# How to Predict Spatial Distribution of Airbnb in Cities: A Machine Learning Method

Mengchen Shi
Advised by: Dr. Richard Evans

M.A. in Computational Social Science
The University of Chicago

April 18, 2019

# Research Question

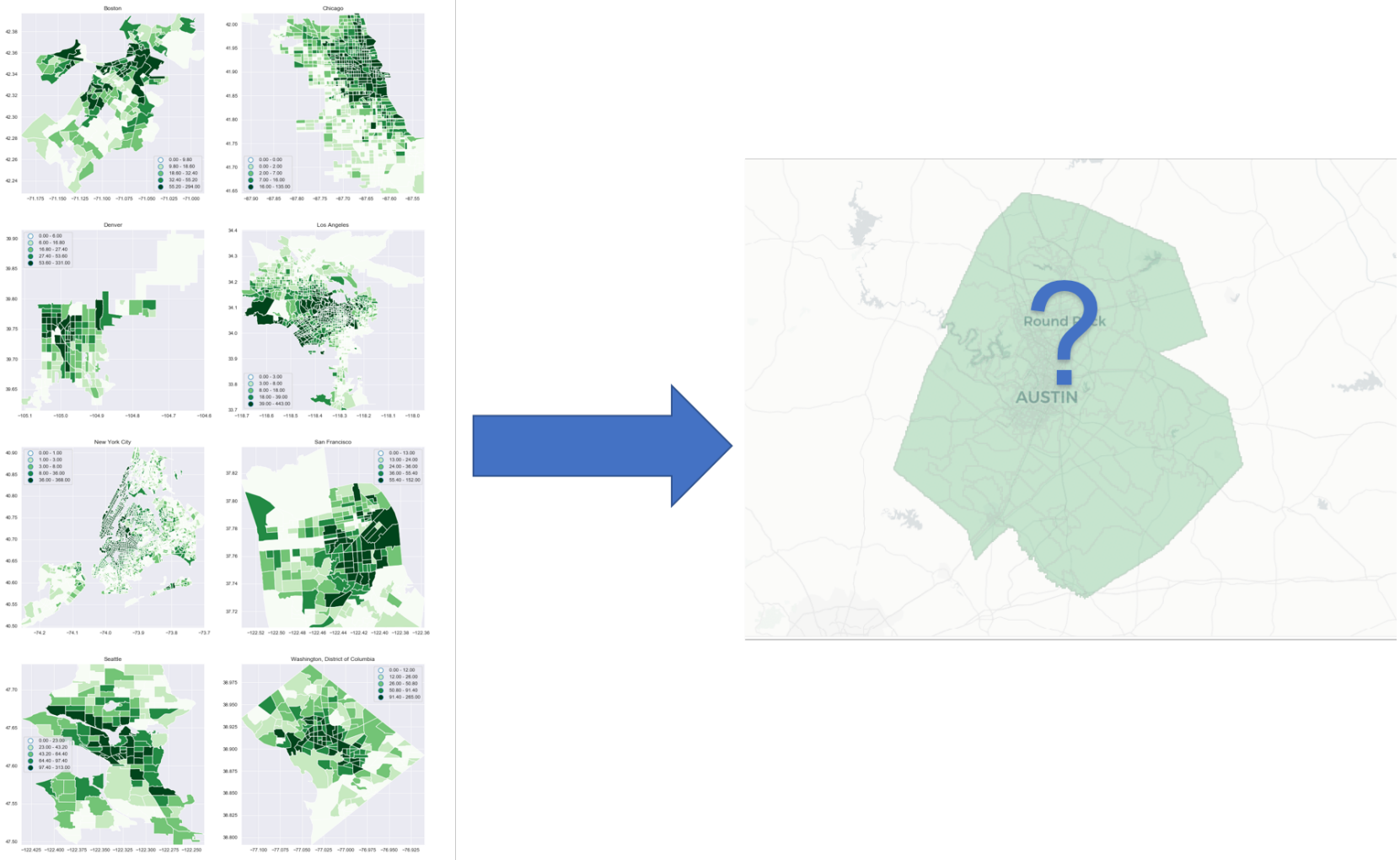
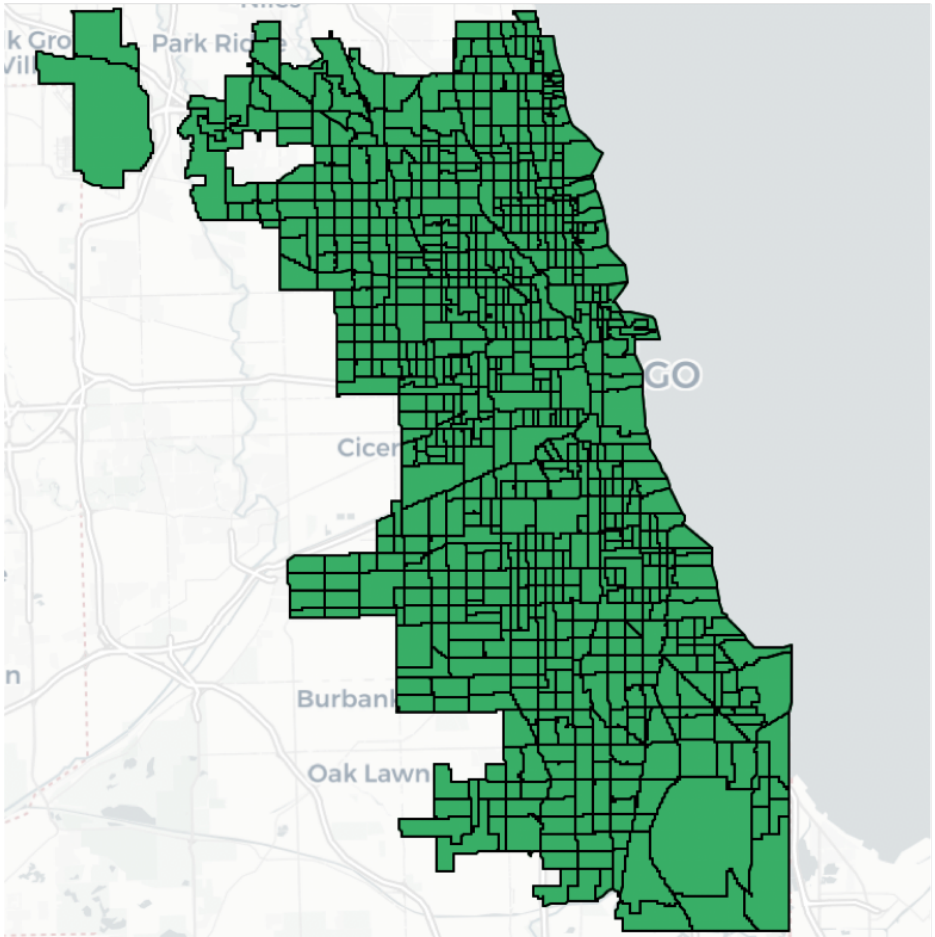


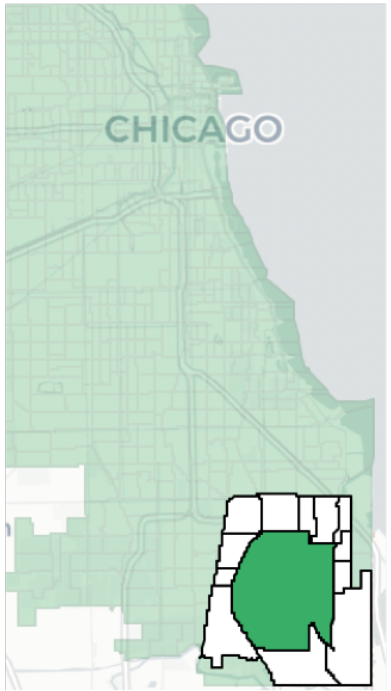

Figure: Predict Airbnb Listings Distribution in a City

# Spatial Units and Neighbors

801 Census Tracts in Chicago (2010 Census)




Queen contiguity-based Neighbors



# Data

- United States Census Bureau: Census data and spatial units
- Inside Airbnb: a website periodically publishes snapshots of Airbnb listings founded by Murray Cox
- OpenStreetMap: a collaborative project to create a free editable map of the world

# Variables of Interest

| Category | Variable | Description |
|---|---|---|
| Airbnb | num_list | Number of Airbnb listings in a given tract |
| Geographic | distance | Distance a tract to downtown |
| | hotel | number of hotels in the area |
| | poi | Number of point of interests in the area. |
| | trans | Number of public transportation infrastructure |
| | pop_den | Population density in the tract |
| Economic | unemp | Proportion of unemployed residents |
| | log_inc | Log(Median of household income in an area) |
| | log_hvalue | Log(Median of housing value in an area) |
| | owner | Proportion of owner-occupied properties |
| | poverty | Proportion below poverty level |
| Social | edu | Proportion of residents with an advanced degree |
| | young | Proportion of people aged between 20 and 34 |
| | race | Race diversity represented by Gini-Simpson index |

# Variables of Interest

## Variables about neighbors

Suppose a census tract has $m$ neighbors, and we have $n$ variables.
Neighbor mean:

$$\frac{\sum_{j=1}^{m_i} X_{ij}}{m}, i = 1, 2, ..., n; j = 1, 2, ..., m$$

Neighbor maximum:

$$Max(X_{ij}), i = 1, 2, ..., n; j = 1, 2, ..., m$$

Neighbor minimum:

$$Min(X_{ij}), i = 1, 2, ..., n; j = 1, 2, ..., m$$

# Spatial Autocorrelation

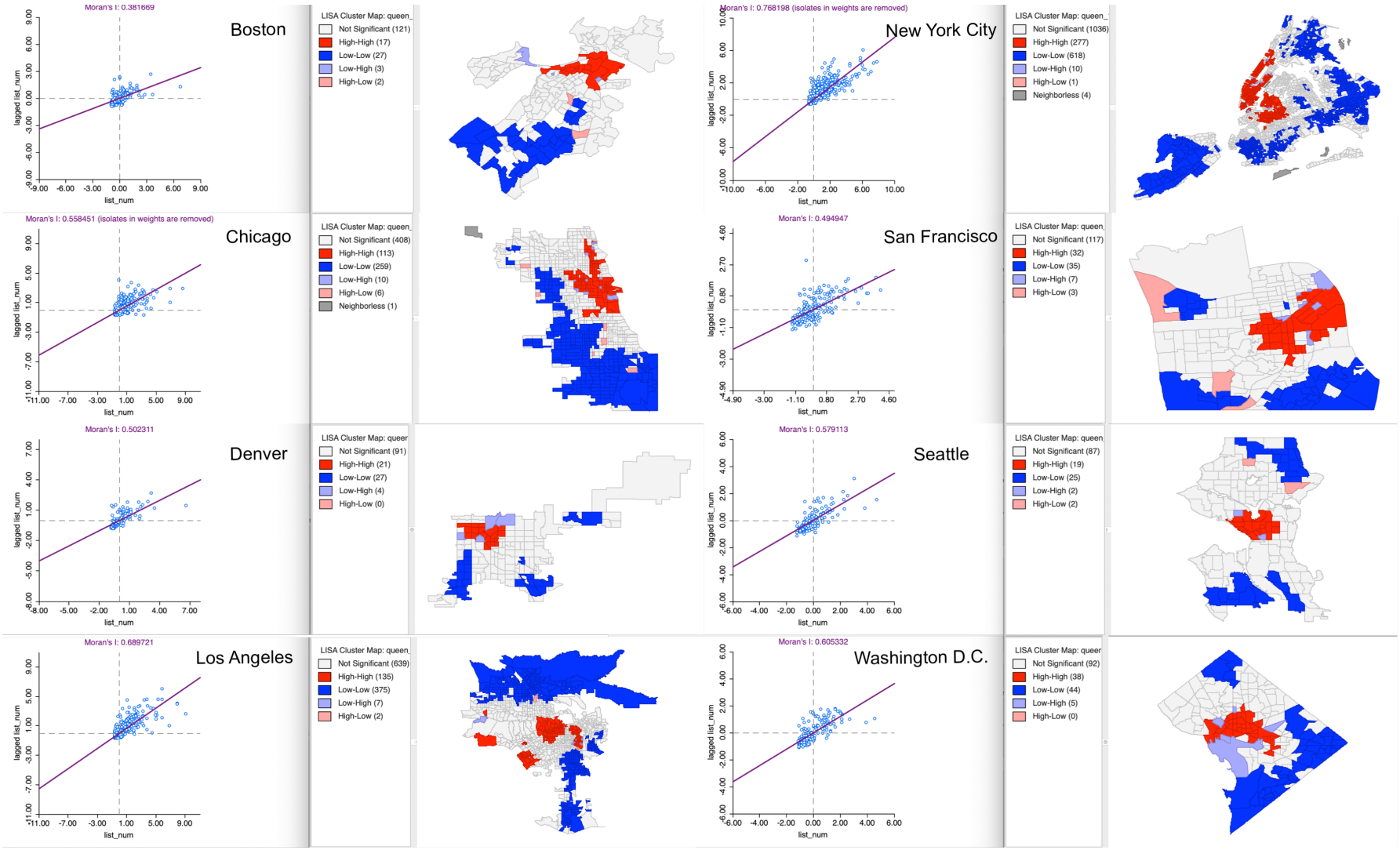



Figure: Local Moran's I and Clusters

# Regression Results

[scale=0.5]

Table: RMSE of Lasso, Random Forest, and XGBoost Regression

| Model | OLS | Random Forest | XGBoost |
|---|---|---|---|
| | Spatial v.s.(non-spatial) | | |
| All 8 cities | 31.49 (32.37) | 26.89 (27.94) | 25.39 (27.62) |
| Boston | 37.69 (26.78) | 32.93 (31.95) | 33.69 (33.61) |
| Chicago | 10.01 (9.20) | 9.76 (9.70) | 9.84 (9.34) |
| Denver | 32.93 (27.10) | 35.04 (35.63) | 37.38 (40.39) |
| Los Angeles | 34.77 (34.66) | 30.69 (33.39) | 32.40 (33.87) |
| New York City | 28.60 (28.73) | 23.01 (24.53) | 21.96 (24.55) |
| San Francisco | 22.49 (19.95) | 21.67 (22.46) | 20.66 (22.88) |
| Seattle | 30.92 (27.90) | 37.35 (37.35) | 36.61 (37.71) |
| Washington, D.C. | 36.18 (27.18) | 30.69 (32.58) | 31.66 (34.12) |

# Feature Importance

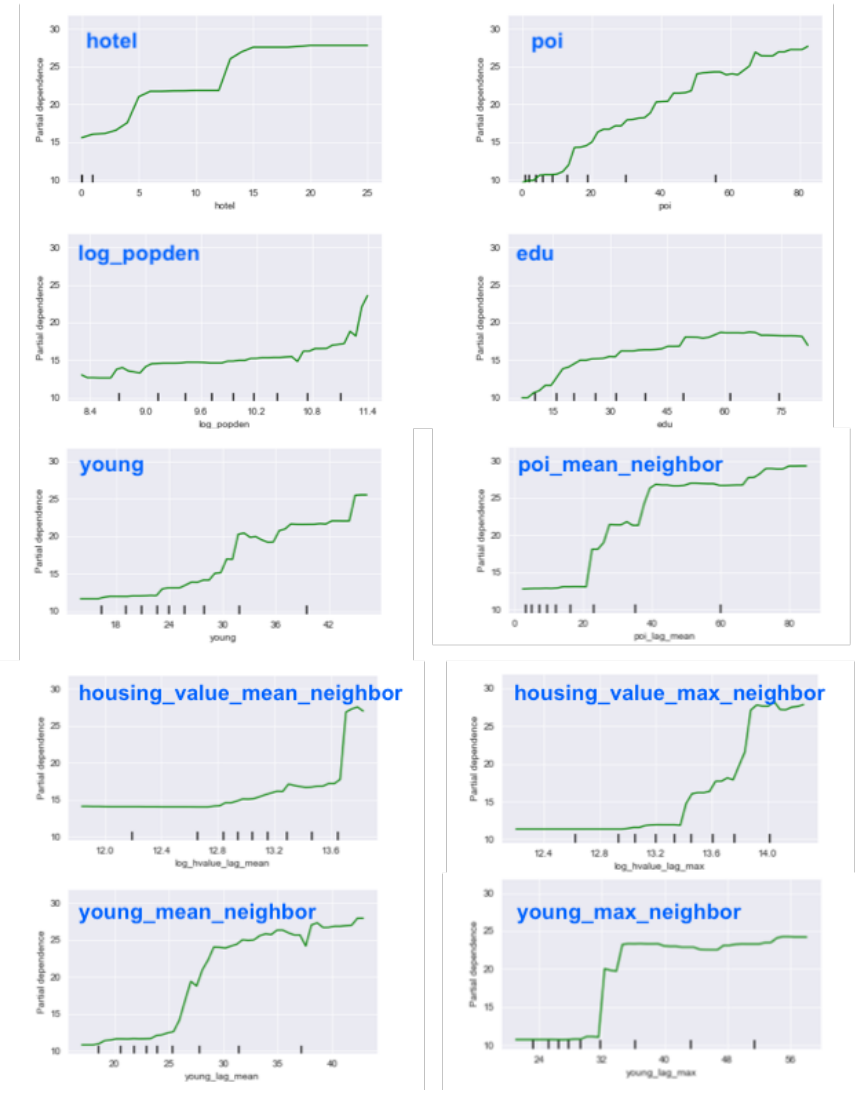
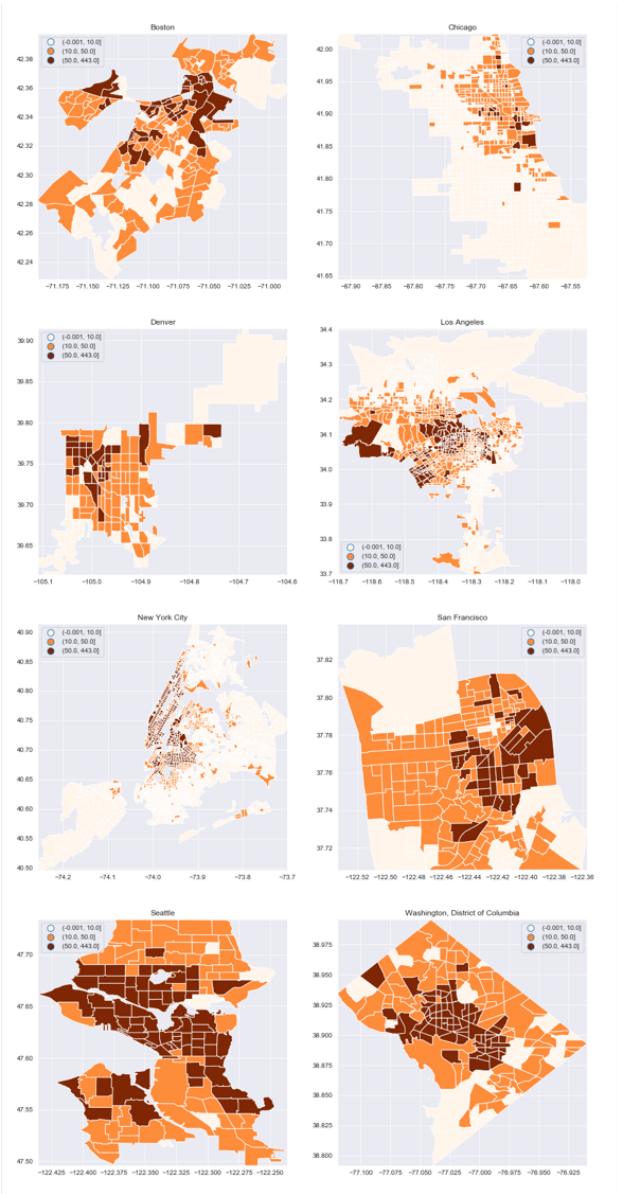



Figure: Most important features selected by XGBoost regression

# XGBoost Prediction Results




Out of sample prediction