Journal of Big Data Analytics in Transportation (2022) 4:23–39
https://doi.org/10.1007/s42421-022-00052-9

ORIGINAL PAPER

Semi-supervised Mode Classification of Inter-city Trips from Cellular Network Data

Nils Breyer[1] · Clas Rydergren[1] · David Gundlegård[1]

Received: 14 July 2021 / Revised: 7 March 2022 / Accepted: 15 May 2022 / Published online: 10 June 2022

Abstract
Good knowledge of travel patterns is essential in transportation planning. Cellular network data as a large-scale passive data source provides billions of daily location updates allowing us to observe human mobility with all travel modes. However, many transport planning applications require an understanding of travel patterns separated by travel mode, requiring the classification of trips by travel mode. Most previous studies have used rule-based or geometric classification, which often fails when the routes for different modes are similar or supervised classification, requiring labelled training trips. Sufficient amounts of labelled training trips are unfortunately often unavailable in practice. We propose semi-supervised classification as a novel approach of classifying large sets of trips extracted from cellular network data in inter-city origin–destination pairs as either using road or rail. Our methods require no labelled trips which is an important advantage as labeled data is often not available in practice. We propose three methods which first label a small share of trips using geometric classification. We then use structures in a large set of unlabelled trips using a supervised classification method (geometric-labelling), iterative semi-supervised training (self-labelling) and by transferring information between origin–destination pairs (continuity-labelling). We apply the semi-supervised classification methods on a dataset of 9545 unlabelled trips in two inter-city origin–destination pairs. We find that the methods can identify structures in the cells used during trips in the unlabelled data corresponding to the available route alternatives. We validate the classification methods using a dataset of 255 manually labelled trips in the two origin–destination pairs. While geometric classification misclassifies 4.2% and 5.6% of the trips in the two origin–destination pairs, all trips can be classified correctly using semi-supervised classification.

Keywords Cellular network data · Travel mode classification · Semi-supervised learning

Introduction

Transportation planning and management on strategic, tactical and operational levels require a good understanding of historic and present mobility patterns. Commonly, travel surveys and traffic counts are used to obtain data on mobility patterns. Unfortunately, these data sources are expensive to collect and thus usually very limited in their volume and variety giving only partial insights. Further, travel surveys are suffering from decreasing response rates (Schulz et al. 2016).

Meanwhile, new large-scale passive data sources providing information on travel patterns have emerged. Cellular network data as a passive data source collected by cellular network operators provides large-scale observations of people's movements and, thus, a comprehensive overview of travel patterns with all travel modes (Calabrese et al. 2011; Gundlegård et al. 2016). Datasets of GPS tracks can provide information on travel patterns as well, but unlike cellular network data these datasets usually cover fewer users (Barbosa et al. 2018) and often not all travel modes. They are usually collected using specific apps and only for users who have activated the data collection, which possibly introduces bias. For example, using data from only one of the cellular network operators in Sweden with around 20% of the market share, we can observe about 1 billion events and extract more than 23 million trips during one week. The data, thus, provide a significant sample of all trips made.

For many applications not only the total mobility patterns are of interest, but also an understanding of the mobility patterns with each travel mode. Examples are the analysis

✉ Nils Breyer
nils.breyer@liu.se

[1] Department of Science and Technology, Linköping University, Linköping, Sweden

of road traffic flows and railway passenger flows. Understanding how different transportation modes are used is also a key to facilitating multi-modal transportation planning. Obtaining the modal split of trips between cities is important to identify potential travel demand that could be shifted to a different mode. Augmenting trips extracted from cellular network data with necessary metadata such as the travel mode has thus been pointed out as a significant challenge for enabling the use of cellular network data as a data source for practical transportation planning (Anda et al. 2017).

Existing mode classification methods for GPS trajectories are usually not applicable to cellular network data. In cellular network data, only an indirect approximate position is given through the cells used and their known antenna locations. Therefore, the data are typically noisier and of lower temporal resolution compared to GPS data. Classification methods need to be adapted to and use cellular network data's specific characteristics to perform well. Using the specific characteristics of cellular network data is generally understudied in the literature that deals with mode classification from cellular network data. Most proposed methods are using simple rule-based or geometric approaches classifying each trip using only that particular trip's data (Huang et al. 2019). These methods perform poorly in cases where the travel modes are hard to distinguish.

Supervised classification methods, which could give an improvement in these cases, have not been considered in the literature, with a few exceptions (Xu et al. 2011; Breyer et al. 2021). A major reason for this is that they require labelled training data, which is rarely available in practice. On the other hand, usually, large amounts of unlabeled data are available. The emerging concepts of semi-supervised learning show that patterns in unlabeled data can be used for solving classification problems (van Engelen and Hoos 2019). In the context of cellular network data, to the best of our knowledge, these techniques have so far only been used by Bachir et al. (2019b), who achieved very promising results with a semi-supervised classification method involving cellular network data combined with travel mode priors from a household survey.

In this paper, we aim to classify inter-city trips as made by rail or road using different semi-supervised learning assumptions. This paper's main contributions are twofold: First, we show how semi-supervised methods can improve mode classification compared to rule-based or geometric methods, which are the most commonly used methods in the literature. We show that the proposed semi-supervised methods can perform better than these methods in challenging cases where the routes of different modes are similar. Second, we advance the practical usability of supervised learning for mode classification. The main reason why supervised and semi-supervised methods are not commonly used is the lack of necessary labeled training data. We show that the need for such labeled training data can be eliminated using semi-supervised labeling methods based on some sensible learning assumptions.

The scope of this paper is to show how semi-supervised methods can be used to classify inter-city trips as made by rail or road. We focus on these modes, as they are the most common modes of inter-city travel. Unlike air traffic, the infrastructure for these modes are often more or less co-located and travel times can be similar. In these cases, it is challenging to classify trips by travel mode.

The remainder of the paper is structured as follows. Sect. 2 introduces the mode classification problem and previous related work. In Sect. 3 a geometric mode classification method is given and Sect. 4 presents methods of mode classification using semi-supervised learning. Using the dataset described in Sect. 5, the methods are compared and validated in Sect. 6. Finally, Sects. 7 and 8 discuss and conclude the findings.

Preliminaries

Cellular network data consist of location updates that are recorded inside the operator's infrastructure. Each location update consists of an anonymised user ID, a timestamp and a cell ID. Several types of events, not only necessarily related to physical movements, can trigger location updates (Gundlegård 2018). Here, we assume that the cellular network data not only contains Call Detail Records (CDR) triggered by phone calls or messages but also other events such as periodic events and location area updates. Given that the antennas' positions and/or the coverage areas of the cells are known, the location updates give an approximation of the user location over time. The data are typically noisy and of much lower and more varying resolution in time and space than GPS tracks.

Cellular network data may contain updates also when users are not moving. To analyse travel patterns, it is, therefore, required to first extract *trips*, that is movements between two stops. We describe a trip by the list of cell IDs used during the trip which we call the *cellpath*

$$T = [c_s, c_1, c_2, \dots, c_e], \tag{1}$$

as well as the trip start time $t_s(T)$ and the trip end time $t_e(T)$. The trips used in this paper have been extracted using a stop-based trip extraction method as described in Breyer et al. (2020); other methods methods are discussed in Alexander et al. (2015); Calabrese et al. (2010); Graells-Garrido et al. (2018); Alexander et al. (2015); Breyer et al. (2017) among others.

Mode Classification of Cellular Network Data

It is generally a feature of cellular network data that trips made with all travel modes and for all purposes or activities can be observed. However, this is also limiting the usefulness for many applications in traffic planning and modelling, which require the data to be enriched by metadata such as travel mode. The problem of travel mode classification has been gaining attention recently. In a systematic review, Huang et al. (2019) found that most of the proposed methods can be described as simple rule-based methods making use of geodata. These methods may perform poorly in challenging OD-pairs where several modes have similar routes. Further, Huang et al. (2019) point towards a lack of validation of the mode classification for individual trips as many studies have only compared aggregated mode share statistics.

Given a set of extracted trips $\boldsymbol{T}_{O,D}$ between city O and city D, we define the mode classification problem as follows: Given a trip $T \in \boldsymbol{T}_{O,D}$, find the mode probabilities $p(m|T)$ for $m \in \{road, rail\}$. The mode probability $p(m|T)$ is the probability that m is the main mode of the trip, meaning that mode m has been used for the major part of the trip's distance. We assume that all trips are made by either rail or road (where the road mode includes private cars, buses, etc.), as they are the only available options for the OD-pairs considered in this paper. For OD-pairs where options such as air or water (ferry) traffic are viable options, the set of modes needs to be extended accordingly. We can distinguish the following four categories of methods to classify a set $\boldsymbol{T}$ of trips by travel mode:

(1) Rule-Based Classification. A trip $T \in \boldsymbol{T}_{O,D}$ is classified using predefined rules based on the characteristics of the particular trip to classify. Kalatian and Shafahi (2016), for example, use the travel speed to distinguish between modes.

(2) Geometric Classification. A trip $T \in \boldsymbol{T}_{O,D}$ is classified by comparing the geometry of the trip with the geometry of infrastructure (such as the railway network) or available route alternatives for each mode. In addition to the cellular network data, these methods also use geometric data describing the transport infrastructure. Examples are found in Qu et al. (2015); Phithakkitnukoon et al. (2017); Breyer et al. (2021) among many others.

(3) Supervised Classification. A trip $T \in \boldsymbol{T}_{O,D}$ is classified using a supervised classification method which has to be trained using a set of labelled trips $\boldsymbol{T}^*_{O,D}$ from the same OD-pair with the correct modes known. Examples are found in Breyer et al. (2021); Xu et al. (2011).

(4) Semi-supervised Classification. A trip $T \in \boldsymbol{T}_{O,D}$ is classified by identifying patterns in the large set of unlabeled trips $\boldsymbol{T}_{O,D}$ together with a small set of labelled trips $\boldsymbol{T}^*_{O,D}$ which is labelled manually or automatically using another method (pseudo-labelling). An example of a semi-supervised method is the cluster-then-label approach used in Bachir et al. (2019b).

Approaches (1) and (2) can be applied using only unlabelled data. However, (1) fails when the trip characteristics are not different enough between trips of different modes or when the data quality is too low. Approach (2) fails when the routes for different modes are very similar. Approach (3) can sometimes achieve better classification results in the above situations (Breyer et al. 2021), but requires enough labelled for the same OD-pair, which is most often not available in practice. This paper presents new methods using approach (4).

Semi-supervised Classification

A lack of sufficient amounts of labelled data to train supervised classification methods is not unique to the classification problem considered in this paper (Zhou 2017). The field of semi-supervised classification methods deals with methods for problems where there is only a small amount of labelled data but a large amount of unlabelled data. According to van Engelen and Hoos (2019), semi-supervised methods may be used if the the distribution of the large set of unlabelled observations $p(x)$ contains some information about the posterior distribution $p(y|x)$, where x is an observation and y the label of x. Different semi-supervised learning methods make different assumptions about which information the unlabelled data contains about the posterior distribution. We present those semi-supervised assumptions in Sect. 4.1 and discuss how they can be used for the mode classification of trips extracted from cellular network data.

Semi-supervised classification methods can be divided into inductive and transductive methods (van Engelen and Hoos 2019). Inductive methods yield a classification model that we can use later to classify new previously unseen observations. A popular type of inductive semi-supervised methods are wrapper methods, which extend supervised classification methods. Self-labelling, for example, is a wrapper method that labels the unlabelled observations iteratively (Triguero et al. 2015). Another approach is building on unsupervised methods as in the cluster-then-label approach (Bachir et al. 2019a). Transductive semi-supervised methods such as graph-based methods obtain labels for a particular set of observations without providing a general classification model (Subramanya and Talukdar 2014).

Semi-supervised classification methods have been successfully used to classify trajectories from GPS data (Yu 2020; Dabiri et al. 2020). These methods are using detailed features, such as speed, acceleration, jerk and turn rates. Kalatian and Farooq (2020) demonstrated a semi-supervised approach to classify trajectories from WiFi data by travel mode involving features specific to this type of data such as signal strengths. These methods can unfortunately not be applied to cellular network data as the used features are not available in a comparable way and as the resolution in space and time is much lower compared to GPS data.

Bachir et al. (2019b) have introduced a first semi-supervised approach for classifying trips extracted from cellular network data. They propose a method that associates mode probabilities with each cell in the cellular network using a semi-supervised approach involving road and rail infrastructure geodata. The probabilities are combined with prior mode probabilities for the different home areas from a household survey using Bayes' theorem to classify individual trips by travel mode. While the aggregated validation of the mode-specific OD-flows is promising, the study does not provide validation of the modes estimated for individual trips. Further, updated household survey data on modal share may not always be available and so far no semi-supervised method only using cellular network data and infrastructure data has been proposed in the literature.

Geometric Mode Classification

This section presents a geometric mode classification method, which can also be used to generate pseudo-labelled trips for semi-supervised classification. The method classifies the mode by comparing a trip T to route alternatives found for each mode.

Route Set Generation

The aim of the route set generation is to find a set of route alternatives $\boldsymbol{R}(T, m)$ for each mode $m \in \{road, rail\}$ that could have been used for trip T. A route $R \in \boldsymbol{R}(T, m)$ is described by its geometry based on the road or rail infrastructure used and its travel time $t(R)$.

To generate the set of rail routes $\boldsymbol{R}(T, rail)$, we use the OpenTripPlanner routing engine[1] using the positions of the first and last cell in T as the origin, respectively, destination. We include all route alternatives within a departure window of from 15 min before to 120 min after the trip's estimated start. This is to account for inaccuracies in the estimated trip stat time as well as in the estimated time for accessing the first bus stop or train station. The routes may include access modes (including bus and car), but at least half of the route's distance must use a train line. The rail route set may be empty or contain up to three route alternatives depending on how many routes OpenTripPlanner could find that fulfil the criteria above.

For the road route set $\boldsymbol{R}(T, road)$, we use a road network of major roads from OpenStreetmap[2] with the free-flow travel time c_l based on the speed limit as the link cost for a link l. For each trip, the road route set consists of two routes. The first of these routes is the fastest route between the origin and destination cell of the trip. As users might not always follow the fastest route alternative, we also generate a second route alternative called the *magnetic route* similar to the method described in Breyer et al. (2018). This route is obtained using a shortest path calculation on the same network but with modified link costs. For a cellpath T, the modified link cost for link l is calculated as $c_m(l, T) = 0.8^{n(l,T)} \cdot c_l$, where $n(l, T)$ is the number of cells in the cellpath T overlapping with link l. The effect of using these modified costs is that the obtained route is likely to follow the cellpath closely. Figure 1 shows an example where the magnetic route follows the cellpath closer while having a significantly longer travel time than the fastest route.

Classification by Proximity to Routes

Using the route sets for a trip T introduced in the previous section, we can classify the travel mode by proximity to routes. The method is based on two assumptions:

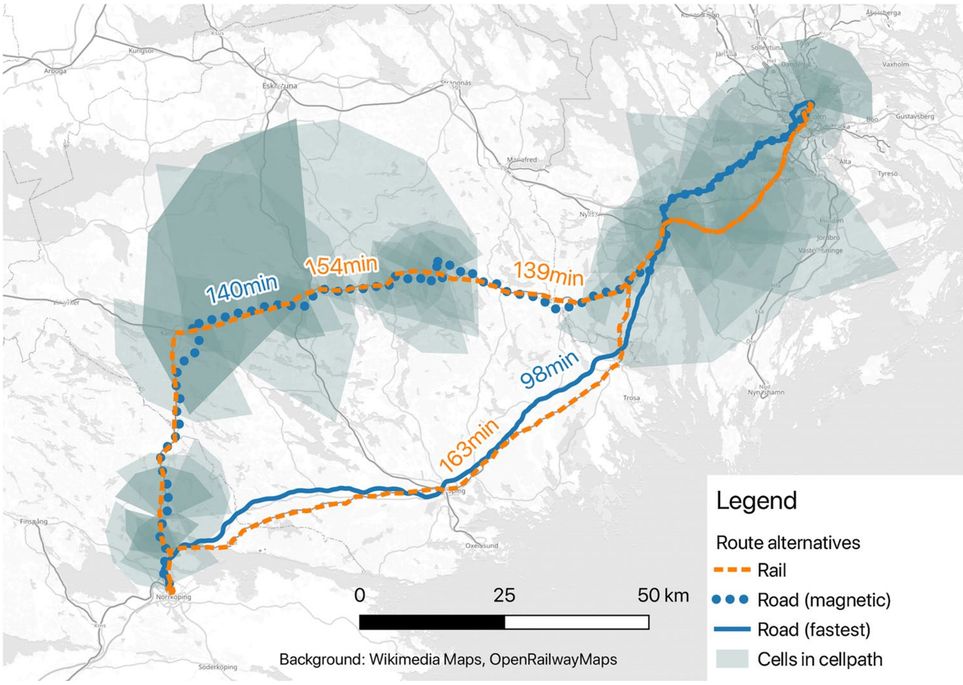
- The likelihood $L(T|R)$ of observing cellpath T when using a given route R decays with the distance between the cellpath and the route.
- The likelihood $L(R)$ that a user uses the route alternative R when travelling with a certain mode decays with the travel time $t(R)$ compared to the travel time of the fastest route alternative of the same mode.

The first assumption is motivated by the cellular network's characteristics, where the observation of a cell whose estimated coverage overlaps with the route is very likely. In contrast, the observation of a cell far away is very unlikely. The second assumption is a behavioural assumption also used in mode choice models (Paulssen et al. 2014). Without this assumption, there would be a risk that a magnetic route using slow minor roads close to the railway would be selected even when the trip was made by rail.

[1] http://www.opentripplanner.org.

[2] http://www.openstreetmap.org.

Fig. 1 Example of route alternatives for a trip from Norrköping to Stockholm with travel time. Note that two of the three rail routes use the same railway line but with different travel times due to different itineraries (different departures and transfer times). The shaded areas show the coverage areas of all cells used during the trip



Background: Wikimedia Maps, OpenRailwayMaps

Algorithm 1: Geometric classification

```
Data: Cellpath $T$, route sets $\boldsymbol{R}(T,m)$ for $m \in \{road, rail\}$
Result: Mode probabilities $p(m|T)$ for $m \in \{road, rail\}$
// First stage: Find best route for each mode
for $m \in \{road, rail\}$ do
    $t_s := \min_{R \in \boldsymbol{R}(T,m)} \mathrm{t(R)}$ ;                // Fastest travel time
    for $R \in \boldsymbol{R}(T,m)$ do
        $L(R) := e^{-\alpha\cdot(\frac{t(R)}{t_s}-1)}$ ;                    // Prior
        $L(T|R) := e^{-\beta \sum_{c \in T} d(c,R)}$ ;          // Observation likelihood
        $L(T \cap R) := L(T|R) \cdot L(R)$ ;                  // Joint likelihood
    end
    $R_m := \arg\max_{R \in \boldsymbol{R}(T,m)} L(T \cap R)$ ;         // Best route for mode
    $L(m|T) := L(T \cap R_m)$ ;                           // Mode likelihood
end
// Second stage: Calculate mode probabilities
for $m \in \{road, rail\}$ do
    $p(m|T) := \frac{L(m|T)}{\sum_{m' \in \{road,rail\}} L(m'|T)}$ ;        // Mode probabilities
end
```

The classification is done in two stages (see Algorithm 1). If the route alternatives of all modes were compared together, a mode with many (even similar) route alternatives would have an unjustified advantage. To avoid this, we only keep the best route for each mode in the algorithm's first stage. For each route alternative, we estimate the posterior likelihood that this route was used by the joint likelihood $L(T \cap R) := L(T|R) \cdot L(R)$, where $L(R) := e^{-\alpha\cdot(\frac{t(R)}{t_s}-1)}$ is the route prior based on the travel time relative to the fastest route for the mode and

Table 1 The values used in the geometric mode classification for the trip shown in Figure 1. In Stage 1 of Algorithm 1, the routes marked with * are selected as the best route for each mode

Route	m	$t(R)$	$\sum_{c\in T} d(c,R)$	$L(R)$	$L(T\|R)$	$L(T\cap R)$
1	Road	98 min	582.2 km	1.0000	$1.18\cdot 10^{-38}$	$1.18\cdot 10^{-38}$
2*	Road	140 min	38.2 km	0.1159	0.0033	$3.77\cdot 10^{-4}$
3*	Rail	139 min	12.7 km	1.0000	0.8269	0.8269
4	Rail	154 min	12.7 km	0.5821	0.8269	0.4813
5	Rail	163 min	600.0 km	0.4152	$7.68\cdot 10^{-40}$	$3,19\cdot 10^{-40}$

$L(T|R) := \prod_{c\in T} e^{-\beta\cdot d(c,R)} = e^{-\beta \sum_{c\in T} d(c,R)}$ is the likelihood to observe the cellpath T given the route with $d(c,R)$ being the Euclidean distance between cell c and the closest point on route R. Note that $d(c,R)=0$ if the route is passing through the cell (described by its estimated coverage area) while otherwise the distance is measured from the cell's boundary. We assume that the observation of the cells are independent events and multiply the likelihood of all observed cells. We use $\alpha = 5$ and $\beta = 0.15$, which give a reasonable decay of the likelihoods with increasing distance. As the cell coverage areas should, in most cases, cover the used route, the likelihood of observing a cell c should decrease fairly quickly when $d(c,R) > 0$, while still allowing smaller error as a tolerance due to the coverage areas being an estimate. In the second stage of Algorithm 1, the mode probabilities $p(m|T)$ are calculated by comparing the best route for each mode.

The calculation according to Algorithm 1 is illustrated in Table 1 for the trip in Fig. 1. The best route for the road mode selected in Stage 1 is the magnetic route 1, as it is much closer to the cellpath T despite its longer travel time compared to the fastest road route (route 1). There are two routes with the same geometry for the rail mode that are equally close to T. Due to the prior $L(R)$, the faster route 3 is selected as the best route for the rail mode. In Stage 2 the best routes for each modes are used to calculate the mode probabilities $p(rail|T) = 0.9995$ and $p(road|T) = 0.0005$.

Semi-supervised Mode Classification

Geometric classification can misclassify trips when the route alternatives are geometrically very close, the used cells have large coverage areas, or few location updates have been made during the trip. It may also fail if the route set generated did not include the used route. The semi-supervised mode classification methods presented in this section aim to work better in these situations. Semi-supervised classification classifies T not only using the information for that particular trip but also patterns from unlabelled trips $\boldsymbol{T}_{O,D}$ in the same OD-pair that are associated with the different travel modes. We present three classification methods using semi-supervised labelling, which are based on different semi-supervised learning assumptions.

Semi-supervised Learning Assumptions

The goal of semi-supervised classification is to improve classification using a large set of unlabelled observations, in our case unlabelled trips (Zhu and Goldberg 2009). This is based on the premise that the unlabelled trips contain patterns which can be related to different travel modes. van Engelen and Hoos (2019) describe three semi-supervised learning assumptions that are used in semi-supervised classification methods. For the mode classification problem we can formulate them as follows:

I. *Smoothness*: If two trips have similar cellpaths they are likely to have the same travel mode.
II. *Manifold assumption*: Even though cellpaths are high-dimensional, they lie on lower-dimensional structures in the feature space (manifolds). Cellpaths on the same manifold usually share the same travel mode.
III. *Low-density*: If there are many trips with similar cellpaths, they likely have the same travel mode. It follows that the decision boundary between travel modes should have few cellpaths close to it.

In addition to a large number of unlabelled trips that fulfil one of the above assumptions, semi-supervised classification also requires a small set of labelled trips. As labelled data are typically often not available in the case of cellular network data. Therefore, the semi-supervised methods presented in Sect. 4.2 make use of additional assumptions specific to cellular network data to obtain pseudo-labelled data:

IV. *Geometric likelihood*: Some trips can be classified likely using geometric classification and can thus be used as a set of pseudo-labelled trips.
V. *Continuity*: A labeled trip passing through cities A–B–C can be used as a training trip for OD-pair A–B.

Assumption IV may not always hold, in particular, in challenging OD-pairs where the routes for different modes are spatially very close and thus no or few trips can be classified with high likelihood. In these cases, using Assumption V can be considered to obtain pseudo-labelled trips. It enables

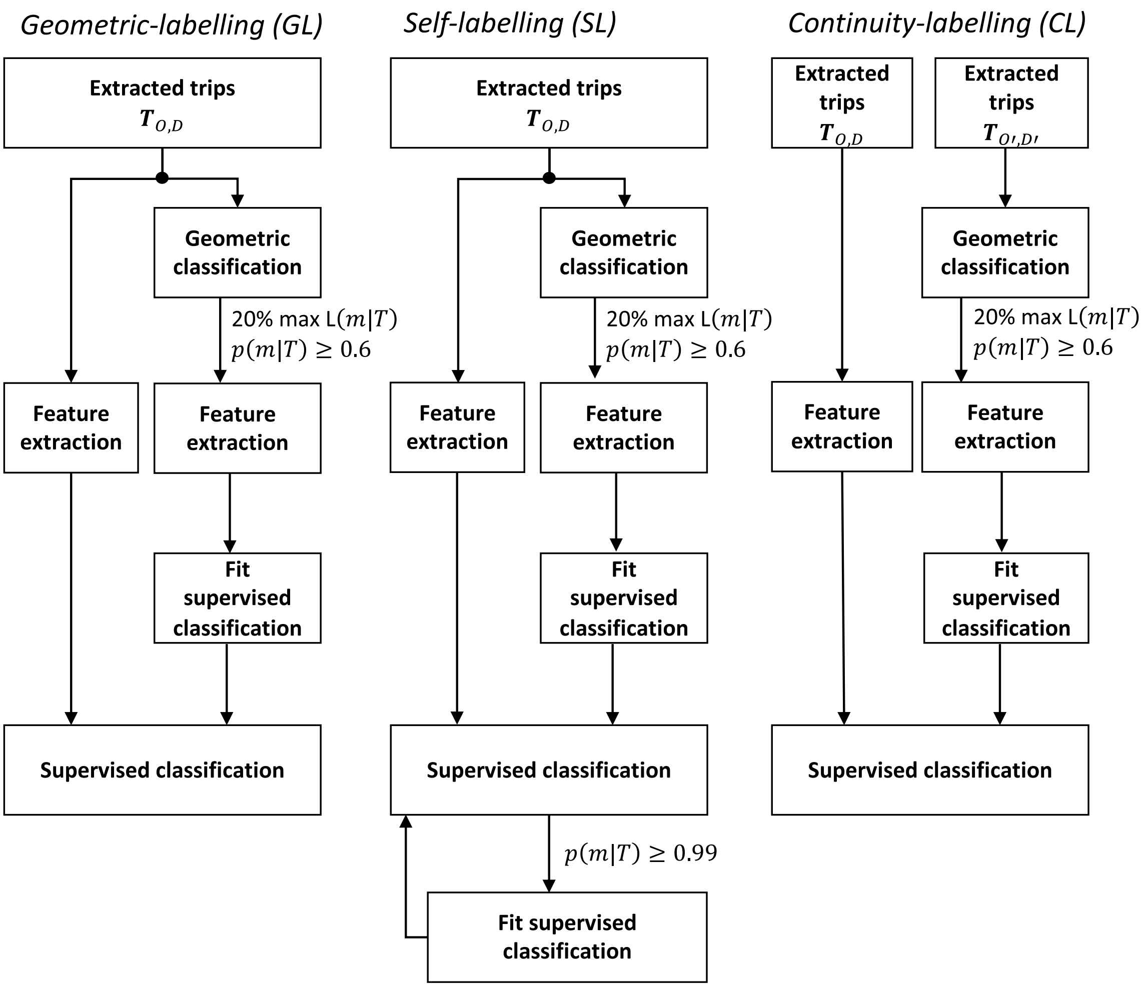


Fig. 2 Overview of the proposed mode classification methods using semi-supervised labelling

learning from trips in a different overlapping OD-pair that is easier to classify by transferring information from another OD-pair.

Semi-supervised Labelling

We present three semi-supervised mode classification approaches of inter-city trips extracted from cellular network data (see Fig. 2) using different semi-supervised labelling techniques. Following the taxonomy introduced by van Engelen and Hoos (2019) all three methods are inductive wrapper methods as they use an existing supervised method as part of the classification process. The methods are also using feature extraction as a method of unsupervised preprocessing.

Geometric-labelling (GL) is using geometric classification (see Sect. 3) to obtain a small set of pseudo-labelled trips out of the set of unlabelled trips $\boldsymbol{T}_{O,D}$. Using Assumption IV, only trips classified likely are pseudo-labelled in this step (see Sect. 4.3). Feature extraction is then used to describe each trip using a small number of features instead of the complete high-dimensional cellpath (see Sect. 4.4) based on Assumption II. A supervised classification method (see Sect. 4.5) is trained exploiting Assumption I using the pseudo-labelled trips transformed to the lower-dimensional feature space. The trained supervised method finally classifies each trip by travel mode.

Self-labelling (SL) starts with the same steps as *Geometric-labelling* and is, thus, also using Assumptions I, II and IV. Additionally, it is also using Assumption III as follows. After the supervised classification method's initial training using the pseudo-labelled trips from geometric classification, the method is trained iteratively again using a new training set. This training set consists of the trips classified with a high probability $p(m|T)$ in the previous iteration. This semi-supervised technique is known as self-labelling

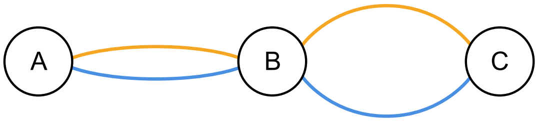


Fig. 3 Schematic example where the classification of trips between cities A and B can be improved using trips between A and C as these can be classified more likely using geometric classification as rail (orange) or road (blue)

(Triguero et al. 2015). Self-labelling is related to Assumption III as it has the effect of moving the decision boundaries towards low-density areas while more trips are labelled. We are using three training iterations for this method with the following trips considered as labelled:

1. pseudo-labelled trips from geometric classification
2. all trips classified with $p(m|T) \geq 0.99$ in iteration 1
3. all trips classified with $p(m|T) \geq 0.99$ in iteration 1 or 2

The previous two methods focus on classifying unlabelled trips in the same OD-pair, which can still be problematic when the routes of several modes are very close, making the initial geometric mode classification unreliable. We propose a third labelling method called *continuity-labelling (CL)* which is using Assumption V (in addition to Assumptions I, II and IV) to transfer information from another overlapping OD-pair which is easier to classify to an OD-pair that is hard to classify. In the schematic example in Fig. 3, pseudo-labelling using geometric classification between A and B is problematic as the route alternatives are close. Instead of only using trips between A and B, however, we can use trips from longer overlapping OD-pairs such as between A and C for training the mode classification between A and B. Thus, we can benefit from the section between B and C, which is easier to classify geometrically due to the clear separation of routes. We illustrate the idea using the simple case of training the classification method for an OD-pair from O to D using trips from another manually selected OD-pair from O' to D'. However, it would be possible to use trips from many different OD-pairs that share parts of their routes with the routes in the OD-pair to classify. Apart from the difference that the geometric pseudo-labelled trips are taken from another OD-pair than the one to classify, continuity-labelling works in the same way as geometric-labelling (see Figure 2). Note that in the feature extraction (see Sect. 4.4), only those antennas common in the OD-pair to classify are considered.

Pseudo-labelling Using Geometric Classification

Semi-supervised classification requires, besides large amounts of unlabelled data, a small set of labelled data. Manually labelled trips are not realistic to obtain for each OD-pair in practice. Therefore, the classification methods presented in Sect. 4.2 are all starting by obtaining a number of pseudo-labelled trips using the geometric classification (see Sect. 3.2) of the set of unlabelled trips $\boldsymbol{T}_{O,D}$. To obtain an adequate set of labelled training trips, we select *likely* classified trips as follows:

1. From $\boldsymbol{T}_{O,D}$ keep trips where a route is found for both modes.
2. From the remaining trips keep those where the mode likelihood $L(m|T)$ is among the top $\lambda = 20\%$ (upper quintile) among the trips with the same predicted mode.
3. From the remaining trips keep only those with mode probability $p(m|T) \geq \tau = 0.6$.

Without (1), there is a risk that trips are included where the route set generation failed to find a route for one of the modes, which then would be prioritised as the other mode will have $p(m|T) = 1.0$. Then, (2) makes sure that only the trips with the best matches with the found route and thus high possibility that this mode was used are kept. Using an absolute threshold for $L(m|T)$ could be considered but might not lead to enough training trips in any OD-pair. Finally, (3) excludes trips that are very close to the decision boundary, causing the classification to be uncertain. The thresholds λ and τ have been set to generate good training data sets. The parameters and selection criteria might need to be adjusted to work for many different OD-pairs.

The reason to not only use trips that can be classified certainly (for example $\tau = 0.6$) is that this systematically excludes routes that are harder to classify as there is a route of another mode that is similar. However, the semi-supervised methods can only work reliably if trips on all manifolds (corresponding to route alternatives) are included in the set of pseudo-labelled trips. Even if the route used was not in the route set, $p(m|T)$ can still be high. This is the case when $L(rail|T)$ and $L(road|T)$ are both low but still different enough for $p(m|T)$ to be high for one of the modes. Using step (2), these trips are excluded from being pseudo-labelled.

Feature Extraction

Each trip needs to be represented by a feature vector to run a supervised classification method. To represent each trip by its cells in the cellpath, we first convert the set of trips $\boldsymbol{T}_{O,D}$ into the binary $n \times m$ matrix X where n is the number of trips in $\boldsymbol{T}_{O,D}$ and m the number of unique cells. We only include cells used by at least 2% of the trips to exclude cells that are not associated with any usual route in the OD-pair. Row i of X corresponds to the feature vector for trip $T_i \in \boldsymbol{T}_{O,D}$ and we define

$$X_{i,j} = \begin{cases} 1 & \text{if } j \in T_i \\ 0 & \text{if } j \notin T_i. \end{cases} \quad (2)$$

The number of unique cells m may be in order of thousands depending on the OD-pair, which is problematic for supervised learning. Training supervised classification with many features requires a large amount of training data. Otherwise, the model will suffer from over-fitting (or even be underdetermined if $n < m$). One approach to reduce dimensionality is feature selection, which is about choosing only a subset of features to use in the model. Here, this would mean selecting the cells that distinguish best between modes. However, even for trips using the same route, the exact cells used may vary considerably, considering that they depend on the interaction between the cellular network and the particular mobile device. Selecting a subset of cells would mean that all information would be lost for trips that do not use any of the cells in the selected subset of cells.

Instead of feature selection, we use feature extraction, which can be seen as a method of unsupervised preprocessing. Methods of feature extraction, which use information from a dataset of unlabelled observations, are one of the tools used in semi-supervised learning (van Engelen and Hoos 2019). The goal of the feature extraction method is to transform the m-dimensional feature vector of all trips to a k-dimensional feature space, that is $X_i \in \{0,1\}^m \mapsto Y_i \in \mathbb{R}^k$, $k < m$. A feature extraction method reduces dimensionality yet preserving significant information about the original observations. This allows revealing lower-dimensional manifolds on which the high-dimensional feature vectors lie based on Assumption II. The feature extraction method should also maintain smoothness (Assumption I), that is, for any $p, q \in 0 \dots n$ if $X_p \approx X_q$, then also $Y_p \approx Y_q$. Methods for feature extraction are embeddings, autoencoders and decomposition methods (van Engelen and Hoos 2019; Vincent et al. 2010). Some methods, for example, the embedding method t-SNE (Maaten and Hinton 2008), transform a given set of observations to the lower-dimensional feature space directly. Other methods, such as Principal Component Analysis (PCA) as a decomposition method, generate a general transformation function. A transformation function can be used later-on transform new unseen observations to the new feature space. A disadvantage of PCA is that it cannot handle large sparse matrices efficiently.

We are using Truncated Singular Value Decomposition (SVD) to extract lower-dimensional features for each trip (Manning et al. 2009, Chapter 18) which can handle sparse matrices. SVD provides a matrix decomposition $X = U\Sigma V^T$, where U is called the left-singular vectors, Σ a matrix that contains the singular values on its diagonal and V the right-singular vectors. In Truncated SVD, we use V_k, the $k \times n$ matrix which contains the first k rows of V. We call V_k the SVD loadings, which if X would be centred, are equal to the first k PCA loadings (Wall et al. 2003). Similar to PCA, we obtain a lower-dimensional matrix $Y = XV_k$ using these loadings. For a given observation X_i, we call the linearly transformed features $Y_i = X_i V_k$ its SVD components. Using the first k values of V as loadings, the low-dimensional representation still contains large parts of the information in the original feature vector. By running the SVD decomposition once for each OD-pair and direction using the full set of unlabelled trips $\boldsymbol{T}_{O,D}$, the decomposition can make use of the manifold assumption (Assumption II). As Truncated SVD tries to approximate the original matrix, trips using similar cells will still be close in the new feature space, and the transformation thus maintains smoothness (Assumption I).

Supervised Classification

All methods in Sect. 4.2 are performing the final classification of all trips using supervised classification. In principle, it is possible to use any supervised classification method for this step. We are using Linear Discriminant Analysis (LDA), which is a standard method for supervised classification (James et al. 2013, Chapter 4). It uses the training trips to obtain linear decision boundaries in the feature space. After training, LDA can be used for classification and estimates mode probabilities $p(m|T)$ for each mode for a given trip T. The LDA model is using the first k features extracted using truncated SVD (see Sect. 4.4). The concept of using dimension reduction of cellpaths using PCA followed by LDA classification was earlier described as a supervised classification method in Breyer et al. (2021). The more features are used, the more detail of the cellpaths is kept. Using too many features, on the other hand, increases the risk of overfitting. The LDA model is trained using the pseudo-labelled trips (in all three labelling methods). In later iterations of self-labelling, training continues using the trips classified with high mode probability $p(m|T)$ in the previous iteration, which moves the decision boundary towards low-density regions based on Assumption III.

Dataset

We have used the semi-supervised classification methods presented in Sect. 4.2 to classify trips in two inter-city OD-pairs in Sweden. In the first OD-pair between Norrköping and Stockholm, the main rail route and the main road route are spatially well separated. However, the OD-pair is challenging to classify by mode as there also is another regional rail route that is very close to the main road route. The second OD-pair is between Norrköping and Linköping, which is shorter and has the main road and rail route separated by at most four kilometres. The cellpaths in this OD-pair contain

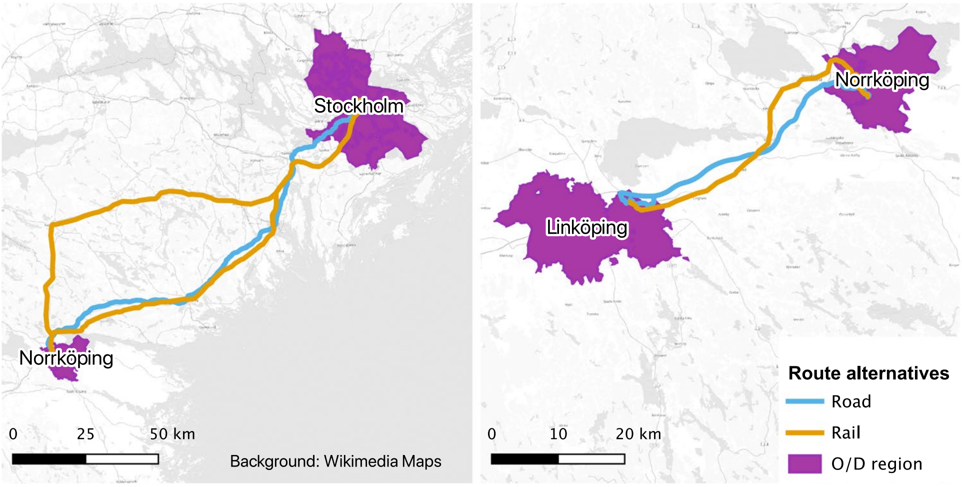


Fig. 4 The two inter-city OD-pairs and their most typical road and rail routes

thus fewer cells and many of these cells cover both the rail and road route.

The raw cellular network data cover three days during fall 2018. The data contains billing data and location updates extracted from the core network following the terminology used by Gundlegård (2018) and includes periodic, location area (LA), routing area (RA), tracking area (TA), and cell updates. We processed the data using a remote-access setup as suggested by de Montjoye et al. (2018), where the code is brought to the data, and only the final results are exported. After an initial data cleaning to remove fast ping-pong events (back and forth between antennas), we extracted trips using a stop-based trip extraction as described in Breyer et al. (2020). A stop is detected when a user stayed inside a two-kilometre radius for at least one hour, and all events between two stops are considered a trip. We have defined origin and destination zones, as shown in Fig. 4. The trips starting and ending in those zones are included in the set of trips in the OD-pair $\boldsymbol{T}_{O,D}$. Table 2 gives the number of trips extracted in each OD-pair after removing a small number of trips (2.8% for Norrköping–Stockholm and 2.7% for Norrköping–Linköping) without any updates between origin and destination. An additional set of 3329 unlabelled trips in the OD-pair between Linköping and Stockholm (via Norrköping) is used to obtain pseudo-labelled trips to train the continuity-labelled classification between Norrköping and Linköping.

In addition to the unlabelled trips, the authors also collected a smaller set of their own trips during 2019 in the two OD-pairs and labelled each trip manually with the actual travel mode used for the trip. This set of labelled trips is only used for validation, that is, to evaluate if the semi-supervised methods trained on the unlabelled data classify the trips in the small set of labelled trips correctly. The labelled set is not representative of the whole population and limited in its variety, but a high test error on those trips indicates that a classification method is not working correctly.

Results

We have used geometric classification (G) (see Sect. 3.2) and the three semi-supervised labelling methods geometric-labelling (GL), self-labelling in two iterations (SL1, SL2) and continuity-labelling (CL) (see Sect. 4.2) to classify the unlabelled dataset described in Sect. 5. We apply training and classification once for all trips from O to D and once for all trips from D to O direction. However, we present the results aggregated for each OD-pair, including both directions. The following sections present the results for feature extraction and classification. Finally, we have used the trained classification methods also to classify the small set of labelled trips (see Sect. 5) to validate the methods.

Feature Extraction

There are in total 204 (Norrköping–Linköping), respectively, 558 (Norrköping–Stockholm) unique cells used by trips in the OD-pairs. All three semi-supervised labelling methods use a representation of the trips by $k = 3$ extracted features instead of the full cellpath for the supervised classification (see Sect. 4.4). While three features are not enough to

Table 2 Number of trips in total for each OD-pair in the dataset both unlabelled and labelled (only used for validation)

OD-pair	Distance	Unlabeled	Labeled	
		Total	Rail	Road
Norrköping ⇔ Stockholm	160 km	2426	17	7
Norrköping ⇔ Linköping	45 km	7119	209	22

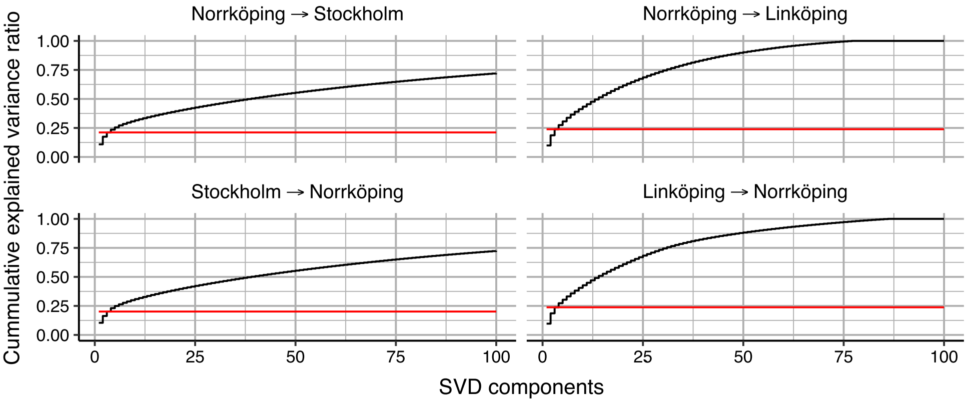


Fig. 5 Cumulative explained variance depending on the number of SVD components. The red line marks the variance explained by three components

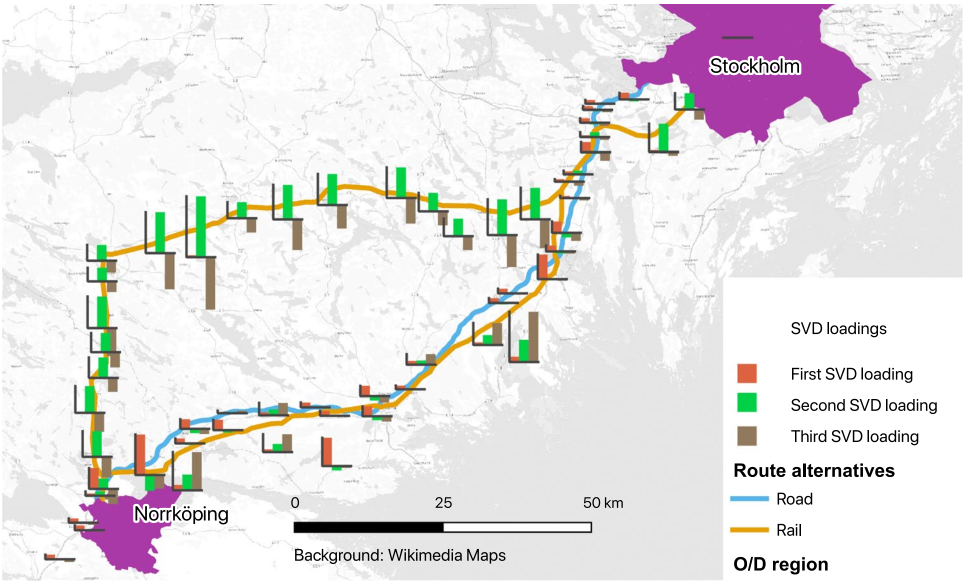


Fig. 6 The SVD loadings V_3 for cells used for trips from Stockholm to Norrköping. For increased visibility, only a sample non-overlapping cells are displayed

recover the exact cellpath information (see Fig. 5), they have shown to be enough to distinguish the different route alternatives as lower-dimensional manifolds. Using too many features has shown to lead to over-fitting and worse classification results. The SVD loadings V_3 (see Sect. 4.4) for each cell are illustrated in Fig. 6 for trips from Stockholm to Norrköping. Cells along the main road route are associated with high values for the first loading component and low values for both the second and third component. Trips on the upper rail line can be associated with high values for the second loading component and negative values for the third component. Trips on the lower rail line are associated with cells with high values for the first component and positive values for the second and third components.

Classification

Figure 7 shows all trips in the dataset by their second versus third SVD component. We show those components as they showed slightly clearer separation of clusters than, for example, the first versus the second component. For each method, Figure 7 shows the predicted mode of each trip. For the semi-supervised labelled methods also the training trips used are shown in the plot. For Norrköping–Stockholm, the trips form three clusters, which can be associated with the three route alternatives. The left cluster corresponds to road trips, the lower right cluster to the upper rail route and the top right cluster can be associated with the lower rail route (compare Fig. 6). For Norrköping–Linköping only two less clearly separated clusters are visible as expected as there is only one reasonable rail and road route, which also are close in space. Hence, Fig. 7 illustrates the use of the manifold assumption (Assumption II) when extracting lower-dimensional features.

Comparing the predicted modes, we find that geometric classification seems to classify trips in the top right cluster almost randomly (see Fig. 7). This might be because the lower rail route and the road route are pretty close in space. Geometric-labelling, in contrast, uses the likely classified trips from geometric classification as training trips and

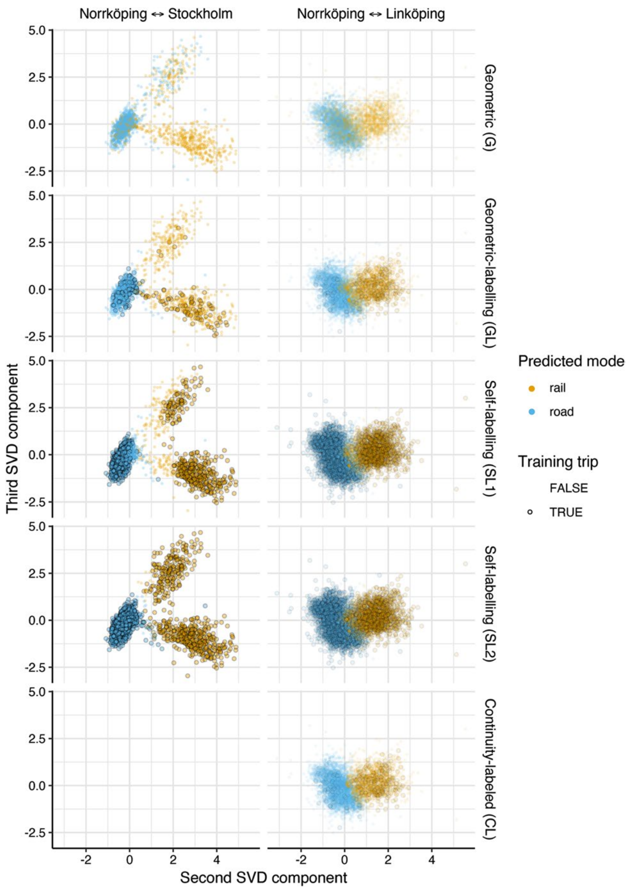


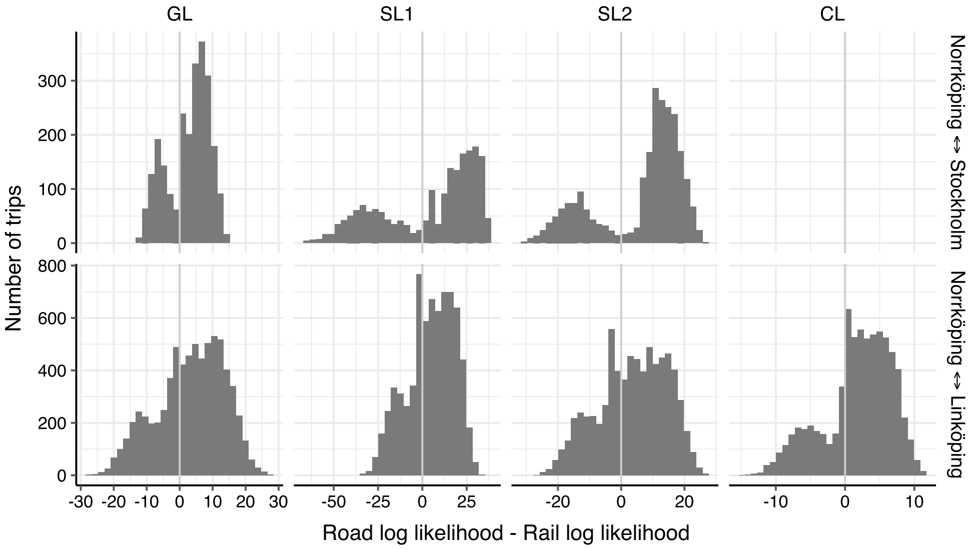
Fig. 7 Predicted mode for trips in the two OD-pairs and the training trips used represented by their *second and third* SVD components. Trips used as pseudo-labelled training trips are marked with a black circle. Note that in the case of continuity-labelling these trips come from the overlapping OD-pair used for training that is Linköping–Stockholm

classifies all trips in the cluster as made by rail. Self-labelling continues then using the trips classified with high probability $p(m|T)$ as training trips. After the second iteration, almost all trips are added to the training set. For continuity-labelling, the training trips used to classify trips between Norrköping and Linköping are instead obtained using trips from the geometric classification of trips between Stockholm and Linköping (which are passing through Norrköping).

The difference in likelihood for the two modes indicates how certain the classification is. For Norrköping–Stockholm, we find a relatively clear separation between modes (see Fig. 8). With each iteration of self-labelling, the separation becomes even clearer. This illustrates how self-labelling uses the low-density assumption (Assumption III). For Norrköping–Linköping, we have a much more unclear separation between the travel modes, which indicates that the classification is more uncertain and the OD-pair more difficult to classify. However, by making use of Assumption V, continuity-labelling seems to be able to separate better between the travel modes than the other methods.

Summarising the classification of all trips, we can estimate the modal split in the OD-pairs (see Fig. 9). Additionally, Fig. 10 shows the share of trips that have been classified with the same mode when comparing a pair of methods. For both OD-pairs, the agreement is lowest between geometric classification and all other methods. For Norrköping–Linköping up to 25% of the trips are classified differently using self-labelling compared to geometric classification. We find only minimal differences between geometric-labelling and self-labelling. For Norrköping–Linköping, classification using continuity-labelling leads to a higher road share. Even though continuity-labelling showed better separation between travel modes (see Fig. 8), this does,

Fig. 8 Histogram of the difference in likelihood $L(road|T) - L(rail|T)$ as estimated by the different methods for all unlabelled trips



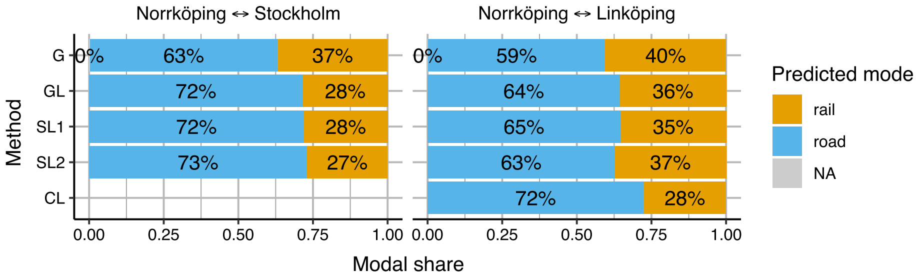


Fig. 9 Modal split resulting from classifying all trips in the set of unlabelled trips using different classification methods

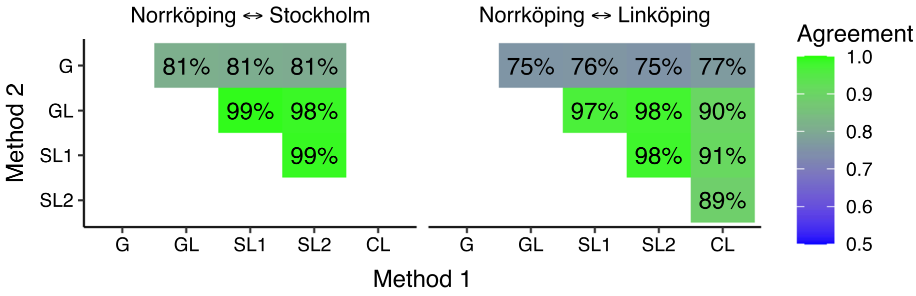


Fig. 10 Share of trips in the set of unlabelled trips classified with the same mode for each pair of methods (agreement)

however, not necessarily mean that the classification is more accurate.

Validation

The classification above was done for unlabelled trips, and therefore, it is not possible to draw definitive conclusions about the quality of the mode classification. To validate the trained methods, we use a smaller dataset of labelled trips (see Sect. 5). We classify the labelled trips using the trained models (that is, the SVD loadings and the LDA decision boundaries) obtained in Sect. 6.2. The labelled trips have not been seen by the models during training and are thus an independent test set. Geometric classification failed to classify 5.6% of the labelled trips Norrköping–Linköping correctly, while the semi-supervised methods classified all trips correctly except for the second iteration of self-labelling which misclassified one trip (0.4%). This shows that the classification improved using semi-supervised learning methods compared to geometric classification. For Norrköping–Stockholm, both geometric classification and self-labelling in iteration 2 failed to classify the only trip on the lower rail route correctly. The fact that self-labelling decreases in performance in the second iteration could indicate a problem of over-fitting when too many trips are added as training trips. Continuity-labelling classified all labelled trips correctly between Linköping and Norrköping, which shows that it is possible to use training data from a different OD-pair using Assumption V. Comparing continuity-labelling and geometric-labelling, we find that the two methods predict different modes for 10% of the unlabelled trips between Norrköping and Linköping (see Fig.10) even though both had no test error for the labelled trips. The labelled trips were, hence, too few to determine which of the two methods performs better. The better separation between travel modes (see Fig. 8) and the fact that the modal split is closer to a travel survey from May 2014, which found a modal split of 24% for rail in the OD-pair (Region Östergötland 2014) indicate that, in fact, the continuity-labelling could be more accurate. It is also reasonable to assume that the vast majority of users are using the same mode when travelling back and forth on the same day. When users made two trips on the same day in the OD-pair (travel back and forth), geometric-labelling predicted the same mode for the two trips in 70% of cases. Continuity-labelling predicted the same mode for the two trips in 87% of cases, which might also indicate that the classification using continuity-labelling was more accurate than geometric-labelling.

All trips misclassified by the semi-supervised methods are very close to the decision boundary (see Fig. 11), indicating that mode probabilities $p(m|T)$ returned by the LDA are a good indicator of the certainty of the classification. Using self-labelling for Norrköping–Linköping, we see that the first iteration increase separation between travel modes compared to geometric-labelling. However, the second iteration introduces an error probably as it over-fits to the large number of training trips used in that iteration (see Fig. 7), which likely also contains some misclassified trips. Similarly to the unlabelled trips (see Fig. 8), also for the labelled trips, none of the methods that we tested can find a very clear separation for the more challenging OD-pair between Norrköping and Linköping.

Discussion

Semi-supervised learning can be used to improve classification when an unlabelled dataset fulfils one or more of the semi-supervised learning assumptions (see Sect. 4.1). The results in Sect. 6 show that the mode classification of inter-city trips extracted from cellular network data can be improved using semi-supervised classification compared to geometric classification.

Geometric classification may fail to classify trips correctly, particularly when the routes for both modes are close to or inside the coverage area of all used cells of a trip such that the distance to the route alternatives does not provide clear evidence for one of the modes. The reason why the semi-supervised labelling methods perform better than geometric classification can be summarised as follows: First, some likely geometric classified trips are selected as pseudo-labelled training trips (Assumption IV). The feature extraction represents each trip such that the lower-dimensional manifolds that correspond to the route alternatives, which are embedded in the cellpath information, are revealed (Assumption II). The remaining trips in the same cluster are classified using the training trips and a supervised classification method using Assumption I. In short, the methods learn which patterns in the cellpath are associated with a particular travel mode.

In the two tested OD-pairs, Assumptions I and II seem to hold well and can be used to improve the classification. Truncated SVD provided a good representation of the cellpath using few features allowing to classify trips as rail and road. For even shorter trips and more detailed modes other methods to extract features such as autoencoders (Vincent et al. 2010) could be investigated that might even better represent the cellpath using few features. When using Assumption IV, we found a trade-off between only including trips as pseudo-labelled where the geometric classification is accurate, but on the other hand, including representatives of all manifolds (routes alternatives) in the training data. We try to achieve that using thresholds for both $L(m|T)$ and $p(m|T)$. However, more OD-pairs need to be tested to find general criteria to select the trips to pseudo-label.

Self-labelling also makes use of Assumption III, which lead to better separation between travel modes in some cases (see Fig. 8). However, adding too many trips as pseudo-labelled can be counterproductive and make the classification worse (see Table 3). Improving the criteria for trips to be pseudo-labelled in each iteration or limiting the number of trips pseudo-labelled in each iteration might help. Instead of a fixed number of iterations, a stop criterion could be used.

With continuity-labelling, we demonstrate that information from another overlapping OD-pair can be used for training. If a section of the other OD-pair is easier to classify geometrically, this may allow discovering cell patterns that could not be discovered otherwise and thus improve the classification. However, this is only true if the continuity assumption V is fulfilled in the particular case. That means that the trips of the same main mode in both OD-pairs share parts of the routes. If there is, for example, separate rail infrastructure for long-distance and short-distance trains, using continuity-labelling may instead worsen the classification. We demonstrated continuity-labelling classification using another manually selected OD-pair for training. However, this could be done automatically by selecting trips for training from all OD-pairs that have many cells in common with the trips in the OD-pair to classify. Further improvement could be made by combining continuity-labelling with self-labelling.

Table 3 Test error on the small set of labelled trips using different classification methods

OD-pair	Test error				
	G	GL	SL1	SL2	CL
Norrköping ⇔ Stockholm	0.042	0	0	0.042	
Norrköping ⇔ Linköping	0.056	0	0	0.004	0

Additional improvement of the classification performance could likely be made by making use of the order and time information of the location updates in addition to which cells have been used. Instead of using LDA as the supervised classification method, for example, a Markov chain could be used to describe how users typically switch between cells given training trips of a specific mode. In this paper, we focused on classifying the main mode of each trip. More attention could be devoted to separately classify the main mode and possible access modes used during a trip. In connection to this, it would be interesting to investigate how other modes than rail and road can be classified, such as air and water-bound traffic.

In this paper, we have mainly used two semi-supervised concepts: feature extraction and self-labelling. However, these are not the only concepts that could be used to classify trips by travel mode. A cluster-then-label approach could, for example, be used to group similar trips into clusters that could then be labelled using pseudo-labelled trips. Another cluster-then-label approach could be to identify clusters of antennas (instead of trips) associated with a specific mode as proposed by Bachir et al. (2019a). Another relevant approach could be using graph-based semi-supervised classification by building a graph based on the similarity of trips (van Engelen and Hoos 2019).

The mode classification problem is closely related to the route inference problem aiming to identify the most likely route in the transportation network used for a given trip.

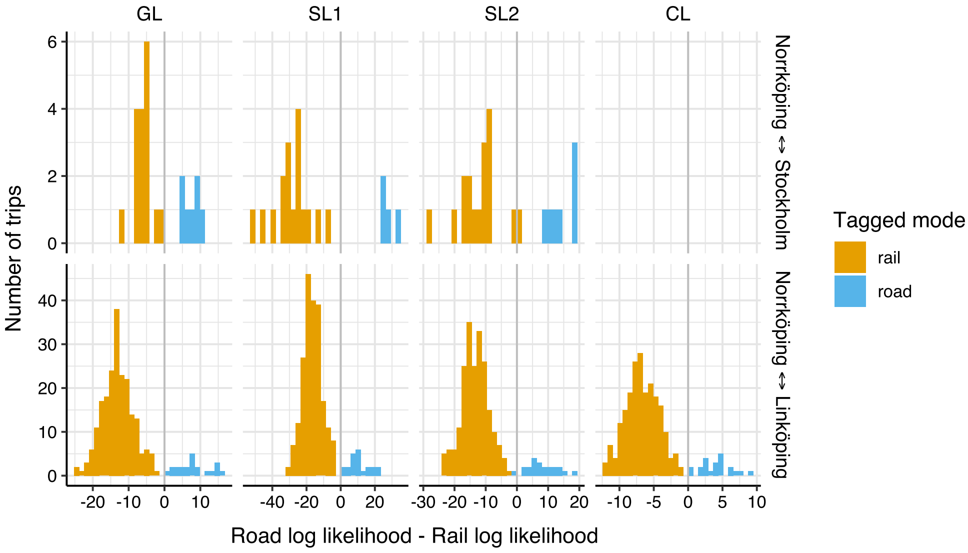


Fig. 11 Histogram of the difference in likelihood $L(road|T) - L(rail|T)$ for the manually tagged validation trips

For one travel mode, there may be several route alternatives (as in the example in Fig. 1), and we found that it is crucial that the geometrically pseudo-labelled trips include representatives of all route alternatives to facilitate the supervised classification. Ensuring that all routes are represented can be difficult, and a better solution might be to consider the two problems jointly. For example, we might obtain route alternatives per OD-pair and classify each trip by the route alternative used, which then could be associated with a travel mode. Besides route and mode classification, semi-supervised learning methods have potential use for other related problems, such as the classification of trip purpose or activities.

The use of cellular network data based on a large sample of users allows to obtain a comprehensive overview of mode-specific travel patterns. It can also be updated much faster than traditional travel surveys. However, there are two main limitations: First, the classification may be unreliable for very short trips and travel modes which use similar or even the same infrastructure. For example, it may be very difficult to distinguish bus, tram and car trips in an urban context. Second, cellular network data does not include any socioeconomic attributes about the users. As suggested by Andersson et al. (2022), this problem could be handled by combining cellular network data and travel survey data.

Conclusions

In this paper, we showed how semi-supervised methods can be used for the classification of trips extracted from cellular network data by travel mode. We proposed three semi-supervised labelling methods based on multiple learning assumptions. The proposed classification methods require no labeled training data. This enables the practical use of the methods in the common situation where such labeled data is unavailable.

The results for the tested OD-pairs indicate that the proposed methods perform better than geometric classification using different semi-supervised learning assumptions. The limited amount of labelled validation data for the two tested OD-pairs was, however, not enough to conclude whether any of the three methods generally performs best. Continuity-labelling demonstrated the potential of propagating information between OD-pairs, which is a concept useful for applications beyond mode classification. A major challenge when using the proposed methods is to define the criteria to select the trips to include in the training set (pseudo-label). On the one hand, only trips where the label is likely to be correct should be included. The results for self-labelling showed that adding too many trips as pseudo-labelled can lead to worse performance due to over-fitting. On the one hand, all different route alternatives should be represented to achieve the best classification accuracy. The criteria that we propose needs to be tested on more OD-pairs to make sure that they can be applied generally.

The proposed classification methods allow classifying trips by travel mode even in challenging OD-pairs where pure geometric classification would often fail. This enables traffic planning applications requiring mode-specific travel patterns to benefit from the large amounts of observations from cellular network data being a faster and less expensive data source compared to travel surveys. In future work, it would be interesting to investigate how semi-supervised methods work for the the classification of even shorter trips and more detailed travel modes than rail and road. Semi-supervised methods also have large potential for solving related problems such as trip extraction, route inference and travel purpose classification of cellular network data.

Acknowledgements Thanks to Telenor and Lars Sköld for enabling the processing of the anonymised mobile network data that are used in this paper. This work was supported by the Swedish Agency for Innovation Systems (Vinnova) (grant number 2013-03077) and by the Swedish Transport Administration (project DEMOPAN).

Funding Open access funding provided by Linköping University.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp Res Part C Emerg Technol 58:240 – 250, https://doi.org/10.1016/j.trc.2015.02.018

Anda C, Erath A, Fourie PJ (2017) Transport modelling in the age of big data. Int J Urban Sci 21(sup1):19–42

Andersson A, Engelson L, Börjesson M, Daly A, Kristoffersson I (2022) Long-distance mode choice model estimation using mobile phone network data. J Choice Model. https://doi.org/10.1016/j.jocm.2021.100337

Bachir D, Khodabandelou G, Gauthier V, El Yacoubi M, Vachon E (2019a) Combining bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data. In: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11053. Springer, Cham. https://doi.org/10.1007/978-3-030-10997-4_35

Bachir D, Khodabandelou G, Gauthier V, El Yacoubi M, Puchinger J (2019b) Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transp Res Part C Emerg Technol 101:254–275

Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2018) Human mobility: models and applications. Phys Rep 734:1–74. https://doi.org/10.1016/j.physrep.2018.01.001

Breyer N, Gundlegård D, Rydergren C, Bäckman J (2017) Trip extraction for traffic analysis using cellular network data. In: 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp 321–326, https://doi.org/10.1109/MTITS.2017.8005688

Breyer N, Gundlegård D, Rydergren C (2018) Cellpath routing and route traffic flow estimation based on cellular network data. J Urban Technol 25(2):85–104. https://doi.org/10.1080/10630732.2017.1386939

Breyer N, Rydergren C, Gundlegård D (2020) Comparative analysis of travel patterns from cellular network data and an urban travel demand model. J Adv Transp. https://doi.org/10.1155/2020/3267474

Breyer N, Gundlegård D, Rydergren C (2021) Travel mode classification of intercity trips using cellular network data. Transp Res Procedia 52:211–218. https://doi.org/10.1016/j.trpro.2021.01.024

Calabrese F, Pereira FC, Di Lorenzo G, Liu L, Ratti C (2010) The geography of taste: Analyzing cell-phone mobility and social events. In: Proceedings of the 8th International Conference on Pervasive Computing, Springer-Verlag, Berlin, Heidelberg, Pervasive'10, pp 22–37. https://doi.org/10.1007/978-3-642-12654-3_2

Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput 10(4):36

Dabiri S, Lu CT, Heaslip K, Reddy CK (2020) Semi-supervised deep learning approach for transportation mode identification using gps trajectory data. IEEE Trans Knowl Data Eng 32(5):1010–1023. https://doi.org/10.1109/TKDE.2019.2896985

de Montjoye YA, Gambs S, Blondel V, Canright G, de Cordes N, Deletaille S, Engø-Monsen K, Garcia-Herranz M, Kendall J, Kerry C, Krings G, Letouzé E, Luengo-Oroz M, Oliver N, Rocher L, Rutherford A, Smoreda Z, Steele J, Wetter E, Pentland AS, Bengtsson L (2018) On the privacy-conscientious use of mobile phone data. Sci Data 5:180286 EP. https://doi.org/10.1038/sdata.2018.286

Graells-Garrido E, Caro D, Parra D (2018) Inferring modes of transportation using mobile phone data. EPJ Data Sci 7(1):49. https://doi.org/10.1140/epjds/s13688-018-0177-1

Gundlegård D (2018) Transport analytics based on cellular network signalling data. PhD thesis, Linköping University, Communications and Transport Systems, Faculty of Science & Engineering, https://doi.org/10.3384/diss.diva-152237

Gundlegård D, Rydergren C, Breyer N, Rajna B (2016) Travel demand estimation and network assignment based on cellular network data. Comput Commun 95:29–42. https://doi.org/10.1016/j.comcom.2016.04.015

Huang H, Cheng Y, Weibel R (2019) Transport mode detection based on mobile phone network data: A systematic review. Transp Res Part C Emerg Technol. https://doi.org/10.1016/j.trc.2019.02.008

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer, Berlin

Kalatian A, Farooq B (2020) A semi-supervised deep residual network for mode detection in wi-fi signals. J Big Data Anal Transp 2(2):167–180. https://doi.org/10.1007/s42421-020-00022-z

Kalatian A, Shafahi Y (2016) Travel mode detection exploiting cellular network data. MATEC Web Conf 81:03008. https://doi.org/10.1051/matecconf/20168103008

Lvd Maaten, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9:2579–2605

Manning CD, Raghavan P, Schütze H (2009) An Introduction to Information Retrieval. Cambridge University Press, Cambridge

Paulssen M, Temme D, Vij A, Walker JL (2014) Values, attitudes and travel behavior: a hierarchical latent variable mixed logit model of travel mode choice. Transportation 41(4):873–888

Phithakkitnukoon S, Sukhvibul T, Demissie M, Smoreda Z, Natwichai J, Bento C (2017) Inferring social influence in transport mode choice using mobile phone data. EPJ Data Sci 6(1):11

Qu Y, Gong H, Wang P (2015) Transportation mode split with mobile phone data. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp 285–289, https://doi.org/10.1109/ITSC.2015.56

Region Östergötland (2014) Region östergötlands resvaneundersökning 2014. Tech Rep, Region Östergötland

Schulz A, Nobis C, Eggs J, Bäumer M (2016) German national travel survey 'mid 2016 – mobility in germany': new challenges – new approaches. In: European Transport Conference 2016, AET Papers Repository, https://elib.dlr.de/109568/

Subramanya A, Talukdar PP (2014) Graph-based semi-supervised learning. Synth Lect Artif Intell Mach Learn 8(4):1–125

Triguero I, García S, Herrera F (2015) Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl Inf Syst 42(2):245–284. https://doi.org/10.1007/s10115-013-0706-y

Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA, Bottou L (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11(12):3371–3408

van Engelen JE, Hoos HH (2019) A survey on semi-supervised learning. Mach Learn 109(2):373–440. https://doi.org/10.1007/s10994-019-05855-6

Wall ME, Rechtsteiner A, Rocha LM (2003) Singular value decomposition and principal component analysis. A practical approach to microarray data analysis. Springer, pp 91–109

Xu D, Song G, Gao P, Cao R, Nie X, Xie K (2011) Transportation modes identification from mobile phone data using probabilistic models. In: International Conference on Advanced Data Mining and Applications, Springer, pp 359–371

Yu JJ (2020) Semi-supervised deep ensemble learning for travel mode identification. Transp Res Part C Emerg Technol 112:120–135. https://doi.org/10.1016/j.trc.2020.01.003

Zhou ZH (2017) A brief introduction to weakly supervised learning. Natl Sci Rev 5(1):44–53. https://doi.org/10.1093/nsr/nwx106

Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 3(1):1–130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006