Scraping rental holiday homes in Finland

Matti Kokkonen / Katja Löytynoja
EFGS 2018 October 17th
What is web scraping?

• Automated retrieval of elements from web site’s source code
• Both visible and invisible for regular users
• Packages available for multiple programs, for example Python and R
• Regular html easy to scrape, JavaScript requires web browser automation
• Fast and relatively easy access to up-to-date data
• Minimizes perceived response burden
• Load on the server, immaterial rights -> scraping ethics
  • Identify yourself and provide contact information (user agent string)
  • Respect robots.txt and terms of use
  • Scrape during quiet periods
  • Do not reuse data commercially
  • Return value
  • Ask a permission
  • Alternatively use API or negotiate a direct access to server
• Statistical legislation and GDPR
Why and what?

- New statistics where target group identified through web scraping
- Largest domestic booking agencies and market platforms
- Started out with 11 scrapers, 5 still in use
- Variables scraped
  - Address
  - Coordinates
  - Construction year
  - Area (m2)
  - Number of beds
  - Rental price (usually per week or weekend) if available
  - Advertiser (used to identify doubles)
  - URL to advertisement (an unique id)
5

19 October 2018

Katja Löytynoja

Coordinate reference system change: From ETRS-TM35FIN to WGS84

Reverse geocoding endpoint
http://api.digitransit.fi/geocoding/v1/reverse
Address verification, correct postal codes

Population Information System
Building information

Nearest coordinate points
Closest size match
Closest construction year
Other?

Rental holiday house database

Import to SAS, conversion to sas7bdat format
Data manipulation: information retrieval from free text fields etc.

digitransit

Statistics Reports
Informant Incentive

Web Scraper(s)

XML
## Raw data

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>Address</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Not available</td>
<td>Not available</td>
<td>0.6 %</td>
<td>14.7 %</td>
</tr>
<tr>
<td>Available</td>
<td>Available</td>
<td>50.5 %</td>
<td>34.2 %</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
<td>51.1 %</td>
<td>48.9 %</td>
</tr>
</tbody>
</table>

## After reverse geocoding

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>Address</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Not available</td>
<td>Not available</td>
<td>0.7 %</td>
<td>1.4 %</td>
</tr>
<tr>
<td>Available</td>
<td>Available</td>
<td>0.1 %</td>
<td>97.8 %</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
<td>0.8 %</td>
<td>99.2 %</td>
</tr>
</tbody>
</table>
Target group?

Classification of Buildings 1994

- 011 One-dwelling houses
- 012 Two-dwelling houses
- 039 Other blocks of flats
- 041 Free-time residential buildings
- 124 Rental holiday cottages and flats
- 811 Cowsheds, pighouses, hen-houses, etc.
- 899 Other buildings in agriculture, forestry and fishing
- 931 Sauna buildings
- 941 Outbuildings
- 999 Buildings n.e.c.
A sample of the data was linked to Population Information System’s Building Registry by:

- Nearest coordinate match
- Closest size match
- Closest construction year match
- Postal code was used for subsetting Building Registry
- The closest match for all of the above was selected
- Subsetting by postal code and street name could be quicker
Quality

- Coordinate precision and accuracy
  - WGS84 vs. building register ESTR-TM35FIN
  - How coordinates are created for the add? Geocoding?
  - From three to seven decimals
  - Coordinates may point to same spot even when street number is different -> address vs. coordinates
  - Address may point to advertiser, not cottage -> How usual is this? In which sites is this common?
  - Reverse geocoding may provide wrong street number
- Are all adds about cottages? There are row houses, tree tents, igloos, even bicycles
- Removing doubles and buildings in Accommodation Statistics
Next steps

- Qualities of reverse geocoding and linking to Building Registry
- Informant incentive
- Optimizing code
  - Reverse geocoding through API takes time
  - Currently linking to Building Registry is highly compute intensive
- Building a strong theoretical background as the data can be considered a non-probability sample