

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

name	California Street & Jones Street
network	Muni
operator	San Francisco Municipal Railway
public_transport	stop_position
railway	tram_stop
short_name	Jones & Beach
tram	yes
wheelchair	no



Apartments

Motivation - Errors in OSM station data

name	California Street & Jones Street
network	Muni
operator	San Francisco Municipal Railway
public_transport	stop_position
railway	tram_stop
short_name	Jones & Beach
tram	yes
wheelchair	no



Motivation - Errors in OSM station data

name	California Street & Jones Street
network	Muni
operator	San Francisco Municipal Railway
public_transport	stop_position
railway	tram_stop
short_name	Jones & Beach
tram	yes
wheelchair	no



- Mainly due to **human error** (outdated data, typos, ...)

Motivation - Errors in OSM station data

name	California Street & Jones Street
network	Muni
operator	San Francisco Municipal Railway
public_transport	stop_position
railway	tram_stop
short_name	Jones & Beach
tram	yes
wheelchair	no



- Mainly due to **human error** (outdated data, typos, ...)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

1. Detect errors and inconsistencies in

- station **naming**

1. Detect errors and inconsistencies in

- station **naming**
- station **grouping**

1. Detect errors and inconsistencies in

- station **naming**
- station **grouping**

1. Detect errors and inconsistencies in

- station **naming**
- station **grouping**

2. Provide mappers with

- tool to **find and analyze** naming errors

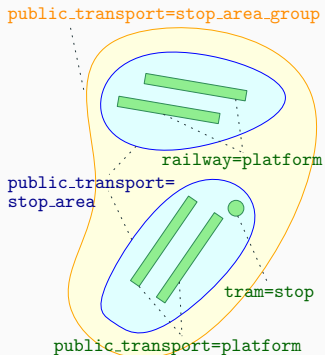
1. Detect errors and inconsistencies in

- 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



lvl	tag	value
2	public_transport	stop_area_group
1	public_transport	stop_area
	public_transport	stop_position, platform, stop, halt, station
0	highway	bus_stop, platform
	railway	halt, tram_stop stop, platform
	tram	stop, platform
	subway	stop, platform

Station identifiers

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**.

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**.

alt_name	Frankfurt Hauptbahnhof
loc_name	Hauptbahnhof
name	Frankfurt (Main) Hauptbahnhof
ref:IFOPT	de:6412:10:1
ref:station	1866
short_name	Frankfurt (Main) Hbf
train	yes
uic_name	Frankfurt(Main)Hbf



(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))

Station similarity classification

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

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).

Station similarity classification

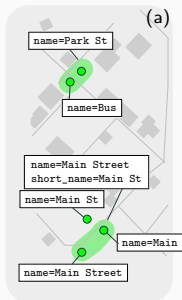
Goal: given two station identifiers s_1 and s_2 , decide whether they describe the same station.

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.

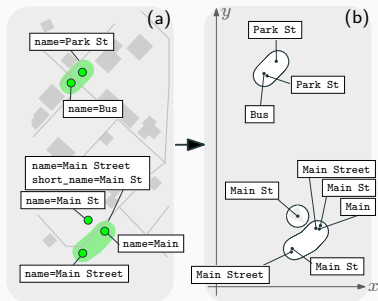
Pipeline

(a) Filter station objects
from OSM



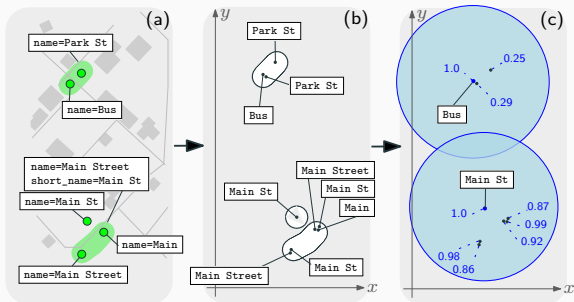
Pipeline

- (a) Filter station objects from OSM
- (b) Extract station identifiers with initial clustering



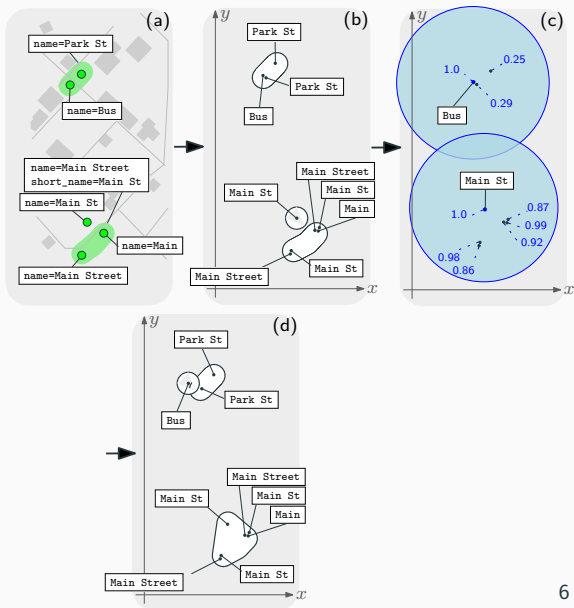
Pipeline

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



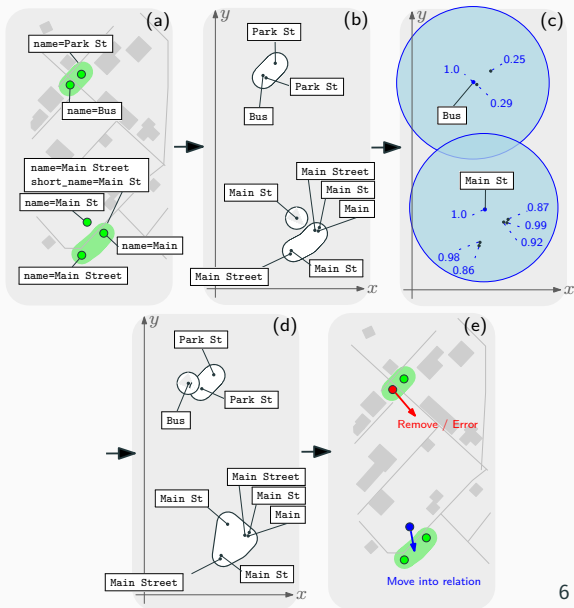
Pipeline

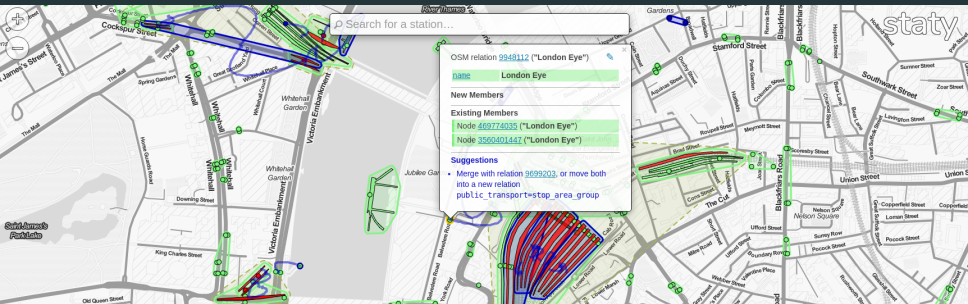
- (a) Filter station objects from OSM
- (b) Extract station identifiers with initial clustering
- (c) Pairwise similarity classification within threshold distance
- (d) Re-cluster based on similarity



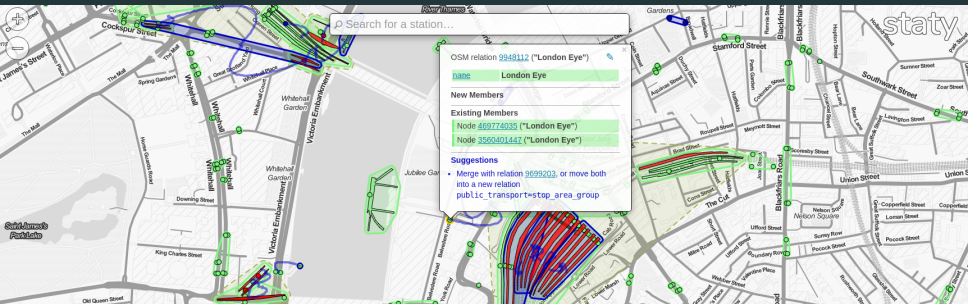
Pipeline

- (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

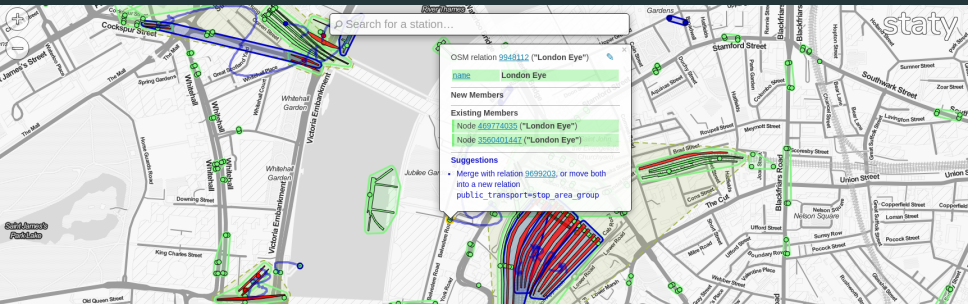




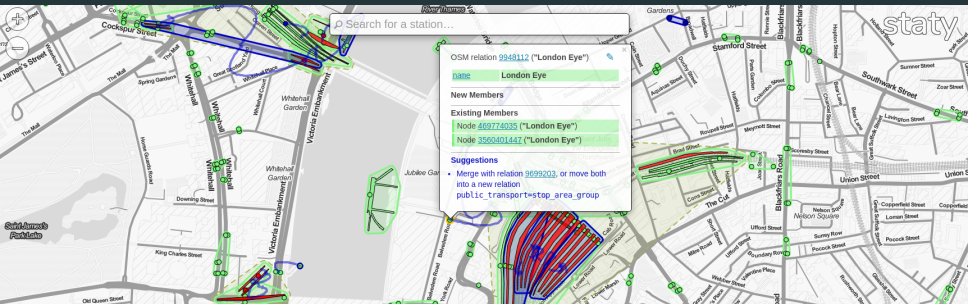
- Search and browse OSM station data for large parts of Europe and North America



- 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>

Thank you!

Thank you!

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