Uncovering the hidden universe of rental units in Surrey
UBC Data Science for Social Good 2018

By: Jocelyn Lee, Andy Fink, Hyeongcheol Park, Zhe Jiang
Overview

• Introduction
• Data Sources and Collection
• Data Processing
• Classification Model Results
• Discussion and Future Work
The Hidden Housing Market

• Surrey is growing at a rapid rate

• Rental unit information for Surrey is incomplete

• Social consequences:
  • School overpopulation
  • Inadequate public transportation availability
  • Lack of available street parking
  • Unsafe secondary suite rentals

• Goal: provide the City of Surrey with up to date information on the type, distribution and amount of secondary suites
Data Sources

Open Sources:
- Statistics Canada
- Surrey's Open Data
- CMHC SCHL

Non-Open Sources:
- City of Surrey
- Craigslist
- VRBO
- Airbnb
- Kijiji
- Quantitative Rhetoric
Data Collection

- Different web crawlers built for different websites:
  - Most postings from Craigslist: 3,000~4,000 raw data monthly
  - Other sources (mainly Kijiji and VRBO) comprise ~300 data monthly
  - Short-term rental very few: VRBO and Airbnb

- Crawler deployed on UBC server and collects data every day

- Current research was mainly based on data collected over the past 3 months
Data Cleaning & Processing

• Excluded non-Surrey region: latitude-longitude (GIS), title, location, url

• Standardization

<table>
<thead>
<tr>
<th>description</th>
<th>housing_type</th>
<th>lat</th>
<th>long</th>
<th>location</th>
<th>price</th>
<th>source</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ 2br - 550ft</td>
<td></td>
<td></td>
<td></td>
<td>(Ladner)</td>
<td>$1,600</td>
<td>Craigslist</td>
<td>2 bedroom su</td>
</tr>
<tr>
<td>/ 4br - 2430f</td>
<td></td>
<td>49.029901</td>
<td>-123.07195</td>
<td>(delta)</td>
<td>$2,600</td>
<td>Craigslist</td>
<td>Delta Tsaww</td>
</tr>
</tbody>
</table>

• Deduplication
  • Set Theory (Deterministic Record Linkage)
  • Fuzzy Matching (Probabilistic RL)

<table>
<thead>
<tr>
<th>lat</th>
<th>long</th>
<th>location</th>
<th>price</th>
<th>source</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.188634</td>
<td>-122.8632</td>
<td>(SURREY CE)</td>
<td>$480</td>
<td>Craigslist</td>
<td>AUG 1* FURNISHED PRIVATE ROOM - SINGLE OCCUPANCY-FEMALE STUDENT</td>
</tr>
<tr>
<td>49.188634</td>
<td>-122.8632</td>
<td>(SURREY CE)</td>
<td>$480</td>
<td>Craigslist</td>
<td>AUG 1* BIG FURNISHED ROOM-ALL INCLUDED-SINGLE FEMALE-SURREY CENTRAL</td>
</tr>
<tr>
<td>49.188634</td>
<td>-122.8632</td>
<td>(SURREY CE)</td>
<td>$480</td>
<td>Craigslist</td>
<td>AUG 1* CLEAN FURNISHED ROOM-ALL INCLUDED-SINGLE FEMALE-SURREY CENTRAL</td>
</tr>
</tbody>
</table>

• Missing value imputation for supervised learning

<table>
<thead>
<tr>
<th>lat</th>
<th>long</th>
<th>location</th>
<th>price</th>
<th>source</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.108673</td>
<td>-122.84547</td>
<td></td>
<td></td>
<td>Craigslist</td>
<td>1 bedroom in 2 bedroom basement only for girls</td>
</tr>
</tbody>
</table>
## Manually Labelled Data and Proportions

<table>
<thead>
<tr>
<th>Categories of Rental</th>
<th>% of Listings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-market Rental</td>
<td>0</td>
</tr>
<tr>
<td>Purpose-built</td>
<td>0.8</td>
</tr>
<tr>
<td><strong>Entire Condo</strong></td>
<td><strong>13.9</strong></td>
</tr>
<tr>
<td><strong>Entire House or Townhouse</strong></td>
<td><strong>25.0</strong></td>
</tr>
<tr>
<td><strong>Basement Secondary Suite</strong></td>
<td><strong>22.1</strong></td>
</tr>
<tr>
<td>Non-basement Secondary Suite</td>
<td>6.8</td>
</tr>
<tr>
<td>Laneway or Coach House</td>
<td>1.4</td>
</tr>
<tr>
<td>Unspecified Secondary Suite</td>
<td>4.5</td>
</tr>
<tr>
<td><strong>Individual Rooms in a Condo or House</strong></td>
<td><strong>19.8</strong></td>
</tr>
<tr>
<td>Non-housing Postings</td>
<td>5.7</td>
</tr>
</tbody>
</table>
“I am a student Punjabi girl. I need someone international Punjabi student to share my one bedroom basement. Internet included no laundry. Available immediately.”
Problems with Such Classification

- It consumes too much to do manual labeling:
- So we built automatic classifiers.

- With the 1000-entry labeled dataset we had:
  - Some of the 10 classes had too few categories;
  - 1000 entries were not supportive enough to train a model to classify 10 categories;

- Shall we condense the current categories into fewer?
3 Category Classification

- Solution: Collapse into 3 categories:
  - 1 - Entire House or Condo 39.7%
  - 2 - Secondary Suites 34.8%
  - 3 - Individual Rooms 19.8%
(Non-housing ads excluded)
Final Classification Results

• From the Random Forest Classifier

<table>
<thead>
<tr>
<th>Category</th>
<th>% Predicted</th>
<th>% Labelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Entire House or Condo</td>
<td>39.2</td>
<td>41.8</td>
</tr>
<tr>
<td>2 - Secondary Suites</td>
<td>37.6</td>
<td>37.0</td>
</tr>
<tr>
<td>3 - Individual Rooms</td>
<td>23.2</td>
<td>21.2</td>
</tr>
</tbody>
</table>

• Prediction Accuracy: **91%** with an out of bag error of 11%
Spatial Distribution of Online Postings

- Maps created using QGIS 3.2.3
- Counts measured using Dissemination Areas
- Highest posting densities in Douglas and City Center, high density in Cloverdale
Spatial Distribution of Online Postings

Private Room
Secondary Suite
Entire Property
Spatial Distribution of Manually Classified Set

• Manually classified set

• Each dot represents an individual posting

• Noticeable clusters in City Centre, Cloverdale and South Surrey
Cluster Examples

- Condos
- Coach/Laneway Houses
- Basement/Private Rooms
- Entire Houses
Discussion

- Current dataset for supervised learning is small:
  - Distribution of categories might be different in real situation;
  - Classifier model possibly overfitting;

- Data was collected over only 3 months;

- Two other models were not ensembled, could have been used to increase accuracy.
Future Work

- Validation and analysis over a time-series;

- Pipeline development: a set of user-friendly automatic tools;

- More robust classifiers with Natural Language Interpretation:
  - Better data imputation: from addresses, descriptions
  - More features generated from titles/descriptions
  - Ensembled methods
Thanks for watching!

Questions?
Final Classification Results
Other Classification Results

- From the Naive Bayes Model (without normalization)

<table>
<thead>
<tr>
<th>Category</th>
<th>% Predicted</th>
<th>% Labelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Entire House or Condo</td>
<td>28.07</td>
<td>39.7</td>
</tr>
<tr>
<td>2 - Secondary Suites</td>
<td>46.04</td>
<td>34.8</td>
</tr>
<tr>
<td>3 - Individual Rooms</td>
<td>20.89</td>
<td>19.8</td>
</tr>
</tbody>
</table>

- Prediction Accuracy: 75%
Other Classification Results

• From the Generalized Additive Model with Majority Voting

<table>
<thead>
<tr>
<th>Category</th>
<th>% Predicted</th>
<th>% Labelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Entire House or Condo</td>
<td>46.0</td>
<td>41.8</td>
</tr>
<tr>
<td>2 - Secondary Suites</td>
<td>37.0</td>
<td>37.0</td>
</tr>
<tr>
<td>3 - Individual Rooms</td>
<td>16.9</td>
<td>21.2</td>
</tr>
</tbody>
</table>

• Prediction Accuracy: 82.53%