Humboldt-Universität zu Berlin

Seminar paper

# Exploratory Data Analysis of airbnb listings in Berlin

*Silvia Ventoruzzo*

Statistical Programming Languages

supervised by

Alla Petukhina

February 24, 2019

# Contents

# List of Abbreviations

| | | | |
|---|---|---|---|
| CPI | Consumer Price Index | ETF | Equity Traded Funds |
| ETH | Eat the Horse | XLM | Xetra Liquidity |

## List of Figures

## List of Tables

# 1 Introduction

- What is the subject of the study? Describe the economic/econometric problem.
- What is the purpose of the study (working hypothesis)?
- What do we already know about the subject (literature review)? Use citations: *Gallant (1987) shows that... Alternative Forms of the Wald test are considered (Breusch and Schmidt, 1988).*
- What is the innovation of the study?
- Provide an overview of your results.
- Outline of the paper:

  *The paper is organized as follows. The next section describes the model under investigation. Section 2 describes the data set and Section 4 presents the results. Finally, Section 5 concludes.*
- The introduction should not be longer than 4 pages.

# 2 Data

For the analysis explained in this paper data was downloaded for a website independent from airbnb itself. insideairbnb (insideairbnb.com) scrapes (????) airbnb to get its data and posts it online for the public to use on own analysis, while also providing some analysis of its own. The data is divided according to cities and for each there is general information about the city's properties and their availability for the next year. The variables that are being kept for this analysis are listing on table 1. (HOW TO MAKE IT CHANGE NUMBER???).

## 2.1 Berlin neighbourhoods and districts

Berlin consists of 96 neighbourhoods (Ortsteile), which are grouped into 12 districts (Bezirke).

| Neighbourhood | District |
|---|---|
| Charlottenburg | Charlottenburg-Wilmersdorf |
| Wilmersdorf | Charlottenburg-Wilmersdorf |
| Grunewald | Charlottenburg-Wilmersdorf |
| Westend | Charlottenburg-Wilmersdorf |
| Schmargendorf | Charlottenburg-Wilmersdorf |
| Charlottenburg-Nord | Charlottenburg-Wilmersdorf |
| Halensee | Charlottenburg-Wilmersdorf |
| Friedrichshain | Friedrichshain-Kreuzberg |
| Kreuzberg | Friedrichshain-Kreuzberg |
| Friedrichsfelde | Lichtenberg |
| Karlshorst | Lichtenberg |
| Malchow | Lichtenberg |
| Wartenberg | Lichtenberg |
| Falkenberg | Lichtenberg |
| Fennpfuhl | Lichtenberg |
| Lichtenberg | Lichtenberg |
| Neu-Hohenschönhausen | Lichtenberg |
| Alt-Hohenschönhausen | Lichtenberg |
| Rummelsburg | Lichtenberg |
| Marzahn | Marzahn-Hellersdorf |
| Biesdorf | Marzahn-Hellersdorf |
| Kaulsdorf | Marzahn-Hellersdorf |
| Mahlsdorf | Marzahn-Hellersdorf |
| Hellersdorf | Marzahn-Hellersdorf |
| Mitte | Mitte |
| Moabit | Mitte |
| Hansaviertel | Mitte |
| Gesundbrunnen | Mitte |
| Tiergarten | Mitte |
| Wedding | Mitte |
| Buckow | Neukölln |
| Buckow | Neukölln |
| Gropiusstadt | Neukölln |
| Neukölln | Neukölln |
| Britz | Neukölln |
| Rudow | Neukölln |

The polygons to plot them are extracted from the relative shapefile which is loaded with the function `st_read` from the `sf` package to have it already as a sf polygon.

**Listing 1: |berlin_districts_neighbourhoods.R|**

```r
# Load shapefiles
berlin = sf::st_read(file.path(getwd(), "spatial_data", "Berlin-Ortsteile-polygon.shp", fsep="/"))
```

The types of objects used by and created with this package come in very handy since they look like data frames and many functions for data frames can be used on them.

| Name | BEZNAME | geometry |
|---|---|---|
| Buckow : 2 | Treptow-Köpenick :15 | POLYGON :97 |
| Adlershof : 1 | Pankow :13 | epsg:4326 : 0 |
| Alt-Hohenschönhausen: 1 | Reinickendorf :11 | +proj=long...: 0 |
| Alt-Treptow : 1 | Lichtenberg :10 | |
| Altglienicke : 1 | Spandau : 9 | |
| Baumschulenweg : 1 | Charlottenburg-Wilmersdorf: 7 | |
| (Other) :90 | (Other) :32 | |

Since the polygons represent the neighbourhoods, we do not need to perform any transformation on this object. Here we keep only the variables of interest, rename them and reorder the rows.

**Listing 2: |berlin_districts_neighbourhoods.R|**

```r
# Object with the neightbourhoods (and respective district)
berlin_neighbourhood_sf = berlin %>%
    dplyr::rename(id    = Name,
                  group = BEZNAME) %>%
    dplyr::select(id, group, geometry) %>%
    dplyr::arrange(group)

# Buckow is composed of two separate parts, so we need to join them
berlin_neighbourhood_singlebuckow_sf = berlin_neighbourhood_sf %>%
    dplyr::group_by(id, group) %>%
    dplyr::summarize(do_union = TRUE)
```

However, we have the problem with the neighbourhood Buckow, is composed of two separate parts. Therefore we need to unite the neighbourhoods according to their name. In this way we obtain an sf object with 96 polygons, the one of Buckow being a list of polygons.

**Listing 3: |berlin_districts_neighbourhoods.R|**

```r
# Object with the districts
berlin_district_sf = berlin_neighbourhood_sf %>%
    dplyr::group_by(group) %>%
    dplyr::summarize(do_union = TRUE) %>%
    dplyr::mutate(id = group)
```

For the districts we perform the same procedure, but this time we unite the polygons only by their district, which are represented here by the group variable.

## 2.2 Berlin VBB Areas

The VBB (Verkehrsverbund Berlin-Brandenburg) is "the public transport authority covering the federal states of Berlin and Brandenburg" (CITATION: VBB Website). The city of Berlin, in particular, is divided in two fare areas: A, covering the center of Berlin up to the Ringbahn (circular line), and B, from the Ringbahn to the border with Brandenburg. After that there is also the area C, which however will not be covered here since we only consider the city of Berlin.

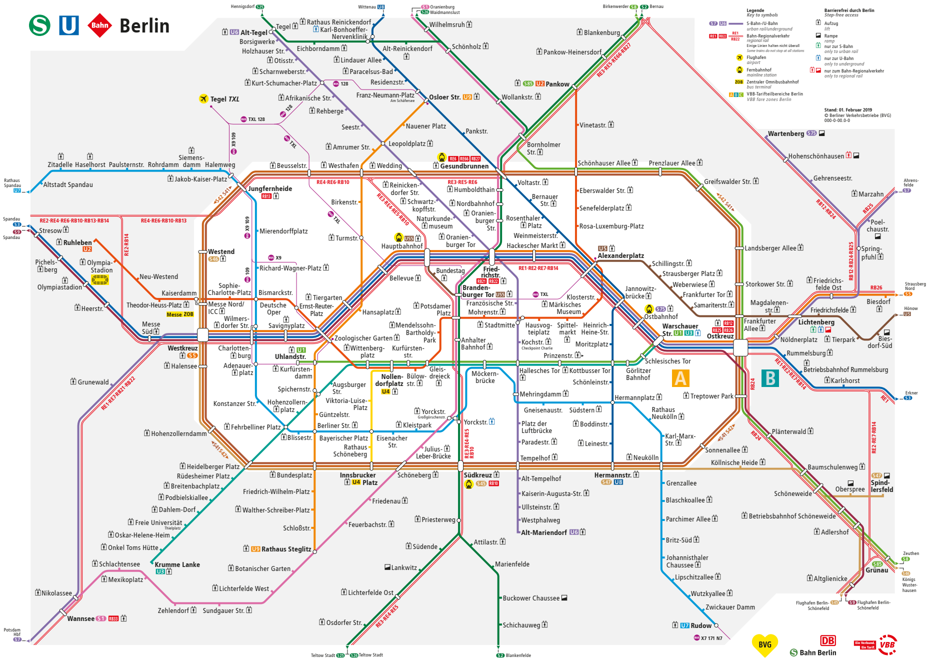

**Figure 1:** Network Map of Berlin Areas A and B (from VBB Website)

We tried to replicate these areas by using the Berlin polygons and the stations points, which are again loaded using `st_read` from the `sf` package.

**Listing 4: |berlin_vbb_areas.R|**

```r
# Load shapefiles
berlin = sf::st_read(file.path(getwd(), "spatial_data", "Berlin-Ortsteile-polygon.shp", fsep="/"))
stations = sf::st_read(file.path(getwd(), "spatial_data", "gis_osm_transport_free_1.shp", fsep="/"))
```

First of all, we need to create the polygon for area A. This is done by joining the points and transforming this into a polygon.

Firstly, we filter the stations that belong to the Ringbahn (border of area A). Since the shapefile does not include the line name for the stations, we need to create our own vector of names. We also add the order in which they need to be connected. The first and last station are the same since the circle need to close.

**Listing 5: |berlin_vbb_areas.R|**

```r
# Create dataframe with names of stations on the Ringbahn (delimits Area A)
ringbahn_names_df = base::data.frame(
      id = c("Südkreuz", "Schöneberg", "Innsbrucker Platz", "Bundesplatz",
          "Heidelberger Platz", "Hohenzollerndamm", "Halensee", "Westkreuz",
          "Messe Nord/ICC", "Westend", "Jungfernheide", "Beusselstraße",
          "Westhafen", "Wedding", "Gesundbrunnen", "Schönhauser Allee",
          "Prenzlauer Allee", "Greifswalder Straße", "Landsberger Allee",
          "Storkower Straße", "Frankfurter Allee", "Ostkreuz", "Treptower Park",
          "Sonnenallee", "Neukölln", "Hermannstraße",
          "Tempelhof", "Südkreuz"),
      stringsAsFactors = FALSE) %>%
    tibble::rownames_to_column(var = "order") %>%
    dplyr::mutate(order = as.numeric(order))
```

# 3 Method/Model/Theory

- How was the data analyzed ?
- Present the underlying economic model/theory and give reasons why it is suitable to answer the given problem.

- Present econometric/statistical estimation method and give reasons why it is suitable to answer the given problem.
- Allows the reader to judge the validity of the study and its findings.
- Depending on the topic this section can also be split up into separate sections.

# 4 Results

- Organize material and present results.

- Use tables, figures (but prefer visual presentation):

  - Tables and figures should supplement (and not duplicate) the text.

  - Tables and figures should be provided with legends.

    *Figure 2 shows how to include and reference graphics. The graphic must be labelled before. Files must be in* `.eps` *format.*

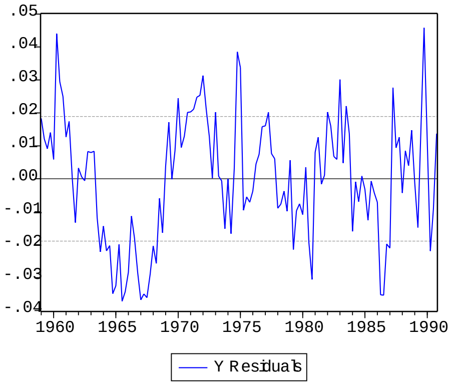

**Figure 2:** Estimated residuals from model XXX. ...

  - Tables and graphics may appear in the text or in the appendix, especially if there are many simulation results tabulated, but is also depends on the study and number of tables resp. figures. The key graphs and tables must appear in the text!

- Latex is really good at rendering formulas:

  *Equation (1) represents the ACs of a stationary stochastic process:*

$$f_y(\lambda) = (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\lambda j} = (2\pi)^{-1} \left( \gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j \cos(\lambda j) \right) \tag{1}$$

  *where $i = \sqrt{-1}$ is the imaginary unit, $\lambda \in [-\pi, \pi]$ is the frequency and the $\gamma_j$ are the autocovariances of $y_t$.*

- Discuss results:

    - Do the results support or do they contradict economic theory ?
    - What does the reader learn from the results?
    - Try to give an intuition for your results.
    - Provide robustness checks.
    - Compare to previous research.

## 5 Conclusions

- Give a short summary of what has been done and what has been found.
- Expose results concisely.
- Draw conclusions about the problem studied. What are the implications of your findings?
- Point out some limitations of study (assist reader in judging validity of findings).
- Suggest issues for future research.

# References

Breusch, T. S. and P. Schmidt (1988): “Alternative Forms of the Wald test: How Long is a Piece of String,” *Communications in Statistics, Theory and Methods*, 17, 2789–2795.

Gallant, A. R. (1987): *Nonlinear Statistical Models*, New York: John Wiley & Sons.

## Declaration of Authorship

I hereby confirm that I have authored this Bachelor's/Master's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, September 30, 2007

your name (and signature, of course)