Airbnb Data Project

Betsy Bersson, Michael Christensen, Evan Knox

Duke University

February 4, 2020

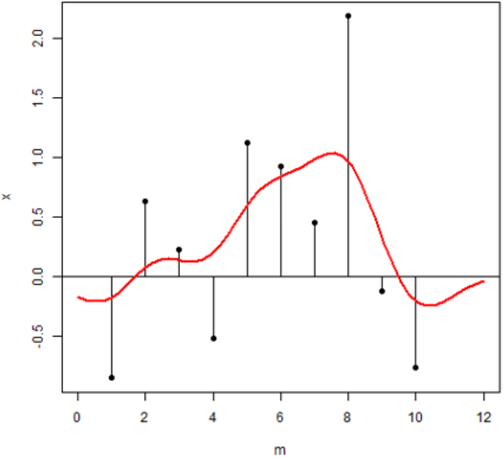
Data Organization

Decided to look at data that was a proxy for a hotel

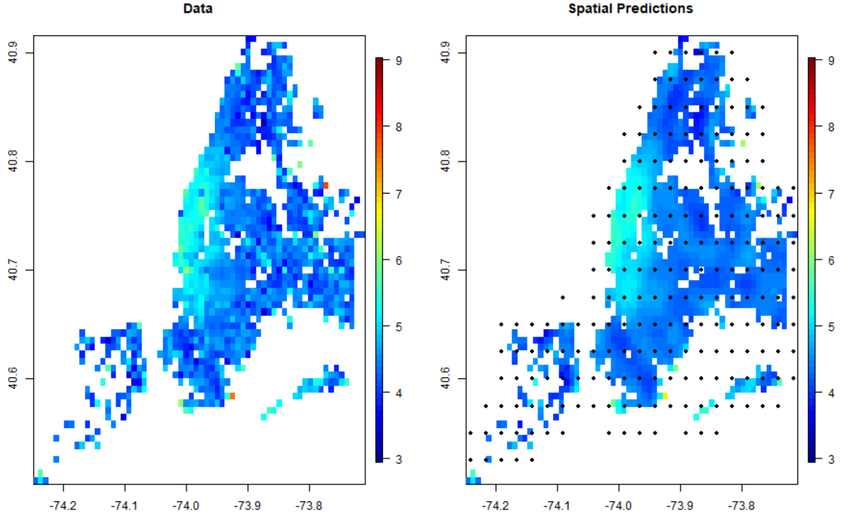
- Limit prices to be between 9 and 9999
- Limit minimum nights to be less than 14
- Remove Airbnbs that are no longer available
- Remove listings with no reviews

Process Convolution Model

- Process Convolution (Higdon, 1998)
- Apply smooth kernel (e.g. normal) to distance function between observed data and discrete collection of knots
- $y(s) = X\beta + K\theta + \epsilon$
- $K_{ij} = k(s_i - \omega_j)$
- K acts as collection of additional covariates in linear model



Price Data and Basic Spatial Model Fit



Left: Log Price Data. Right: Predicted values using only spatial component of model.

Hierarchical Neighborhood Structure

- 5 Boroughs containing 217 neighborhoods
- Fit Bayesian hierarchical model
- $\beta_{b_k} \sim N(\mu_0, \sigma_b^2)$
- $\beta_{n_{kl}} \sim N(\beta_{b_k}, \sigma_n^2)$

Model for Airbnb Price

- $\log(\text{price}) = \beta_{\boldsymbol{n}} + X\beta + K\theta + \epsilon$
- Neighborhood effect acts as random intercept
- X matrix contains terms for room type, minimum nights, time since last review, reviews per month, host listings, and availability
- $K\theta$ adds spatial effect
- Model implemented in JAGS (slow to fit, results to come)

Price Model Results

- Still computing, but EDA confirms that there are variables that impact price, believe it or not

Assessing Airbnb Popularity

- Popularity Metric: Number of reviews with an offset of number of reviews (backed out from total number of reviews and reviews per month)
- Model: Negative-Binomial with a log link

$$\log \frac{\mu_i}{t_i} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Popularity Model Results

Variable	e^{β_i}	
Price	1.00	***
Private room	0.93	***
Shared room	0.79	***
Host listing count	1.00	**
Minimum nights	0.86	***
Name length	1.03	***
Availability	1.00	***
Last review year	2.16	***
Brooklyn	0.96	
Manhattan	1.04	
Queens	1.12	***
Staten Island	0.96	

Room Listings by Neighborhood and Borough

Question 3: does room type vary by neighborhood?
Answer: Yes, but too many neighborhoods to examine in-depth. Most variation is ratio of whole home/apt to private room, usually between 0.5 and 2.
Corona and Port Morris have more shared rooms listings than either other category; Harlem, East Harlem, and Hell's Kitchen also have lots of shared rooms
By borough:

Borough	Entire home/apt	Private room	Shared room
Bronx	0.356	0.592	0.051
Brooklyn	0.505	0.475	0.019
Manhattan	0.560	0.411	0.028
Queens	0.372	0.591	0.037
Staten Island	0.469	0.516	0.014

Shared rooms rare; entire home/apt much more common in Manhattan and Brooklyn

Text Analysis

Methods: median price/reviews per month by word, subject to number of appearances

LDA was used, but limited in value

Excluding stopping words and borough/room type, main important word types for price and reviews per month were:

1. positive adjectives - spectacular, stunning, designer, luxury, charming, amazing, perfect. High on AFINN sentiment scale.
2. Ways to decrease price off-listing - No cleaning/service fee
3. Hotel names - Wyndham, Sonder, Incentra
4. Numbers of bedrooms/bathrooms
5. Location markers - Some borough/neighborhood, also stock exchange and (for reviews) nearby airport/subway

Most expensive/popular listing: entire home in Manhattan/Brooklyn, large number of beds/baths (5+), near Subway/JFK, no cleaning fees, heavy use of adjectives from above

The End