Pair Assignment No. 3 - Data Snapshot

Daniel J Murphy & Paulo H Kalkhake

November 9, 2016

Introduction

While the effects of Airbnb on Berlin apartment prices has garnered a great deal of focus, significantly less attention has been paid to Airbnb's effect on the hotel industry. Airbnb claims that they are largely complementary to hotels, since most Airbnb listings are found outside of the main hotel district in a given city. In Berlin for example, 77% of Airbnb listings are outside of the major hotel districts.

However, a paper by Zervas, Proserpio, and Byers (2016) found that the rise of Airbnb had a negative effect on hotel revenue in the state of Texas. This shows that, while Airbnb may be complentary, they also compete with the incumbent hotel industry. In this paper we will seek to illustrate the magnitude of the "Airbnb effect" on hotels in Berlin. To that end, two hypotheses will guide our thinking.

> *H1: The higher the Airbnb supply in a given district in Berlin, the lower the hotel occupancy rate will be in that same district.*

> *H2: Since Airbnb's listings are concentrated in districts with low hotel density, the effect of substitution will be more pronounced in those districts.*

Data & Variables

Our data comes from three different sources, the Statistical Information System Berlin/Brandenburg (SBB) (StatIS-BBB 2016), the Federal Statistical Office and the statistical offices of the Länder (FSO)[1] (Germany 2015), and *InsideAirbnb.com* (Cox 2016). The data covers the time period between 2010 and 2014.

From the Statistical Information System Berlin/Brandenburg, we collected monthly data on the number of overnight stays from guests and guest arriving at accommodations in the reporting period in Berlin (StatIS-BBB 2016). The surveys are carried out at the beginning of each month and refer to the reporting period of the previous month. The results are organized regionally according to districts and municipalities, allowing us to have specific data for each of the twelve districts in Berlin. Further, we gathered data for monthly household income and the number of employed and umemployed persons per district.

[1] Both databases use JAVA-based website, which did not allow direct web scraping. The data was manually downloaded

From the Regional Database of Germany, we collected data on the supply of hotel beds in each district in a given year. Unfortunately this data is only recorded annually. Hence, for further anylisis, we aassume that the number of beds in each hotel stay constant throughout the year.

From *InsideAirbnb.com*, we scraped data on listings and reviews on all Airbnb listings for Berlin between 2010 and 2014. The most important variables for our model are the following (Cox 2016).

1. The neighbourhood of each listing
2. The date that an Airbnb host signed up
3. The date of the last review of each listing
4. The date of each inividual review of each listing

Data Cleaning & Merging

In the original detailed listings data set from *InsideAirbnb.com*, there were 92 variables covering topics ranging from room price to information on the host. Most of these variables were superfluous, however, leading us to drop all but six variables. Those were listing ID, host ID, listing Neighbourhood, room type, number of beds, date of host registration, date of the first review, and the date of the last review.

Based on the data we computed a yearly unemployment rate per district, assuming that it is constant over the period.

Methodology

We will use a Regression Discontinuity Model to evaluate whether the emergence of Airbnb in Berlin has affected hotel occupancy rates. By introducing a binary variable that is equal to one for all observations after Airbnb's market entry in Berlin and interacting that variable with all regressors in our model, we will account for all the changes in hotel occupancy rates before and after Airbnb's market entry in Berlin, i. e. the dicontinuity at market entry (Jacob et al. 2012).

Taking the arguments above into account, our main specification for our proposal would look like this:

$$(1) \quad OccupRate_{it} = \beta_i * \log Abb_{it} + \tau_i + \beta_j * X_j + \varepsilon_{it}$$

Here, *OccupRate* is the occupancy rate for all hotels in district i at time t. Log *Abb* is the log of the approximated total number of Airbnb listings available in district i at time t. We take the log of each of these variables in order to normalize their distribution. Next, we control for economic conditions specific to each neighbourhood (unemployment rate and income) and a district-specific linear time trend to account for unobserved heterogeneous variation across districts by using a month-year time dummy τ.

Analysis

Descriptive Statistics

Upon plotting Airbnb supply against hotel room occupancy, we were surprised to see a positive correlation. However, upon taking a second look at the Zervas, Proserpio, and Byers (2016) paper, we realized the importance of accounting for general demand in an area as it becomes more attractive to tourists. Zervas, Proserpio, and Byers (2016) accomplished this by controlling for passengers listing the local airport as their final destination. We will incorporate a similar statistic in our final analysis, as tourism in Berlin has increased dramatically in recent years.

(both descriptive and inferential statistics)

Bibliography

Cox, Murray. 2016. "Inside Airbnb." http://insideairbnb.com/get-the-data.html.

Germany, Regional Database. 2015. https://www.regionalstatistik.de.

Jacob, Robin Tepper, Pei Zhu, Marie-Andrée Somers, and Howard S Bloom. 2012. *A Practical Guide to Regression Discontinuity.* Citeseer.

StatIS-BBB. 2016. https://www.statistik-berlin-brandenburg.de/webapi/jsf/dataCatalogueExplorer.xhtml.

Zervas, Georgios, Davide Proserpio, and John Byers. 2016. "The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry." *Boston U. School of Management Research Paper*, no. 2013-16.