

Initialisierung eines Werkzeuges zur Haltestellengruppierungen auf Basis von Koordinaten und Namen

Aufgabenstellung

Ziel der Thesis ist es, ein Programm zu schreiben, das Folgendes beinhaltet:

- Automatisches Lesen und sinnvolles Zusammenführen der verschiedenen Haltestellendaten auf Basis von Namen und Koordinaten
- Gruppierung der Haltestellen zu Metahaltestellen
- Abspeichern der Dateien im NetTEx-Format

Dabei liegt der Fokus auf der Erforschung verschiedener Methoden zur sinnvollen Gruppierung.

12. Juni 2024 Autor: Marosch Novák, Exterminatorin: Prof. Dr. Pia Bereuter, Betreuer: Stefan Eberlein, Experten: Matthias Günter / Stefan de Konink © IGEO FHNW

Ausgangslage

Entwicklung in den letzten Jahren:

- Immer mehr Haltestellen aufgrund neuer Formen des ÖV (z.B. Flexibus)
- Herkömmliche Definitionen von Haltestellen ändern sich immer mehr durch:
- Vielfalt der Betreiber
- Größe der Haltestellen (Wo setzt man den Koordinatenpunkt für ein ganzes Gebäude?)
- Unterschiede in Haltestellenbezeichnungen durch Region oder Betreiber

Verschiedene Betreiber bieten außerdem die Daten in verschiedenen Formaten an, darunter GTFS, HRDF und NetTEx.



12. Juni 2024 Autor: Marosch Novák, Exterminatorin: Prof. Dr. Pia Bereuter, Betreuer: Stefan Eberlein, Experten: Matthias Günter / Stefan de Konink © IGEO FHNW

Die Methoden

Problematik «einfacher» Gruppierungen:
Gruppierungen auf Basis von nur einem Kriterium (Name oder Koordinaten) erweisen sich als nicht sinnvoll, da:

- Die Gruppierung nach Namen zu Metahaltestellen führen kann, die über Ortsgrenzen hinaus ähnliche Stationen findet.
- Die Gruppierung nach Koordinaten nicht die örtlichen Eigenheiten berücksichtigt. (z.B. Mattheil in Bern)



Gruppierung der Haltestellen Mattheil und Mattheil in Bern
Quelle: <https://www.sbb.ch/de>

12. Juni 2024 Autor: Marosch Novák, Exterminatorin: Prof. Dr. Pia Bereuter, Betreuer: Stefan Eberlein, Experten: Matthias Günter / Stefan de Konink © IGEO FHNW

Diskussion und Fazit

- Es konnten bereits gute Gruppierungsergebnisse erzielt werden.
- Weitere Verfeinerungen der Methoden wären nach wie vor möglich, so z.B.:
 - Eine Anpassung der erlaubten Distanzen basierend auf der Dichte der Haltestellen
 - Das Definieren von Schlagwörtern, die bei der Namensähnlichkeit aufgrund hoher Häufigkeit zu Fehlern führen können («Strasse», «Platz», etc.)
- Eine weitere Ausarbeitung des Gesamtprogrammes würde weitere Vorteile bringen.
 - Insbesondere das Zusammenführen der Daten könnte verbessert werden, um dem Entstehen von Duplikaten durch mehrfache Auflistung vorzubeugen.

Metastelle	Prüfer	Metastelle	Prüfer
Kommunale Haltestelle	Metastelle	Metastelle	Metastelle
Metastelle (z.B. von der Haltestelle)	Metastelle	Metastelle	Metastelle
Metastelle (z.B. von der Haltestelle)	Metastelle	Metastelle	Metastelle
Metastelle (z.B. von der Haltestelle)	Metastelle	Metastelle	Metastelle
Metastelle (z.B. von der Haltestelle)	Metastelle	Metastelle	Metastelle
Metastelle (z.B. von der Haltestelle)	Metastelle	Metastelle	Metastelle
Metastelle (z.B. von der Haltestelle)	Metastelle	Metastelle	Metastelle

Metastelle und Metastelle von der Haltestelle in Metastelle

12. Juni 2024 Autor: Marosch Novák, Exterminatorin: Prof. Dr. Pia Bereuter, Betreuer: Stefan Eberlein, Experten: Matthias Günter / Stefan de Konink © IGEO FHNW

Aufgabenstellung

Ziel der Thesis ist es, ein Programm zu schreiben, das Folgendes beinhaltet:

- Automatisches Lesen und sinnvolles Zusammenführen der verschiedenen Haltestellendaten auf Basis von Namen und Koordinaten
- Gruppierung der Haltestellen zu Metahaltestellen
- Abspeichern der Dateien im NeTEx-Format

Dabei liegt der Fokus auf der Erforschung verschiedener Methoden zur sinnvollen Gruppierung.

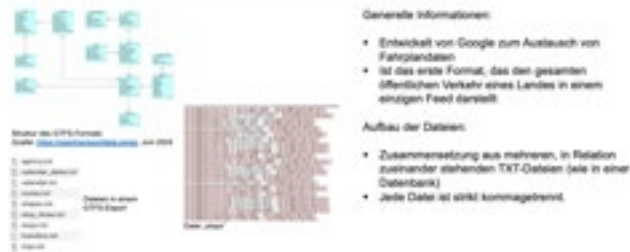
Ausgangslage

Entwicklung in den letzten Jahren:

- Immer mehr Haltestellen aufgrund neuer Formen des ÖV (z.B. FlixBus)
- Herkömmliche Definitionen von Haltestellen ändern sich immer mehr durch:
 - Vielfalt der Betreiber
 - Größe der Haltestellen (Wo setzt man den Koordinatenpunkt für ein ganzes Gebäude?)
 - Unterschiede in Haltestellenbezeichnungen durch Region oder Betreiber

Verschiedene Betreiber bieten außerdem die Daten in verschiedenen Formaten an, darunter GTFS, HRDF und NeTEx.

GTFS (General Transit Feed Specification)



HRDF (Hafas Rohdaten Format)



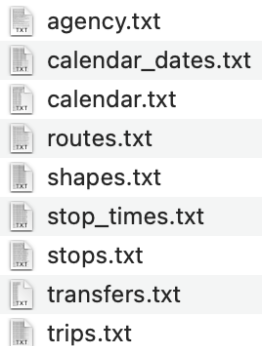
NeTEx (Network Timetable Exchange)



```

classDiagram
    class stop {
        stop_id int
        stop_datetime datetime
        stop_name varchar
        stop_desc varchar
        stop_lat float
        stop_lon float
        stop_name varchar
        stop_desc varchar
        location_type int
        parent_station int
    }
    class stop_time {
        stop_id int
        arrival_time datetime
        departure_time datetime
        stop_sequence int
        stop_headsign varchar
        pickup_status int
        drop_off_status int
        shape_dist_traveled float
    }
    class fare_attribute {
        fare_id int
        fare_name varchar
        currency_type int
        payment_method int
        transferable int
        transfer_distaveled float
    }
    class fare_rule {
        fare_id int
        route_id int
        route_short_name varchar
        agency_id int
        fare_basis int
        contains_wagon int
    }
    class agency {
        agency_id int
        agency_name varchar
        agency_desc varchar
        agency_length int
        agency_phone varchar
    }
    class route {
        route_id int
        route_short_name varchar
        route_desc varchar
        route_type int
        route_status int
        route_color int
        route_text_color int
    }
    class calendar_dates {
        service_id int
        date datetime
        exception_type int
    }
    class calendar {
        service_id int
        monday boolean
        tuesday boolean
        wednesday boolean
        thursday boolean
        friday boolean
        saturday boolean
        sunday boolean
        start_date datetime
        end_date datetime
    }
    class shape {
        shape_id int
        shape_pt_wkt geometry
        shape_pt_sequence int
        shape_dist_traveled float
    }
    class trip {
        route_id int
        trip_id int
        headsign varchar
        stop_sequence int
        stop_headsign varchar
        block_id int
        shape_id int
    }
    class frequency {
        trip_id int
        start_time datetime
        end_time datetime
        headway_sec int
    }
    stop "1" -- "0..*" stop_time
    stop_time "0..*" -- "0..*" fare_attribute
    fare_attribute "1" -- "0..*" fare_rule
    fare_rule "0..*" -- "0..*" route
    agency "1" -- "0..*" route
    route "1" -- "0..*" calendar
    calendar "0..*" -- "0..*" calendar_dates
    calendar "1" -- "0..*" trip
    trip "1" -- "0..*" frequency
    trip "0..*" -- "0..*" shape
    
```

Quelle: <https://opentransportdata.swiss>, Juni 2024

[illegible]

Datei „stops“

- Entwickelt von Google zum Austausch von Fahrplandaten
- Ist das erste Format, das den gesamten öffentlichen Verkehr eines Landes in einem einzigen Feed darstellt

- Zusammensetzung aus mehreren, in Relation zueinander stehenden TXT-Dateien (wie in einer Datenbank)
- Jede Datei ist strikt kommasetrennt.

[illegible]

Quelle: <https://opentransportdata.swiss>, Juni 2024

Dateien BFKOORD_LV95, BAHNHOF und BHFART_60 (links nach rechts)

ATTRIBUT
ATTRIBUT_DE
ATTRIBUT_EN
ATTRIBUT_FR
ATTRIBUT_IT
BAHNHOF
BETRIEB_DE
BETRIEB_EN
BETRIEB_FR
BETRIEB_IT
BFKCOORD_LV95
BFKCOORD_WGS
BFPRIOS
BHFART
BHFART_60
BITFELD
DURCHBI
ECKDATEN
FEIERTAG
FPLAN
GLEIS
GLEIS_LV95
GLEIS_WGS
GRENZHLT
INFOTEXT_DE
INFOTEXT_EN
INFOTEXT_FR
INFOTEXT_IT
KMINFO
LINIE
METABHF
RICHTUNG
UMSTEIGB
UMSTEIGL
UMSTEIGV
UMSTEIGZ
ZEITVS
ZUGART

- Ursprünglich von der Firma Hacon entwickelt für den Datenaustausch mit der Software Hacon Fahrplan-Auskunfts-System (HAFAS)
- Inzwischen Standard in der Branche

- Zusammensetzung aus mehreren, in Relation zueinander stehenden Dateien (wie in einer Datenbank)
- Jede Datei hat ihre eigene Struktur
 - Mit streng festgelegten Zeichenbreiten
 - Immer Leerschlag-getrennt

NeTEx (Network Timetable Exchange)

id	ch:1:CompositeFrame:j22
version	any
ValidBetween	
FrameDefaults	
frames	
SiteFrame	
id	ch:1:SiteFrame:j22
version	any
topographicPlaces	TopographicPlace (1725)
stopPlaces	StopPlace (10036)

Struktur der SITE.xml-Datei aus dem NeTEx-Format mit Fokus auf Haltestellen

PROD_NETEX_TT_1.10_CHE_SKI_2024_OEV-SCHWEIZ_COMMON_1_1_202402210602.xml
PROD_NETEX_TT_1.10_CHE_SKI_2024_OEV-SCHWEIZ_RESOURCE_1_1_202402210602.xml
PROD_NETEX_TT_1.10_CHE_SKI_2024_OEV-SCHWEIZ_SERVICE_1_1_202402210602.xml
PROD_NETEX_TT_1.10_CHE_SKI_2024_OEV-SCHWEIZ_SERVICECALENDAR_1_1_202402210602.xml
PROD_NETEX_TT_1.10_CHE_SKI_2024_OEV-SCHWEIZ_SITE_1_1_202402210602.xml
PROD_NETEX_TT_1.10_CHE_SKI_2024_OEV-SCHWEIZ_TIMETABLE_1_231_202402210602.xml

Dateien eines NeTEx-Exports

Generelle Informationen:

- Ist ein XML-basiertes Datenformat für den Austausch von Fahrplan-, Tarif- und Infrastrukturinformationen
- Hat sich als europäischer Standard etabliert
- Ermöglicht umfassende und strukturierte Darstellung von Transitinformationen

Aufbau der Dateien:

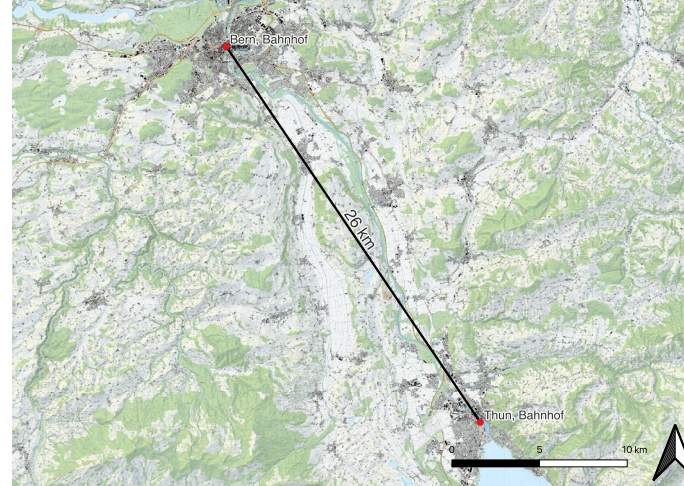
- Zusammensetzung aus mehreren, in Relation zueinander stehenden XML-Dateien (wie in einer Datenbank)

Dieses Dateiformat soll auch zur Abspeicherung der Resultate dienen.

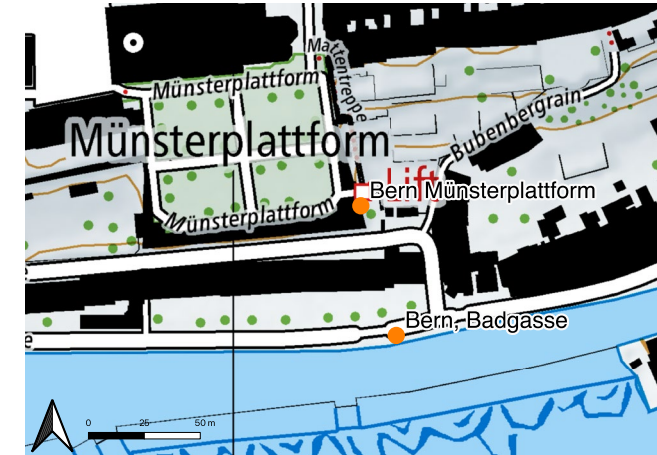
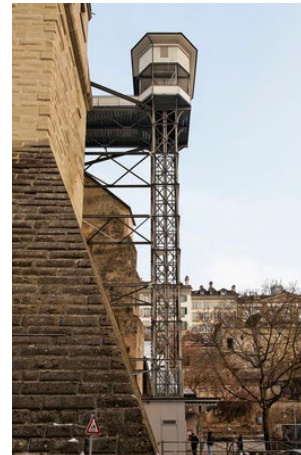
Die Methoden

Problematik «einfacher» Gruppierungen:
Gruppierungen auf Basis von nur einem Kriterium
(Name oder Koordinaten) erweisen sich als nicht
sinnvoll, da:

- Die Gruppierung nach Namen zu Metahaltstellen führen kann, die über Ortsgrenzen hinaus ähnliche Stationen findet.
- Die Gruppierung nach Koordinaten nicht die örtlichen Eigenheiten berücksichtigt. (z.B. Mattelift in Bern)



Gruppierung der Bahnhöfe Bern und Thun trotz großer Distanz



Gruppierung der Stationen Münsterplattform und Badgasse trotz großem Höhenunterschied (und kostenpflichtigem Aufzug)

Quelle: <https://www.mattelift.ch>

Die Methoden

Gewählte Lösungsansätze sind daher:

- Gruppierung aufgrund Namensähnlichkeit in Verbindung mit Puffern
- Gruppierung aufgrund Namensähnlichkeit und nächster Nachbarn
- Gruppierung, basierend auf Fuzzy-Matching und Distanzkriterien

Gruppierung auf Basis von Puffern

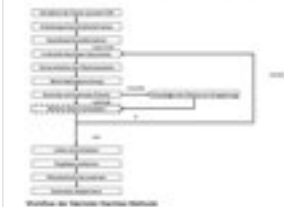


Ansatz:

- Verknüpfung dreier verschieden großer Puffer mit unterschiedlichen Namensähnlichkeitsanforderungen als Gruppierungskriterien



Gruppierung auf Basis von Nächsten Nachbarn

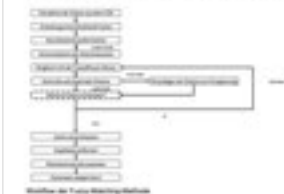


Ansatz:

- Berechnung der k-nächsten Nachbarn für jede Station
- Anschließende Prüfung auf Maximaldistanz und Namensähnlichkeit



Gruppierung auf Basis von Fuzzy Matching

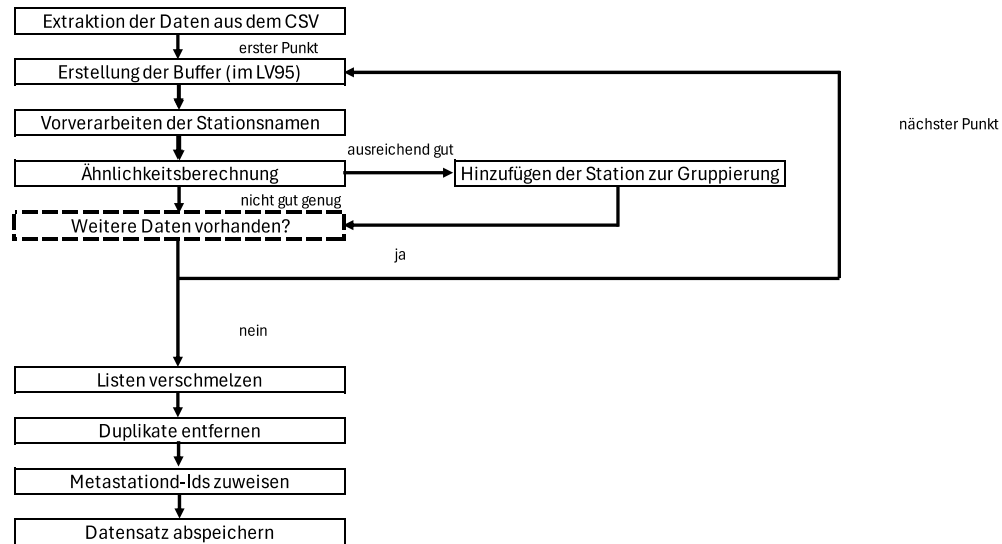


Ansatz:

- Fokus auf die Überprüfung der Namensähnlichkeit
- Anschließende Filterung nach Distanz



Gruppierung auf Basis von Puffern



Workflow der Puffer-Methode

Ansatz:

- Verknüpfung dreier verschieden großer Puffer mit unterschiedlichen Namensähnlichkeitsanforderungen als Gruppierungskriterien

Resultat der Puffer-Methode



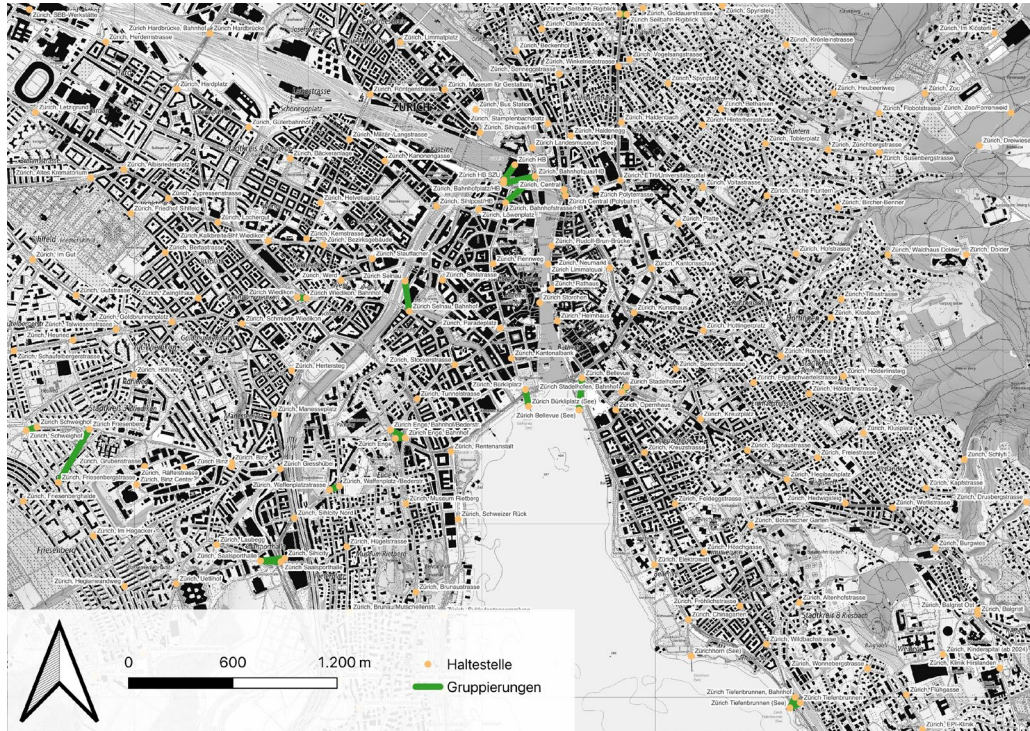
Vorteile:

- Einfache Implementierung der Methode
- Leicht verständliche Vorgehensweise
- Direkte Kontrollierbarkeit der Distanzen in Relation zu den Ähnlichkeitsschwellen
 - Dadurch schnelle und einfache Änderungen an den Variablen vornehmbar

Nachteile:

- Hohe Rechenintensivität durch die Erstellung dreier Puffer pro Haltestelle
 - Dies führt zu langen Rechenzeiten bei großen Datensätzen.
- Keine dynamische Anpassung an die tatsächliche Verteilung der Stationen gegeben

Resultat der Puffer-Methode



Haltestellen der Innenstadt von Zürich, gruppiert mit der Puffer-Methode

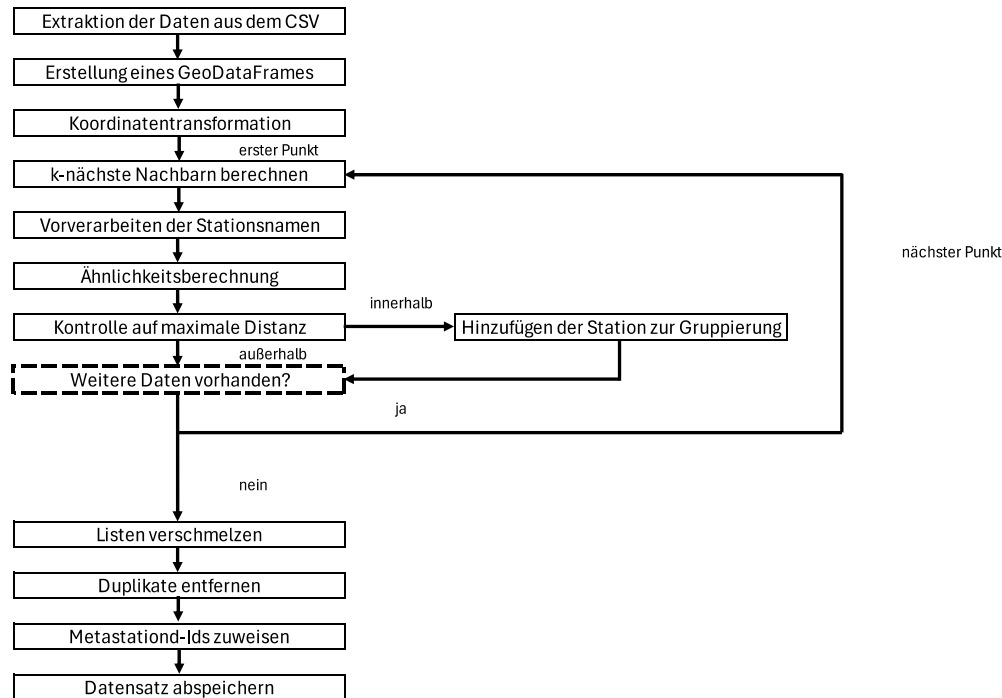
Vorteile:

- Einfache Implementierung der Methode
- Leicht verständliche Vorgehensweise
- Direkte Kontrollierbarkeit der Distanzen in Relation zu den Ähnlichkeitsschwellen
 - Dadurch schnelle und einfache Änderungen an den Variablen vornehmbar

Nachteile:

- Hohe Rechenintensivität durch die Erstellung dreier Puffer pro Haltestelle
 - Dies führt zu langen Rechenzeiten bei großen Datensätzen.
- Keine dynamische Anpassung an die tatsächliche Verteilung der Stationen gegeben

Gruppierung auf Basis von Nächsten Nachbarn



Workflow der Nächster-Nachbar-Methode

Ansatz:

- Berechnung der k-nächsten Nachbarn für jede Station
- Anschließende Prüfung auf Maximaldistanz und Namensähnlichkeit

Resultat der Nächster-Nachbar-Methode



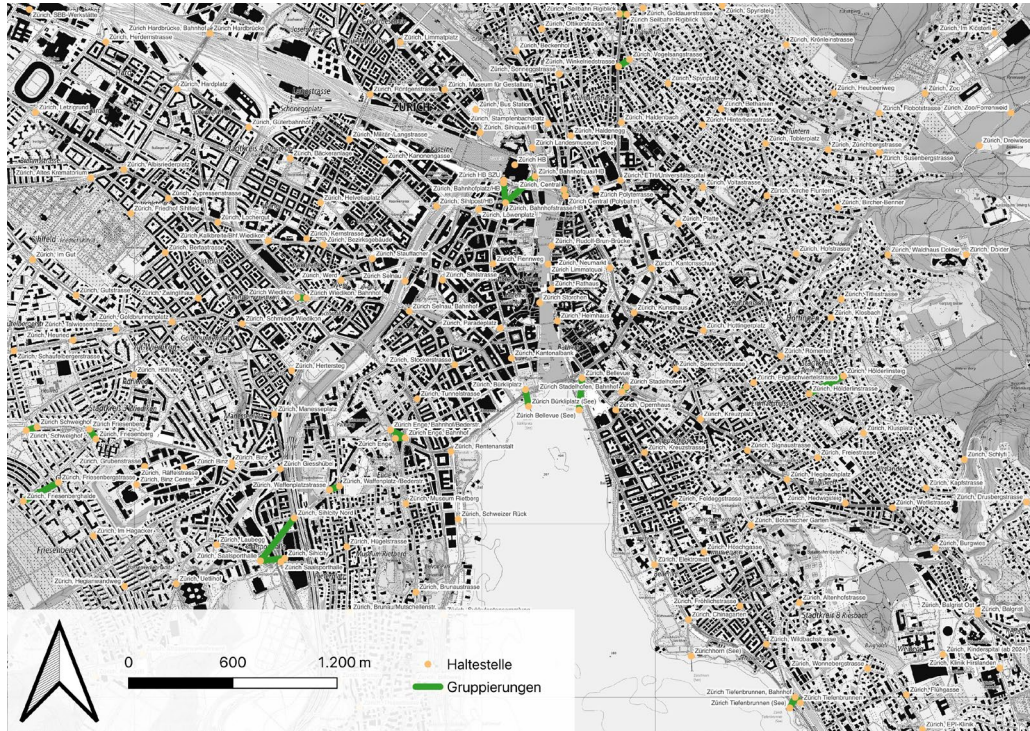
Vorteile:

- Schafft einen effizienten Algorithmus, der auch bei großen Datenmengen mit hoher Geschwindigkeit die Gruppierungen durchführen kann
- Durch die Berechnung der k-nächsten Nachbarn ist eine dynamischere Anpassung an die tatsächliche Verteilung der Stationen möglich.

Nachteile:

- Komplexität der Implementierung
- Tieferes Verständnis von Geodaten und maschinellen Lernmethoden zum Verständnis des Codes nötig
- Evtl. werden nicht alle möglichen Stationen für die Gruppierung berücksichtigt.

Resultat der Nächster-Nachbar-Methode



Haltestellen der Innenstadt von Zürich, gruppiert mit der Nächster-Nachbar-Methode

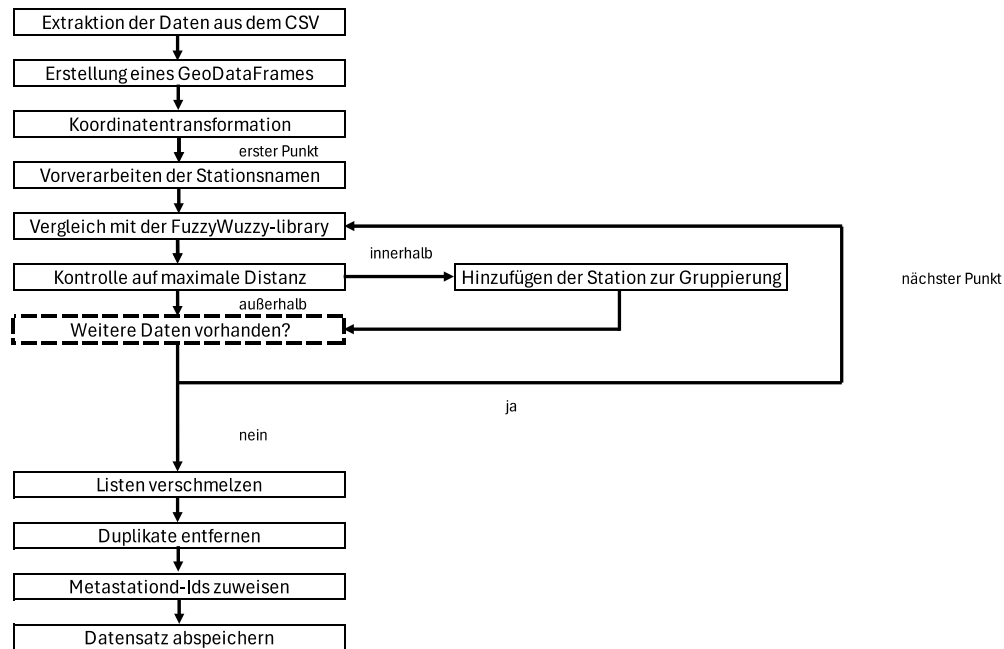
Vorteile:

- Schafft einen effizienten Algorithmus, der auch bei großen Datenmengen mit hoher Geschwindigkeit die Gruppierungen durchführen kann
- Durch die Berechnung der k-nächsten Nachbarn ist eine dynamischere Anpassung an die tatsächliche Verteilung der Stationen möglich.

Nachteile:

- Komplexität der Implementierung
- Tieferes Verständnis von Geodaten und maschinellen Lernmethoden zum Verständnis des Codes nötig
- Evtl. werden nicht alle möglichen Stationen für die Gruppierung berücksichtigt.

Gruppierung auf Basis von Fuzzy Matching

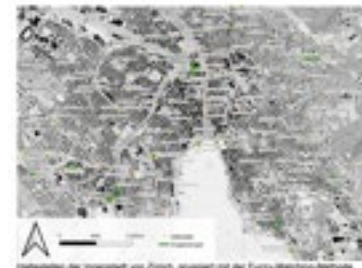


Workflow der Fuzzy-Matching-Methode

Ansatz:

- Fokus auf die Überprüfung der Namensähnlichkeit
- Anschließende Filterung nach Distanz

Resultat der Fuzzy-Matching-Methode



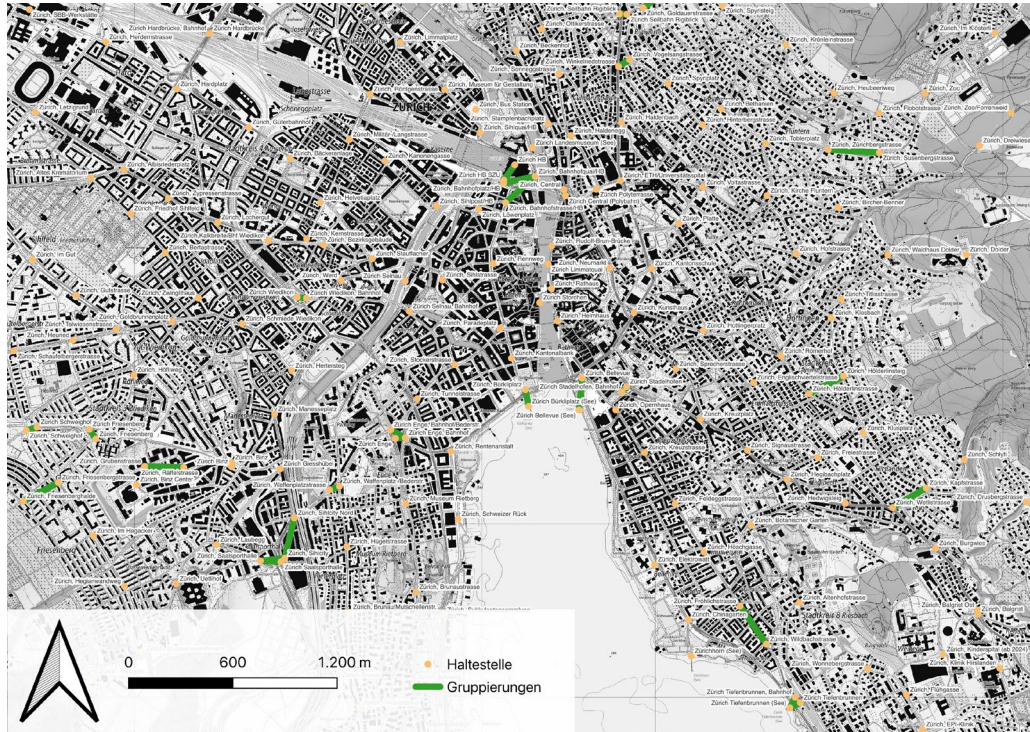
Vorteile:

- Besserer Umgang mit Dateninkonsistenzen durch Fuzzy-Matching

Nachteile:

- Empfindlichkeit gegenüber Schreibweisen
- Hoher Berechnungsaufwand durch Überprüfung der Namen aller Stationen
 - Dies kann bei großen Datenmengen zu langen Rechenzeiten führen

Resultat der Fuzzy-Matching-Methode



Haltestellen der Innenstadt von Zürich, gruppiert mit der Fuzzy-Matching-Methode

Vorteile:

- Besserer Umgang mit Dateninkonsistenzen durch Fuzzy-Matching

Nachteile:

- Empfindlichkeit gegenüber Schreibweisen
- Hoher Berechnungsaufwand durch Überprüfung der Namen aller Stationen
 - Dies kann bei großen Datenmengen zu langen Rechenzeiten führen.

Diskussion und Fazit

- Es konnten bereits gute Gruppierungsergebnisse erzielt werden.
- Weitere Verfeinerungen der Methoden wären nach wie vor möglich/nötig, so z.B.:
 - Eine Anpassung der erlaubten Distanzen basierend auf der Dichte der Haltestellen
 - Das Definieren von Schlagwörtern, die bei der Namensähnlichkeit aufgrund hoher Häufigkeit zu Fehlern führen können («Strasse», «Platz», etc.)
- Eine weitere Ausarbeitung des Gesamtprogrammes würde weitere Vorteile bringen.
 - Insbesondere das Zusammenführen der Daten könnte verbessert werden, um dem Entstehen von Duplikaten durch mehrfache Auflistung vorzubeugen.

Bibliothek	Puffer	<u>Nearest Neighbors</u>	Fuzzy-Matching
Implementierungsaufwand	Niedrig	Hoch	Hoch
Abhängigkeit (Anzahl der libraries)	Wenige	Mehrere	Mehrere
Effizienz bei großen Daten	Mittel-Niedrig	Hoch	Niedrig
Flexibilität	Mittel	Hoch	Hoch
Präzision der Gruppierung	Hoch	Mittel	Hoch
Anpassbarkeit	Hoch	Mittel	Hoch

Vor- und Nachteile der drei Methoden im Überblick

Diskussion und Fazit

Jede Methode hat ihre Stärken und Schwächen. Die Wahl, welche Methode verwendet werden sollte, hängt stark von den spezifischen Anforderungen ab.

Bibliothek	Puffer	Nearest Neighbors	Fuzzy-Matching
Implementierungs-aufwand	Niedrig	Hoch	Hoch
Abhängigkeit (An-zahl der libraries)	Wenige	Mehrere	Mehrere
Effizienz bei großen Daten	Mittel-Niedrig	Hoch	Niedrig
Flexibilität	Mittel	Hoch	Hoch
Präzision der Grup-pierung	Hoch	Mittel	Hoch
Anpassbarkeit	Hoch	Mittel	Hoch

Vor- und Nachteile der drei Methoden im Überblick

Quellen und Literatur

https://opentransportdata.swiss/de/cookbook/hafas-rohdaten-format-hrdf/#Fachliche_Beschreibung

<https://opentransportdata.swiss/de/cookbook/gtfs/>

Mattelif. (s.d.). Mattelif. URL: <https://www.mattelif.ch> [Stand: Mai 2024]