staty: Quality Assurance for Public Transit Stations in OpenStreetMap

Hannah Bast, Patrick Brosi and Markus Näther

University of Freiburg

ACM SIGSPATIAL 2020 - Seattle, Washington, USA
Motivation - Errors in OSM station data

- Mainly due to human error (outdated data, typos, ...)
- Correct station data is necessary e.g. for route planning, station search, transit graph drawing, ...
Motivation - Errors in OSM station data

- Mainly due to human error (outdated data, typos, ...)
- Correct station data is necessary e.g. for route planning, station search, transit graph drawing, ...

<table>
<thead>
<tr>
<th>name</th>
<th>California Street &amp; Jones Street</th>
</tr>
</thead>
<tbody>
<tr>
<td>network</td>
<td>Muni</td>
</tr>
<tr>
<td>operator</td>
<td>San Francisco Municipal Railway</td>
</tr>
<tr>
<td>public_transport</td>
<td>stop_position</td>
</tr>
<tr>
<td>railway</td>
<td>tram_stop</td>
</tr>
<tr>
<td>short_name</td>
<td>Jones &amp; Beach</td>
</tr>
<tr>
<td>tram</td>
<td>yes</td>
</tr>
<tr>
<td>wheelchair</td>
<td>no</td>
</tr>
</tbody>
</table>
Motivation - Errors in OSM station data

- Mainly due to human error (outdated data, typos, ...)

<table>
<thead>
<tr>
<th>name</th>
<th>California Street &amp; Jones Street</th>
</tr>
</thead>
<tbody>
<tr>
<td>network</td>
<td>Muni</td>
</tr>
<tr>
<td>operator</td>
<td>San Francisco Municipal Railway</td>
</tr>
<tr>
<td>public_transport</td>
<td>stop_position</td>
</tr>
<tr>
<td>railway</td>
<td>tram_stop</td>
</tr>
<tr>
<td>short_name</td>
<td>Jones &amp; Beach</td>
</tr>
<tr>
<td>tram</td>
<td>yes</td>
</tr>
<tr>
<td>wheelchair</td>
<td>no</td>
</tr>
</tbody>
</table>
Motivation - Errors in OSM station data

• Mainly due to human error (outdated data, typos, ...)
• Correct station data is necessary e.g. for route planning, station search, transit graph drawing, ...
Goals

1. Detect errors and inconsistencies in
   - station naming
1. Detect errors and inconsistencies in
   - station *naming*
   - station *grouping*
Goals

1. Detect errors and inconsistencies in
   - station **naming**
   - station **grouping**
Goals

1. Detect errors and inconsistencies in
   • station **naming**
   • station **grouping**

2. Provide mappers with
   • tool to **find and analyze** naming errors
Goals

1. Detect errors and inconsistencies in
   - station *naming*
   - station *grouping*

2. Provide mappers with
   - tool to **find and analyze** naming errors
   - suggestions how to *(re-)* **group stations**
Simplified station hierarchy model

<table>
<thead>
<tr>
<th>lvl</th>
<th>tag</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>public_transport stop_area_group</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>public_transport stop_area</td>
<td></td>
</tr>
<tr>
<td></td>
<td>public_transport</td>
<td>stop_position, platform, stop, halt, station</td>
</tr>
<tr>
<td>0</td>
<td>highway</td>
<td>bus_stop, platform</td>
</tr>
<tr>
<td></td>
<td>railway</td>
<td>halt, tram_stop, platform</td>
</tr>
<tr>
<td></td>
<td>tram</td>
<td>stop, platform</td>
</tr>
<tr>
<td></td>
<td>subway</td>
<td>stop, platform</td>
</tr>
</tbody>
</table>
Abstraction: station identifiers are tuples \( s = (n, p) \), where \( n \) is a station label, and \( p \) is a station position.
Station identifiers

**Abstraction:** station identifiers are tuples $s = (n, p)$, where $n$ is a station label, and $p$ is a station position.

Multiple labels (name, alt_name, ref_name) yield multiple station identifiers.
**Abstraction:** station identifiers are tuples \( s = (n, p) \), where \( n \) is a station label, and \( p \) is a station position.

Multiple labels (name, alt_name, ref_name) yield multiple station identifiers.

<table>
<thead>
<tr>
<th>alt_name</th>
<th>Frankfurt Hauptbahnhof</th>
</tr>
</thead>
<tbody>
<tr>
<td>loc_name</td>
<td>Hauptbahnhof</td>
</tr>
<tr>
<td>name</td>
<td>Frankfurt (Main) Hauptbahnhof</td>
</tr>
<tr>
<td>refiftOPT</td>
<td>de:6412:10:1</td>
</tr>
<tr>
<td>refistation</td>
<td>1866</td>
</tr>
<tr>
<td>short_name</td>
<td>Frankfurt (Main) Hbf</td>
</tr>
<tr>
<td>train</td>
<td>yes</td>
</tr>
<tr>
<td>uic_name</td>
<td>Frankfurt(Main)Hbf</td>
</tr>
</tbody>
</table>

→

(Frankfurt Hauptbahnhof, (50.1067, 8.6627))
(Hauptbahnhof, (50.1067, 8.6627))
(Frankfurt (Main) Hauptbahnhof, (50.1067, 8.6627))
(Frankfurt (Main) Hbf, (50.1067, 8.6627))
(Frankfurt(Main)Hbf, (50.1067, 8.6627))
**Goal:** given two station identifiers \( s_1 \) and \( s_2 \), decide whether they describe the same station.
**Goal:** given two station identifiers $s_1$ and $s_2$, decide whether they describe the same station.

We tried a lot and ultimately trained a random forest classifier on common 3-grams, meter distance and the station position on multiple offsetted grids (to capture regional label characteristics).
Station similarity classification

**Goal:** given two station identifiers $s_1$ and $s_2$, decide whether they describe the same station.

We tried a lot and ultimately trained a random forest classifier on common 3-grams, meter distance and the station position on multiple offsetted grids (to capture regional label characteristics).

**F1 score on an international dataset for Germany, Austria and Switzerland:** > 0.99.
(a) Filter station objects from OSM
Pipeline

(a) Filter station objects from OSM
(b) Extract station identifiers with initial clustering
(a) Filter station objects from OSM
(b) Extract station identifiers with initial clustering
(c) Pairwise similarity classification within threshold distance
(a) Filter station objects from OSM
(b) Extract station identifiers with initial clustering
(c) Pairwise similarity classification within threshold distance
(d) Re-cluster based on similarity
(a) Filter station objects from OSM

(b) Extract station identifiers with initial clustering

(c) Pairwise similarity classification within threshold distance

(d) Re-cluster based on similarity

(e) Derive errors and suggestions
• Search and browse OSM station data for large parts of Europe and North America
• Search and browse OSM station data for large parts of Europe and North America
• Station name errors and suggestions are highlighted
• Search and browse OSM station data for large parts of Europe and North America
• Station name **errors** and **suggestions** are highlighted
• **Grouping suggestions** and **correct groups** are shown
• Search and browse OSM station data for large parts of Europe and North America
• Station name **errors** and **suggestions** are highlighted
• **Grouping suggestions** and **correct groups** are shown
• [https://staty.cs.uni-freiburg.de](https://staty.cs.uni-freiburg.de)
Thank you!

https://staty.cs.uni-freiburg.de