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). Before conducting our analysis we cleaned, merged and manipulated the data sources.

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

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

yearly household income groups and the number of employed and umemployed persons per district. Based on the data we calculated a yearly average household income and unemployment rate per district, assuming that both stay constant in each period.

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 assume that the number of beds in each hotel stay constant throughout the year. For our main dependent variable, we used data on the number of beds generally available in each district per day in a given year and the monthly data on the number of overnight stays from guests to compute a variable as a proxy for hotel occupancy (in per cent).

From *InsideAirbnb.com*, we scraped data on 15,368 listings, i.e. apartments or rooms, and reviews for Berlin from August, 2008 until October, 2015. Amongst 92 variables for each listing covering topics ranging from room price to information on the host, the data includes (1) the neighbourhood of each listing, (2) the date that an Airbnb host signed up, and (3) the date of the last review of each listing.

Our most significant methodological challenge the absence of precise listing availability during our period of interest, as it is not directily available in the data. However, in keeping with the Zervas, Proserpio, and Byers (2016) methodology, we used data on when the host became Airbnb member as a proxy for market entry. We then construct a variable for cumulative supply based on this information, i.e. the total number of users with listings in a district that have signed up on Airbnb in prior to that month. This is not ideal, as this proxy does not take into account whether or not a listing was available in a given month. However, Airbnb itself is unable to produce exact supply data, because owners do not accurately or deliberately update the availability.

The final data set covers 720 monthly obversations across tewlve districts in Berlin and covers the time period between 2010 and 2014.

Descriptive Analysis

Upon plotting Airbnb supply against hotel room occupancy, we were surprised to see a positive correlation.

However, 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.

Inferential Analysis

Thus for an appropriate analysis of the data, tools of multilevel modeling should be used. As stated above, it is reasonable to assume a different intercept for each district.

Zeileis and Grothendieck (2005), Wickham and Francois (2016), Dowle et al. (2015), Wickham (2016b), Gandrud (2016), Wickham and Hester (2016), Wickham (2011), Wickham (2016a), Chan et al. (2016), Ooms (2016), Grosjean and Ibanez (2014), Harrell Jr, Charles Dupont, and others. (2016), Xie (2014), Grolemund and Wickham (2011), Wickham (2009)}

Bibliography

Chan, Chung-hong, Geoffrey CH Chan, Thomas J. Leeper, and Jason Becker. 2016. *Rio: A Swiss-Army Knife for Data File I/O.*

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

Dowle, M, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta, and E Antonyan. 2015. *Data.table: Extension of Data.frame.* https://CRAN.R-project.org/package=data.table.

Gandrud, Christopher. 2016. *DataCombine: Tools for Easily Combining and Cleaning Data Sets.* https://CRAN.R-project.org/package=DataCombine.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. http://www.jstatsoft.org/v40/i03/.

Grosjean, Philippe, and Frederic Ibanez. 2014. *Pastecs: Package for Analysis of Space-Time Ecological Series.* https://CRAN.R-project.org/package=pastecs.

Harrell Jr, Frank E, with contributions from Charles Dupont, and many others. 2016. *Hmisc: Harrell Miscellaneous.* https://CRAN.R-project.org/package=Hmisc.

Ooms, Jeroen. 2016. *Curl: A Modern and Flexible Web Client for R.* https://CRAN.R-project.org/package=curl.

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

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. http://ggplot2.org.

———. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

———. 2016a. *Stringr: Simple, Consistent Wrappers for Common String Operations.* https://CRAN.R-project.org/package=stringr.

———. 2016b. *Tidyr: Easily Tidy Data with 'Spread()' and 'Gather()' Functions.* https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, and Romain Francois. 2016. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.

R-project.org/package=dplyr.

Wickham, Hadley, and James Hester. 2016. *Xml2: Parse Xml*. https://CRAN.R-project.org/package=xml2.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zeileis, Achim, and Gabor Grothendieck. 2005. "Zoo: S3 Infrastructure for Regular and Irregular Time Series." *Journal of Statistical Software* 14 (6): 1–27. doi:10.18637/jss.v014.i06.

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.