. Barcelona.

Gabinet Tècnic de Programació. (2016). *Preu De Lloguer Dels Habitatges a Barcelona* (tech. rep.). Ajuntament de Barcelona. Barcelona.

Garcia, A., Marfà, R., Camprodon, M., Logan, I., & Funollet, M. (2021). Com puja el preu del lloguer on vius? Retrieved May 27, 2021, from https://interactius.ara.cat/lloguers/indicadors/barcelona

Garcia-López, M.-À., Jofre-Monseny, J., Martínez-Mazza, R., & Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, *119*(96131), 103278. https://doi.org/10.1016/j.jue.2020.103278

Gazoni, E., & Clark, C. (2021). Openpyxl - a python library to read/write excel 2010 xlsx/xlsm files. https://openpyxl.readthedocs.io/en/stable/

GitHub's Comunity of Collaborators. (2021). Beautiful soup 4. https://github.com/SeleniumHQ/selenium/

Google. (2021). Python client for google maps services. https://github.com/googlemaps/google-maps-services-python

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (2nd). Springer. https://doi.org/10.1111/j.1751-5823.2009.00095_18.x

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: the np package. *Journal of Statistical Software*, *27*(5). http://www.jstatsoft.org/v27/i05/

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media:An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy*, *82*(129), 657–673. https://doi.org/10.1016/j.landusepol.2018.12.030

INCASÒL. (2021). Lloguers Barcelona per districtes i barris. Retrieved May 21, 2021, from https://habitatge.gencat.cat/ca/dades/estadistiques_publicacions/indicadors_estadistiques/estadistiques_de_construccio_i_mercat_immobiliari/mercat_de_lloguer/lloguers-barcelona-per-districtes-i-barris/#bloc4

INE. (2019). *Hogares por régimen de tenencia de la vivienda y edad y sexo de la persona de referencia* (tech. rep.). Instituto Nacional de Estadística. Madrid. https://ine.es/jaxiT3/Datos.htm?t=9994#!tabs-grafico

Jiang, B. (2015). Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity. *GeoJournal*, *80*(1), 1–13. https://doi.org/10.1007/s10708-014-9537-y

Jiang, B., & Thill, J. C. (2015). Volunteered Geographic Information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems*, *53*(November 2018), 1–3. https://doi.org/10.1016/j.compenvurbsys.2015.09.011

Kahle, D., & Wickham, H. (2013). Ggmap: spatial visualization with ggplot2. *The R Journal*, *5*(1), 144–161. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Kyriakidou, V., Michalakelis, C., & Varoutas, D. (2011). Applying Zipf's Power Law Over Population Density and Growth as Network Deployment Indicator. *Journal of Service Science and Management*, *04*(02), 132–140. https://doi.org/10.4236/jssm.2011.42017

Li, Q., & Racine, S. (2007). *Nonparametric Econometrics: Theory and Practice* (1st). Princeton University Press.

Loberto, M., Luciani, A., & Panagallo, M. (2018). *The potential of big housing data: an application to the Italian real-estate market*, Banca d'Italia, Eurosistema. http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4616.2010.00331.x/full

Mitchell, R. (2018). *Web Scraping with Python* (2nd). O'Reilly Media, Inc.

Padgham, M., Rudis, B., Lovelace, R., & Salmon, M. (2017). Osmdata. *The Journal of Open Source Software*, *2*(14). https://doi.org/10.21105/joss.00305

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

R Core Team. (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rae, A. (2015). Online Housing Search and the Geography of Submarkets. *Housing Studies*, *30*(3), 453–472. https://doi.org/10.1080/02673037.2014.974142

Richardson, L. (2020). Beautiful soup 4. https://www.crummy.com/software/BeautifulSoup/

Romero, T. (2020). Home ownership rate in Spain from 2004 to 2019. Retrieved May 21, 2021, from https://www.statista.com/statistics/1185377/housing-ownership-ratein-spain-by-type/

Saiz, A. (2010). The Geographic Determinants of Housing Supply *. *Quarterly Journal of Economics*, *125*(3), 1253–1296. https://doi.org/10.1162/qjec.2010.125.3.1253

Schernthanner, H., Steppan, S., Kuntzsch, C., Borg, E., & Asche, H. (2017). Automated Web-Based Geoprocessing of Rental Prices. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 512–524). https://doi.org/10.1007/978-3-319-62401-3_37

Solc, T. (2021). Unidecode 1.2.0. https://pypi.org/project/Unidecode/

Sweeney, O. (2021). Spain's Love of Home Ownership Set to Decline. Retrieved May 21, 2021, from https://www.euroweeklynews.com/2021/01/20/spains-love-of-home-ownership-set-to-decline/

Sweigart, A. (2019). *Automate the Boring Stuff with Python, Practical Programming for Total Beginners*. No Starch Press.

Thomschke, L. (2015). Changes in the distribution of rental prices in Berlin. *Regional Science and Urban Economics*, *51*, 88–100. https://doi.org/10.1016/j.regsciurbeco.2015.01.001

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(sup1), 234–240. http://www.jstor.org/stable/143141

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer. https://doi.org/10.1007/978-0-387-79052-7

van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*, 111–122. https://CRAN.R-project.org/package=stringdist

Van Rossum, G. (2020). *The python library reference, release 3.8.2*. Python Software Foundation.

Warnes, G. R., Bolker, B., & Lumley, T. (2020). *Gtools: various r programming tools* [R package version 3.8.2]. https://CRAN.R-project.org/package=gtools

Watts, D. (2007). A twenty-first century science. *Nature*, 489.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Seidel, D. (2020). *Scales: scale functions for visualization* [R package version 1.1.1]. https://CRAN.R-project.org/package=scales

A Scraping Procedure

Since each OHRW required a different program, each one is explained separately. The sources consulted to write the programs are [Mitchell, 2018; Sweigart, 2019].

A.1 Habitaclia

First, the program asks the user how many result pages of advertisements are to be scrapped. Then, for every page of results, the program sends the server a request to download the result page in HTML. The downloaded file is then interpreted. Once interpreted, the program reads the page and fetches every individual advertisement URL that it finds in the result page and copies them in a list. Then, for every URL of an advertisement page, the program sends a request to the server to download the page in HTML format. Once this is done and the file interpreted, the program searches in the file some structure patterns that I beforehand specified, and that contain the qualities of the advertisement (e.g., the rent, surface, last update, and so on). Once the procedure has been run on every result page, the data is stored into an Excel file.

The geocoding is done taking the accommodation's address, and if it is not provided, taking the neighbourhood.

A.2 Fotocasa

The procedure is very similar as with Habitaclia, except that

1. First *all* of the advertisement URLs are retrieved from all the result pages to be processed, and only then the program starts to analyse every individual advertisement.
2. The result pages generate themselves as the user scrolls them down. This means that if the HTML is retrieved using the usual procedure, only a fraction of URLs can be obtained because the rest of the page is not yet loaded. The solution is to program the default browser to act automatically and gradually scroll down the page, "tricking" the server to believe a human user is accessing the page. Only then the whole page is generated and the URLs retrieved.
3. The geolocation is based on the address, which is always provided.

A.3 Pisos

Exactly as with Fotocasa, except that it was not necessary to scroll gradually down through the page and it was enough to jump directly to the end for the server to generate the whole contents of the webpage.

A.4 Idealista

This was the most peculiar scraping process. The Terms of Use of Idealista explicitly forbid automated use of its webpage. Upon request to the administrators, I was granted limited access to their database through the use of their API. The process needed to be automated nevertheless, because even though it did not imply requesting access a page and interpreting its HTML code, it did imply sending requests to their servers in cURL format via command-line interface instructions. The access I was given was limited to 100 monthly requests of a maximum of 50 advertisements per request. During the month of April I used the 100 requests to understand how the process worked, and the 3rd of May, after some small trials, I ran the whole program. Out of the 100 requests of the month of May, around 5 were trials. This leaves out 95 requests. Since there were more advertisements (around 130 pages of 50 advertisements per page) than I could access, the procedure was designed as follows: first, the program sends a request to access the first page of results. Among the response data that the request yields there is the number of total pages of results. Since analysing the 95 first pages would imply a risk of incurring in a sampling bias, the program randomly selected 94 elements of the set $\mathbb{N} \cap [2, \text{Last Page}]$. Then it analysed the pages that had the randomly selected index.

B Duplicate Detection

Duplicate detection is of extreme importance when dealing with user-generated data. It is also a very sensitive matter. To understand why, consider the following example. A landlord publishes an advertisement in an OHRW. She later realises that she has made a small mistake. She has, for instance, written that the flat has 3 bedrooms, when in reality it has but 2. Instead of modifying the advertisement, she simply creates a new one, without deleting the old one. As a result, the same physical accommodation is advertised twice, and moreover they appear to be different observations because they have different qualities. Even worse, imagine that a landlord publishes twice an advertisement, but on the second version, she adds a single line to the description saying that pets are not allowed. Now the two advertisements appear as two different accommodations because of a single line in the description.

If the probability an advertisement has of having a duplicate is correlated with some quality of the accommodation (and [Loberto, Luciani, and Panagallo, 2018] indicate that this might indeed be the case), then all of the estimations might be biased. Hence the necessity of having a meditated protocol to detect and deal with duplicates. Absolute duplicates, that is, observations that are absolutely identical to any other, are deleted straight away[1]. To weed out duplicates as the ones described in the paragraph above it was necessary to write a specific algorithm. Algorithm 1 displays a pedagogic simplification of the procedure used in reality. The algorithm clusters all observations that share the same coordinates and rent per m^2. Then, in each cluster, it does every possible pairwise comparison. For those pairwise comparisons such that the rate of similarity between their description texts is high enough to consider them to be the same text but with minor modifications, one of the observations is marked as a duplicate. All observations marked as duplicates are afterwards dropped.

The coefficient of similarity between the pairs of descriptions is calculated using the package Stringdist [van der Loo, 2014]. According to the package documentation, the similarity between two strings *a* and *b* is "*calculated by first calculating the distance using* stringdist[2]*, dividing the distance by the maximum possible distance, and substracting the result from 1. This results in a score between 0 and 1, with 1 corresponding to complete similarity and 0 to complete dissimilarity.*". The distance is calculated by counting "*the number of deletions,* [one-position transpositions,] *insertions and substitutions necessary to turn to turn* b *into* a".

[1] More correctly, if n observations are absolute duplicates, $n - 1$ of them are deleted
[2] A function that calculates the distance between strings of text.

Algorithm 1 Simplified Duplicate Detection Algorithm

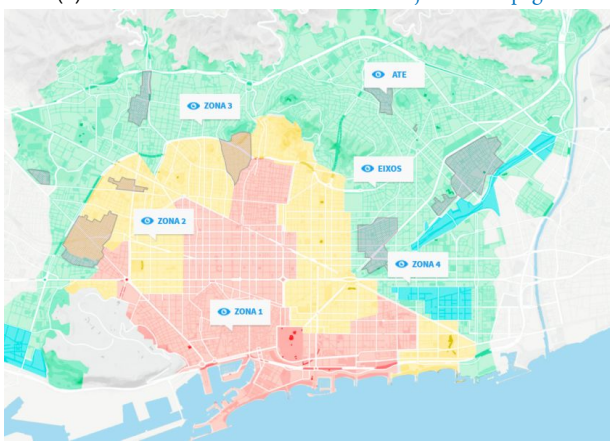
```
1:  procedure DUPLICATE PROCEDURE
2:      dataset ← dataset from one sample
3:      suspects ← obs. from dataset such that latitude, longitude, rent per m^2 combination is repeated
4:      uniques ← unique combinations of latitude, longitude and rent per m^2 in dataset
5:      function TEXT_SIM(text1, text2)
6:          return the similarity coefficient (∈ [0, 1]) between text1 and text2
7:      end function
8:      indicator ← 0
9:      for unique in uniques do                    ▷ Clusters obs. with same characteristics
10:         for suspect in suspects do
11:             bool1 ← unique[latitude] == suspect[latitude]
12:             bool2 ← unique[longitude] == suspect[longitude]
13:             bool3 ← unique[rent per m^2] == suspect[rent per m^2]
14:             if bool1, bool2, bool3 == TRUE then
15:                 suspect[cluster] ← indicator
16:             end if
17:         end for
18:         indicator ← indicator +1
19:     end for
20:     confirmed ← empty dataset
21:     for group in 0, ..., indicator do      ▷ Marks one element of every comparison if too similar
22:         for each perm. of obs. suspect in suspects such that suspect[cluster] == group do
23:             similarity ← TEXT_SIM(suspect1[description], suspect2[description])
24:             if similarity > 0.6 then
25:                 add suspect2 to confirmed
26:             end if
27:         end for
28:     end for
29:     for observation in dataset do                          ▷ Drops obs. if marked
30:         if observation is in confirmed then
31:             drop observation
32:         end if
33:     end for
34: end procedure
```

C Correspondence with the PEUAT Model

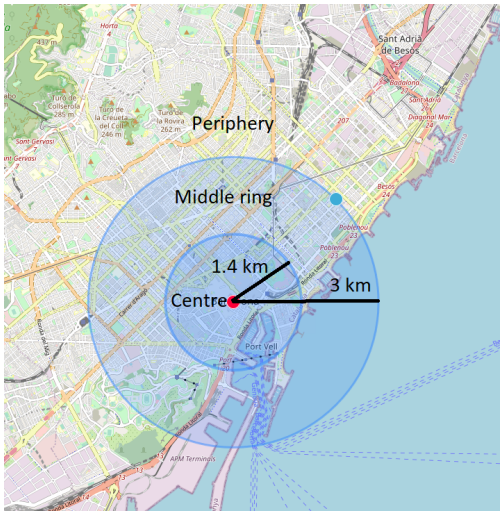
Figure C.1 shows the relationship between the divisions made by the City Council in the context of the PEUAT model and the divisions defined in this project, which are characterised by the simplicity of their definition.

Figure C.1: Comparison between city divisions.

(a) PEUAT zones. Source: PEUAT Project Homepage.



(b) Concentric circumferences. Map and circumferences created with the Map Radius Calculator, edited with MS Paint.



D Statistical Summary of the Data

Table D.1 summarises the final datasets after the cleaning procedure, on which all the tests and analyses of the paper are conducted. Note that not all variables are instantly retrieved from the OHRWs; for instance, rent per m^2 and distance are artificially created by using and transforming other pieces of information available in the advertisements.

Table D.1: Summary statistics of the datasets.

Variable	**Fotocasa**	**Habitaclia**	**Idealista**	**Pisos**
	$n = 3958$	$n = 8109$	$n = 4287$	$n = 2236$
URL	$n = 3958$	$n = 8109$	$n = 4287$	$n = 2236$
Rent (€)	$n = 3958$ $\max = 32747$ $\min = 300$ $\hat{\mu} = 1249.63$ $\hat{\sigma} = 1009.88$	$n = 8109$ $\max = 32747$ $\min = 300$ $\hat{\mu} = 1411.39$ $\hat{\sigma} = 1208.89$	$n = 4287$ $\max = 12000$ $\min = 400$ $\hat{\mu} = 1257.39$ $\hat{\sigma} = 857.58$	$n = 2236$ $\max = 12000$ $\min = 450$ $\hat{\mu} = 1291.3$ $\hat{\sigma} = 880.68$
Surface (m^2)	$n = 3958$ $\max = 895$ $\min = 11$ $\hat{\mu} = 79.81$ $\hat{\sigma} = 44.53$	$n = 8109$ $\max = 900$ $\min = 10$ $\hat{\mu} = 84.97$ $\hat{\sigma} = 52.4$	$n = 4287$ $\max = 500$ $\min = 15$ $\hat{\mu} = 82.22$ $\hat{\sigma} = 44.44$	$n = 2236$ $\max = 613$ $\min = 19$ $\hat{\mu} = 84.66$ $\hat{\sigma} = 46.69$
Bedrooms	$n = 3874$ $\max = 10$ $\min = 1$ $\hat{\mu} = 2.42$ $\hat{\sigma} = 1.08$	$n = 7924$ $\max = 11$ $\min = 1$ $\hat{\mu} = 2.47$ $\hat{\sigma} = 1.11$	$n = 4287$ $\max = 10$ $\min = 0$ $\hat{\mu} = 2.36$ $\hat{\sigma} = 1.16$	$n = 2184$ $\max = 9$ $\min = 1$ $\hat{\mu} = 2.46$ $\hat{\sigma} = 1.09$
Bathrooms	$n = 3926$ $\max = 8$ $\min = 1$ $\hat{\mu} = 1.42$ $\hat{\sigma} = 0.68$	$n = 8021$ $\max = 11$ $\min = 1$ $\hat{\mu} = 1.51$ $\hat{\sigma} = 0.75$	$n = 4287$ $\max = 7$ $\min = 1$ $\hat{\mu} = 1.43$ $\hat{\sigma} = 0.67$	$n = 2230$ $\max = 9$ $\min = 1$ $\hat{\mu} = 1.46$ $\hat{\sigma} = 0.71$
Municipality	NI	NI	$n = 4287$	NI
Neighborhood	NI	$n = 8109$	$n = 4287$	$n = 2236$
Address	$n = 3958$	$n = 5775$	NI	$n = 2236$
Property Subtype	$n = 3958$	$n = 8270$	$n = 4287$	$n = 2236$

Variable	**Fotocasa** $n = 3958$	**Habitaclia** $n = 8109$	**Idealista** $n = 4287$	**Pisos** $n = 2236$
Description	$n = 3901$	$n = 8081$	NI	$n = 2236$
Latitude	$n = 3958$ $\max = 41.46$ $\min = 41.33$ $\hat{\mu} = 41.4$ $\hat{\sigma} = 0.02$	$n = 8109$ $\max = 41.46$ $\min = 41.33$ $\hat{\mu} = 41.4$ $\hat{\sigma} = 0.01$	$n = 4286$ $\max = 41.45$ $\min = 41.33$ $\hat{\mu} = 41.4$ $\hat{\sigma} = 0.01$	$n = 2236$ $\max = 41.45$ $\min = 41.35$ $\hat{\mu} = 41.4$ $\hat{\sigma} = 0.02$
Longitude	$n = 3958$ $\max = 2.22$ $\min = 2.09$ $\hat{\mu} = 2.16$ $\hat{\sigma} = 0.02$	$n = 8109$ $\max = 2.23$ $\min = 2.1$ $\hat{\mu} = 2.16$ $\hat{\sigma} = 0.02$	$n = 4287$ $\max = 2.22$ $\min = 2.1$ $\hat{\mu} = 2.16$ $\hat{\sigma} = 0.02$	$n = 2236$ $\max = 2.22$ $\min = 2.09$ $\hat{\mu} = 2.16$ $\hat{\sigma} = 0.02$
Distance (km)	$n = 3958$ $\max = 8.69$ $\min = 0.07$ $\hat{\mu} = 2.81$ $\hat{\sigma} = 1.53$	$n = 8109$ $\max = 8.94$ $\min = 0.02$ $\hat{\mu} = 2.71$ $\hat{\sigma} = 1.56$	$n = 4287$ $\max = 8.03$ $\min = 0.04$ $\hat{\mu} = 2.75$ $\hat{\sigma} = 1.46$	$n = 2236$ $\max = 8.78$ $\min = 0.07$ $\hat{\mu} = 2.76$ $\hat{\sigma} = 1.65$
Floor	NI	$n = 4596$ $\max = 9$ $\min = 1$ $\hat{\mu} = 2.82$ $\hat{\sigma} = 1.77$	$n = 3562$ $\max = 60$ $\min = 1$ $\hat{\mu} = 3.14$ $\hat{\sigma} = 2.38$	$n = 1182$ $\max = 9$ $\min = 1$ $\hat{\mu} = 2.94$ $\hat{\sigma} = 1.76$
Rent per m^2	$n = 3958$ $\max = 131.82$ $\min = 5.56$ $\hat{\mu} = 16.33$ $\hat{\sigma} = 7.04$	$n = 8109$ $\max = 185$ $\min = 5$ $\hat{\mu} = 17.21$ $\hat{\sigma} = 8.45$	$n = 4287$ $\max = 85.38$ $\min = 5$ $\hat{\mu} = 15.89$ $\hat{\sigma} = 5.79$	$n = 2236$ $\max = 120$ $\min = 6.67$ $\hat{\mu} = 15.77$ $\hat{\sigma} = 5.78$
Last Update	$n = 3958$	$n = 8109$	NI	$n = 2236$
Professional	$n = 3958$ $\max = 1$ $\min = 0$ $\hat{\mu} = 0.88$ $\hat{\sigma} = 0.32$	$n = 8109$ $\max = 1$ $\min = 0$ $\hat{\mu} = 0.99$ $\hat{\sigma} = 0.11$	NI	$n = 2236$ $\max = 1$ $\min = 0$ $\hat{\mu} = 0.9$ $\hat{\sigma} = 0.3$

Remarks:

1. NI: Not included in the dataset.
2. Distance (km): Distance to city centre, Plaça Sant Jaume.
3. Professional: Dummy variable. 1 if user is a real estate agency, 0 otherwise.

E Code

Both the web scraping and the statistical methods described in this paper have been implemented by myself on Python and R, respectively. I have published the code in an ordered and detailed fashion in my GitHub page, in the publicly available repository COrtsJosep/TFG-2021.

The R packages I have used for this paper are Ggmap [Kahle and Wickham, 2013], Gtools [Warnes, Bolker, and Lumley, 2020], Scales [Wickham and Seidel, 2020], Stringdist [van der Loo, 2014], OSMData [Padgham, Rudis, Lovelace, and Salmon, 2017], the marvellous Tidyverse collection [Wickham, Averick, Bryan, Chang, McGowan, François, Grolemund, Hayes, Henry, Hester, Kuhn, Pedersen, Miller, Bache, Müller, Ooms, Robinson, Seidel, Spinu, Takahashi, Vaughan, Wilke, Woo, and Yutani, 2019], and naturally, the core functions [R Core Team, 2020]. Additionally, for exploratory analysis and validation I have used the packages MapView [Appelhans, Detsch, Reudenbach, and Woellauer, 2020], NP [Hayfield and Racine, 2008] and SF [Pebesma, 2018].

There is less of a convention of how to properly cite Python packages, so I will do it to the best of my knowledge. I have used the packages Beautiful Soup [Richardson, 2020], GoogleMaps [Google, 2021], OpenPyExcel [Gazoni and Clark, 2021], Requests [Chandra and Varanasi, 2015], Selenium [GitHub's Comunity of Collaborators, 2021] and Unidecode [Solc, 2021]; the standard packages [Van Rossum, 2020] Json, Math, OS, Regular Expressions, Random, Shelve, Subprocess, System and Time; and naturally, the core functions [Van Rossum, 2020].

F Erratum

Appendix D claimed that the distance with respect to city centre was expressed in meters, when it is expressed in kilometres.