GTFS Bus Stop Mapping to the OSM Network

Peer-reviewed author version

VUURSTAEK, Jan; CICH, Glenn; KNAPEN, Luk; ECTORS, Wim; YASAR, Ansar; BELLEMANS, Tom & JANSSENS, Davy (2020) GTFS Bus Stop Mapping to the OSM Network. In: Future generation computer systems, 110, p. 393-406.

DOI: 10.1016/j.future.2018.02.020
Handle: http://hdl.handle.net/1942/26214

GTFS Bus Stop Mapping to the OSM Network

Jan Vuurstaek[a], Glenn Cich[a], Luk Knapen[a], Wim Ectors[a], Ansar-Ul-Haque Yasar[a], Tom Bellemans[a], Davy Janssens[a]

[a]*Hasselt University, Transportation Research Institute (IMOB), Agoralaan, 3590 Diepenbeek, Belgium*

Abstract

Due to budget constraints *public transportation* (PT) can no longer be deployed in regions where it attracts insufficient customers. Novel techniques such as *demand-responsive collective transportation* (DRT) are evaluated to cut costs. This requires detailed simulations that are able to predict travel demand and include trip execution. Simulating facilities acting as feeder services to timetable based PT services requires detailed and accurate information about the PT infrastructure on a network. However, there are no public data sources that combine network and PT infrastructure data with the preferred level of detail. This led to the development of a new bus stop mapping technique that combines the *OpenStreetMap* (OSM) and *General Transit Feed Specification* (GTFS) open data sources, which are maintained independently. Merging the data into a single database requires alignment. Developing bus stop mapping algorithms is challenging due to (i) inaccurate location data, (ii) inconsistent data sources and (iii) the vastly interconnected PT network and services. Due to the inaccuracy in the GTFS stop locations and in the OSM network, pure geometric considerations might lead to multiple candidate solutions to map a stop to the network. The new technique handles all GTFS trips *at once* and operates under the assumption that PT operators minimize the total distance driven to complete all trips.

Keywords: OpenStreetMap (OSM), General Transit Feed Specification

*Corresponding author
Email address: jan.vuurstaek@uhasselt.be (Jan Vuurstaek)

(GTFS), Micro-Simulation, Public Transport, Mapping Algorithm
2010 MSC: 00-01, 99-00

1. Introduction

In Belgium a *bus stop* is a pole in the ground, at a specific side of the road, where people can board or alight a transit vehicle. These bus stops are serviced by a Public Transportation (PT) provider, which implies that if a road is serviced in both directions, each side of the road gets its own bus stop. The research presented in this paper attempts to match the bus stops of the only bus PT provider in Flanders (Belgium), called "De Lijn", to a road network. The bus stops are extracted from a dataset that is made publicly available by "De Lijn". This dataset follows the General Transit Feed Specification (GTFS)[1]. As for the road network, the publicly available OpenStreetMap (OSM)[2] data source was used.

The OSM database is fairly complete but the number and size of positional errors cannot be ignored, as stated in the research of Haklay et al.[1] The accuracy on GPS recordings is in the range of [15, 30] meters (example given in Figure 1). Each bus stop in GTFS is represented by exactly one coordinate pair which is assumed to have been determined by a GPS recording. As a consequence it is not always clear to which road segment a bus stop needs to be assigned. Assigning a bus stop to the wrong side of a road can induce large distance and travel time errors in reconstructed bus routes. Such errors are not acceptable in simulations of demand-responsive (private and public) collective transportation (DRT). These simulations are characterized by mutual coordination (cooperation), low volumes and hence small capacity vehicles. Reasoning about average flows is infeasible due to large quantisation effects. Solutions to such problems require combinatorial optimization and are very sensitive to small changes (errors) in the input.

[1] `https://developers.google.com/transit/gtfs/`
[2] `https://www.openstreetmap.org/`

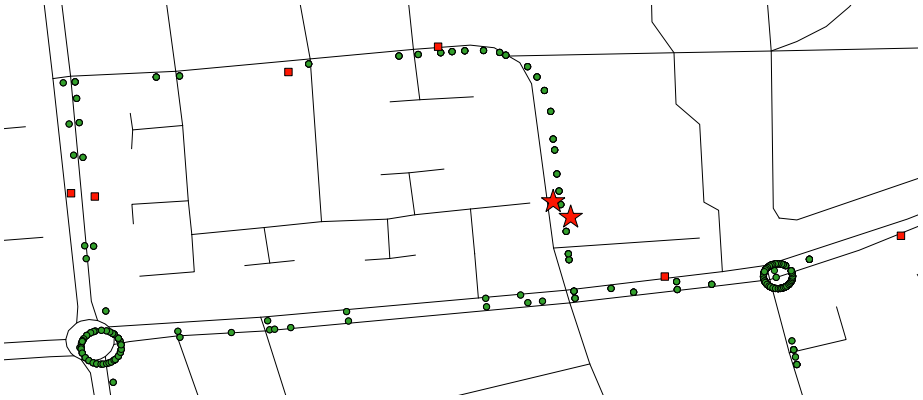


Figure 1: Israel, Kiryat Gat: example of inaccuracy. Shapes (green circles) and bus stops (red squares/stars) in GTFS displayed on an OSM network. The stops represented by the red stars show the *wrong side of the road problem*. Note the missing roundabout at the right side and the positional errors.

DRT services are used as feeders to PT services. In this context, accurate assignment of bus stops to the road network is required for modelling.

Many published GTFS datasets do not contain the optional shape file.[3] Furthermore, the shape file only provides information on the geometry but not the topology. The `shapes.txt` file in GTFS contains tuples $\langle s, x, y, o \rangle$ where s denotes the sequence identifier, x and y denote longitude and latitude respectively and o is the offset of the point in the sequence. It is observed in the Israeli GTFS data that exactly the same $\langle x, y \rangle$ pair can occur in multiple sequences. It is easy to find pairs of locations p_0 and p_1 that do occur consecutively and in both orders in multiple sequences. This means that sequences of $\langle x, y \rangle$ locations were recorded once and then used in both orders to represent trips driven in opposite directions. Hence, the shape files cannot be used to determine the bus stop locations from geometry and ordering.

[3]Following GTFS data sources where checked: De Lijn (Belgium), RATP (France), Fritz Behrendt OHG (Germany), Günter Anger Güterverkehrs GmbH & Co. Omnibusvermietung KG (Germany), Haru Reisen OHG (Germany). These data sources are available on `http://www.gtfs-data-exchange.com/`.

Several research efforts to assign GTFS bus stops to OSM and other road networks have been reported in literature. A first category makes use of pure geometric methods which can lead to bus stops matched to the wrong side of the road due to positional errors. A second category starts from the assumption that the bus travels the minimum distance required to serve consecutive stops; those methods process a single bus trip at a time. For bus stops used in multiple trips, inconsistent results are reported by independently handling individual trips (routes). This paper describes a newly developed technique to handle all GTFS trips *at once*.

The paper is organized as follows. Section 2 provides a discussion of related work. Next, the data preparation is described in Section 3, followed by the bus stop mapping algorithm in Section 4. Then, in Section 5, a case study for Flanders (Belgium) is presented. This case study is used to validate the bus stop mapping algorithm and is described in Section 6. Section 9 concludes the paper and discusses future work.

This paper is an extended version of the work published by Vuurstaek et al.[2] The amount of experiments was increased and a new extensive validation method was added.

2. Related Work

A handful solutions have been provided in the past. Lektauers et al.[3] mention a "*tool to associate stop facilities to network nodes*" without any details (in the context of an integrated MATSim-based simulation). Li[4] exploits the relational information between adjacent bus stops. Several possible projections (candidates) of GTFS stops on the road network are considered by calculating the distance to the nodes of nearby road links, instead of the links themselves. This could lead to mismatches when the correct location of a bus stop is on a long straight link that is parallel with a nearby short link. The shortest path linking all stops for a trip in the specified order is considered to be the real one. Their model only considers *single* bus trips. In case of overlapping bus trips, manual

checking is required. Perrine et al.[5] propose a two step method that uses the optional shape file. In the first step the shape points are map-matched in order to determine a link sequence for each track using a multi-hypothesis technique (MHT). This is done by considering possible successive extensions of the link sequence and retaining the solution that results in the lowest distance driven. While mapping the GTFS shape points, the curvature of the road segments is not considered, which may lead to mismatches. In the second step, the links identified by the map matcher are retained as a new network. Then the sequence of bus stops is used as a virtual GPS track and map-matched against the reduced network. This is done independently on a *per route basis* so that a particular GTFS bus stop used in several routes can lead to different results. Finally, bus stops located near crossings are reported to be matched against the wrong link. The issues related to the correct road side selection and overlapping bus trips are not mentioned. Ordóñez et al.[6] developed a semi-automatic procedure to combine public bus routes information with a high resolution network as an extension to MATSim. First, a simple map-matching algorithm is applied on a per route basis. This algorithm makes use of the GPS points that are present in the optional shape file. It does not assume that the nearest link to a stop is always the correct one. Second, an automatic verification procedure including following checks is performed: (i) is the path joined, (ii) is the path without U-turns, (iii) is the path without repeated links, (iv) does every stop of the route have a stop-link relationship, (v) does every link related to a stop belong to the path, (vi) does the related links order in the path comply with the corresponding stops order in the route profile, (vii) is the nearest point between the stop point and the infinite line defined by the link inside its line segment for every stop and (viii) are the first and last links of the path related to the first and last stops of the route profile? Steps (ii), (iii) and (vii) can be disabled. If the automatic verification fails, manual editing is required. The authors report that if routes are not processed in an appropriate order, some re-work might be necessary. It required ten days to process the public transport data of Singapore.

3. Data Preparation

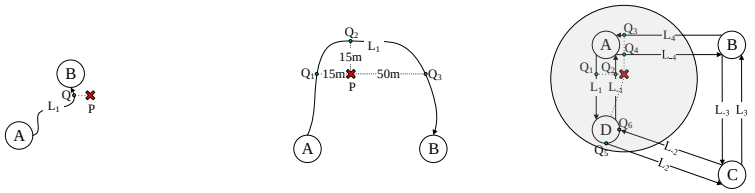
Data preparation from the publicly available OSM and GTFS sources is described by Cich et al.[7] The following procedures were developed: (i) build a model for the road network derived from OSM suitable for simulations, (ii) extract bus stops from GTFS while removing anomalies (details can be found in Cich et al.[7]) and (iii) for each bus stop specified in GTFS, find a set of candidate network links to attach it.

The road network model consists of a directed graph. Each link in the road network is represented by one or two edges in this graph so that each driving direction is represented by a separate edge. Note that in the road network road curvature defining shape (geometry) nodes are also included. This ensures accurate distance calculations. The actual attachment of the candidate GTFS bus stops to the OSM road network makes use of the following concepts:

1. **Projection of a point on a curve**: A point Q on the curve C_L representing a network link (road segment) L is a (geometric) projection of a point P if and only if the euclidean distance $d(P, Q)$ is minimal (Figure 2a). Note (i) that the projection can coincide with one of the end nodes of L (e.g. Q_6 in Figure 2c), (ii) that a point can have multiple projections on a single link (Figure 2b) and (iii) that $\overline{PQ}$ is not necessarily perpendicular to C_L (e.g. Q_6 in Figure 2c).

2. **Projection of a point on network**: Each GTFS bus stop P is projected to the set of all network links $L \in \mathcal{L}$ for which the distance between the GTFS stop and its projection $d(P, Q(P, L))$ does not exceed a given threshold. A particular GTFS stop can have a projection on multiple links. Remember that all links are unidirectional. The projections of a single point P on multiple links can geometrically coincide; this occurs when the respective projections all map to a node shared by those links. (Figure 2c)

Multiple projections of a point P to the geometry of a particular link L can



(a) A projection of a point on a (curve) link.

(b) Point P has three projections on link L_1. Projection Q_3 is discarded. Since both projections Q_1 and Q_2 are at the minimum distance to P, one of them is randomly selected.

(c) A projection of a point on a network. The threshold is indicated by the grey circle. The projection of point P on link L_2 and L_{-2} coincides with link node D.

Figure 2: Examples of projections of a point.

exist (an example is shown in Figure 2b). Both the algorithm described in this paper and the simulations into which the bus stops are fed are indifferent to which of the alternative projections of P to L is chosen. Only the projections at the minimum distance to P are considered because those are assumed to have the highest probability to be the ones searched for. One of them is chosen randomly.

A projection is a candidate to represent the GTFS bus stop from which it was derived (hence the name *projected stop* used in the remainder of the paper). Since a projection is associated with a unidirectional link (road segment) the next node in the network reached by a bus serving the projected stop is unambiguously known. Hence, it determines the side of the street to which the candidate applies.

4. Bus Stop Mapping

4.1. Principle of Operation

A technique to handle all GTFS trips *at once* has been developed. Let S_G denote the stops found in GTFS and let S_P denote the set of projected stops. A GTFS stop $g \in S_G$ can result in several projections $p \in S_P(g)$ located on multiple links. For each $g \in S_G$, exactly one $p \in S_P(g)$ has to be assigned to g.

The assignment is computed by optimization. It is assumed that the PT operator minimizes the total distance driven to complete all trips. Hence, the number of times a particular trip is serviced during a day is used as a multiplicative weight factor.

Each $p \in S_P(g)$ constitutes a *degree of freedom* (DOF) for g, i.e. the degree of freedom for g equals $|S_P(g)|$. The number of possible injections $S_G \Rightarrow S_P : g \mapsto p$ is huge (see Section 5). Assigning a projection $p \in S_P(g)$ to a GTFS stop g is called *fixing* g.

Let $G(V, E)$ denote the bus stop connection graph defined by the GTFS trips. Then $V = S_G$ and two vertices are connected by an edge if and only if they appear as consecutively serviced stops in GTFS. G is built by adding edges one after another. Finding an optimal assignment is computationally efficient for an acyclic digraph. Hence, the algorithm decomposes G into mutually independent acyclic components. Whenever the addition of an edge introduces one or more cycles, the set C of vertices that *individually* can break all newly induced cycles is determined. Exactly one of those is to be chosen as a cycleBreaker: $v \in C$ is chosen so that $|S_P(v)|$ is minimal (see Figure 5 for an example). In several stages of the algorithm cycleBreakers can become redundant (in which case they are removed).

The basic idea of the solution is to fix stops in mutually independent bounded acyclic subgraphs of G. Such subgraphs are isolated by boundary vertices. A vertex is a *boundary vertex* only if it is (i) a *source*, (ii) a *sink*, (iii) a previously *fixed* stop or (iv) a *cycleBreaker*. All possible assignments for the boundary vertices need to be examined; hence it is important to select the boundary $B \subseteq$

S_G so that $\prod_{b \in B} DOF(b)$ is small. For each assignment $\mathcal{A}_B$ to the boundary set B, the lowest cost assignment $\mathcal{A}_I$ for the set of *internal* vertices I in the acyclic component is computed. If a particular $p \in S_P(g)$ occurs in each assignment $a \in \mathcal{A}_I$, then g can be fixed to p. On the other hand, if p is not used in any $a \in \mathcal{A}_I$ then p can be discarded as a candidate for the final assignment. Smart selection of the bounded subgraphs results in an efficient procedure. E.g. if the subgraph is a linear sequence of GTFS stops, B contains the two end nodes g_0 and g_1. The shortest paths between each pair $\langle p_0, p_1 \rangle \in S_P(g_0) \times S_P(g_1)$ represents an assignment $\mathcal{A}_B$ and can be computed very efficiently because the corresponding stop projections constitute a layered graph.

4.2. Algorithm

The bus stop mapping procedure is summarized in Algorithm 4.1. The concepts presented in the algorithm will be explained using the interconnected trips that are shown in Figure 3.

Line 3 builds a graph G_G using GTFS stops based on the sequences found in GTFS. Furthermore, it builds a graph G_P using projected stops so that each $p \in S_P(g)$ for a given $g \in S_G$ is connected to all the projected stops of the neighbours of g. Hence, each pair $\langle g_0, g_1 \rangle \in S_G \times S_G$ of neighbours defines a distance matrix (between the respective stop projections). This is graphically represented in Figure 4 where $P(g_C)$ and $P(g_D)$ make up the distance matrix induced by GTFS stops g_C and g_D. This step includes the marking of some GTFS stops as *cycle breakers*. Figure 5 shows an example of cycle breaking. Subgraphs delimited by cycle breakers are directed acyclic graphs (DAG). Cycle breakers contribute to the DOF size of the irreducible components generated in line 11 and therefore need to be chosen carefully.

Line 6 considers *triples* $\langle g_0, g_1, g_2 \rangle$ for which the set of neighbours of g_1 is $\{g_0, g_2\} = B$ (the boundary of the subgraph). Determine all shortest paths between $p_{0,i} \in S_P(g_0)$ and $p_{2,j} \in S_P(g_2)$ for all i and all j. Drop the projected stops $p_{1,k}$ for all k that are *not* part in *any* shortest path. If and only if for a given k the projected stop $p_{1,k}$ appears in *all* shortest paths, assign it to g_1.

Algorithm 4.1 Determination of optimal assignment of projected stops S_P to GTFS stops S_G.

1: $S_P \leftarrow \mathit{projStops}(S_G)$
2: $\mathit{fixTrivial}(S_G)$ ▷ Assign GTFS stops having a single projection
3: $\langle G_G, G_P \rangle \leftarrow \mathit{graphFromBusStopSequences}()$ ▷ Introduces *cycleBreakers*
4: **repeat**
5: $\quad \mathit{removedAtLeastOneCandidate} \leftarrow$ **false**
6: $\quad \mathit{handleTriples}(\langle G_G, G_P \rangle)$ ▷ Vertex in the middle has $\mathit{inDegree} = \mathit{outDegree} = 1$
7: $\quad \mathit{handleNonBifurcatingMaximalSequences}(\langle G_G, G_P \rangle)$ ▷ Internal vertices have $\mathit{inDegree} = \mathit{outDegree} = 1$
8: $\quad \mathit{handleStars}(\langle G_G, G_P \rangle)$
9: $\quad \mathit{reduceCycleBreakers}(\langle G_G, G_P \rangle)$ ▷ Removes *cycleBreakers* that became redundant by *fixing* some vertices
10: **until** $\neg \mathit{removedAtLeastOneCandidate}$ ▷ Either by *explicit discarding* or as a *consequence of fixing*
11: $\mathit{components} \leftarrow \mathit{decompose}(\langle G_G, G_P \rangle)$
12: **for all** $c \in \mathit{components}$ **do**
13: $\quad \mathit{assign}(c)$ ▷ Assignment by solution enumeration
14: **end for**

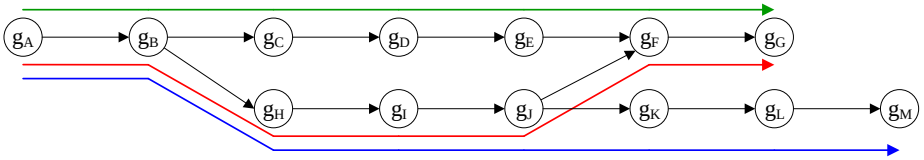


Figure 3: A collection of three (red, green, blue) interconnected bus trips. Each vertex represents a bus stop on a trip.

Then g_1 is said to be fixed. Figure 6 shows an example of a triple.

Line 7 performs a similar technique on *maximal non bifurcating sequences*. Actually, *handleTriples* is a special case of this one that is implemented to increase efficiency. Figure 7 shows an example of a maximal non bifurcating sequence.

Line 8 again performs a similar technique on *stars* (i.e. a GTFS stop having an inDegree and/or outDegree larger than one). Let $S_P(c)$ denote the projected stops for the *core* of the star. If $p \in S_P(c)$ is used in each possible assignment for the neighbours, c is fixed to p. If p is not used in any assignment, it is discarded as a candidate. Figure 8 shows an example of a star.

Line 9 removes *cycleBreaker* markings that became redundant by stop fixations. If an assignment was found for GTFS stop g_B in Figure 5, a cycle breaker is no longer needed, since g_B then acts as a boundary vertex.

Line 11 decomposes the graph so that each component is maximal and delimited by GTFS stops that are sources, sinks, assigned (fixed) or cycleBreakers. Figure 9 shows an example of a component.

After assigning a stop projection to each GTFS stop, the bus trips were reconstructed by connecting consecutive stops in trips by the shortest path in the road network.

5. Results

The accuracy $\overline{A}$ of GPS recording devices is reported to be in the range from 15 meters till 30 meters as stated by Haklay et al.[1] This is interpreted as follows: the positional error in both directions is normally distributed with zero

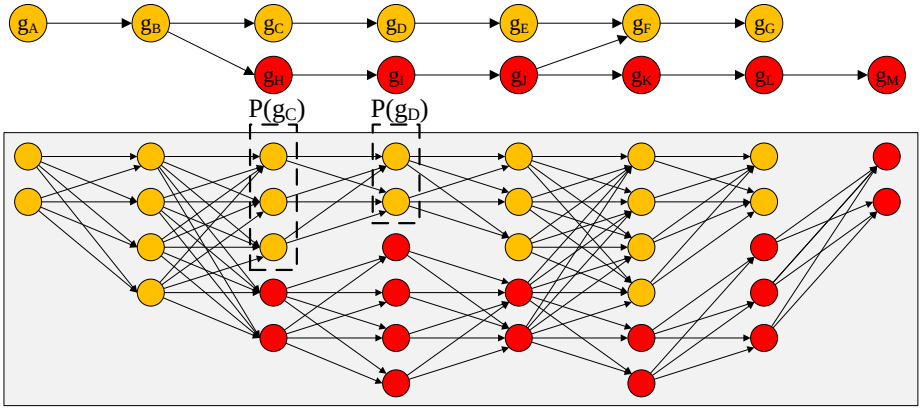


Figure 4: GTFS stops (upper) and associated stop projections (lower) graphs. The orange/red *string* in the upper part corresponds to the orange/red *layered graph* in the lower part. Each pair of consecutive vertices $\langle g_0, g_1 \rangle \in S_G \times S_G$ corresponds to a complete bipartite subgraph in the lower part.

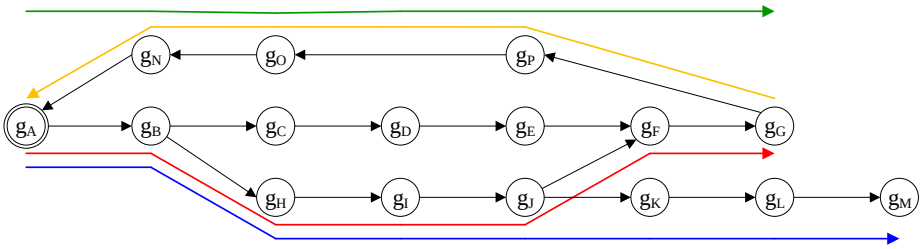


Figure 5: Example of a cycle breaker. Cycle breaking makes a graph acyclic. Cycle breakers are indicated by double circle vertices. In this example GTFS stop g_A is chosen as a cycle breaker. Other candidates to break the same cycle are: g_B, g_F, g_G, g_N, g_O and g_P.

mean and its absolute value does not exceed the specified value $\overline{A}$ in 95% of the cases. The standard deviation then is estimated by equation (1). This led to the creation of two experiments based on the two ends of the range. The device accuracy $\overline{A}$ affects the projected stop search radius, i.e. when an accuracy of $\overline{A}$ meters is assumed for GPS recording devices, candidate projected stops will be searched for within a radius of $\overline{A}$ meters around the GTFS stop locations. In both experiments the number of projections for any particular GTFS stop is limited to 10. Characteristics of the cleaned GTFS and OSM data for the Flanders case study are summarized in Table 1. Computation results for the 15

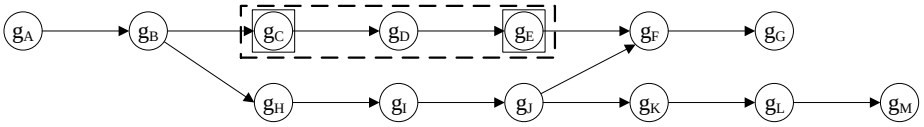


Figure 6: Example of a triple. Each vertex that has only one incoming and one outgoing edge can be considered as the centre of a triple. The vertex at the incoming and outgoing edge are the boundary vertices of the triple. Together they constitute a subgraph. The figure depicts only one triple. Boundary vertices are indicated by a square around the vertex. The subgraph is indicated by a dashed box. All triples present in this figure are: $\{\langle g_B, g_C, g_D\rangle, \langle g_C, g_D, g_E\rangle, \langle g_D, g_E, g_F\rangle, \langle g_B, g_H, g_I\rangle, \langle g_H, g_I, g_J\rangle, \langle g_J, g_K, g_L\rangle, \langle g_K, g_L, g_M\rangle\}$.

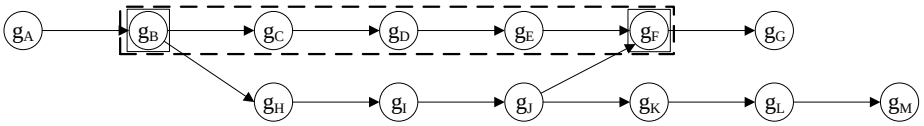


Figure 7: Example of a maximal non bifurcating sequence. A maximal non bifurcating sequence is a sequence that is delimited by sources, sinks, cycle breakers, assigned vertices or vertices that have more than one incoming or more than one outgoing edge. The intermediate vertices of the sequence have exactly one incoming and one outgoing edge. Together they constitute a subgraph. The figure depicts only one maximal non bifurcating sequence. Boundary vertices are indicated by a square around the vertex. The subgraph is indicated by a dashed box. All maximal non bifurcating sequences present in this figure are: $\{\langle g_A, g_B\rangle, \langle g_B, g_C, g_D, g_E, g_F\rangle, \langle g_F, g_G\rangle, \langle g_B, g_H, g_I, g_J\rangle, \langle g_J, g_K, g_L, g_M\rangle\}$

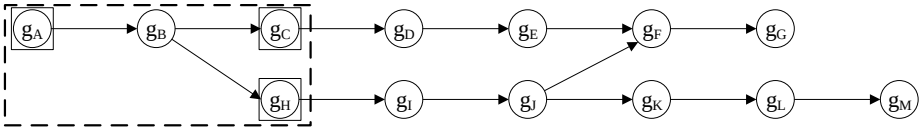


Figure 8: Example of a star. Each vertex that has more than one incoming and/or more than one outgoing edge can be considered as the centre of a star. The vertices at the incoming and outgoing edges are the boundary vertices of the star. Together they constitute a subgraph. The figure depicts only one star. Boundary vertices are indicated by a square around the vertex. The subgraph is indicated by a dashed box. All stars present in this figure are: $\{\langle g_A, g_B, g_C, g_H\rangle, \langle g_E, g_J, g_F, g_G\rangle, \langle g_I, g_J, g_F, g_K\rangle\}$.

meter and 30 meter experiments are shown in Table 2. The run times shown in Table 3 hold for Ubuntu Linux and Java7 on Intel(R) Xeon(R) CPU X5570 @

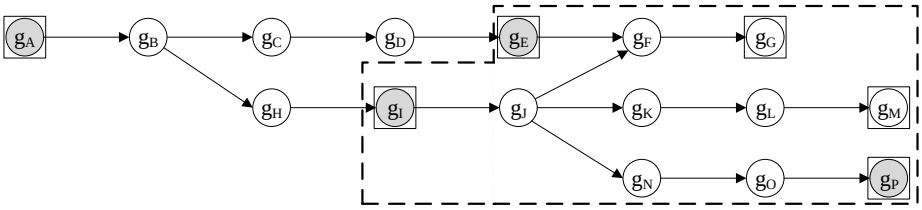


Figure 9: Example of a component. A component is the largest possible subgraph that can be extracted without including any assigned vertices or cycle breakers as intermediate vertices. The sources, sinks, cycle breakers and assigned vertices make up the boundary vertices of the component. The figure depicts only one component. Boundary vertices are indicated by a square around the vertex. Assigned vertices are indicated by a grey colour fill. The subgraph is indicated by a dashed box. All components present in this figure are: $\{\langle g_A, g_B, g_C, g_D, g_E, g_H, g_I\rangle, \langle g_E, g_F, g_G, g_I, g_J, g_K, g_L, g_M, g_N, g_O, g_P\rangle\}$.

2.93GHz. This computation server was used for the data preparation algorithm as well as the bus stop mapping (assignment) algorithm. The *data preparation* stage comprises (i) the determination of the projections of GTFS stop locations on the link geometries using `PostGIS` functions and (ii) the computation of the distance matrices between the stop projections for consecutive GTFS stops. Computation of the distance matrices consumes most of the time of the data preparation stage. The distance matrices computation time grows quadratic with the number of projections per GTFS stop (hence with a factor $(\frac{3.95}{2.62})^2 = 2.27$).

$$\sigma = \frac{\overline{A}}{\sqrt{-2 * \ln(1 - 0.95)}} \tag{1}$$

Comparing the results from the 15 meter experiment with the results from the 30 meter experiment shows that 91 % of the GTFS stop assignments is the same. The different assignments lead to different reconstructed bus trips. This was observed when comparing the road network link sequences for the reconstructed GTFS trips of both scenarios by means of a Levenshtein distance similarity score, which shows that 82 % of the reconstructed GTFS trips are more than 95 % similar and 28 % of the reconstructed GTFS trips is exactly

Table 1: Characteristics of the cleaned GTFS and OSM data of Flanders.

Quantity	Value
# Network Nodes	692 259
# Network Links	1 671 428
# GTFS stops	30 008
# GTFS trips (sequences)	215 426
# unique GTFS trips (sequences)	6 231
# Stop pairs (dist. matrices)	35 698

the same. Figure 10 shows both the distribution of the similarity scores and the cumulative distribution. When looking at the distribution of the trip length percentage change, which is shown in Figure 11, it can be seen that 85 % of the reconstructed GTFS trips have a a percentage change of 0 %. The trip length percentage change ranges from −34 % to +32 % with 91 % of the reconstructed GTFS trips having a trip length percentage change of less than 1 % in one of either direction. A very small decrease (0.16 %) in total distance driven between the 30 meter experiment and 15 meter experiment can be noticed. This result was expected since the 30 meter analysis has a wider search area which led to more candidates, which are possibly better, to choose from.

6. Validation

In order to validate the correctness of the bus stop mapping algorithm, three validation methods where applied: (i) a visual inspection, (ii) a speed distribution analysis and (iii) a perturbation analysis.

6.1. Visual Inspection

Validation of a small part of the resulting assignments of the experiments was done by visual inspection of known bus stops and reconstructed bus trips simi-

Table 2: Results for the 15 meter and 30 meter search radius experiments.

	15 m case	30 m case
# Projected stops	78 207	118 625
Max projStop / GtfsStop	10	10
Avg projStop / GtfsStop	2.61	3.95
# Trivial stops (DOF=1)	1 583	398
# Cycle breakers	549	777
$\log_{10}$(complexity)	10 975.095	15 443.095
$\log_{10}$(complexityCycleBreakers)	180.813	301.325
# Iterations	35	142
# Components	1 353	3 388
Duration iterations	21 s	34 s
Duration components solving	8 s	20 s
Max # DOFs for component	336	4 000

Table 3: Run times for the 15 meter and 30 meter search radius experiments.

	15 m case	30 m case
Read cleaned GTFS data	02 min 48 s	02 min 48 s
Read cleaned OSM data	01 min 16 s	01 min 16 s
Data preparation	08 min 06 s	19 min 24 s
Assignment	16 min 55 s	84 min 52 s
Trip reconstruct	01 min 10 s	01 min 24 s

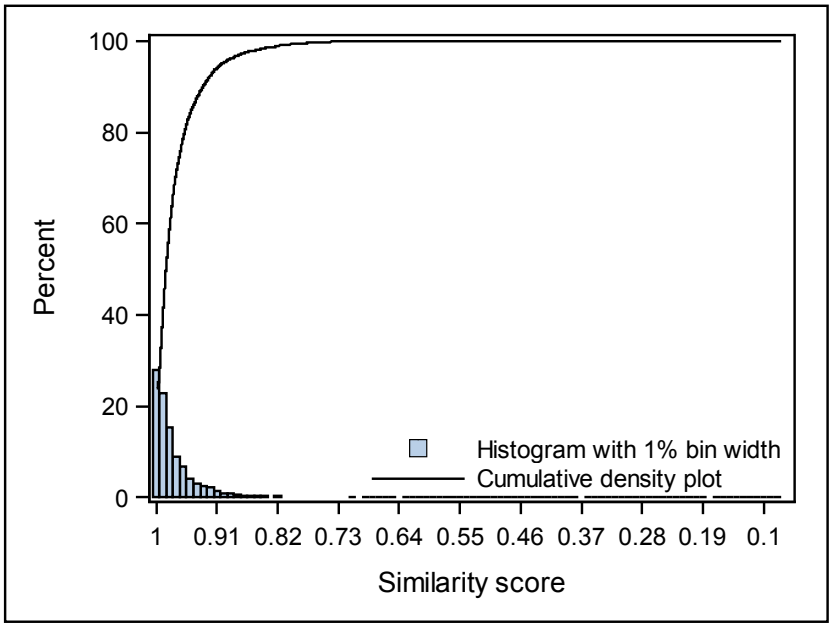


Figure 10: Distribution of the similarity score between reconstructed trips of the 15 meter and 30 meter experiments.

lar to the validation described by both Li[4] and Perrine et al.[5] All inspected bus stops were assigned correctly (correct link and side of the road) with the exception of bus stops near railway stations. However, these mismatches are not harmful for the objective of the algorithm (see Section 7.1 for an elaborate discussion). Figure 12 shows an example of a visual representation of a reconstructed bus trip.

6.2. Speed Distribution Analysis

Because interactive visual inspection only covers part of the results, a simple automated validation method was developed. Timing for the trips is taken from GTFS and trip length is taken from the reconstructed bus trips. For each experiment the following was computed: (i) the average speed of each complete trip (so-called *commercial speed*) and (ii) the speed between consecutive

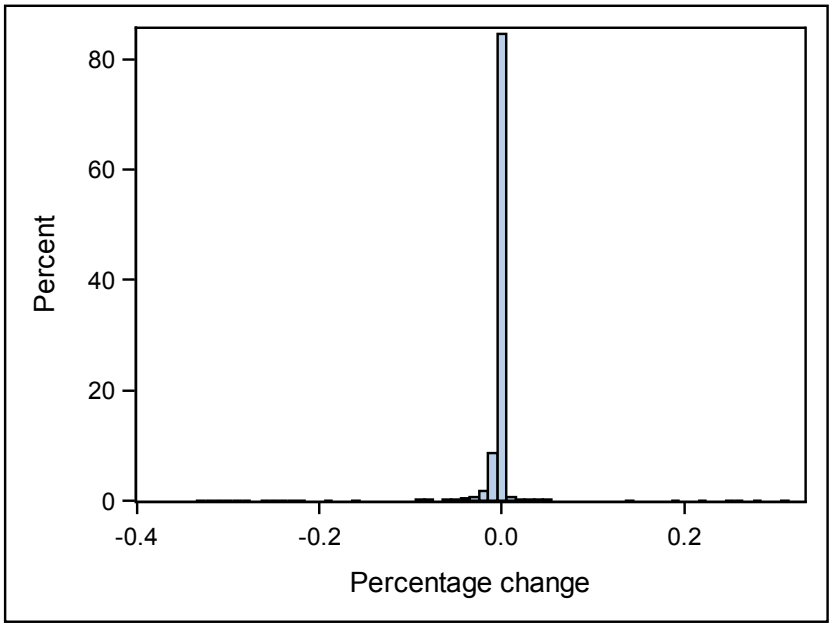


Figure 11: Distribution of the trip length percentage change between reconstructed trips of the 15 meter and 30 meter experiments.

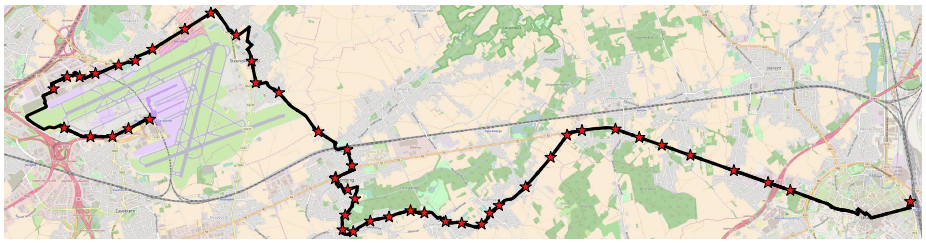


Figure 12: Visual inspection of a line going from Leuven (right) to Zaventem Airport (left). The black line represents the followed path to serve the GTFS bus stops (red stars).

stops. Outliers designate errors in the assignment; they are harmful to travel simulations.

6.2.1. The 15 meter experiment

Speed distributions are shown in Figure 13 and Figure 14.

The GTFS dataset contains 6 231 unique trips. One percent of these trips have a commercial speed lower than 16.18 km/h and one percent of these trips have a speed higher than 52.50 km/h. The average speed equals 31.85 km/h. The target commercial speed values for the buses of "De Lijn" are separated into 4 categories: 25 km/h for suburb and city services, 30 km/h for feeder services, 35 km/h for main services and 50 km/h for express services[8] The obtained average speed values for the individual trips in the 15 meter experiment fit well in these categories.

The GTFS dataset contains 7 016 328 *trip segments*. A trip segment is the passage of two consecutive stops. Several of these passages can belong to different executions of the same trip. For 859 440 trip segments the duration equals zero. This is a result of the time resolution of one minute used by "De Lijn". In other words, some bus stops are separated by a small distance so that the traversal time is less than a minute, which results in a zero duration. The zero duration segments were removed from our data. One percent of the segments have a speed lower than 8.47 km/h and one percent of the segments have a speed higher than 66.98 km/h. The average speed equals 27.20 km/h. The number of unique segments that have a speed lower than 8.47 km/h or higher than 66.98 km/h is equal to 1 411. The number of unique stops involved in these segments equals 2 492, i.e. 8.30 % of all GTFS stops. The segment speed distribution graph shows only those segments that have a speed lower than or equal to 100 km/h. The 8 771 segments with a speed higher than 100 km/h were removed since they are obvious outliers.

6.2.2. The 30 meter experiment

Speed distributions are shown in Figure 15 and Figure 16.

The GTFS dataset contains 6 231 unique trips. One percent of these trips have a commercial speed lower than 16.23 km/h and one percent of these trips have a speed higher than 52.28 km/h. The average speed equals 31.73 km/h. The obtained average speed values for the individual trips fit well in commercial speed categories of "De Lijn", which are 25 km/h for suburb and city services,

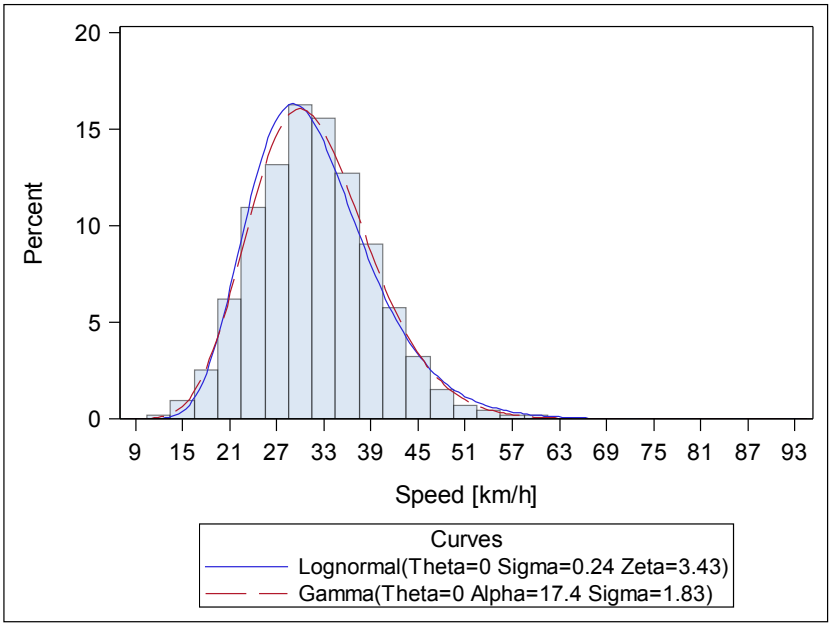


Figure 13: Distribution of the commercial speed value for all unique trips based on the results of the 15 meter experiment.

30 km/h for feeder services, 35 km/h for main services and 50 km/h for express services[8]

The GTFS dataset contains 7 016 328 trip segments. There are 659 trip segments with a speed equal to 0 km/h and there are 859 440 trip segments that have a zero duration. Speeds equal to 0 km/h occur in segments which have a zero distance. This is the result of coinciding projections of different GTFS stops. The time resolution used by "De Lijn" is one minute. Some bus stops are separated by a small distance so that the traversal time is less than a minute, which results in a zero duration. Both these segment groups were removed from our data. One percent of the segments have a speed lower than 8.37 km/h and one percent of the segments have a speed higher than 68.48 km/h. The average speed equals 27.23 km/h. The number of unique segments that have a speed

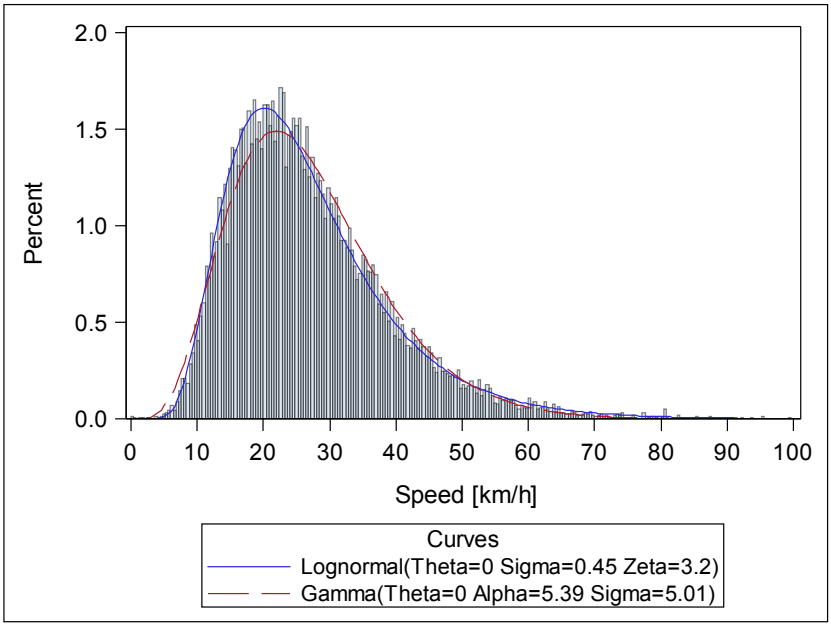


Figure 14: Distribution of the speed values of the individual segments of all trips based on the results of the 15 meter experiment. Only values lower than or equal to 100 km/h are shown.

lower than 8.37 km/h or higher than 68.48 km/h is equal to 1 282. The number of unique stops involved in these segments equals 2 232, i.e. 7.44 % of all GTFS stops. The segment speed distribution graph shows only those segments that have a speed lower than 100 km/h. The 9 950 segments with a speed higher than 100 km/h were removed since they are obvious outliers.

The 30 meter experiment has a slightly fewer outliers than the 15 meter experiment. Further research includes the investigation of the outliers. The source of the outliers will be investigated, i.e. do they emerge due to our optimization assumption, due to the underlying network or due to the data contained in the GTFS dataset. The latter is a possible candidate because the publicly available

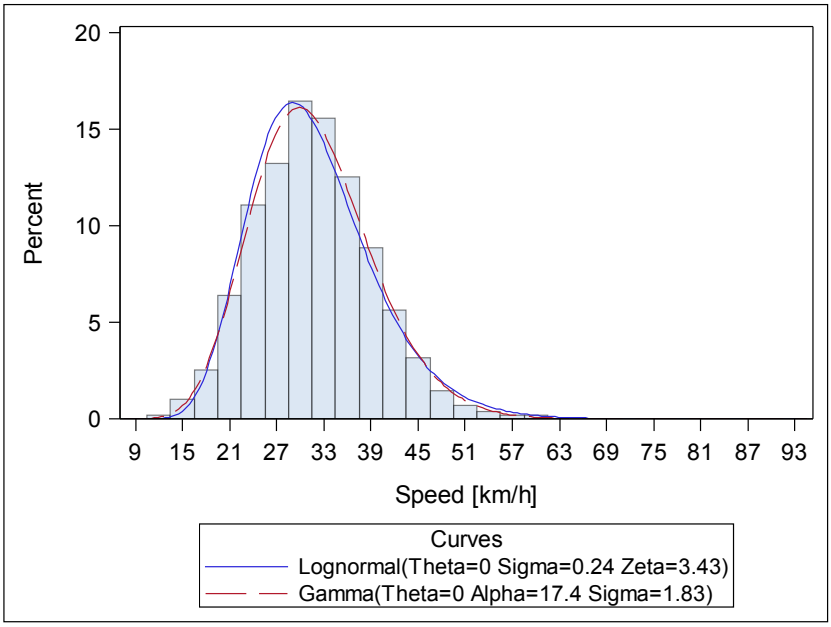


Figure 15: Distribution of the commercial speed value for all unique trips based on the results of the 30 meter experiment.

GTFS transit feed validator[4] reports an excessive travel speed ($\geq 100\,\mathrm{km/h}$) on 445 of all trip executions in the GTFS dataset.

6.3. Perturbation Analysis

The last validation method will test to what extent the bus stop mapping algorithm is able to provide the correct solution. In order to check the reproduction of the correct solution, the results of a previously executed experiment are assumed to constitute the *stated ground truth*. This output is translated back into a GTFS dataset and will be used as the new input for the algorithm.

For both the 15 meter and 30 meter analysis, three scenarios have been executed. The first scenario takes the original GTFS data as input and results

[4] `https://github.com/google/transitfeed/wiki/FeedValidator`

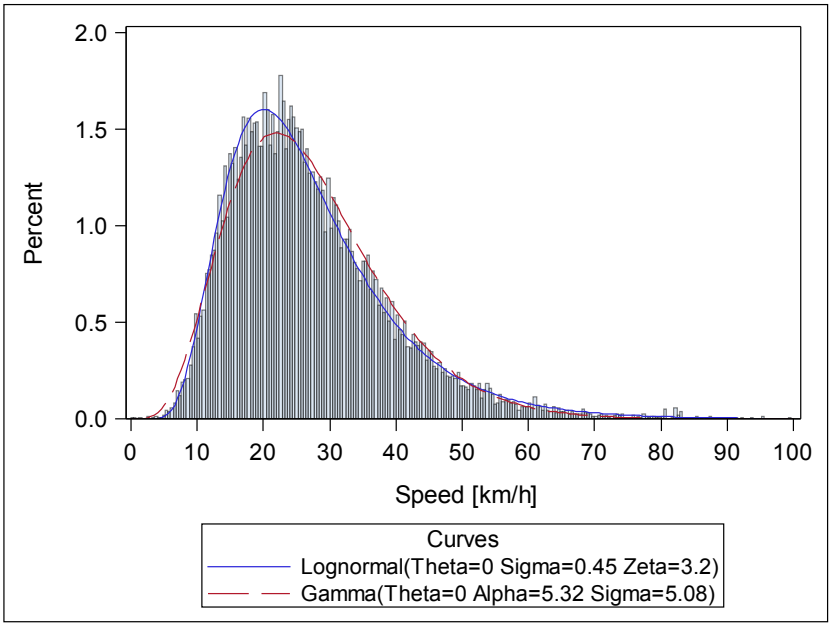


Figure 16: Distribution of the speed values of the individual segments of all trips based on the results of the 30 meter experiment. Only values lower than or equal to 100 km/h are shown.

in the stated ground truth solution as output. The second scenario takes the stated ground truth solution as input and its result describes the stability of the mapping. The third and final scenario for each analysis is based on a disturbed stated ground truth and is the core of our perturbation analysis. The three scenarios will be referred to as *original*, *stated ground truth* and *disturbed* scenario respectively. Figure 17 provides a visual overview of the three scenarios.

6.3.1. The 15 meter experiment

The first analysis assumes that the accuracy of GPS devices is approximately 15 meters. This assumption has two implications. The first implication is that the data preparation algorithm only searches for projected stops in a radius of 15 meters around the GTFS stop. The second implication is that the disturbance in polar coordinates is generated by sampling a distance d from a Rayleigh

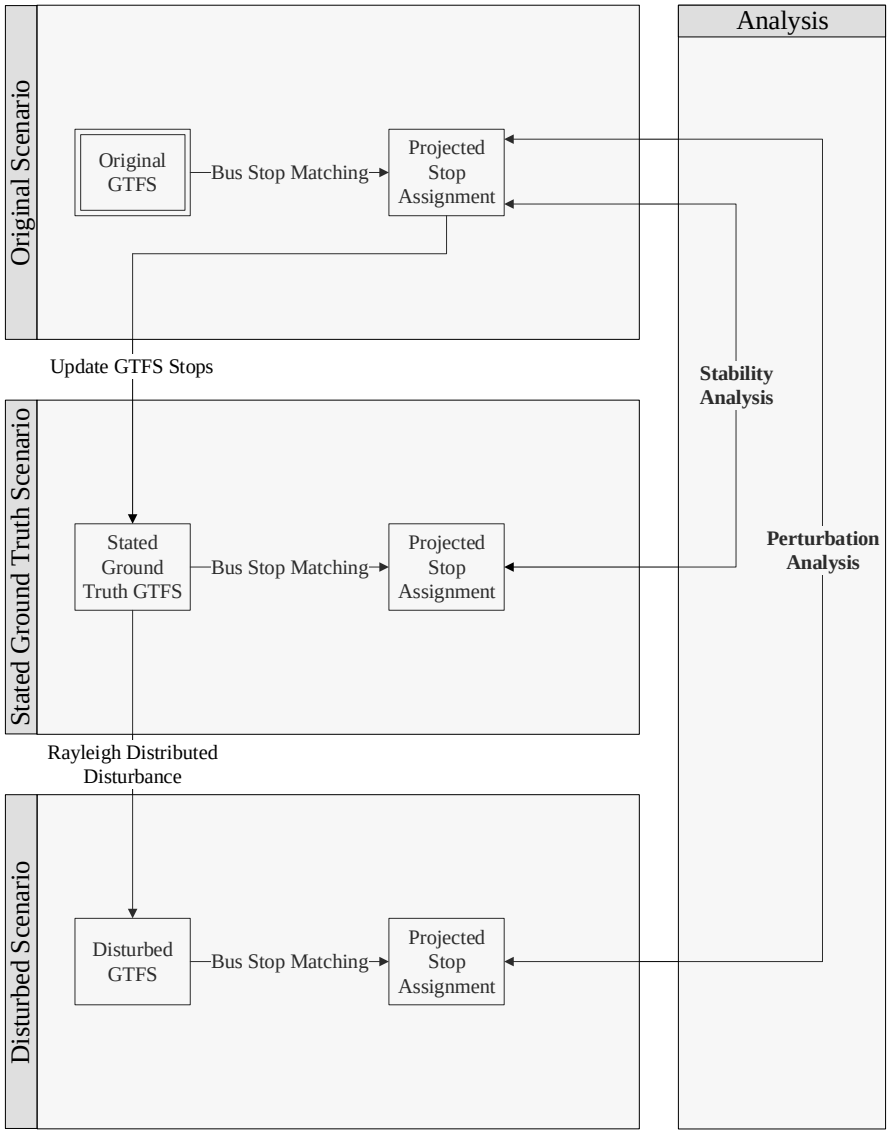


Figure 17: Overview showing how the three scenarios of a perturbation analysis are intertwined.

distribution and an angle $\alpha \in [0, 2\pi[$ from a uniform distribution. The Rayleigh distribution is the distribution of the square root of the sum of the squares of two identical and independent normally distributed variables. The inverse

cumulative distribution function used for distance sampling is: $Q(F;\sigma) = \sigma \cdot \sqrt{-2 \cdot \ln 1 - F}$ with σ estimated by equation (1) (see Section 5).

Comparing the results from the original scenario with the results of the stated ground truth scenario describes the stability of the stated ground truth solution. In the stated ground truth scenario each GTFS stop has a candidate which is equal to the assigned projected stop from the original scenario. This result was expected since the GTFS stops of the stated ground truth scenario lie on top of the assigned projected stops of the original scenario. Despite the same assignment always being possible, only 96 % percent of the assignments is the same. However, this does not necessarily imply that the reconstructed GTFS trip is affected. A different assignment could mean two things: (i) a new projected stop is chosen along the same path or (ii) a new projected stop is chosen along a different path. Investigation of the reconstructed GTFS trips is required in order to determine the stability of the solution. Comparing the road network link sequences for the reconstructed GTFS trips of both scenarios by means of a Levenshtein distance similarity score shows that 90 % of the reconstructed GTFS trips are more than 95 % similar and only 45 % of the reconstructed GTFS trips are exactly the same. This can be seen in Figure 18 which shows both the distribution of the similarity scores and the cumulative distribution. The result of different assigned projected stops does affect the paths of the reconstructed GTFS trips quite heavily, i.e. the newly assigned projected stops are along different paths. When looking at the distribution of the trip length percentage change, which is shown in Figure 19, it can be seen that 94 % of the reconstructed GTFS trips have a percentage change of 0 %. The trip length percentage change ranges from -18 % to +29 % with 96 % of the reconstructed GTFS trips having a trip length percentage change of less than 1 % in one of either direction. While the reconstructed GTFS trips differ quite a lot in followed path, the distance of the paths stays in most cases nearly the same. This is caused by the fact that projected stops for the i-th bus stop are now searched within a given radius around position P_I^S (stated) instead of around position P_i^O (original). When comparing the total distance driven of

all reconstructed GTFS trips of the original scenario, 129 346 638 meters, and the stated ground truth scenario, 129 313 053 meters, an improvement of 33 585 meters (0.03 %) can be noticed. This means the stated ground truth solution is not perfectly stable, but an improvement can be found, which is also an effect of the movement of the search area. This instability will also affect the disturbed scenario since the updated GTFS stop locations (stated ground truth GTFS data) are disturbed.

The results of the comparison between the original scenario and the stated ground truth scenario can be summarized as follows. Most GTFS stops receive the same assignment in the stated ground truth scenario as in the original scenario. The few different assignments lead to different reconstructed GTFS trips that can either be an improvement or deterioration with regards to the *individual* reconstructed GTFS trip length. A decrease happens slightly more often that an increase. Despite some exceptional cases where the reconstructed GTFS trip length experiences a larger increase or decrease, the total distance driven over all reconstructed GTFS trips slightly decreases.

Comparing the results of the original scenario with the results of the disturbed scenario shows how well the bus stop mapping algorithm is able to find back the stated ground truth solution when a disturbance is applied. As mentioned earlier, this disturbance complies with the accuracy reported for GPS recording devices. In the disturbed scenario 99 % of the GTFS stops have a candidate on a link that is equal to the link on which the assigned projected stop from the original scenario is located. In other words, 1 % of the GTFS stops have no candidate on the link on which the assigned projected stop from the original scenario is located. One might expect this number to be larger since a 95 % confidence interval is used which results in 5 % of the GTFS stops having a disturbance larger than 15 meters. However, the perturbation in general is not perpendicular to the link containing the bus stop, resulting in a higher percentage of GTFS stops having the assigned link among its candidates. Despite the same assignment being possible for 99 % of the GTFS stops, only 92 % of the assignments is the same: this may be caused by the occurrence of very

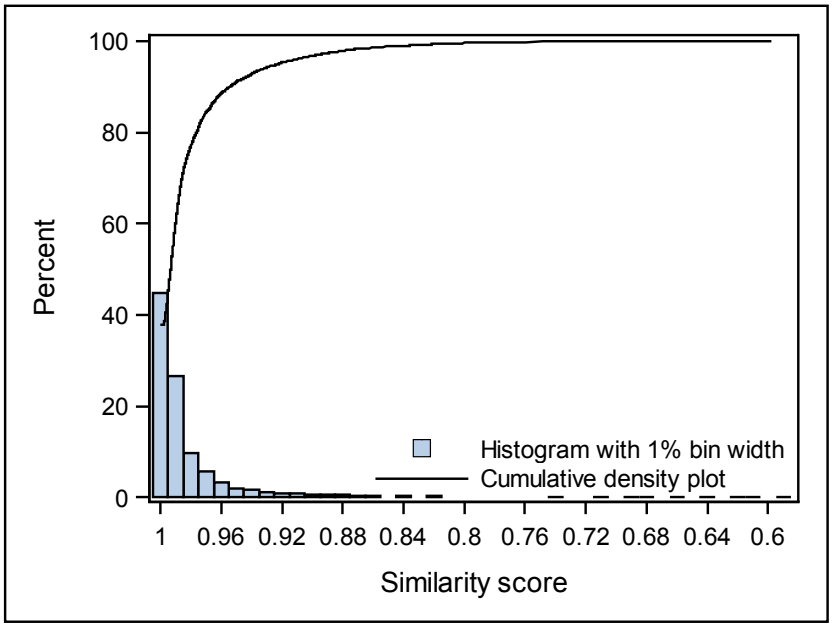


Figure 18: Distribution of the similarity score between reconstructed trips of the original and stated ground truth scenario for the 15 meters analysis.

short links in urban areas in particular near train stations where bus stops are clustered in a small area. Comparing the road network link sequences for the reconstructed GTFS trips of both scenarios by means of a Levenshtein distance similarity score shows that only 77 % of the reconstructed GTFS trips are more than 95 % similar and only 28 % of the reconstructed GTFS trips is exactly the same. This can be seen in Figure 20 which shows both the distribution of the similarity scores and the cumulative distribution. The result of different assigned projected stops does affect the paths of the reconstructed GTFS trips quite heavily. When looking at the distribution of the trip length percentage change, which is shown in Figure 21, it can be seen that 78 % of the reconstructed GTFS trips have a percentage change of 0 %. The trip length percentage change ranges from -20 % to +33 % with 86 % of the reconstructed

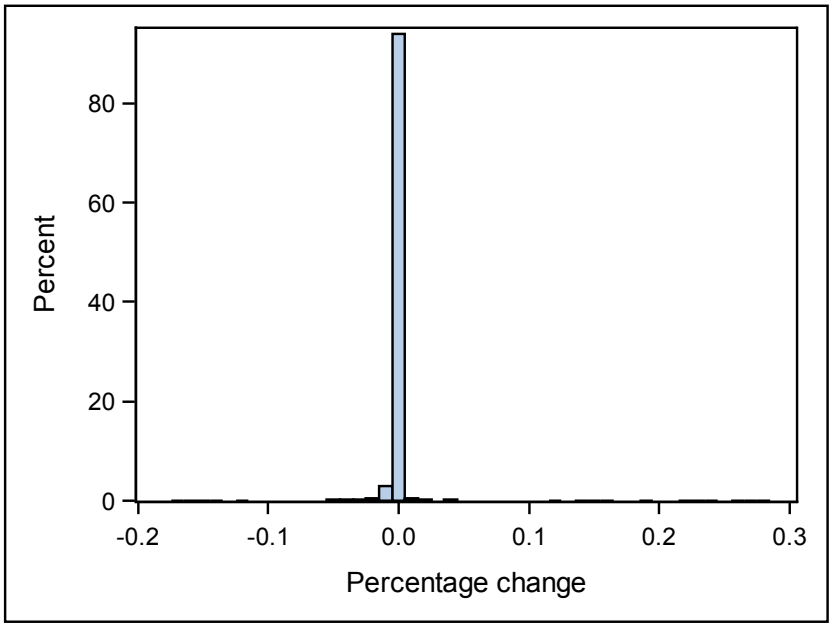


Figure 19: Distribution of the trip length percentage change between reconstructed trips of the original and stated ground truth scenario for the 15 meters analysis.

GTFS trips having a trip length percentage change of less than 1 % in one of either direction and with 92 % of the reconstructed GTFS trips having a trip length percentage change of less than 2 % in one of either direction. While the reconstructed GTFS trips differ quite a lot in followed path, the distance of the paths stays in most cases nearly the same. When comparing the total distance driven of all reconstructed the GTFS trips of the original scenario, 129 346 638 meters, and the disturbed scenario, 129 770 893 meters, a deterioration 424 255 meters can be noticed. In other words there is an increase of 0.3 % in the total distance driven over all reconstructed GTFS trips.

The results of the comparison between the original scenario and the disturbed scenario can be summarized as follows. Most GTFS stops receive the same assignment in the disturbed scenario as in the original scenario. The

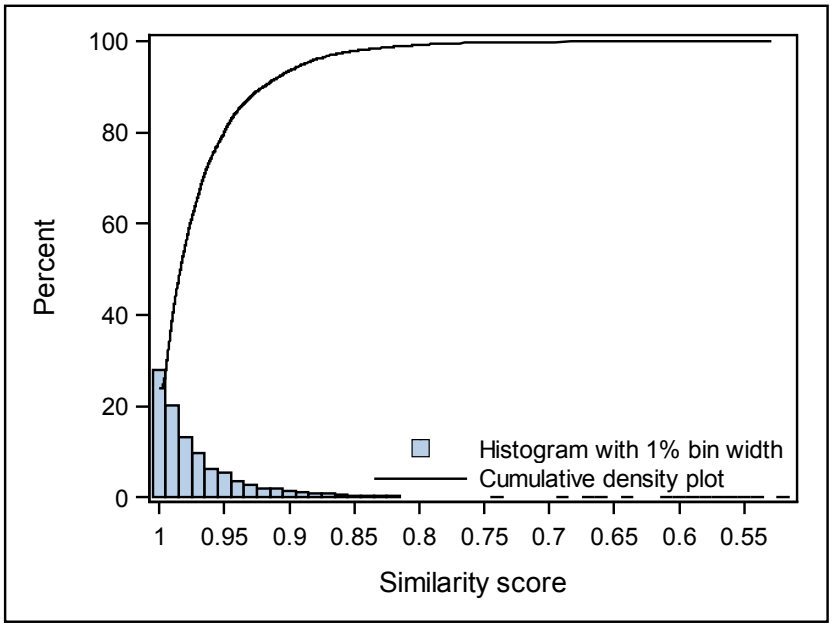


Figure 20: Distribution of the similarity score between reconstructed trips of the original and disturbed scenario for the 15 meters analysis.

few different assignments lead to different reconstructed GTFS trips that can either be an improvement or deterioration with regards to the *individual* reconstructed GTFS trip length. An increase happens slightly more often that a decrease. Despite some exceptional cases where the reconstructed GTFS trip length experiences a larger increase or decrease, the total distance driven over all reconstructed GTFS trips stays almost the same.

6.3.2. The 30 meter experiment

The second analysis assumes that the accuracy on GPS devices is approximately 30 meters. This assumption has two implications. The first implications is that the data preparation algorithm only searches for projected stops in a radius of 30 meters around the GTFS stop. The second implications is similar to the second implication of the 15 meter experiment.

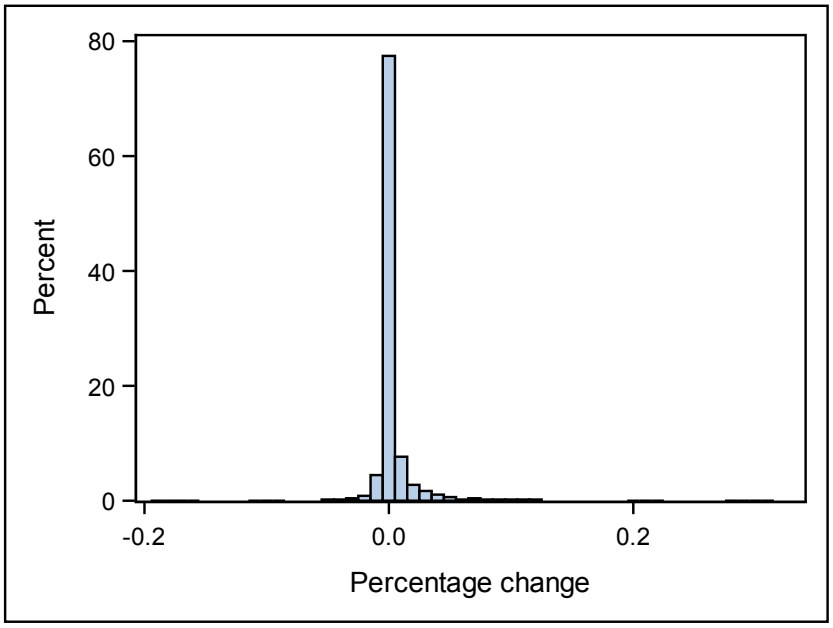


Figure 21: Distribution of the trip length percentage change between reconstructed trips of the original and disturbed scenario for the 15 meters analysis.

Comparing the results from the original scenario with the results of the stated ground truth scenario describes the stability of the stated ground truth solution. In the stated ground truth scenario each GTFS stop has a candidate which is equal to the assigned projected stop from the original scenario. This result was also observed in the 15 meter perturbation analysis when comparing the original scenario with the stated ground truth scenario and was due to the fact that the GTFS stops of the stated ground truth scenario lie on top of the assigned projected stops of the original scenario. Despite the same assignment always being possible, only 82 % percent of the assignments is the same. This is a decrease of 14 percent with regards to the 15 meter analysis.

Comparing the road network link sequences for the reconstructed GTFS trips of both scenarios by means of a Levenshtein distance similarity score shows that

87 % of the reconstructed GTFS trips are more than 95 % similar and 31 % of the reconstructed GTFS trips is exactly the same. This can be seen in Figure 22 which shows both the distribution of the similarity scores and the cumulative distribution. The result of different assigned projected stops does affect the paths of the reconstructed GTFS trips quite heavily, meaning that new assigned projected stops are along different paths. When looking at the distribution of the trip length percentage change, which is shown in Figure 23, it can be seen that 90 % of the reconstructed GTFS trips have a a percentage change of 0 %. The trip length percentage change ranges from -32 % to +25 % with 95 % of the reconstructed GTFS trips having a trip length percentage change of less than 1 % in one of either direction. While the reconstructed GTFS trips differ quite a lot in followed path, the distance of the paths stays in most cases nearly the same. This also an effect of the movement of the search area. When comparing the total distance driven of all reconstructed GTFS trips of the original scenario, 129 139 706 meters, and the stated ground truth scenario, 128 922 524 meters, an improvement of 217 182 meters (0.17 %) can be noticed. This means the stated ground truth solution is not perfectly stable, but an improvement can be found, which was expected due the movement of the GTFS stop locations which results in movements of the search areas. This instability will also affect the disturbed scenario since the updated GTFS stop locations (stated ground truth GTFS data) are disturbed.

The results of the comparison between the original scenario and the disturbed scenario can be summarized as follows. Most GTFS stops receive the same assignment in the disturbed scenario as in the original scenario. The few different assignments lead to different reconstructed GTFS trips that can either be an improvement or deterioration with regards to the *individual* reconstructed GTFS trip length. A decrease happens slightly more often that an increase. Despite some exceptional cases where the reconstructed GTFS trip length experiences a larger increase or decrease, the total distance driven over all reconstructed GTFS trips slightly decreases.

The stated ground truth solution in the 30 meter analysis is a little less

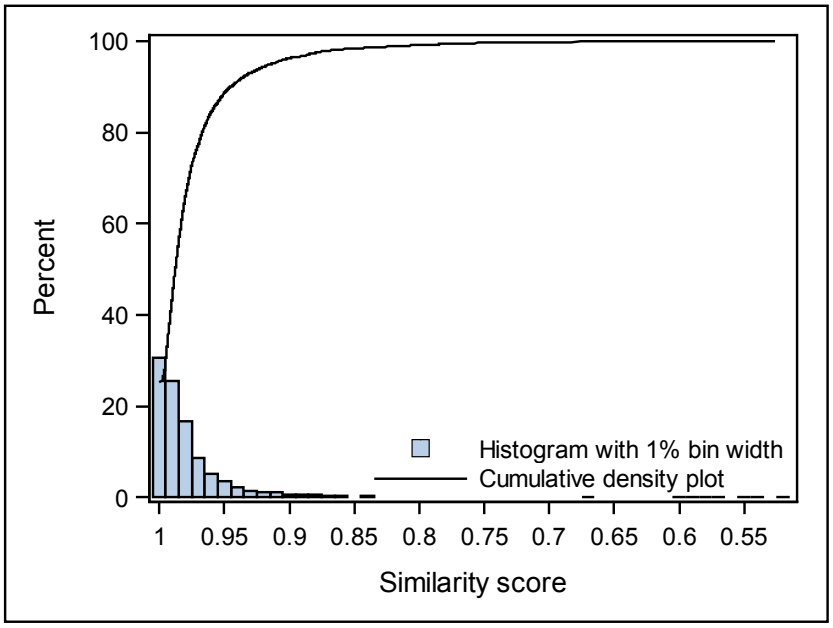


Figure 22: Distribution of the similarity score between reconstructed trips of the original and stated ground truth scenario for the 30 meters analysis.

stable than in the 15 meter analysis which is expected since the search range for projected stops is larger which provides more candidates to choose from.

Comparing the results of the original scenario with the results of the disturbed scenario shows how well the bus stop mapping algorithm is able to find back the stated ground truth solution when a disturbance is applied. As mentioned earlier, this disturbance corresponds to the error on GPS recording devices. In the disturbed scenario 96 % of the GTFS stops have a candidate on a link that is equal to the link on which the assigned projected stop from the original scenario is located. In other words, 4 % of the GTFS stops have no candidate on the link on which the assigned projected stop from the original scenario is located. Despite the same assignment being possible for 96 % of the GTFS stops, only 82 % of the assignments is the same. Comparing the road

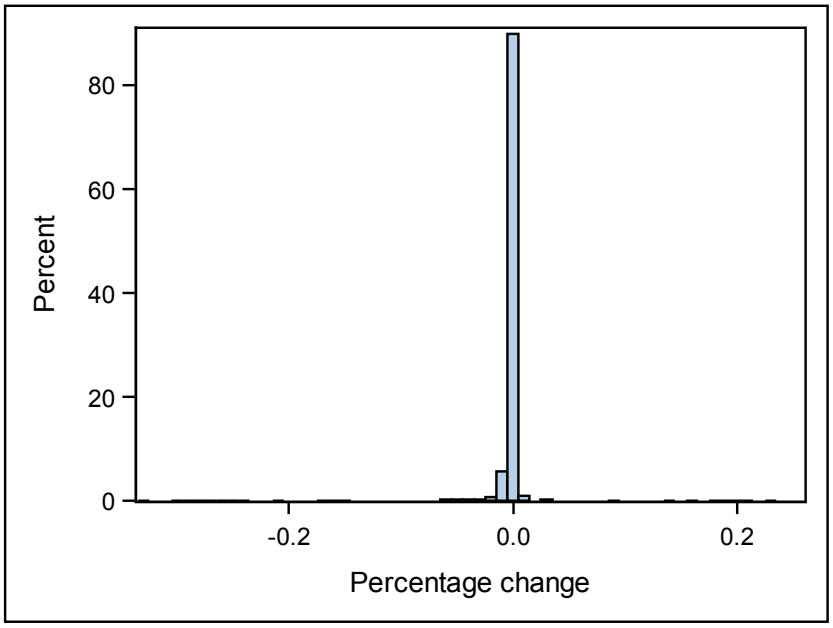


Figure 23: Distribution of the trip length percentage change between reconstructed trips of the original and stated ground truth scenario for the 30 meters analysis.

network link sequences for the reconstructed GTFS trips of both scenarios by means of a Levenshtein distance similarity score shows that only 59 % of the reconstructed GTFS trips are more than 95 % similar and only 11 % of the reconstructed GTFS trips is exactly the same. This can be seen in Figure 24 which shows both the distribution of the similarity scores and the cumulative distribution. The result of different assigned projected stops does affect the paths of the reconstructed GTFS trips quite heavily. When looking at the distribution of the trip length percentage change, which is shown in Figure 25, it can be seen that 59 % of the reconstructed GTFS trips have a a percentage change of 0 %. The trip length percentage change ranges from -32 % to +77 % with 74 % of the reconstructed GTFS trips having a trip length percentage change of less than 1 % in one of either direction and with 90 % of the reconstructed GTFS

trips having a trip length percentage change of less than 3 % in one of either direction. While the reconstructed GTFS trips differ quite a lot in followed path, the distance of the paths stays in most cases nearly the same. When comparing the total distance driven of all reconstructed the GTFS trips of the original scenario, 129 139 706 meters, and the disturbed scenario, 129 912 016 meters, a deterioration 772 310 meters can be noticed. In other words there is an increase of 0.6 % in the total distance driven over all reconstructed GTFS trips.

The results of the comparison between the original scenario and the disturbed scenario can be summarized as follows. Many GTFS stops receive the same assignment in the disturbed scenario as in the original scenario. However, the small number of different assignments leads to many different reconstructed GTFS trips that can either be an improvement or deterioration with regards to the reconstructed GTFS trip length. An increase happens slightly more often that a decrease. Despite the large number of reconstructed GTFS trips that experience a change in their length, the total distance driven over all reconstructed GTFS trips stays almost the same.

The disturbed scenario of the 30 meter analysis is worse than the disturbed scenario of the 15 meter analysis. The followed paths of the reconstructed GTFS trips differ more often and more reconstructed GTFS trips suffer from a change in their length. This length change has also a larger range in the 30 meter analysis. This can be explained by the larger disturbance and the larger radius used to search for projected stops.

7. Discussion

7.1. Areas having a High Bus Stop Density

Areas near railway stations in Belgium tend to have a higher bus stop density, i.e. the number of bus stops is very high compared to the size of the area. Consider the situation near the Hasselt (Flanders, Belgium) railway station, shown in Figure 26. It consists of a large array of parallel platforms sharing a common entry lane and a common exit lane. The area contains 15 bus stops in

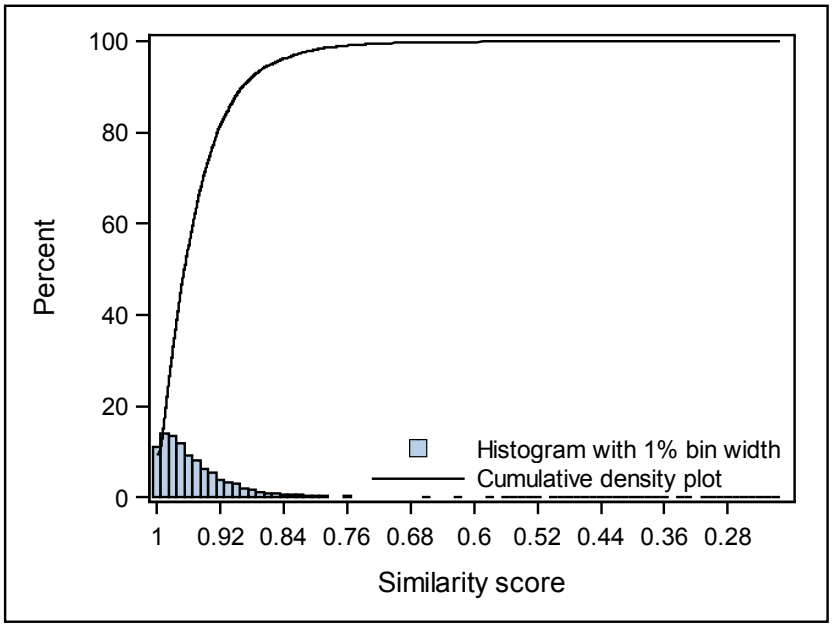


Figure 24: Distribution of the similarity score between reconstructed trips of the original and disturbed scenario for the 30 meters analysis.

a rectangle of 30 meters × 150 meters. The distance between adjacent parallel platforms is about two times the width of a bus. Though the spatial density of the bus stops is high, the complexity of the algorithm is not affected. Computational complexity depends mainly on the number of services using a particular stop because (in general) it increases the number of neighbours the stop depends on. Another important impact factor is the number of projected stops that are found for each GTFS stop. This number is higher in areas with a higher network density. Despite the fact that the computational complexity is not affected by the density of the bus stops, visual inspection showed that the bus stops near railway stations are still assigned to the wrong link. The reason for these mismatches is threefold: (i) the positional error of bus stops near stations is often higher than the search radius, (ii) the density of the network (number

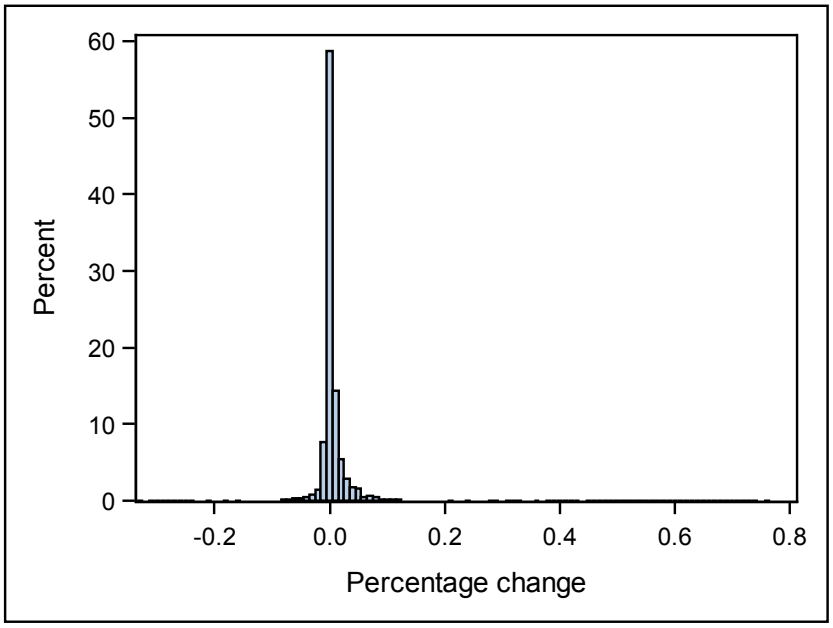


Figure 25: Distribution of the trip length percentage change between reconstructed trips of the original and disturbed scenario for the 30 meters analysis.

of links) is very high for a relatively small area and (iii) most bus trips start or stop at a station. The first reason, which can be observed in Figure 26, is very important because it prevents the correct solution from being among the candidates. The origin of this higher positional error is unknown. The second reason impacts the number of candidates, i.e. the number of candidates in the search radius is higher than the upper limit, which results in keeping only the closest X candidates, where X represents the upper limit. This is done to reduce the complexity of the algorithm. The third reason is a problem because the algorithm tries to minimize the total distance driven, which results in selecting a link that is closer to the entry/exit of the station area. Though bus stops are assigned to the wrong link near railway stations, the objective of the algorithm is not harmed. Bus stops are assigned to the correct side of the road while

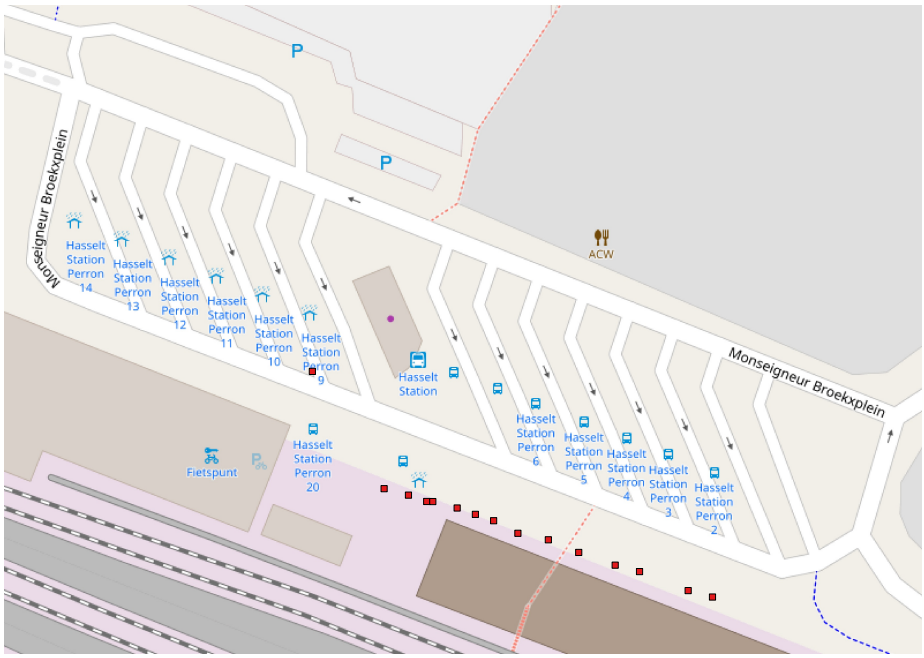


Figure 26: Location of the GTFS stops (red squares) of De Lijn near Hasselt station shown on a OSM map.

obtaining a sufficiently accurate bus trip distance in order to perform accurate microsimulations.

7.2. Multi-Threading

GTFS stops are assigned using a method that consists of a single loop that applies several techniques (handling triplets, stars, etc.) to isolated smaller problems. Problem decomposition seems to happen gradually in concrete cases. The small isolated problems may be solved in parallel: up to now we did not yet implement this because the cost to set up a separate thread and starting it may turn out to be of the same order of magnitude as the cost to solve the problem. We did not perform this experiment. The software to set up data structures (parsing GTFS data, preparing graphs, etc.) cannot be parallelised.

7.3. Effects of Link Geometry Errors

According to Haklay et al.[1] the positional error for a road is well below 6 meter in regions where 15 or more OSM contributors are active. In so called *complete* areas (which cover all of Europe) the average error is estimated at 9.57 meter. The distribution in Figure 27 shows that the large majority of GTFS stop coordinates are within a narrow band around the link geometries. If (many) geometry points were missing and hence many straight links would replace curved roads, the distribution would look different (i.e. the variance would be larger) because the geometries and GTFS coordinates come from different independent datasets. The goal of the bus stop mapping algorithm is to assign bus stops to the road network in order to enable accurate simulation of bus based public transport. It operates under the assumption that public transport operators minimize the total distance driven to complete all trips. An error in geometry of a road segment can result in a GTFS stop being assigned to (i) the correct road segment but the wrong side of the street or (ii) an incorrect road segment. In case of the former error (i), the correct street will still be served but the route length may change. Taking into account the typical length of a street, the change is small unless the majority of the streets in the neighbourhood constitute one-way streets. In case of the latter error (ii), the service level of the area containing the GTFS stop will be inaccurate. This is a local effect. Note that the positional errors in the geometry need to be large in order to provoke this effect. The incorrect assignment of a bus stop can have an impact on the length of the trips that pass by that bus stop, which in turn has an impact on the travel speeds of the buses driving those trips. The latter impact is a result of the fixed service times specified in the GTFS dataset. Bus speeds may become either unrealistically small if the computed trip distance is shorter than the actual trip distance or unrealistically large if the computed trip distance is longer than the actual trip distance.

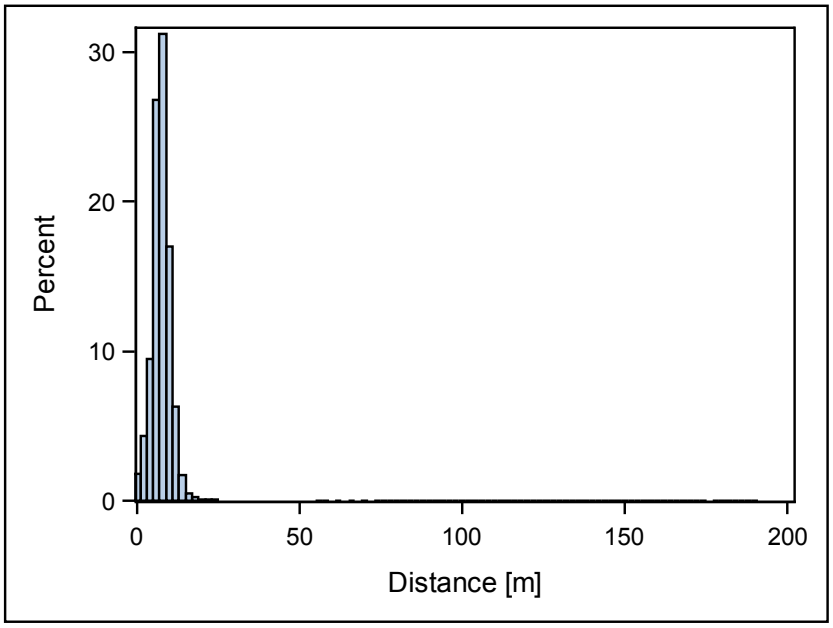


Figure 27: Distribution of the distance between GTFS stops locations and their closest link projections.

7.4. Data Quality: Outliers

The GTFS of De Lijn contains 30 008 bus stops that are part of a route. In case of the 15 meter experiment, the algorithm starts by searching candidate links for projected stops in a radius of 15 meter around each GTFS stop. This search radius is not large enough for 389 GTFS stops and will be increased in steps of 15 meter until at least one candidate link is found. Table 4 lists the number of GTFS stops that find their first candidate link in a given radius for the 15 meter experiment. The largest distance between a GTFS stop and a candidate link is 192.68 meter. The smallest distance is 0.001 meter. The average distance is 13.37 meter. The 30 meter experiment has the same largest, smallest and average distance. However, only in 32 cases the search radius needed to be increased. This 30 meter experiment increased the search radius

Table 4: Experiment with 15 meter search radius: number of GTFS stops that find a first candidate link in a given radius.

Radius	GTFS stops
0 m to 15 m	29619
15 m to 30 m	352
30 m to 45 m	24
45 m to 60 m	8
60 m to 75 m	3
75 m to 90 m	0
90 m to 105 m	0
105 m to 120 m	0
120 m to 135 m	0
135 m to 150 m	0
150 m to 165 m	0
165 m to 180 m	1
180 m to 195 m	1

in steps of 30 meters. Table 5 lists the number of GTFS stops that find their first candidate link in a given radius for the 30 meter experiment.

8. Software Availability

According to the policy defined by the Hasselt University Transportation Research Institute (IMOB), the software is made available at no cost for non-commercial use by research institutes.

Table 5: Experiment with 30 meter search radius: number of GTFS stops that find a first candidate link in a given radius.

Radius	GTFS stops
0 m to 30 m	29971
30 m to 60 m	32
60 m to 90 m	3
90 m to 120 m	0
120 m to 150 m	0
150 m to 180 m	1
180 m to 210 m	1

9. Conclusion

This paper presented a bus stop mapping algorithm that is capable of handling all bus trips *at once* and assigns each bus stop to a particular side of the road. This was a challenging task since the PT network is vastly interconnected. The assignment of the bus stops is computed by optimization, under the assumption that the PT operators minimize the total distance driven to complete all trips. Visual inspection of known bus stops shows correct assignment of these stops. Validation based on average segment speeds shows that less than 9 % of the stops require interactive attention and those stops can be identified automatically in order to correct the data. The most important quantity for bus operation simulations is the trip distance. Validation was done by disturbing the positions for presumably known bus stops using position errors derived from GPS device accuracy and successively running the matching algorithm. This shows that the trip distances of the disturbed scenario differ only slightly from the trip distances of the original scenario, which indicates that the algorithm is able to deal with a disturbance.

Acknowledgement

The research reported was partially funded by the IWT 135026 Smart-PT: Smart Adaptive Public Transport (ERA-NET Transport III Flagship Call 2013 "Future Traveling").

References

[1] M. Haklay, S. Basiouka, V. Antoniou, A. Ather, How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information, The Cartographic Journal 47 (4) (2010) 315 – 322. doi:10.1179/000870410X12911304958827.

[2] J. Vuurstaek, G. Cich, L. Knapen, A.-U.-H. Yasar, T. Bellemans, D. Janssens, GTFS Bus Stop Mapping to the OSM Network, Procedia Computer Science 109 (2017) 50–58, 8th International Conference on Ambient Systems, Networks and Technologies, ANT-2017 and the 7th International Conference on Sustainable Energy Information Technology, SEIT 2017, 16-19 May 2017, Madeira, Portugal. doi:10.1016/j.procs.2017.05.294.

[3] A. Lektauers, J. Petuhova, A. Teilans, A. Kleins, The Development of an Integrated Geosimulation Environment for Public Transit Analysis and Planning, Information Technology and Management Science 15 (1) (2012) 200 – 205. doi:10.2478/v10313-012-0026-3.

[4] J.-Q. Li, Match bus stops to a digital road network by the shortest path model, Transportation Research Part C: Emerging Technologies 22 (2012) 119 – 131. doi:10.1016/j.trc.2012.01.002.

[5] K. Perrine, A. Khani, N. Ruiz-Juri, Map-Matching Algorithm for Applications in Multimodal Transportation Network Modeling, TRB Research Record 2537 (2015) 62 – 70. doi:10.3141/2537-07.

[6] S. A. Ordóñez, Semi-Automatic Tool for Bus Route Map Matching, in: A. Horni, K. Nagel, K. W. Axhausen (Eds.), The Multi-Agent Transport

Simulation MATSim, Ubiquity Press, London, 2016, Ch. 18, pp. 115–122. doi:10.5334/baw.

[7] G. Cich, J. Vuurstaek, L. Knapen, A.-U.-H. Yasar, T. Bellemans, D. Janssens, Data Preparation to Simulate Public Transport in Micro-Simulations Using OSM and GTFS, Procedia Computer Science 83 (2016) 50 – 57, the 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops. doi:10.1016/j.procs.2016.04.098.

[8] De Lijn, Tritel, Goudappel-Cofeng, Mint, Mobiliteitsvisie De Lijn 2020 (2009).