A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures

Antonin Danalet

Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Station 18, CH-1015, Lausanne, Switzerland, +41-21-6932532, antonin.danalet@epfl.ch

Bilal Farooq

Département des Génies Civil, Géologique et des Mines, Polytechnique Montréal, École Polytechnique Montréal, 2500 Ch. Polytechnique Montréal, H3T 1J4 Montréal, Canada, bilal.farooq@polymtl.ca

Michel Bierlaire

Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, michel.bierlaire@epfl.ch

Abstract

In this paper, we propose a methodology to use the communication network infrastructure, in particular WiFi traces, to detect the sequence of activity episodes visited by pedestrians. Due to the poor quality of WiFi localization, a probabilistic method is proposed that infers activity-episode locations based on WiFi traces and calculates the likelihood of observing these traces in the pedestrian network, taking into account prior knowledge. The output of the method consists of candidates of activity-episodes sequences associated with the likelihood to be the true one. The methodology is validated on traces generated by a known sequence of activities, while the performance being evaluated on a set of anonymous users. Results show that it is possible to predict the number of episodes and the activity-episodes locations and durations, by merging information about the activity locations on the map, WiFi measurements and prior information about schedules and the attractivity in pedestrian infrastructure. The ambiguity of each activity episode in the sequence is explicitly measured.

Keywords: Network traces, activity choice modeling, pedestrians, semantically-enriched routing graph (SERG), potential attractivity measure, activity-episode sequence

1. Introduction

In recent years, interest in crowd dynamics and pedestrian modeling is reviving due to urban growth and its pressure on urban infrastructure (Bierlaire and Robin, 2009; Duives et al., 2013; Kasemsuppakorn and Karimi, 2013; Kneidl et al., 2013; Weidmann et al., 2014). Management of congestion is the main issue for pedestrian infrastructure. Crowd and pedestrian simulation is emerging as a tool for designing new infrastructures and optimizing the use of current infrastructures. Innovative data collection techniques and realistic experiments are vital in estimating the demand for these infrastructures.

In order to predict the total demand within a given area, activity choice models need to be developed at the scale of pedestrian infrastructure. Pedestrian demand is driven by a need to perform activities in different locations. The existence of time-space constraints in pedestrian infrastructures on the one hand and the spontaneous choice of en route destinations on the other hand ask for explicit modeling of activity scheduling decisions. Such models are traditionally used for car trips as an important source of information for strategic planning, and management or optimization of transportation networks (Ben-Akiva et al., 1996; Bowman and Ben-Akiva, 2001; Arentze and Timmermans, 2004; Balmer et al., 2006; Roorda et al., 2008, among others). For pedestrians, they will be useful in describing the congestion, for the efficient design of new facilities, and travel guidance and information systems.

Individual mobility traces are becoming available from pervasive systems, such as cellular networks (González et al., 2008) or WiFi hotspots (Section 2.1). In many cases, cost and privacy issues prohibit from installing high precision sensors such as cameras covering an entire pedestrian infrastructure. The large size of an airport or a railway station implies either precise sensors with incomplete coverage (e.g., cameras or bluetooth sensors in intersections), or full coverage with imprecise long range sensors (e.g., cellular network data, traces from WiFi infrastructures). As a result, localization data are either scarce, fuzzy, or both. We propose a methodology exploiting scarce data with an explicit modeling of the imprecision in the measure, and using prior knowledge of the infrastructure.

Section 2 reviews existing works about traces from communication infrastructure, pedestrian maps and activity-based modeling. Section 3 describes the necessary data for detecting pedestrians, while Section 4 describes the methodology to merge these data. A case study on the Ecole Polytechnique Fédérale de Lausanne (EPFL) campus is described in Section 5, with results of this case study, together with validation and sensitivity analysis. Finally, we conclude and discuss future work in Sections 6.

2. Literature review

This paper focuses on using existing localization data for modeling macroscopic behavior such as destination or activity choice. The following literature review is divided in three parts, corresponding to the three challenges we meet in pedestrian demand modeling: data collection (Section 2.1), representation of space (Section 2.2) and modeling (Section 2.3).

2.1. Collecting data from digital footprints

The recent developments in detection technologies open doors to new researches about pedestrian behavior. In the field of trajectory detection, Taniguchi et al. (2013) are using Bayesian estimation on binary sensors located at the border of a cell, while Alahi (2011) and Alahi et al. (2014) use networks of cameras to track and analyze pedestrian trajectories. Alahi's main motivation is the number of already installed cameras generating large datasets. Smartphones are sharing this characteristic: a majority of people are carrying a mobile device such as a smartphone, and they generate data. Several data collection techniques about smartphones are device-centric. We focus here on data from communication network infrastructure ("network traces").

Using traces from communication network infrastructure has several advantages on data from the smartphone. First, full coverage of the facility is usually cheap and allows for an estimation of the overall demand. The communication infrastructure sometimes already exists, and increasing its density has a positive side effect. Smartphone users do not need to install anything on their device, and so the access to sensitive information such as emails or address book is limited for the analyst, which ensure privacy for the users. Finally, traces from communication network infrastructure are related to the infrastructure and not to the individual: we are tracking all individual smartphones going through a facility and not all places visited by the same individuals. It allows the analyst to focus on the pedestrian facility covered by the communication network.

There are few drawbacks to network traces as well. Socio-economic and demographic attributes are difficult to collect due to both privacy concerns (if the data already exist) and to the difficulty to survey the tracked person from the infrastructure side (if the data does not exist). Additionally, smartphone users are not necessarily representative of the full population.

Several applications using data from communication infrastructure, both with WiFi and GSM traces (Bekhor et al., 2013; Calabrese et al., 2011), have been developed to study mobility behavior. These new data collections are motivated by the needs for calibrated agent-based models. Post-processing methods are needed to transform these raw observations into data adapted for modeling purpose to overcome imprecision and missing observations in the data: detection of stops points, activity purpose detection through land-use information and spatial matching (Rieser-Schüssler, 2012). With GSM traces, Bekhor et al. (2013) mention the elimination of "unreasonable movements performed in short time periods between antennas located far apart" without more details. Calabrese et al. (2011) does not consider the underlying transportation network to correct for anisotropy.

A large literature exists about WiFi traces from a computer communication point of view. A complete review can be found in Aschenbruck et al. (2011). All references in this paper define mobility trace-based models as a tool to improve the quality of the WiFi. Field studies have been done (Tang and Baker, 2000; Balachandran et al., 2002; Balazinska and Castro, 2003; Yoon et al., 2006; Sevtsuk et al., 2009; Zola and Barcelo-Arroyo, 2011; Wanalertlak et al., 2011; Meneses and Moreira, 2012). The main results are the prediction of changes in access points (APs). The main problem reported in these articles is the *ping pong* effect, when the device has similar signal strengths from different APs and changes regularly from one to another. This is a problem from a network viewpoint, and also for modeling pedestrian origins and destinations. Yoon et al. (2006) propose to use a moving average weighted by time spent at destination to remove the extra AP logs. A general solution presented in Aschenbruck et al. (2011) consists in aggregation of data over different APs. Most studies about WiFi are focusing on network performance and management and not on human mobility. In Yoon et al. (2006), contrarily to all other papers cited here, an OD matrix is estimated at the building level in Dartmouth college. Variations in time/day are not considered, as Aschenbruck et al. (2011) noticed.

In the literature about mobility models for WiFi infrastructure from a computer communication point of view, the most common model, Random Waypoint model (RWP), is often criticized as not representing real human mobility (Conti and Giordano, 2007). One of the problems with RWP consists in using straight lines between two signals in different APs, even if this path is not physically possible. This is the main reason why trace-based mobility models were developed in this domain of research. A key challenge in building a realistic model is to define a pedestrian network and the corresponding possible paths that the user with a device can follow. This process of constructing and using the pedestrian network in order to improve the mobility model is not explicitly presented in Aschenbruck et al. (2011) in their large review of trace-based mobility models. The need for a more complex approach is emphasized in Rojas et al. (2005).

2.2. Pedestrian network

The pedestrian network depends on the scale of the study area and on the definition of destinations: buildings (Yoon et al., 2006), APs (Wanalertlak et al., 2011; Tuduce and Gross, 2005), rooms (Sevtsuk et al., 2009), or, at an urban scale, subzones of the motorized regional zone system as nonmotorized destinations (Eash, 1999).

The walking distance between the destinations is usually not available (Kasemsuppakorn and Karimi, 2013). Yoon et al. (2006) converted a map to a graph between buildings and limited themselves to major roads. In the extension of a Chicago model for nonmotorized trips, the Manhattan distance is motivated by the grid plan and the absence of a pedestrian network (Eash, 1999).

Kasemsuppakorn and Karimi (2013) propose to build it from GPS traces, which does not work indoor. Kang et al. (2004) are using WiFi to cluster places of interest and label them, but APs serve different kind of locations surrounding them (Calabrese et al., 2010).

Without a model based on a pedestrian network, changes in the pedestrian facilities such as pedestrian bridges or underpasses cannot be tested.

Indoor networks of pedestrian facilities allowing for computation of the shortest path between two destinations are increasingly available for airports, museums, campuses, hospitals and malls due to the complexity of path finding (Goetz and Zipf, 2011). Crowdsourced geodata such as OpenStreetMap are extending to indoor spaces (Goetz, 2012).

2.3. Activity-based modeling: from driving to walking

The premise of activity-based approach involves considering activity as a choice and trips as a way to complete the chosen activity. In other words, modeling the daily activity patterns allows the development of behavioral travel demand models that are sensitive to changes in policy. Complete and general reviews not related to pedestrians can be found in Roorda (2005), Habib (2007) and Feil (2010).

With respect to pedestrians, Borgers and Timmermans (1986) develop a destination choice model as part of a system of models to predict the total demand for retail facilities within inner-city shopping areas. Timmermans et al. (1992) provide a review of models existing in 1992 and of a few applications to urban and transportation planning in The Netherlands. Zhu and Timmermans (2005) focus on shopping decision processes, using bio-inspired heuristics to mimic the decision process. Eash (1999) developed an extension of a regional travel model of Chicago for nonmotorized trips (bicycle and pedestrian modes), including destination and mode choice.

Influenced by traditional practice in travel demand analysis, several models are derived from origin-destination matrices (Nagel and Barrett, 1997; Antonini et al., 2006), where the set of potential origins and destinations is predefined, and flows between origins and destinations is estimated. In a disaggregate context, the choice of the destination can be modeled conditional to a given activity, or as a joint choice of an activity and a destination. In both cases, the choice set is typically large and difficult to characterize (Bierlaire and Robin, 2009).

In some circumstances, it is possible that no destination is explicitly chosen by a pedestrian. It is typical when walking is the activity as such, or in shopping and touristic activities. In these cases, an itinerary is undertaken without a known target, trying to maximize the chances to reach attractive places along the way (Borst et al., 2008). This type of behavior is clearly difficult to formalize, but relates to behavior that highly characterizes pedestrian movements.

The current state of the literature on the one hand comprises very detailed conventional activity-based models for multimodal urban areas and on the other end few specific models related to pedestrians. This big gap between these two different contexts is explained by the lack of data. A pedestrian facility such as a transportation hub shares some characteristics with a city, but surveying and counting pedestrians is more challenging.

There are three problems with surveying tools, such as MAGIC (Ettema et al., 1994). First, the data is collected at one point in time for a single person. Second, there could be a difference between what people plan and answer to the survey and their real behavior.

Finally, the vast amount of information needed to understand the pattern of a day could create response burden.

In multimodal urban systems, solutions include using localization-aware technologies for longitudinal surveys. A typical example is GPS-based prompted recall activity-travel survey as in Frignani et al. (2010), where respondent carried GPS devices for 14 days. CHASE (Computerized Household Activity Scheduling Elicitor) has been implemented on mobile devices (Rindfüser et al., 2003) and is similar to some extent to MAGIC. Asking people to confirm their GPS traces on the internet or answer questions directly on their mobile device is a way to reduce the cognitive incongruence with actual behavior and to avoid response burden. Etter et al. (2012) also show that it is possible to predict up to 60% of next visited places from passive smartphone data.

For pedestrian infrastructure, GPS devices cannot precisely detect users indoor and do not specifically focus in infrastructure users. Cui et al. (2012) propose a paper-and-pencil survey to determine usage patterns, with the same problems as in a larger scale. We propose to use network traces and land use data. Network traces allow to track the real location of several pedestrians for several days, without response burden. Land use data, existing at the pedestrian scale as shown in Section 2.2, allows for a better spatial resolution of activity-based models (Goulias et al., 2013).

3. Data requirement

To apply our methodology, we assume two kinds of data sources: network traces and semantically-enriched routing graph to which we can associate the network traces.

3.1. Network traces

An input of this probabilistic method consists of timestamps and localization data coming from network traces: WiFi traces, GSM traces, Bluetooth tracking or RFID localization. We define a measurement as $\hat{m} = (\hat{x}, \hat{t})$, where $\hat{x} \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ is the position of the measurement (x-y coordinates in a coordinate system, and floor or altitude in a multi-floor environment) and $\hat{t}$ the measurement timestamp. In data from access points (APs) (WiFi APs, cell towers, ...), $\hat{x}$ is the position of an AP and is discrete; in multilateration data, the measurement $\hat{x}$ is continuous in space, and not only at AP locations. For a given individual i, we assume a chronologically ordered sequence $(\hat{m}_{1,i}, , ..., \hat{m}_{j_i,i} ..., \hat{m}_{J_i,i})$, which is abbreviated as $\hat{m}_{1:J_i,i}$, where J_i is the total number of measurements. The measurement timestamp $\hat{t}$ is continuous.

Accuracy ξ is also needed for each measurement $\hat{x}$. It is defined as the distribution of the Euclidean distance between the location estimate $\hat{x}$ and the actual location $\mathring{x}$, $\hat{x} = \mathring{x} + \xi$. It can either be constructed by the information provided by the localization tool (e.g. level of confidence, attenuation rate, etc.) or by the analyst itself. In the second case, one has to design experiments and calibrate the error distribution based on already known locations in the WiFi coverage area.

Different levels of anonymity are possible with these data. Originally, the MAC address is collected which uniquely identify the device. This MAC address may be processed in two different ways. First, it may be associated with a username through identification in the system and thus to the identity or socioeconomic information if available, such as gender, age or income. Second, it may be anonymized to guarantee anonymity. The anonymization can be total or keeping some socioeconomic information. This way, i can correspond to a unique ID; or associated with some socioeconomic characteristics; or i can correspond to the device or personal identifier.

3.2. Pedestrian semantically-enriched routing graph

The following methodology needs a *semantically-enriched routing graph* (SERG). We define SERG as a set of nodes $\mathcal{N}$ and a set of edges $\mathcal{E}$. SERG allows for routing pedestrians from origins to destinations through an optimal path, and contains information such as the name of the room, or the type of the room. In order to link localization measurements $\hat{x}$ to the graph, each node $n \in \mathcal{N}$ must be associated with a coordinate system.

In a centerline approach as defined in Goetz and Zipf (2011) for a corridor, some nodes correspond to intersections and not to possible pedestrian destinations. Nodes are defined as destinations if they correspond to a room, a shop or a restaurant, i.e. if they are points of interest (POI) in the pedestrian infrastructure. POI is a subset of $\mathcal{N}$.

Formally, $\text{SERG} := (\mathcal{N}, \mathcal{E}, \mathcal{L}, f, g, \text{POI})$, where $\mathcal{L}$ is a set of relevant labels for rooms, restaurants, shops, etc., $f : \mathcal{N} \to \mathcal{L}$ is the labeling function, and $g : \mathcal{N} \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ a function associating nodes with coordinates in a coordinate system. $\text{POI} \subset \mathcal{N}$.

More information can be added to SERG. The path realism (both physically and behaviorally) can be improved by adding information to the graph and using a generalized cost for the shortest path algorithm. A solution to balance between the shortest path and the simplest path is to give each edge of the pedestrian network a weight. It represents the aversion to floor changes and less important walkways. Goetz and Zipf (2011) propose a *weighted indoor routing graph*, which is an enriched version of SERG. Adding information to edges $\mathcal{E}$ allows for a more realistic shortest path algorithm. Adding information to nodes $\mathcal{N}$ gives the opportunity to associate other data, such as schedules, opening hours, or door access.

4. Methodology

We are proposing a modeling approach to extract the possible activity-episode sequences performed by pedestrians from digital traces in a communication networks. This Bayesian approach merges measured network traces (continuous in space) (Section 3.1) and pedestrian semantically-enriched routing graph (Section 3.2) to compute the likelihood that a given sequence of activity episodes (discrete in space) has actually generated the observed traces.

We define an activity episode $a = (x, t^-, t^+)$ as a POI where the user is spending time, where x is the episode location, t^- the episode start time, and t^+ the episode end time. The episode location x is a POI in SERG, $x \in \text{POI}$, and is labeled, $f(x) \in \mathcal{L}$. t^- and t^+ are continuous random variables and define the time spent at destination, $t^+ - t^-$. We impose that $t^+ - t^- \geqslant T_{min}$, a minimum threshold (typically 5 min in a pedestrian context, similar to Bekhor et al. (2013), where a stop is assumed when data are collected 20 min in the same antenna location). The output of the probabilistic method consists of a set of L candidate activity-episode sequences $(a_1, ..., a_{k_i}, ..., a_{K_i})$, which is abbreviated as $a_{1:K_i}$, where K_i is the total number of episodes. K_i is individual specific and unknown to the analyst. In the following developments, both the number of measurements in the sequence J_i and the number of episodes K_i are individual specific, but the i subscript is omitted to make the notation light. Each candidate activity-episode sequence $a_{1:K}$ is associated with the probability of being the actual one.

In the next section, we propose a probabilistic measurement model associated with an activity-episode sequence. Then, in Section 4.2, it is described how to generate candidate activity-episode sequences. Figure 1 shows the plate model (see Koller and Friedman (2009), Section 6.4.1) of the link between the activity episodes and the measurements.

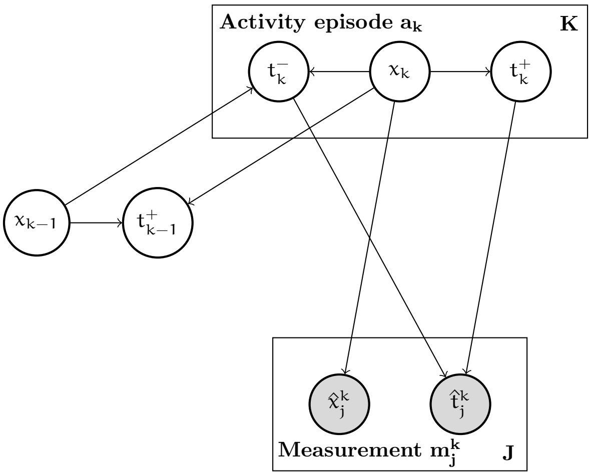

Figure 1: Plate model for the probabilistic measurement model. It represents the generation process of network traces. While being in point of interest x_k between times t_k^- and t_k^+, users generate measurements $\hat{x}_j^k$ at time $\hat{t}_j^k$. Dark shaded nodes represent the observed variables. The arrows represent the dependencies between the variables. Boxes express the multiple iterations of the conceptual object: there are K activity episodes a_k, and J measurements m_j^k in total.

4.1. Probabilistic measurement model: a Bayesian approach

A probability is associated with each activity-episode sequence. It takes into account the inaccuracy in the network traces based on the measurements and some prior knowledge about the potential activity-episode locations. The activity probability $P(a_{1:K}|\hat{m}_{1:J})$ that $a_{1:K}$ is the actual activity-episode sequence given the measurement $\hat{m}_{1:J}$ is decomposed as:

$$P(a_{1:K}|\hat{m}_{1:J}) \quad \propto \quad P(\hat{m}_{1:J}|a_{1:K}) \cdot P(a_{1:K}) \tag{1}$$

where $P(\hat{m}_{1:J}|a_{1:K})$ is the measurement likelihood and $P(a_{1:K})$ is a prior knowledge about the activity episodes.

4.1.1. Measurement likelihood

For each activity-episode sequence, our goal is to compute the probability that the performed episodes generated the observed measurement sequence:

$$P(\hat{m}_{1:J}|a_{1:K}). \tag{2}$$

We assume that a measurement $\hat{m}_j$ always corresponds to an activity episode a_k. We denote $\hat{m}_j^k = (\hat{x}_j^k, \hat{t}_j^k)$ the measurement in $\hat{m}_{1:J}$ corresponding to $a_k = (x_k, t_k^-, t_k^+)$, i.e. when $t_k^- \leqslant \hat{t}^k \leqslant t_k^+$. As a result, $\hat{m}_{1:J} = \cup_k \hat{m}_{1:J}^k$. If a measurement is generated while walking, the model will consider it as a very short activity episode (that can be eliminated later).

If the device's owner is performing activity episode a, the probability that it will generate a measurement $\hat{m}$ is a function of the location of the episode location x and the measurement location $\hat{x}$ (e.g., the distance). Thus we can decompose Equation 2 as:

$$\begin{aligned} P(\hat{m}_{1:J}|a_{1:K}) &= \prod_{k=1}^{K} P(\hat{m}_{1:J}^k|a_k) && (3)\\ &= \prod_{k=1}^{K}\prod_{j=1}^{J} P(\hat{m}_j^k|a_k) && (4)\\ &= \prod_{k=1}^{K}\prod_{j=1}^{J} P(\hat{x}_j^k|x_k) && (5)\end{aligned}$$

Equality in Equation 3 assumes measurement independence between activities, i.e. measurement error in the sequence is only related to the corresponding activity episode in time. Equality in Equation 4 assumes independence between measurements, i.e. error is the same for different measurements while in the same location x_k and time interval t_k^-, t_k^+. Equality in Equation 5 assumes no measurement error in time, e.g. measurement error is only a localization error.

4.1.2. Prior: potential attractivity measure

Introducing prior knowledge may be needed when localization is weak and pedestrian map is dense. Moreover, the prior gives the possibility to add information from available data. In this section, we propose the first formal definition of attractivity for pedestrian infrastructures to our knowledge. It is built on existing literature for urban context and allows to coherently merge different data sources.

Space-time accessibility is a very common concept in land use planning (Miller, 2010). It measures availability of activities for individuals given temporal and spatial constraints. Several definitions have been proposed. Hansen (1959) defines accessibility as a "potential of opportunities for interaction":

$$A_i = \sum_j \frac{S_j}{tt_{ij}^{\alpha}}$$

where A_i is the accessibility of place i, S_j is a measure of the "size of the activity" at j, such as the number of jobs, the annual retail sales or the population in a residential area, and tt_{ij} represents the travel time between i and j. The α parameter, defining the weight of travel time in accessibility measure, is evaluated based on the urban growth, assuming it is directly proportional to accessibility. Weibull (1980) developed a rigorous axiomatic framework defining attraction-accessibility measure based on distance and attractivity (also called supply capacity). No clear definition of what exactly is attractivity is given but in an example about labor market, attractivity is defined as a function of the number of jobs and a demand potential for each zone (Weibull, 1976). He mentions that attractivity may be described as "offer", and gives the examples of places at day-nurseries and hospital beds.

A definition of accessibility merging the attractivity-accessibility measures (Weibull, 1980) and the contraints-oriented approach (see Section 2.3 and Hägerstraand (1970)) is proposed by Miller (2010). He emphasizes that a pure constraints-oriented approach gives each opportunity an equal weight, and, conversely, an attractivity-accessibility approach does not take into account temporal constraints. We propose to similarly define a potential attractivity measure by merging attractivity and time constraints for the pedestrian context.

Formally, we define the potential attractivity measure as a model of aggregated occupation per point of interest (POI). The unit of attractivity is the number of persons. The potential attractivity measure $S_{x,i}(t^-, t^+)$ between a start time t^- and an end time t^+ for $x \in$ POI and individual i is time dependent and may differ across individuals. It depends on the instantaneous potential attractivity measure $S_{x,i}(t)$ at a given time t:

$$S_{x,i}(t^-, t^+) = \int_{t=t^-}^{t^+} S_{x,i}(t)dt \tag{6}$$

The instantaneous potential attractivity measure depends on time-constraints and attractivity:

$$S_{x,i}(t) = \delta_{x,i}(t) \cdot \mathcal{A}_i(x, t)$$

where $\delta_{x,i}(t)$ is a dummy variable for time-constraints such as schedules or opening hours, with value 1 if the POI is open or scheduled and 0 otherwise: opening hours of shops and restaurants, or timetables in the case of conferences, campuses, or public transport infrastructures. Timetable are individual-specific. Their availability depends on the level of anonymity for localization data (see Section 3.1).

Attractivity $\mathcal{A}_i(x, t)$ is context-specific, as seen in the land use literature: number of jobs, annual retail sale, population per zone, places at day-nurseries, hospital beds. In the pedestrian facility context, data sources could be checkouts in supermarkets, metro card swapping data, concert tickets data, number of seats in a restaurant, number of employees per office, number of students in class, capacity of different zones in a stadium or a public transport infrastructure.

As a general guideline, the potential attractivity measure depends on the available information:

- If the attractivity is stable in time for a given POI x (e.g., an office on campus with a given number of employees and no explicit office hours), $\delta_{x,i}(t) = 1\ \forall t$ and $S_{x,i}(t) = \mathcal{A}(x)$;
- If the POI has opening hours (e.g. a shop on campus), $\delta_{x,i}(t) = 1$ for t in the opening hours and 0 otherwise, and consequently $S_{x,i}(t) = \mathcal{A}(x)$ for t in the opening hours and 0 otherwise;
- If the POI has varying attractivity in time, $S_{x,i}(t) = \mathcal{A}(x,t)$ with $\mathcal{A}(x,t)$ being a step function (e.g. for classrooms with different numbers of students at different periods of the day) or any function representing the number of people in the POI per time (e.g., point-of-sale data for restaurants);
- If the attractivity varies for different people or categories of people, $S_{x,i}(t) = \mathcal{A}_i(x,t)$ with different attractivity functions $\mathcal{A}_i(x,t)$ for different individuals i (e.g., a classroom has different attractivities for employees and students on a campus).

The prior $P(a_{1:K})$ is derived assuming that, for each k, activity episode a_{k-1} is independent from activity episode a_k.

$$\begin{aligned} P(a_{1:K}) &= \prod_{k=1}^{K} P(a_k) && (7)\\ &= \prod_{k=1}^{K} P(x_k, t_k^-, t_k^+) && (8)\\ &= \prod_{k=1}^{K} \frac{S_{x_k,i}(t_k^-, t_k^+)}{\sum_{x \in POI} S_{x,i}(t_k^-, t_k^+)} && (9) \end{aligned}$$

The prior probability is proportional to the attractivity. As an extension, attractivity measure can be replaced by a probability model such as a logit model predicting the number of persons in a given POI at a given time. The proposed model assumes independence between consecutive activity episodes. An activity choice model could be included as a prior.

We define four specifications of the prior based on different assumptions on the available data: uniform, aggregate, disaggregate and diary. If no information about the attractivity is available, a default assumption has to be used, and attractivity is fixed for all POI. The corresponding prior is called “uniform”. If information about attractivity and schedule is available, the quality of the prior depends on the level of anonimity of the network traces. Without personal information, a single aggregate prior is defined using the same time constraints for all individuals to define the potential attractivity measure for each location. Disaggregate information about schedules may be available without knowing the identity of the individual: travelers (with trip schedules) and non-travelers in a transport hub, employees (with working hours) and visitors in a shop, etc. These information define one “disaggregate” prior per group. They come either directly from the network traces (see Section 3.1) or from pattern recognition (e.g., in a railway station, individuals directly arriving on a platform are automatically travelers coming from a train). Finally, individual schedules can be used to define a “diary” prior. Due to respondent burden, individual schedules from activity schedule surveys are particularly difficult to collect (see, e.g., Chen et al. (2010)). This prior is important for establishing the consistency of our approach.

4.2. Generation of activity-episode sequences

The probabilistic measurement model computes the likelihood of a given activity-episode sequence $a_{1:K}$. This section focuses on the generation of candidate activity-episode sequences. An algorithm is proposed to generate candidates from localization data and pedestrian semantically-enriched routing graph. At each new measurement of $\hat{m}_{1:J}$, we build a list of candidates to be the true location x and for each of them start and end times t^- and t^+.

4.2.1. Generating episode location

Inspired by the methodology developed by Bierlaire et al. (2010) for smartphone GPS data, we generate candidate episode locations for each measurement using the concept of *domain of data relevance* (DDR) originally introduced by Bierlaire and Frejinger (2008).

We define the DDR as a physical area in space where a measurement location is relevant. The definition of the area can be different depending on the precision of the measurement, i.e., the DDR depends on the type of data. In the pedestrian context, this area could even be in 3D, covering several floors. A point of interest $x \in \mathrm{POI}$ is considered to be in the DDR of measurement location $\hat{x}$ if the probability $P(\hat{x}|x) \geqslant \theta$, with θ a given threshold. This probability is a function of the location x of the POI and the measurement location $\hat{x}$, similar to the one in Equation 5.

Using the domain of data relevance DDR_j for each measurement $\hat{m}_j$, we generate all possible activity sequences for each individual. Each element of the sequence represents a possible episode location. It is connected with all possible next episode locations contained in the domain of data relevance DDR_{j+1} of the following measurement in time, $\hat{m}_{j+1}$ (a simple example with two DDRs containing respectively 3 and 2 points of interest is presented in Figure 2). For a list $\hat{m}_1, \hat{m}_2, ..., \hat{m}_J$ of measurements associated with a given individual, the result of this process is a network structure with path in this network with length J. Each path in the network corresponds to a sequence $x_{1:J}$ of potential episode locations. This network is built recursively. For each measurement $\hat{m}_j, j = 1, ..., J$ for a particular individual in chronological order, we consider all possible episode locations, i.e., all POI in SERG, in the domain of data relevance DDR_j. At each new measurement $\hat{m}_j$, the network structure of activity episodes is extended with all locations associated with $\hat{m}_j$ (Figure B.20 in Appendix B).

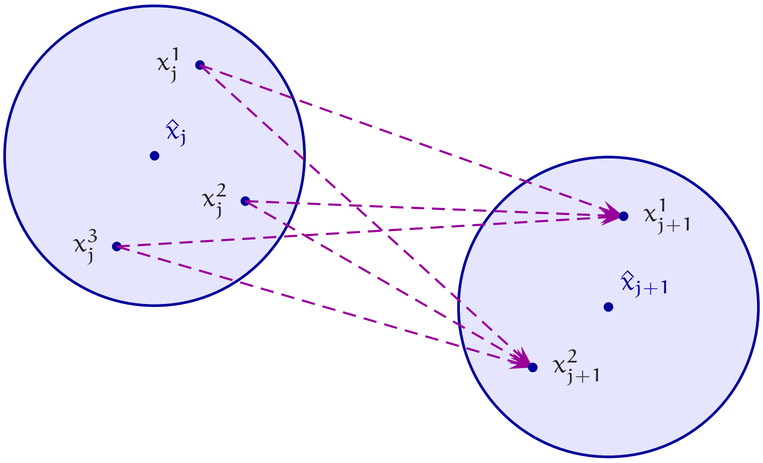

Figure 2: The symbolic representation of two domains of data relevance DDR_j and DDR_{j+1} corresponding to measurements $\hat{m}_j$ and $\hat{m}_{j+1}$ following each other chronologically. In this simple example, we assume DDR_j contains 3 possible episode locations x_j^1, x_j^2, x_j^3, and DDR_{j+1} contains 2 possible episode locations x_{j+1}^1, x_{j+1}^2 ($x_j^1, x_j^2, x_j^3, x_{j+1}^1, x_{j+1}^2 \in \mathrm{POI}$).

If a measurement is imprecise and the corresponding DDR is huge (e.g., in an area with low WiFi coverage the size of the confidence interval can be of the order of magnitude of the whole pedestrian infrastructure), all points of interest have a high measurement likelihood. Here, the prior is defining alone the location as the point of interest with the highest attractivity in the pedestrian infrastructure. An upper bound for the size of the DDR is needed in these cases. The activity episodes corresponding to these measurements containing no information will then be eliminated (see Section 4.3).

4.2.2. Generating episode start and end times

Once a sequence $x_{1:J}$ of potential episode locations is defined, the episode start and end times t^- and t^+ at these locations need to be generated.

Given two consecutive measurements $\hat{m}_j$ and $\hat{m}_{j+1}$ and their corresponding timestamps $\hat{t}_j$ and $\hat{t}_{j+1}$, a trip between the two generated positions x_j and x_{j+1} of the consecutive activity episodes a_j and a_{j+1} is assumed to take place. This trip defines both the end time t_j^+ from episode a_j and the start time t_{j+1}^- of episode a_{j+1}. The departure of the trip occurs after measurement $\hat{m}_j$ and before the latest possible departure time, i.e., the time that allows to reach the next episode location through the shortest path. Similarly, the arrival occurs before the next measurement $\hat{m}_{j+1}$ and after the trip from the episode location of the previous measurement (Figure 3).

The travel time used by a pedestrian to walk from x_j to x_{j+1} is approximated by the ratio between the shortest path distance between x_j and x_{j+1}, and the speed of 1.34 m/s (see Buchmüller and Weidmann (2006)). In this way, the episode end time

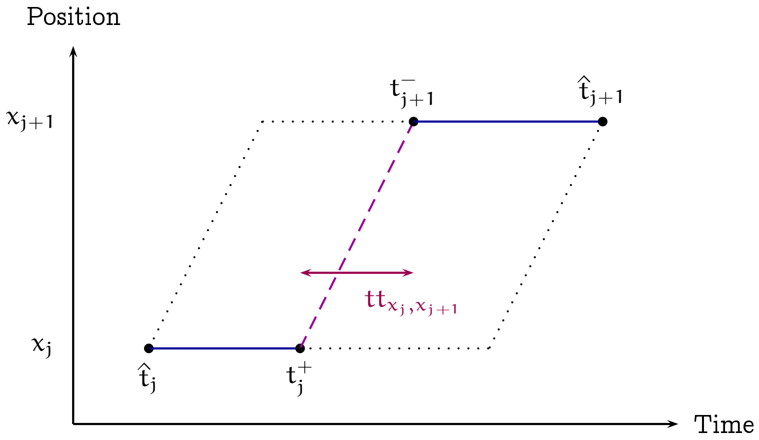

Figure 3: Time-space representation of two consecutive activity episodes j and j + 1.

$t^+_j \in [\hat{t}_j, \max(\hat{t}_j, \hat{t}_{j+1} - tt_{x_j,x_{j+1}})]$ and the next episode start time $t^-_{j+1} \in [\min(t^+_j + tt_{x_j,x_{j+1}}, \hat{t}_{j+1}), \hat{t}_{j+1}]$. The maximum and the minimum in the bounds of the intervals manage the situation when $\hat{t}_j \geqslant \hat{t}_{j+1} - tt_{x_j,x_{j+1}}$ or $t^+_j + tt_{x_j,x_{j+1}} \geqslant \hat{t}_{j+1}$. This may happen when the pedestrian was much faster than what we assume, or when a measurement was generated while walking (no stop, thus no time spent at this location).

No information is available about the exact time when the trip actually happens between the bounds for start and end times, and so a uniform distribution is used. The end time t^+_j is uniformly distributed, $t^+_j \sim U(\hat{t}_j, \hat{t}_{j+1} - tt_{x_j,x_{j+1}})$, with density function

$$f(x) = \frac{1}{\hat{t}_{j+1} - tt_{x_j,x_{j+1}} - \hat{t}_j}.$$

The start time t^-_{j+1} is uniformly distributed between $t^+_j + tt_{x_j,x_{j+1}}$ and $\hat{t}_{j+1}$. Since t^+_j is itself normally distributed, the density function of t^-_{j+1} is

$$f(t^-_{j+1}) = \frac{1}{\hat{t}_{j+1} - tt_{x_j,x_{j+1}} - \hat{t}_j} \ln \frac{\hat{t}_{j+1} - tt_{x_j,x_{j+1}} - \hat{t}_j}{\hat{t}_{j+1} - t^-_{k+1}}$$

(and expected value is $E(t^-_{j+1}) = \frac{\hat{t}_j + tt_{x_j,x_{j+1}} + 3\cdot\hat{t}_{j+1}}{4}$) (see Appendix A for a derivation). In the cases when $\hat{t}_j \geqslant \hat{t}_{j+1} - tt_{x_j,x_{j+1}}$ or $t^+_j + tt_{x_j,x_{j+1}} \geqslant \hat{t}_{j+1}$, t^+_j and t^-_{j+1} are fixed with value $\hat{t}_j$ and $\hat{t}_{j+1}$ respectively.

4.3. Intermediary measurements

The duration of activity is assumed to have a lower bound T_{min}. Any episode with an expected duration lower than T_{min} is rejected. It is assumed that the corresponding measurement has been generated while the pedestrian was walking, and therefore does not correspond to an activity.

In the case of very imprecise measurements, the DDR is bounded (see Section 4.2.1). It avoids accumulated activity probability on a location with a strong prior based on no geographical evidence (very weak measurement). It creates false activity episodes, artificially close to the measurement location. Then, if there is no confirmation from another measurement in this area, the time spent at this activity episode will be very short and thus this activity episode will be eliminated.

4.4. Sequence elimination procedure

The number of path in the network grows exponentially with the number of measurement. For each measurement $\hat{m}_j, j = 1, ..., J$, all the elements of the corresponding DDR, $|DDR_j|$, have to be connected with all the previous candidates, resulting in $\prod_{j=1}^{J} |DDR_j|$ candidates. In practice, it is not possible to consider all possible combinations. Therefore, the proposed implementation of the procedure imposes an upper bound L on the number of candidates. Whenever the number of candidates exceeds L, the least likely candidates (according to Eq(1)) are eliminated to enforce the maximum number of paths in the network (Figure B.21 in Appendix B). This procedure performs better deterministically (keeping the L most likely candidates) than stochastically (drawing L candidates based on the activity probability). Indeed, accumulation over several measurements generates an activity episode. When randomly picking a candidate, there is a risk to have several activity episodes with a small duration. They will be considered as intermediary measurements and eliminated.

The processes described in Section 4.1, 4.2, 4.3 and 4.4 define Algorithm 1. It is illustrated in Figure C.22 in Appendix C. The algorithm runs in $O(J \cdot |DDR| \cdot L \cdot |\mathcal{E}| \cdot |\mathcal{N}| \cdot \log(|\mathcal{N}|))$. Computational burden mainly comes from the shortest path algorithm. The number of shortest paths computations depends on the size of the DDR (controlled by the modeler) and on the size of the list $a_{1:k}$ (L, also in control of the modeler).

The network traces bring the dynamics in the process by allowing to track a pedestrian during all the journey in the pedestrian infrastructure. The prior is a way to add information about time constraints and attractivity. Finally, the pedestrian semantically-enriched routing graph (SERG) has two roles in the process. First, it allows to link the network traces (coordinates in a continuous space) to time constraints and attractivity of POI in the prior (places and landmarks in discrete space). Second, shortest path in SERG being bigger than Euclidean distance between two POI, it corrects for anisotropy in the pedestrian infrastructure. It impacts the elimination procedure through the computation of shortest paths.

The proposed probabilistic measurement model computes the probability of performing an activity-episode sequence while generating measurements. It assumes that each measurement corresponds to an activity episode. In reality, some measurements are generated while walking. While it makes sense in order to simplify the model and handle a lot of measurements in a reasonable time, an extension of the proposed model consists in applying a probabilistic map matching approach such as Bierlaire et al. (2013) for measurements not related to an activity episode.

Algorithm 1: Generation of activity-episodes sequences.

```
for each ID do
    for measurement $\hat{m}_j = (\hat{x}_j, \hat{t}_j)$, $j = 1, ..., J$ $O(J)$ do
        Define the corresponding Domain of Data Relevance, $DDR_j$ $O(|DDR|\log(|\mathcal{N}|))$;
        for each $x \in DDR_j$ $O(|DDR|)$ do
            Compute the measurement likelihood ;
            if T empty then
                Initialize the network structure for activity-episodes sequences T with $x_1 = x$,
                $t_1^- = t_1^+ = \hat{t}_1$ ;
                Update the activity probability with the measurement likelihood and the prior ;
            else
                for each $a_{1:k}$ of T $O(L)$ do
                    if $x_k = x$ then
                        Update the definition of the episode end time: $t_k^+ = \hat{t}_j$ ;
                        Update the prior for $a_k$ ;
                        Update the activity probability of $a_{1:k}$ with the measurement
                        likelihood and the prior ;
                    else
                        Compute the shortest path between $x_k$ and $x$, and the travel time
                        $tt_{x_k,x}$ $O((|\mathcal{E}| + |\mathcal{N}|)\log(|\mathcal{N}|))$;
                        Define the last episode end time: $t_k^+ \sim U(\hat{t}_{j-1}, \hat{t}_j - tt_{x_k,x})$ ;
                        Define the new episode start time: $t_{k+1}^- \sim U(\hat{t}_{j-1} + tt_{x_k,x}, \hat{t}_j)$ ;
                        $a_{k+1} = (x, t_{k+1}^-, \hat{t}_j)$ ;
                        if $a_k$ is an intermediary measurement then
                            Connect $a_{k-1}$ with $a_{k+1}$ in T ;
                            Compute the prior for $a_{k+1}$ ;
                            Update the activity probability of $a_{1:k+1}$ with the new
                            measurement likelihood and prior, but without the prior for $a_k$ ;
                        else
                            Connect $a_k$ with $a_{k+1}$ in T ;
                            Update the prior for $a_k$ and compute it for $a_{k+1}$ ;
                            Update the activity probability of $a_{1:k+1}$ with the new
                            measurement likelihood and priors ;
        Sequence elimination procedure: keep the L most likely paths of the network T
        $O(L|DDR|\log(L|DDR|))$
```

5. A case study on EPFL campus

We conduct an experiment on the EPFL campus. We assume here that the only mode on campus is walking, even if some people outside of the campus could be detected, either within a car on the road or within public transportation.

In Section 5.1 and 5.2, localization data and a pedestrian semantically-enriched routing graph of the campus are presented. We show how they comply with the data requirement defined in Section 3. Then, in Section 5.3, the potential attractivity measure used to generate the prior distribution is described. Finally, in Section 5.5, sensitivity analysis is performed on the different parameters, in particular the ones defining the DDR, the prior and the density of measurements.

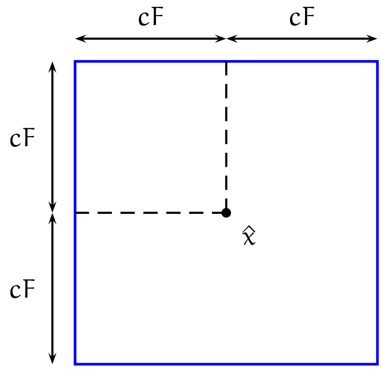

Figure 4: 95% confidence square for localization tool by Cisco

5.1. EPFL WiFi data

The data used for this case study have been collected with the Cisco Context Aware Mobility API with the Cisco Mobility Services Engine (MSE) (Cisco, 2011). It uses triangulation based on signal strength from the access points to generate a measurement $\hat{m} = (\hat{x}, \hat{t})$. Therefore, in this case study, the location of the device $\hat{x}$ is continuous in space. A confidence factor cF defines a square around each x-y coordinates (see Figure 4). The device is estimated to be inside this confidence square centered at the measurement $\hat{x}$ with sides $2 \cdot cF$ with 95% probability. cF is calculated assuming that the device is located on the correct floor (Cisco, 2011). These data correspond to the localization data requirement defined in Section 3.1.

For the measurement equation as defined in Section 4.1.1, we assume that the errors in latitude and longitude are independently and normally distributed. We decompose both the measurement $\hat{x}$ and the activity location x in latitude and longitude $\hat{x}_{lat}, \hat{x}_{long}, x_{lat}$ and x_{long}. Assuming the errors in latitude and longitude are independent, $P(\hat{x}|x) = P(\hat{x}_{lat}|x_{lat}) \cdot P(\hat{x}_{long}|x_{long})$ with:

$$P(\hat{x}_{lat}|x_{lat}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\hat{x}_{lat} - x_{lat})^2}{2\sigma^2} \right) \tag{10}$$

$$P(\hat{x}_{long}|x_{long}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\hat{x}_{long} - x_{long})^2}{2\sigma^2} \right) \tag{11}$$

where $\sigma = \frac{cF}{2}$. This is equivalent to assume a Rayleigh distribution for the distance between the measurement $\hat{x}$ and the activity location x (Chen, 2013).

The domain of data relevance used for the generation of episode location (Section 4.2.1) is computed for each measurement based on the confidence square.

5.2. EPFL pedestrian network

The EPFL website proposes an orientation tool for the campus, `http://map.epfl.ch`. It provides locations of offices and points of interest (such as restaurants and classrooms) on campus. It also generates itineraries between two such locations. It consists of a semantically-enriched graph (SERG) as defined in Section 3.2, containing $|\mathcal{N}| = 50131$ nodes, $|\mathcal{E}| = 56655$ edges, and $|\mathrm{POI}| = 5387$ points of interest.

The network as described above corresponds to the minimum data requirement as defined in Section 3.2. However, more information is provided. Similarly to road networks for car driving, each edge is associated with a hierarchical status. Based on this hierarchical status, weights are defined in the routing tool of EPFL website as shown in Algorithm 2 in Appendix D. The higher the weight is, the less likely the link is to be selected for the shortest path.

5.3. Potential attractivity measure on campus

On campus, each point of interest, $x \in \mathrm{POI}$, belongs to one of six categories: offices, classrooms, restaurants, shops, library, and other points of interest. For each POI, we define attractivity $\mathcal{A}(x, t)$ depending on the category it belongs to.

For each office, attractivity is equal to the aggregate work rates of employees provided by the human resources management software. For classrooms, attractivity equals the number of students who subscribed for a course at the beginning of the semester. This number varies with the time of the day. For restaurants and the library, we use the number of seats as a proxy. For shops on campus, no information is available and we arbitrarily assume that attractivity corresponds to a capacity of 20 people. Finally, for all other points of interest, we arbitrarily assume an attractivity of one, since we have no information about it.

Time constraints $\delta_{x,i}(t)$ as defined in Section 4.1.2 are based on class schedules for classrooms, and opening hours for restaurants, the library and shops. For offices, we assume no time constraint, and thus $\delta_{x,i}(t) = 1 \; \forall t$.

5.4. Results

5.4.1. One of the authors knowing the real activity-episode sequence

The methodology presented in Section 4 is tested with traces from one of the authors. 76 measurements were generated on Monday May 14, 2012 (Figure 5).

With EPFL WiFi data, a confidence square is defined assuming that the device is located on the correct floor. In order to account for floor error, we also consider the below and top floors, using a square with side $2 \cdot r$ on these floors. We define F as the probability of being in the detected floor, and $\frac{1-F}{2}$ the probability of being on the below or top floor. Both r and F are not provided and must be fixed by the modeler.

With this definition and given the high density of potential episode locations in the pedestrian network (in particular for offices), the number of locations in DDR is large (a mean of 712.0 with $r = 25m$). The most distant episode locations of the DDR have a very low measurement likelihood. Decreasing the size of the DDR decreases the computation time. Moreover, very weak measurements generate huge DDRs and create a risk of giving too much importance to the prior. We define a maximum distance R in meters for taking potential episode locations in consideration, $cF_{trunc} = \min(cF, R)$.

The sequence of activity episodes has been recorded manually by the tracked author and is shown in Table 1 and Figure 6. He first went in a classroom from 8.32 to 10.30 am

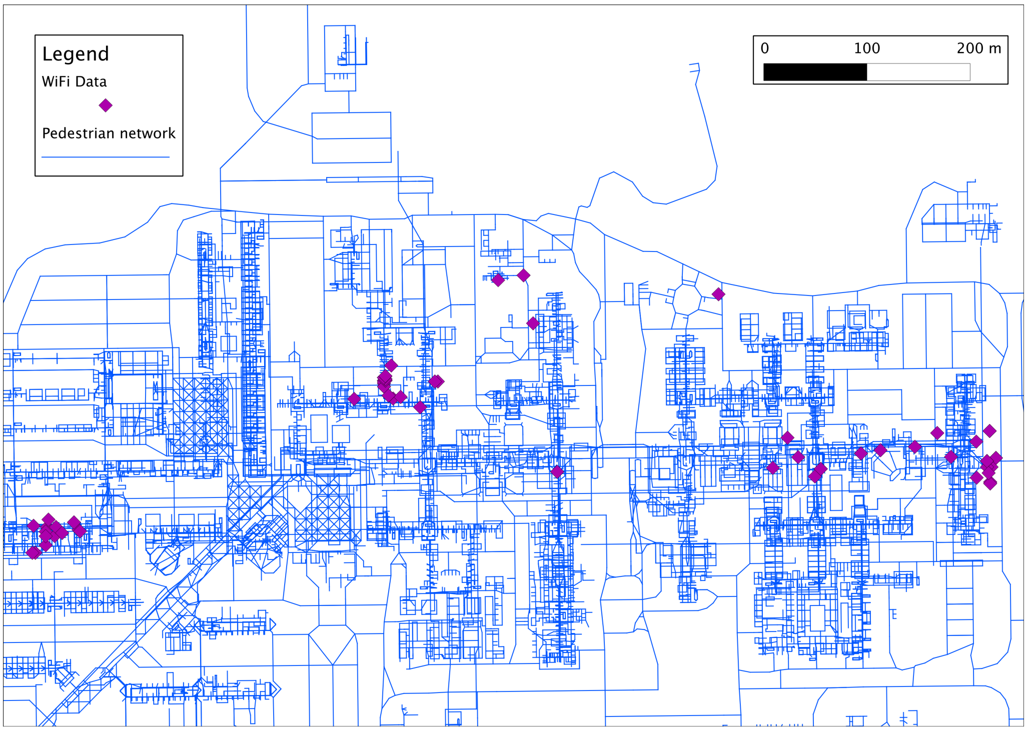

Figure 5: WiFi traces generated by one of the authors on Monday May 14, 2012 (violet) and the pedestrian network (blue).

for a course, then in his office until 11.47 am. For lunch break, he arrived in a restaurant on campus at 11.55 am. He came back to his office around 1 pm and went for a coffee around 2 pm. Finally he came back in his office until the end of his working day, around 7.45 pm.

Activity log

Time spent	*Floor*	*Location*
8.32am-10.30am	1	Classroom
Until 11.47am	3	Author's office
From 11.55 am	1	Restaurant
Around 1pm	3	Author's office
Around 2pm	2	Cafeteria
Until around 7.45pm	3	Author's office

Table 1: Sequence of activity episodes as reported by one author. It contains 6 activity episodes.

The results are presented in Table 2, with $R = 80m$, $r = 25m$ and $F = 0.6$. At each iteration, only the best candidate is kept ($L = 1$). The potential attractivity measure is using the individual disaggregate class schedules. Δx is the shortest walking path between the episode location from the model and the one from the activity log in the semantically-enriched routing graph.

Compared to the mean confidence factor $\overline{cF} = 124.2m$, the spatial error (Δx) is

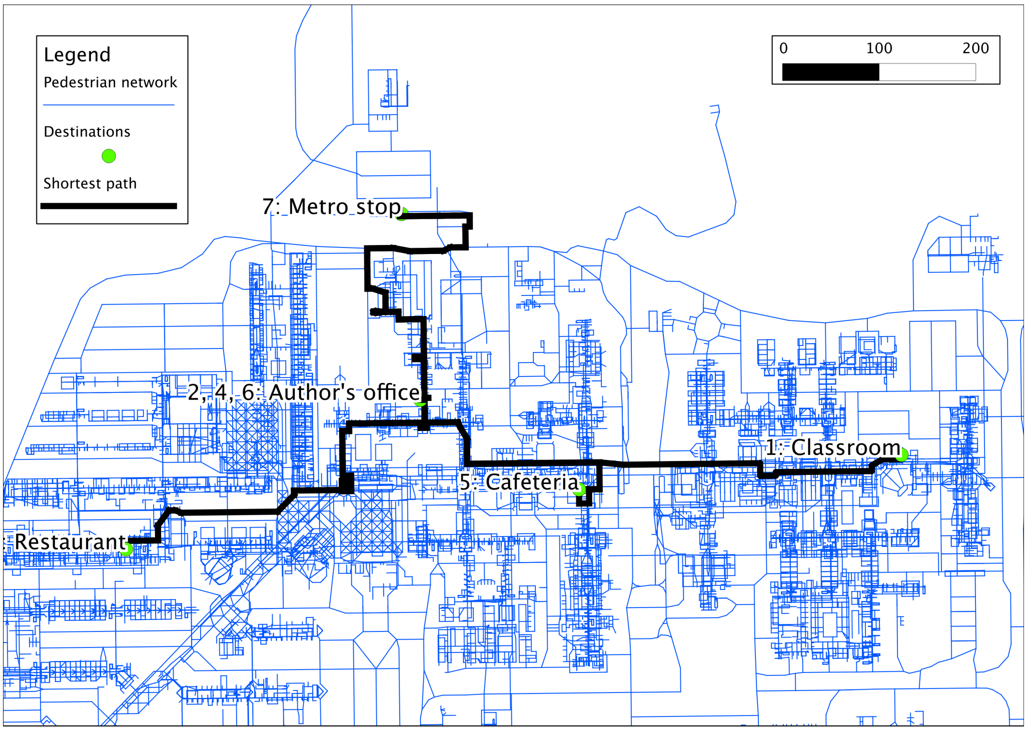

Figure 6: Sequence of activity episodes as reported by one author. The dots represent activity episodes. The black thick lines represent the weighted shortest paths presented in Section 5.2. They use the pedestrian network and connect the activity episodes.

Model with disaggregate prior				Activity log			Δx
Arrival time	*Departure time*	*Floor*	*Location*	*Time spent*	*Floor*	*Location*	*(in m.)*
8:35-8:35	10:38-10:38	1	Classroom	8.32am-10.30am	1	Classroom	0
10:40-10:40	11:51-11:51	3	Office	Until 11.47am	3	Author's office	9
11:54-11:54	12:47-12:53	1	Restaurant	From 11.55 am	1	Restaurant	0
12:51-12:58	13:03-13:44	3	Office	Around 1pm	3	Author's office	9
13:06-13:47	13:53-14:02	2	Cafeteria	Around 2pm	2	Cafeteria	0
13:55-14:04	19:40-19:44	3	Office	Until around 7.45pm	3	Author's office	9

Table 2: Comparison between the most likely output of the model and the activity log as reported by one author.

low. The last activity episode, the metro stop, is not covered by WiFi. It is at the border of the campus. Thus, the error is big in this case. 3 out of 7 activity episodes are perfectly detected, and 3 more have a correct category. The number of episodes is correctly detected, as well as the floor of each activity episode. The temporal precision seems coherent with the diary. Results are presented on a map in Figure 7.

As the precision of the WiFi data is low and attractivity measure does not perfectly correct for this imprecision, the number of candidates L in the sequence elimination procedure (Section 4.4) can be increased to represent this uncertainty. L must be defined by the analyst to balance between algorithm speed and representation of uncertainty in

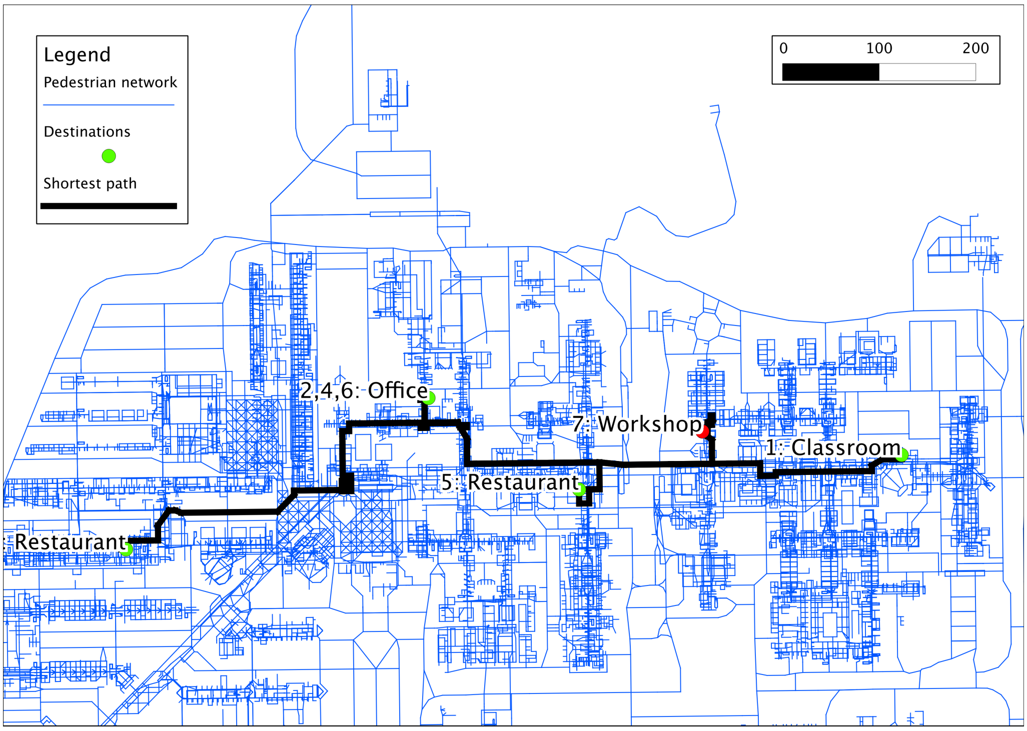

Figure 7: Activity-episode sequence of the most likely output of the model with disaggregate prior on EPFL Campus pedestrian network. Episode locations are connected with the weighted shortest path presented in Section 5.2. Destinations are represented in green if in the correct category of POI and red otherwise. Only the last destination, the metro stop, not covered by WiFi, does not have the correct category.

the data. Figure 8 shows results with $L = 100$ candidates. Some activity episodes are present in each of the 100 candidates, expressing the absence of ambiguity at this time of the day (episodes 3 and 5, restaurants and cafeteria). In other cases, a strong ambiguity, both in horizontal error and activity-episode category, is present (episode 1, classroom). Measuring this uncertainty allows for corrections in further analysis.

5.4.2. Individual and aggregate results for anonymous members of campus

The same methodology was applied to 3490 employees and 767 students of campus (with $L = 20$ and $F = 0.99$). Data were collected for 5 weekdays, between May 17 and May 23, 2012. Campus users authenticate themselves on the WiFi network through WPA (WiFi Protected Access) using a Radius server. Accounting is one of the process on the Radius server. It allows to associate a MAC address with a username (Koo et al., 2003). The username was associated with employee or class attribute through LDAP (Lightweight Directory Access Protocol) requests. Then, both the MAC address and the username were deleted to guarantee privacy. This process generates anonymized network traces with known category of users on campus.

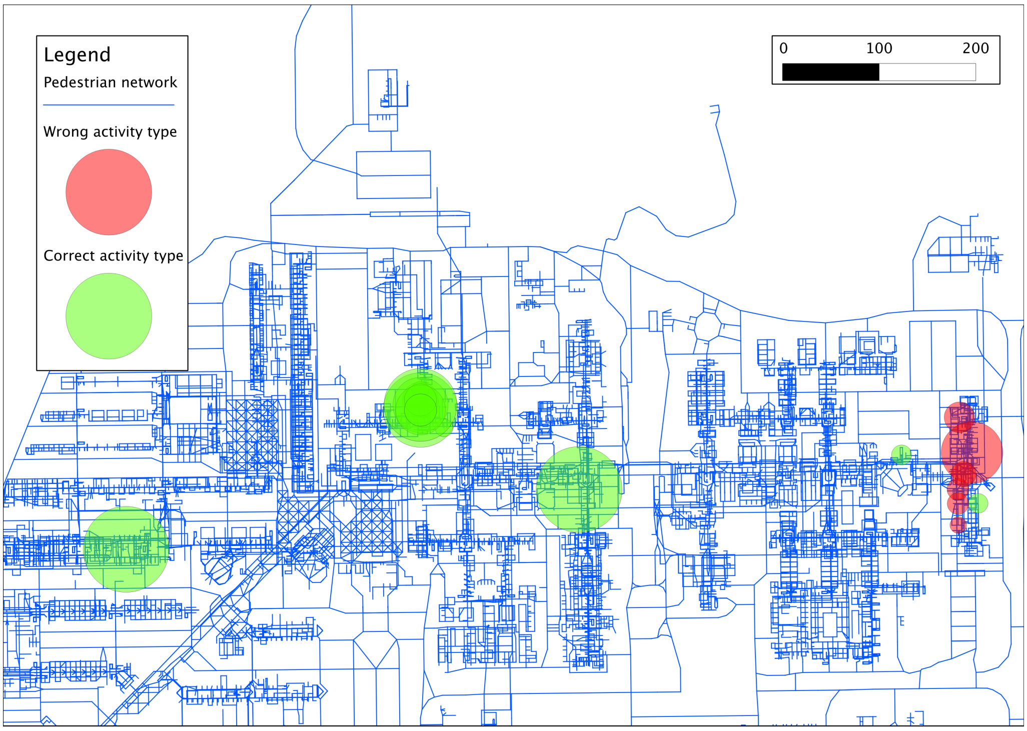

Figure 8: Activity-episode sequence of the $L = 100$ most likely output of the model with disaggregate prior on EPFL Campus pedestrian network. Destinations are represented in green if in the correct category of POI and red otherwise. The surface of each point represents the normalized probability of this destination being the correct one (Eq. 1). The two restaurants are detected in all 100 activity-episode sequences. The author's office is not always perfectly detected and variations can be observed, but the category is always correct. The classroom, in the beginning of the day, is not correctly detected, and the destination category is wrong in some cases. The actual classroom is detected in a minority of cases. In some of the 100 sequences, there is a seventh episode, but their likelihood is too small to be seen on the picture.

Figures 9, 10 and 11 show the activity patterns of two employees and one student. The POIs are aggregated per category in these figures. Figure 9 shows an arrival on campus between 8h05 and 8h10. The employee visits two offices first, then a restaurant, then an office and a lab in the morning. During lunch break, the employee visits two different restaurants. Then the employee visits a lab again, a restaurant, and finally the last episode is a lab with probability around 80% and an office with probability around 20%. Between 13h36 and 14h01, there is no destination where the measurements are stable for more than 5 min. This output seems realistic.

Figure 10 shows that WiFi devices are not necessarily mobile. Here, the device is accessing the WiFi all day long and not moving from one office. It is likely to be a fixed equipment.

A student's device activity pattern is shown on Figure 11. The student's device was in a restaurant during lunch break and following courses in the afternoon. The

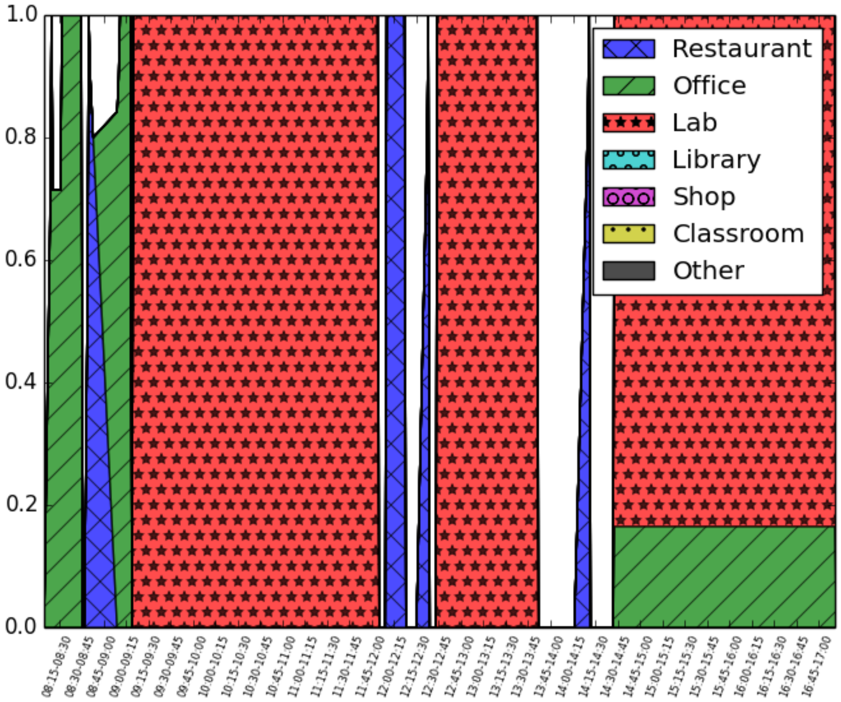

Figure 9: Activity pattern for one employee's device on May 23, 2012. The x-axis represents the time of the day. The colors/patterns represent the different categories of the points of interest. The y-axis is the probability to be the correct point of interest based on Eq. 1.

"other" activity type in the morning represents here the campus bike service. It is very likely to be a measurement error since there is a class two floors up and 10 m away. Our measurement equation does not take into account more than one floor error, so the actual classroom is not in the domain of data relevance. The limitation of the size of the domain of data relevance increases the speed of the algorithm but also excludes some points of interest that could be realistic.

At a more aggregate level, we can observe from the output of Algorithm 1 that people on campus are performing 3 activity episodes on average. At an average they spend 1 h and 37 min on each activity. Focusing on the restaurants, Figure 12 shows the number of devices detected in restaurants per quarter of an hour during the 5 weekdays. We observe a peak of transactions around noon, which is expected.

There are no data about the real behavior of people for validating the number of episodes, their duration or the proportion of people going to restaurants.

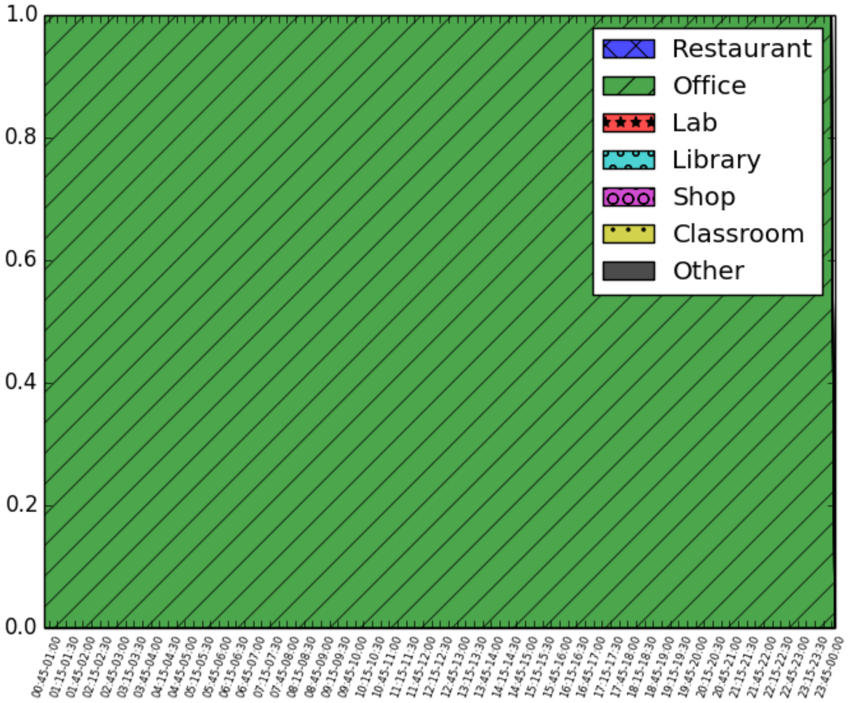

Figure 10: Activity pattern for one employee's device on May 23, 2012. The x-axis represents the time of the day. The colors/patterns represent the different categories of the points of interest. The y-axis is the probability to be the correct point of interest based on Eq. 1.

5.5. Sensitivity analysis

Based on results from one of the authors (Section 5.4.1), the sensitivity of the results to the parameters, the prior and the density of measurements are measured in terms of spatial and temporal precisions of each activity episode and more globally at the sequence level, quantitatively and qualitatively. Four criteria of stability are defined to evaluate the impact of the changes: the number of episodes that are detected by the algorithm ("Nb episodes"), the walking distance between the episode location from the model and the activity log one ("Delta dist.", in meters), the mean absolute difference between the activity log schedule and the schedule defined by the model ("Delta duration", in minutes), and the number of correct destinations categories ("Nb OK"). The reported walking distance, "Delta dist.", is the shortest path between the episode location from the model and the activity log one in the semantically-enriched routing graph. This criteria is more relevant than Euclidean distance since a small difference in localization may have a big impact on the actual distance in the pedestrian graph for the tracked individual. The reported episode start and end times are not very precise, in particular in the afternoon, and thus we only consider here the 5 start and end times with a sufficient

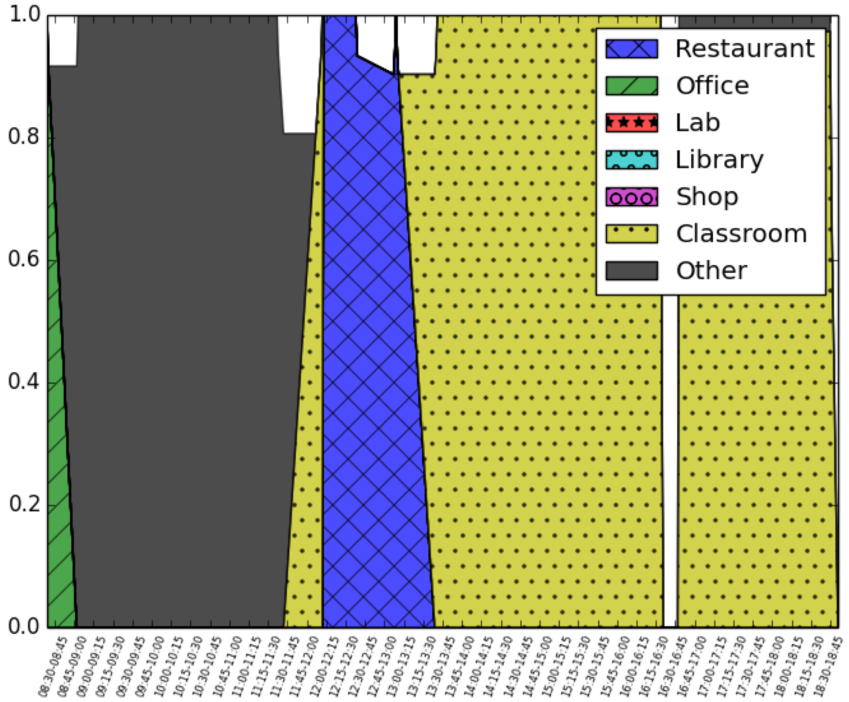

Figure 11: Activity pattern for one computer science master student's device on May 23, 2012. The x-axis represents the time of the day. The colors/patterns represent the different categories of the points of interest. The y-axis is the probability to be the correct point of interest based on Eq. 1.

precision in the activity log schedule: $t_1^-, t_1^+, t_2^+, t_3^-, t_6^+$. The number of correct categories is important, since the detection of the exact office is not necessary for understanding mobility patterns, while knowing the kind of destination is crucial.

In the next sections, the impact of the changes in different parameters (Section 5.5.1), the impact of the prior (Section 5.5.2) and the impact of the density of measurements (Section 5.5.3) are shown.

5.5.1. Sensitivity to the parameters

There are mainly four parameters that need to be defined in the model. First, the maximum radius R of the DDR, allowing to limit the computational burden related to some very imprecise measurements; second, the probability F of being in the detected floor. Since the precision is expressed in the horizontal plan, a vertical precision needs to be set up; third, we also define the minimum time spent at destination to T_{min} in minutes; and finally the number L of candidates that are kept during the sequence elimination procedure.

We use as a base case $R = 80m$, $F = 1.0$ and $T_{min} = 5min$, as in the previous results,

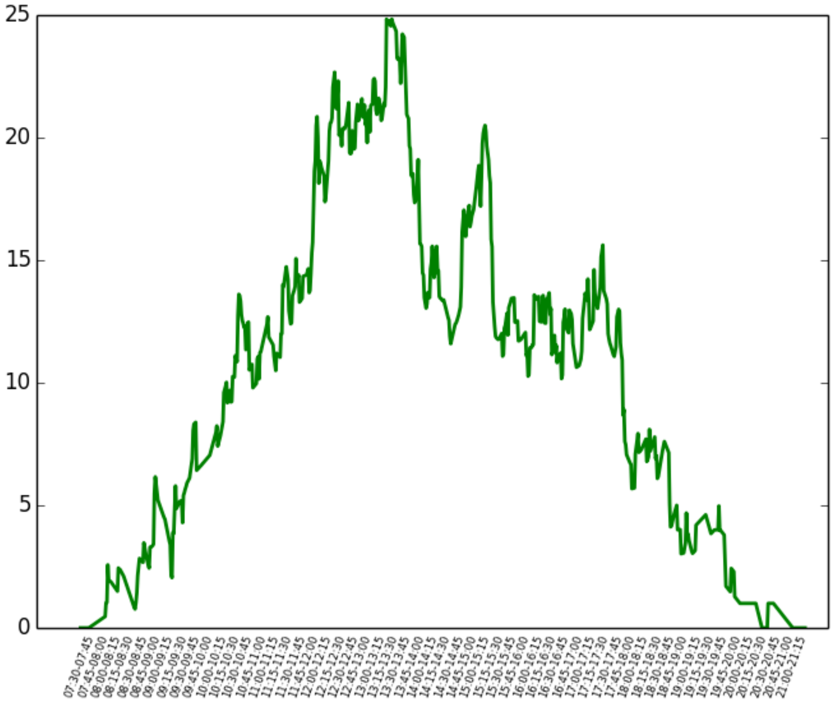

Figure 12: The x-axis represents the time of the day (quarters of hour). The y-axis represents the cumulated number of people detected in the restaurants of the campus over 5 days from the WiFi traces (devices/quarter). Since 20 activity-episode sequences are generated per individual, each one is weighted by its probability to be the correct one based on Eq. 1.

and we fix $L = 40$. The expected values for the criteria of stability are used based on the normalized activity probability from Eq 1.

Figure C.22 showed the impact of L on a small illustration, with $L = 1$ and 2. With real data, the same effect is appearing for the last episode. It shows than $L = 1$ should be avoided (Figure 13). By fixing the number L of candidates to 1 at each iteration, only the most likely last activity episode is kept at each measurement. This does not allow for explicit management of ambiguity of the measurement and does not provide memory to the process. With $L > 1$, results are stable.

Varying T_{min} defines the time length of intermediary measurements. For some large values of T_{min} (9, 10, 11, 12 min), it is possible that performing several activity episodes of less than T_{min} is more likely than performing the actual activity episode. It explains the low values for "Nb episodes" and "Nb OK" in Figure 14. Reasonable values for T_{min} represent the expected error in the travel time between two episodes, because of slower walking speed or longer distance than the shortest path (see Figure 15, i.e., less than 9 min. In these cases, results are stable. T_{min} should also be bigger than 1. In this

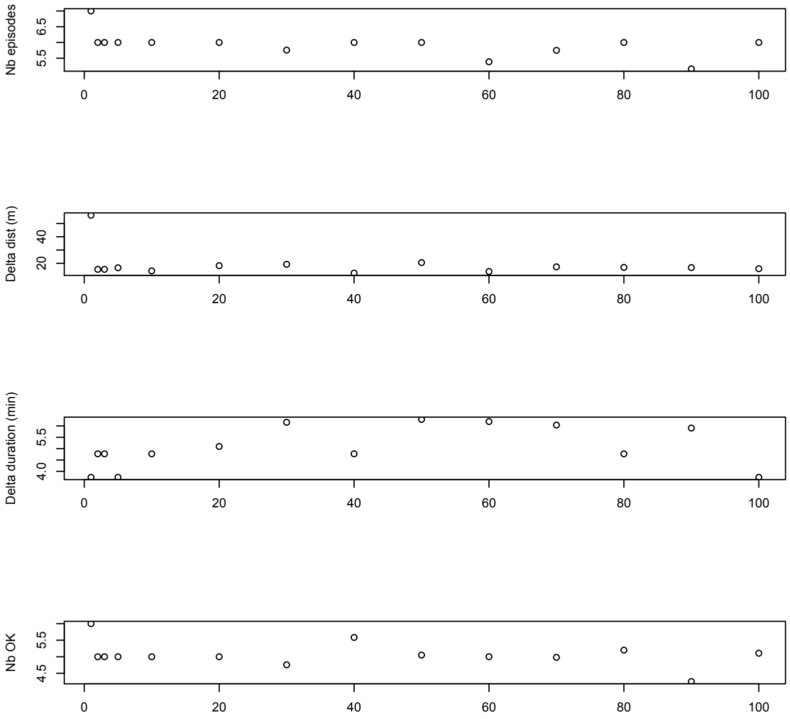

Figure 13: Sensitivity to the number L of candidates kept between each measurement, L = 1, 2, 3, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

case, very short and unrealistic activity episodes on the way between two actual activity episodes are more likely than the actual ones.

In our particular example, if R is small (R = 30, 40), the ambiguity for the first episode (see Figure 8) disappeared. It is case specific. In general, with small R, destinations are missed by the algorithm. For large values (R = 90, 100) or no limitation ($R = \infty$), the geographical information provided by the WiFi measurement is almost flat. In this case, the activity probability depends on prior only and the prior is bigger for less activity episodes. It explains the low number of activity episodes in Figure 16. The output is stable for R = 60, 70 and 80 m.

In our example, 10 of the 76 measurement are not on the correct floor (13%). Only one of them corresponds to an activity episode (the 9 others are measurements related to the metro stop, not covered). Figure 17 shows that the interfloor probability F has a very small impact in this specific example, if F is in the order of magnitude of the error (0.9). Still, the vertical imprecision in a multifloor environment must be taken into account. In particular, when the device is next to windows, stairwells or mezzanines, the

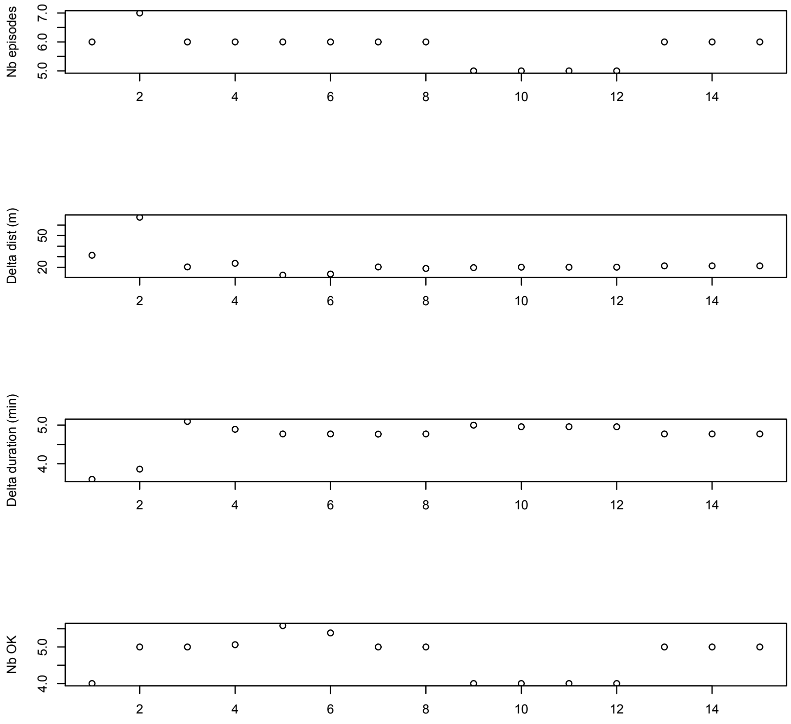

Figure 14: Sensitivity to the minimum time spent at destination T_{min}, $T_{min} = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15$ minutes.

signal could cross the floor separation.

As an extra example, one of the authors had a class on March 27, 2012. 2 of the 14 measurements corresponding to this activity episode are detected on the floor below the actual episode location. These two consecutive measurements happened in more than 5 min difference and thus are not considered as intermediary measurements. Moreover, they happen after the beginning of the activity episode. With $F = 1$, three activity episodes are detected: on the correct floor, then downstairs, then on the correct floor again. With $F = 0.9$, only one activity episode is detected.

5.5.2. Effect of prior

In Figure 18, we present results with the different priors defined in Section 4.1.2 to show their effects: (1) uniform; (2) aggregate for all campus members, using the same class attractivity for all students; (3) disaggregate for a class, meaning that we know the exact class schedule for the tracked pedestrian; and (4) diary, based on the recorded sequence presented in Table 1.

We observe first that the total number of episodes in the day is estimated correctly

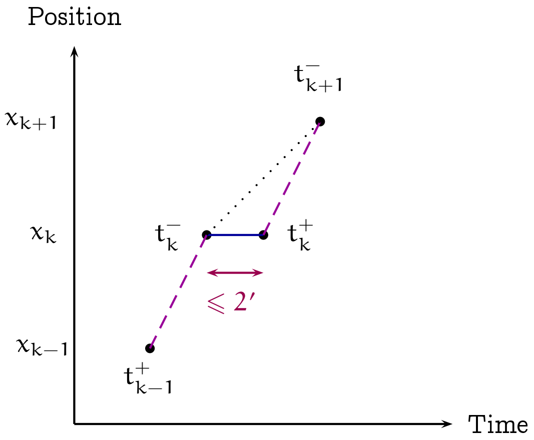

Figure 15: Time-space representation of one activity episode k with short time spent at it, $t_k^+ - t_k^- \leqslant 2$ minutes. The purple dashed line represents the assumed trip, with mean speed and shortest path. The dotted black line is the actual trip, with slower walking speed or longer distance than the shortest path.

from the 76 measurements with each different prior (Figure 18). Using a threshold of 5 min spent at episode locations, we reach the same number of episodes as the activity log. It means that without extra information, with a uniform prior, the WiFi data are already providing information about the number of episodes in the day. On the other hand, with a uniform prior, only 4 out of 7 activity episodes have correct category. This information is crucial for understanding and modeling activity choice.

We can observe that the aggregate prior is not precise enough (Figure 18). The number of episodes is stable, as well as the number of correct categories ("Nb OK") compared to a uniform prior, and spatial precision is worse. The aggregate prior does not improve the results. A deeper analysis of the results shows that including class schedules for all pedestrians, even those to whom these schedules are not relevant, is giving too much importance to classrooms compared to offices and other points of interest. It creates a bias towards classes by applying the same time constraints to everyone, even when these schedules are wrong for a particular individual.

Applying the correct time constraints needs class schedules, and thus lower anonymity level of the WiFi data. The disaggregate prior does not require the student identity but to which class the student belongs to. It detects almost all destinations perfectly, with correct categories. The individual anonymity is kept, while the attractivity and time constraints allow to correctly detect the category of the episode. Spatial error ("Delta dist.") is almost as good as the diary prior.

The diary prior allows to correctly detect only 6 out of the 7 activity episodes since the metro stop is not covered by WiFi and out of the confidence square. It is interesting to notice that exactly same results are reached with an attractivity of 3 for the visited points of interest for the diary prior and 1 for the non-visited POI. It shows the needed order of magnitude of the prior to overcome WiFi data imprecision in a pedestrian infrastructure with dense points of interest and detect correct categories.

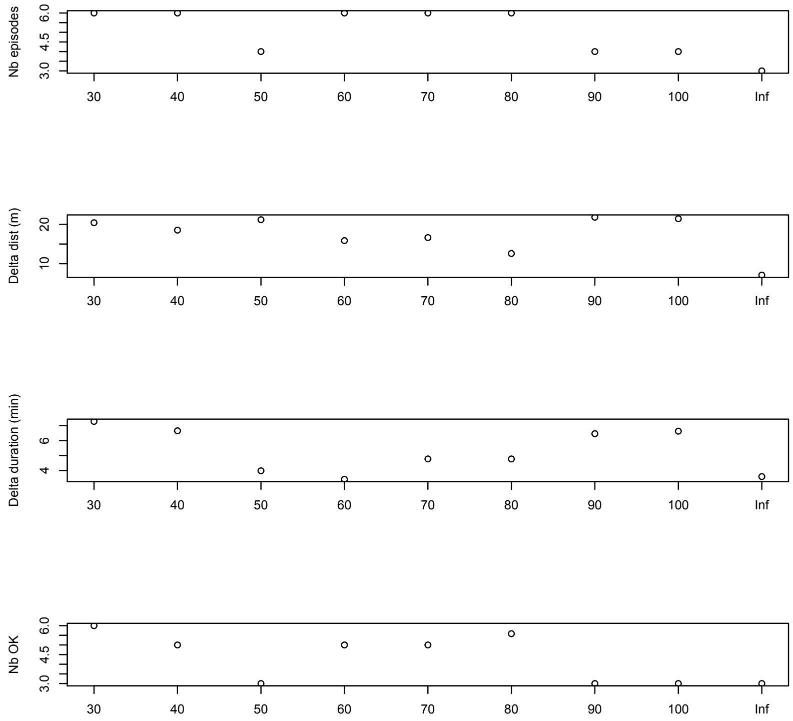

Figure 16: Sensitivity to the maximum radius R of the DDR, $R = 30, 40, 50, 60, 70, 80, 90, 100$ meters and $R = \infty$.

A prior with more information does not necessarily improve the results. Individual attractivity and time constraints allow to detect the correct categories of activity episodes and to reach a better spatial precision, while maintaining anonymity of the tracked pedestrians.

5.5.3. Sensitivity to the density of measurements

To evaluate the sensitivity of the model to the density of measurements, some data were artificially removed. 5%, 10%, 15%, 20%, 25% and 30% of the data were randomly removed. The model ran 100 times for each case. Results are shown in Figure 19. We observe that the model is very stable until -15% and then start to have variations, particularly in the difference of walking distance between the activity log and the model. This variation more frequently improves the difference of distance, which is good since the precision is better, but is less stable. Results with -25% and -30% of the full dataset show less stability, with variations in distances between the activity log and the models and also in the number of episodes and the number of correct destination categories.

As a general recommendation, 76 measurements cover properly an almost 12-h jour-

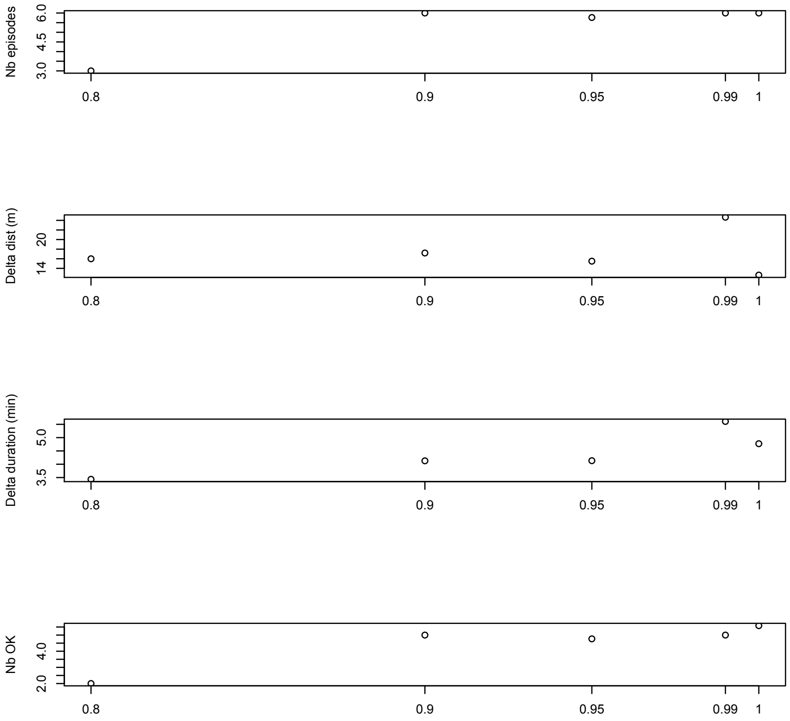

Figure 17: Sensitivity to the probability of being in the detected floor F, and not in the upper or lower floor, $F = 0.8, 0.9, 0.95, 0.99, 1.0$. F=1.0 means that only the detected floor is considered.

ney on a campus. Results are still stable and trustworthy with -15% of measurements, i.e. 65 measurements, which corresponds to a mean of 5.4 measurements per hour.

6. Conclusion

We propose a methodology detecting the different activity locations visited by a device using its network traces supported by knowledge of the underlying pedestrian map and attractivity, in particular time constraints. We present an empirical study on a campus.

Our approach accounts for the fact that pedestrian networks are traditionally denser than other mobility networks and localization is often sparse, in particular indoor. The methodology presented here is flexible and tunable. It allows for introducing *a priori* knowledge on the activities and information on the pedestrian map structure. In particular, time constraints (such as schedules for trains in a railway station, for planes in an airport, or for classes on a campus, or opening hours for shops or restaurants) can be added in the model. Moreover, the usage of a pedestrian network corrects for anisotropy

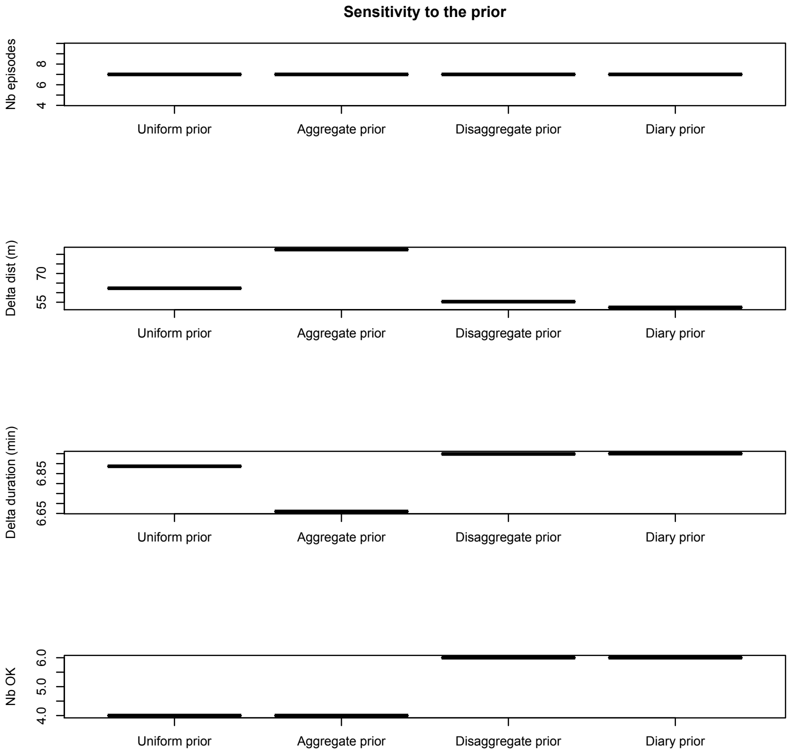

Figure 18: Sensitivity to the prior, with uniform, aggregate, disaggregate and diary prior.

in the facility. This methodology also uses the concept of domain of data relevance and avoids the traditional pingpong effect observed in other studies. Here, if access points are changing very often from one to another while the device is in fact static, the true activity location is contained in both domains of data relevance and does not change. This methodology is robust for low density measurements. Finally, ambiguity is explicitly stated through the likelihood of each activity-episode sequence.

This approach has limitations. First, it works in pedestrian facilities and does not account for mode detection. Also, we emphasize the importance of a good knowledge of the map behind the technical infrastructure. As results show, more data in the prior does not necessarily mean better results, and a careful definition of attractivity and time constraints is needed. Finally, R, the bound for the size of the DDR, must be fixed by the analyst and may cause a wrong number of detected episodes.

Future works involve applying this methodology with different sensors and in different contexts. It can be used with other network traces, such as Bluetooth tracking. In other contexts, such as train stations, hospitals, festivals or airports, attractivity measures and time constraints are different. Also, more data can be included in the model,

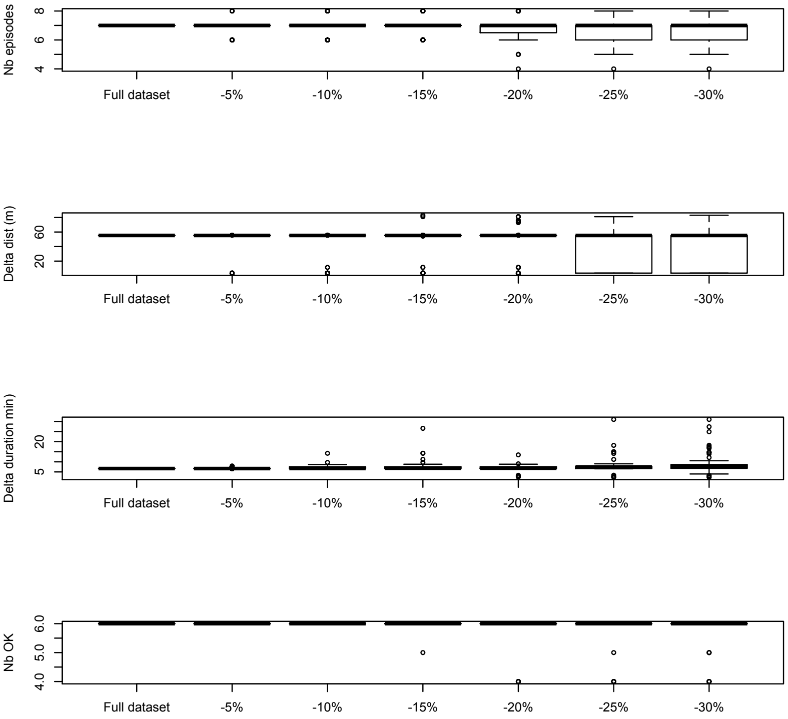

Figure 19: Sensitivity to datasets with less data, with the full dataset as a base case.

regarding the measurement equation, the prior, the shortest path, the first and last activity episodes, and the estimation of the parameters. The measurement equation may be improved by determining the source of propagation errors such as obstacles or walls. The prior could be further extended with models on activity choice or with more precise data about attractivity. The shortest path algorithm may describe big obstacles to increase length-optimality of the shortest path algorithm, or consider one-way paths similarly to street networks. The first and last episodes in the studied area are particular in the sense that they represent the access to the area. In our experiment, access to campus can be detected using prior knowledge, like studies about mode choice to access campus. The methodology can be improved by using the probabilistic map matching method developed by Bierlaire et al. (2013) for intermediary measurements. If exact activity-episode sequences are available for a large sample, e.g., from surveying pedestrians or from cameras, they can be used for Bayesian estimation of parameters such as F and r.

Merging different sources of data with our Bayesian approach can be used to estimate a destination choice model, e.g., by modeling activity pattern (Bowman and Ben-Akiva, 2001) or by using a dynamic discrete continuous choice model (e.g., Habib, 2011), taking into account the ambiguity in type, location and duration in the activity-episode sequence. Such models lead up to decision-aid tools for the evaluation and planning of pedestrian facilities.

Acknowledgment

This research is supported by the Swiss National Science Foundation Grant 200021-141099 “Pedestrian dynamics: flows and behavior”.

On the WiFi data collection side, the authors would like to thank Richard Timsit from the Network services team (DIT-TI) of EPFL’s Information Technology Domain (DIT), for his tremendous work and for his enthusiasm. We also would like to thank François Pujol, from Cisco, for his interest and for lending us the Context Aware Mobility API. On the map and pedestrian network side, we would like to thank Yves Bolognini, from Camptocamp, and Florent Deseneux, from EPFL’s Real Estate and Infrastructures Department (DII), in charge of map.epfl.ch, for sharing their knowledge of EPFL map.

References

Alahi, A., 2011. Vision-Based Scene Understanding with Sparsity Promoting Priors. Ph.D. thesis, EPFL.
URL http://dx.doi.org/10.5075/epfl-thesis-5070

Alahi, A., Bierlaire, M., Vandergheynst, P., Feb. 2014. Robust Real-time Pedestrians Detection in Urban Environments with a Network of Low Resolution Cameras. Transportation Research Part C 39, 113–128.
URL http://dx.doi.org/10.1016/j.trc.2013.11.019

Antonini, G., Bierlaire, M., Weber, M., Sep. 2006. Discrete choice models of pedestrian walking behavior. Transportation Research Part B 40 (8), 667–687.
URL http://dx.doi.org/10.1016/j.trb.2005.09.006

Arentze, T. A., Timmermans, H. J., Aug. 2004. A learning-based transportation oriented simulation system. Transportation Research Part B 38 (7), 613–633.
URL http://dx.doi.org/10.1016/j.trb.2002.10.001

Aschenbruck, N., Munjal, A., Camp, T., May 2011. Trace-based mobility modeling for multi-hop wireless networks. Computer Communications 34 (6), 704–714.
URL http://dx.doi.org/10.1016/j.comcom.2010.11.002

Balachandran, A., Voelker, G. M., Bahl, P., Rangan, P. V., Jun. 2002. Characterizing user behavior and network performance in a public wireless LAN. ACM SIGMETRICS Performance Evaluation Review 30 (1), 195.
URL http://dx.doi.org/10.1145/511399.511359

Balazinska, M., Castro, P., 2003. Characterizing mobility and network usage in a corporate wireless local-area network. In: Proceedings of the 1st international conference on Mobile systems, applications and services - MobiSys ’03. MobiSys ’03. ACM Press, New York, New York, USA, pp. 303–316.
URL http://dx.doi.org/10.1145/1066116.1066127

Balmer, M., Axhausen, K., Nagel, K., Jan. 2006. Agent-Based Demand-Modeling Framework for Large-Scale Microsimulations. Transportation Research Record 1985 (1), 125–134.
URL http://dx.doi.org/10.3141/1985-14

Bekhor, S., Cohen, Y., Solomon, C., Jun. 2013. Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. Journal of Advanced Transportation 47 (4), 435–446.
URL http://dx.doi.org/10.1002/atr.170

Ben-Akiva, M., Bowman, J., Gopinath, D., Aug. 1996. Travel demand model system for the information era. Transportation 23 (3), 241–266.
URL http://dx.doi.org/10.1007/BF00165704

Bierlaire, M., Chen, J., Newman, J., Jan. 2013. A probabilistic map matching method for smartphone GPS data. Transportation Research Part C 26, 78–98.
URL http://dx.doi.org/10.1016/j.trc.2012.08.001

Bierlaire, M., Chen, J., Newman, J. P., 2010. Modeling Route Choice Behavior From Smartphone GPS data. Tech. Rep. TRANSP-OR 101016, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.
URL http://transp-or.epfl.ch/documents/technicalReports/BierChenNewm10.pdf

Bierlaire, M., Frejinger, E., Apr. 2008. Route choice modeling with network-free data. Transportation Research Part C 16 (2), 187–198.
URL http://dx.doi.org/10.1016/j.trc.2007.07.007

Bierlaire, M., Robin, T., 2009. Pedestrians Choices. In: Timmermans, H. (Ed.), Pedestrian Behavior. Models, Data Collection and Applications. Emerald Group Publishing Limited, pp. 1–26.

Borgers, A., Timmermans, H., Sep. 1986. A Model of Pedestrian Route Choice and Demand for Retail Facilities within Inner-City Shopping Areas. Geographical Analysis 18 (2), 115–128.
URL http://dx.doi.org/10.1111/j.1538-4632.1986.tb00086.x

Borst, H. C., Miedema, H. M., de Vries, S. I., Graham, J. M., van Dongen, J. E., Dec. 2008. Relationships between street characteristics and perceived attractiveness for walking reported by elderly people. Journal of Environmental Psychology 28 (4), 353–361.
URL http://dx.doi.org/10.1016/j.jenvp.2008.02.010

Bowman, J., Ben-Akiva, M., Jan. 2001. Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A 35 (1), 1–28.
URL http://dx.doi.org/10.1016/S0965-8564(99)00043-9

Buchmüller, S., Weidmann, 2006. Parameters of pedestrians, pedestrian traffic and walking facilities. Tech. rep., IVT-Report Nr. 132, Institut for Transport Planning and Systems (IVT), ETHZ.

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., Apr. 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing 10 (4), 36–44.
URL http://dx.doi.org/10.1109/MPRV.2011.41

Calabrese, F., Reades, J., Ratti, C., Jan. 2010. Eigenplaces: Segmenting Space through Digital Signatures. IEEE Pervasive Computing 9 (1), 78–84.
URL http://dx.doi.org/10.1109/MPRV.2009.62

Chen, C., Gong, H., Lawson, C., Bialostozky, E., Dec. 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A 44 (10), 830–840.
URL http://dx.doi.org/10.1016/j.tra.2010.08.004

Chen, J., 2013. Modeling route choice behavior using smartphone data. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
URL http://dx.doi.org/10.5075/epfl-thesis-5649

Cisco, Apr. 2011. Cisco MSE API Specification Guide - Location Service, Release 7.1. Tech. rep., Cisco.

Conti, M., Giordano, S., Apr. 2007. Multihop Ad Hoc Networking: The Theory. IEEE Communications Magazine 45 (4), 78–86.
URL http://dx.doi.org/10.1109/MCOM.2007.343616

Cui, J., Allan, A., a.P. Taylor, M., Lin, D., Nov. 2012. The perception of accessibility and ease of orientation of underground pedestrian systems: a survey in Shanghai. International Journal of Urban Sciences 16 (3), 301–320.

Duives, D. C., Daamen, W., Hoogendoorn, S. P., Mar. 2013. State-of-the-art crowd motion simulation models. Transportation Research Part C 31.
URL http://dx.doi.org/10.1016/j.trc.2013.02.005

Eash, R., 1999. Destination and Mode Choice Models for Nonmotorized Travel. Transportation Research Record: Journal of the Transportation Research Board 1674, 1–8.

Ettema, D., Borgers, A., Timmermans, H. J. P., 1994. Using interacting computer experiments for identifying activity scheduling heuristics. In: 7th IATBR Conference.

Etter, V., Kafsi, M., Kazemi, E., 2012. Been There , Done That: What Your Mobility Traces Reveal about Your Behavior. In: Nokia Mobile Data Challenge 2012 Workshop. June 18-19, Newcastle, UK, pp. 1–6.
URL http://pixlr.ch/wp-content/uploads/2012/12/been_there_done_that.pdf

Feil, M., 2010. Choosing the Daily Schedule: Expanding Activity-Based Travel Demand Modelling.

Ph.D. thesis, ETH.
URL http://dx.doi.org/10.3929/ethz-a-006200573

Frignani, M. Z., Auld, J., Mohammadian, A. K., Williams, C., Nelson, P., 2010. Urban Travel Route and Activity Choice Survey (UTRACS): An Internet-Based Prompted Recall Activity Travel Survey using GPS Data. Transportation Research Record: Journal of the Transportation Research Board 2183, 19–28.

Goetz, M., Jun. 2012. Using Crowdsourced Indoor Geodata for the Creation of a Three-Dimensional Indoor Routing Web Application. Future Internet 4 (4), 575–591.
URL http://dx.doi.org/10.3390/fi4020575

Goetz, M., Zipf, A., Jan. 2011. Formal definition of a user-adaptive and length-optimal routing graph for complex indoor environments. Geo-spatial Information Science 14 (2), 119–128.
URL http://dx.doi.org/10.1007/s11806-011-0474-3

González, M. C., Hidalgo, C. A., Barabási, A.-L., Jun. 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–82.
URL http://dx.doi.org/10.1038/nature06958

Goulias, K. G., Pendyala, R. M., Bhat, C. R., 2013. KeynoteTotal Design Data Needs for the New Generation Large-Scale Activity Microsimulation Models. Transport Survey Methods: Best Practice for Decision Making, 21–45.

Habib, K. M. N., 2007. Modelling activity generation processes. Ph.D. thesis, University of Toronto.

Habib, K. M. N., Jul. 2011. A random utility maximization (RUM) based dynamic activity scheduling model: Application in weekend activity scheduling. Transportation 38 (1), 123–151.
URL http://dx.doi.org/10.1007/s11116-010-9294-9

Hägerstraand, T., Jan. 1970. What about people in regional science. Papers in Regional Science 24 (1), 7–24.
URL http://dx.doi.org/10.1111/j.1435-5597.1970.tb01464.x

Hansen, W. G., May 1959. How Accessibility Shapes Land Use. Journal of the American Institute of Planners 25 (2), 73–76.
URL http://dx.doi.org/10.1080/01944365908978307

Kang, J. H., Welbourne, W., Stewart, B., Borriello, G., 2004. Extracting places from traces of locations. In: Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots - WMASH '04. WMASH '04. ACM Press, New York, New York, USA, p. 110.
URL http://dx.doi.org/10.1145/1024733.1024748

Kasemsuppakorn, P., Karimi, H. A., Jan. 2013. A pedestrian network construction algorithm based on multiple GPS traces. Transportation Research Part C 26 (0), 285–300.
URL http://dx.doi.org/10.1016/j.trc.2012.09.007

Kneidl, A., Hartmann, D., Borrmann, A., Apr. 2013. A hybrid multi-scale approach for simulation of pedestrian dynamics. Transportation Research Part C.
URL http://dx.doi.org/10.1016/j.trc.2013.03.005

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. The MIT Press, Cambridge, MA.

Koo, S. G. M., Rosenberg, C., Chan, H.-H., Lee, Y. C., 2003. Location discovery in enterprise-based wireless networks: case studies and applications. Annales des T{é}l{é}communications (Annals of Telecommunications) 58 (3-4), 531–552.

Meneses, F., Moreira, A., Nov. 2012. Large scale movement analysis from WiFi based location data. In: 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE, pp. 1–9.
URL http://dx.doi.org/10.1109/IPIN.2012.6418885

Miller, H. J., Sep. 2010. Measuring Space-Time Accessibility Benefits within Transportation Networks: Basic Theory and Computational Procedures. Geographical Analysis 31 (1), 1–26.
URL http://dx.doi.org/10.1111/j.1538-4632.1999.tb00408.x

Nagel, K., Barrett, C. L., 1997. Using Microsimulation Feedback For Trip Adaptation For Realistic Traffic In Dallas. International Journal of Modern Physics C 8 (3), 505–525.
URL http://dx.doi.org/10.1142/S0129183197000412

Rieser-Schüssler, N., Apr. 2012. Capitalising modern data sources for observing and modelling transport behaviour. Transportation Letters: The International Journal of Transportation Research 4 (2), 115–128.
URL http://dx.doi.org/10.3328/TL.2012.04.02.115-128

Rindfüser, G., Mühlhans, H., Doherty, S. T., Beackmann, K. J., 2003. Tracing the planning and execution of activities and their attributes: Design and application of a hand-held scheduling process survey. In: 10th International Conference on Travel Behaviour Research. August 10-14, Lucerne, Switzerland, pp. 1–31.

Rojas, A., Branch, P., Armitage, G., 2005. Experimental validation of the random waypoint mobility model through a real world mobility trace for large geographical areas. In: Proceedings of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems - MSWiM '05. MSWiM '05. ACM Press, New York, New York, USA, p. 174.
URL http://dx.doi.org/10.1145/1089444.1089474

Roorda, M. J., 2005. Activity-based modelling of household travel. Ph.D. thesis, University of Toronto.

Roorda, M. J., Miller, E. J., Habib, K. M., Feb. 2008. Validation of TASHA: A 24-h activity scheduling microsimulation model. Transportation Research Part A 42 (2), 360–375.
URL http://dx.doi.org/10.1016/j.tra.2007.10.004

Sevtsuk, A., Huang, S., Calabrese, F., Ratti, C., 2009. Mapping the MIT campus in real time using WiFi. IGI Global, Hershey, PA, Ch. XXII, pp. 326–338.
URL http://dx.doi.org/10.4018/978-1-60566-152-0.ch022

Tang, D., Baker, M., 2000. Analysis of a local-area wireless network. In: Proceedings of the 6th annual international conference on Mobile computing and networking. MobiCom '00. ACM, New York, NY, USA, pp. 1–10.
URL http://dx.doi.org/10.1145/345910.345912

Taniguchi, Y., Sasabe, M., Aihara, S., Nakano, H., Jan. 2013. Bayesian estimation-based pedestrian tracking in microcells. The Scientific World Journal 2013, 1–10.
URL http://dx.doi.org/10.1155/2013/187479

Timmermans, H., der Hagen, X. V., Borgers, A., 1992. Transportation systems, retail environments and pedestrian trip chaining behaviour: modelling issues and applications. Transportation Research Part B 26B (1), 45–59.
URL http://dx.doi.org/10.1016/0191-2615(92)90019-S

Tuduce, C., Gross, T., Mar. 2005. A mobility model based on WLAN traces and its validation. In: INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. Vol. 1. pp. 664 – 674 vol. 1.
URL http://dx.doi.org/10.1109/INFCOM.2005.1497932

Wanalertlak, W., Lee, B., Yu, C., Kim, M., Park, S.-M., Kim, W.-T., 2011. Behavior-based mobility prediction for seamless handoffs in mobile wireless networks. Wireless Networks 17 (3), 645–658.
URL http://dx.doi.org/10.1007/s11276-010-0303-x

Weibull, J. W., Dec. 1976. An axiomatic approach to the measurement of accessibility. Regional Science and Urban Economics 6 (4), 357–379.
URL http://dx.doi.org/10.1016/0166-0462(76)90031-4

Weibull, J. W., 1980. On the numerical measurement of accessibility. Environment and Planning A 12 (1), 53–67.
URL http://dx.doi.org/10.1068/a120053

Weidmann, U., Kirsch, U., Schreckenberg, M. E. (Eds.), 2014. Pedestrian and Evacuation Dynamics 2012. Springer.

Yoon, J., Noble, B. D., Liu, M., 2006. Building realistic mobility models from coarse-grained traces. In: in Proc. MobiSys. ACM Press, pp. 936–5983.
URL http://dx.doi.org/10.1145/1134680.1134699

Zhu, W., Timmermans, H., 2005. Exploring Pedestrian Shopping Decision Processes - an Application of Gene Expression Programming. In: Waldau, N., Gatterman, P., Knoflacher, H., Schreckenberg, M. (Eds.), Pedestrian and Evacuation Dynamics. Springer, pp. 145–154.
URL http://dx.doi.org/10.1007/978-3-540-47064-9_13

Zola, E., Barcelo-Arroyo, F., 2011. A comparative analysis of the user behavior in academic WiFi networks. In: Proceedings of the 6th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks. PM2HW2N '11. ACM, New York, NY, USA, pp. 59–66.
URL http://dx.doi.org/10.1145/2069087.2069096

Appendix A. Derivation of the distribution of t^-_{i+1}

Let's first assume $\hat{t}_i \leqslant \hat{t}_{i+1} - tt_{x_i,x_{i+1}}$ and $t^+_i + tt_{x_i,x_{i+1}} \leqslant \hat{t}_{i+1}$ to get rid of the maximum and the minimum in the bounds of the intervals, and thus simplify the notation (if these two conditions are not met, the two random variables t^+_i and t^-_{i+1} are fixed and the derivation is obvious).

The end time t^+_i is uniformly distributed, $t^+_i \sim U(\hat{t}_i, \hat{t}_{i+1} - tt_{x_i,x_{i+1}})$, with density function $f_{t^+_i}(x) = \frac{1}{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i}$ for $x \in [\hat{t}_i, \hat{t}_{i+1} - tt_{x_i,x_{i+1}}]$ and 0 otherwise. The start time t^-_{i+1} is uniformly distributed between $t^+_i + tt_{x_i,x_{i+1}}$ and $\hat{t}_{i+1}$. Its density for a given value of t^+_i is $f_{t^-_{i+1}|t^+_i=x}(y) = \frac{1}{\hat{t}_{i+1} - t^+_i - tt_{x_i,x_{i+1}}}$ for $y \in [x + tt_{x_i,x_{i+1}}, \hat{t}_{i+1}]$ and 0 otherwise. Now, the density of t^-_{i+1} is:

$$\begin{aligned}
f_{t^-_{i+1}}(y) &= \int_{x=\hat{t}_i}^{\hat{t}_{i+1} - tt_{x_i,x_{i+1}}} f_{t^-_{i+1}|t^+_i=x}(y) \cdot f_{t^+_i}(x)dx && \text{(A.1)}\\
&= \int_{x=\hat{t}_i}^{y - tt_{x_i,x_{i+1}}} f_{t^-_{i+1}|t^+_i=x}(y) \cdot f_{t^+_i}(x)dx && \text{(A.2)}\\
&= \int_{x=\hat{t}_i}^{y - tt_{x_i,x_{i+1}}} \frac{1}{\hat{t}_{i+1} - x - tt_{x_i,x_{i+1}}} \cdot \frac{1}{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i} dx && \text{(A.3)}\\
&= \frac{1}{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i} \ln\left(\frac{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i}{\hat{t}_{i+1} - t^-_{i+1}}\right) && \text{(A.4)}
\end{aligned}$$

The modification of the upper bound of the integral between Eq. A.1 and Eq. A.2 is explained by the support of y: $y \in [x + tt_{x_i,x_{i+1}}, \hat{t}_{i+1}]$, i.e., $x + tt_{x_i,x_{i+1}} \leqslant y$. Note that $x \leqslant y - tt_{x_i,x_{i+1}} \leqslant \hat{t}_{i+1} - tt_{x_i,x_{i+1}}$.

Expected value is $E(t^-_{i+1}) = \frac{\hat{t}_i + tt_{x_i,x_{i+1}} + 3 \cdot \hat{t}_{i+1}}{4}$.

Appendix B. Tree representation of the generation of activity-episode sequences and of the sequence elimination procedure

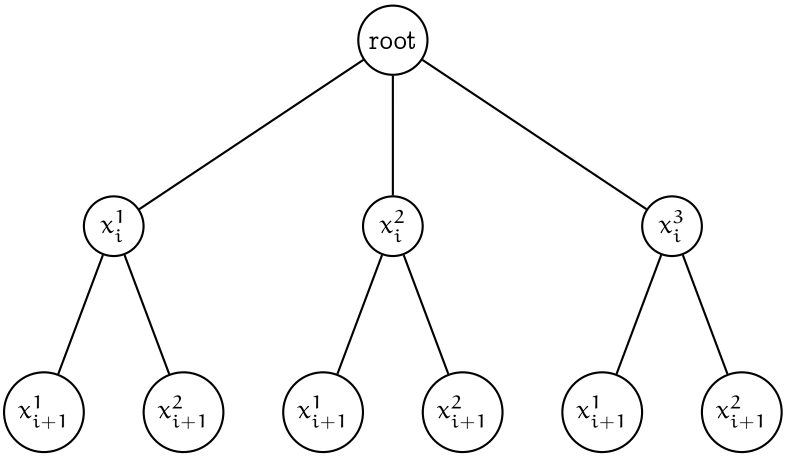

Figure B.20: A tree representation of the network corresponding to the two measurements in Figure 2. Each path from the root to a leaf of the tree represents a possible activity-episode sequence.

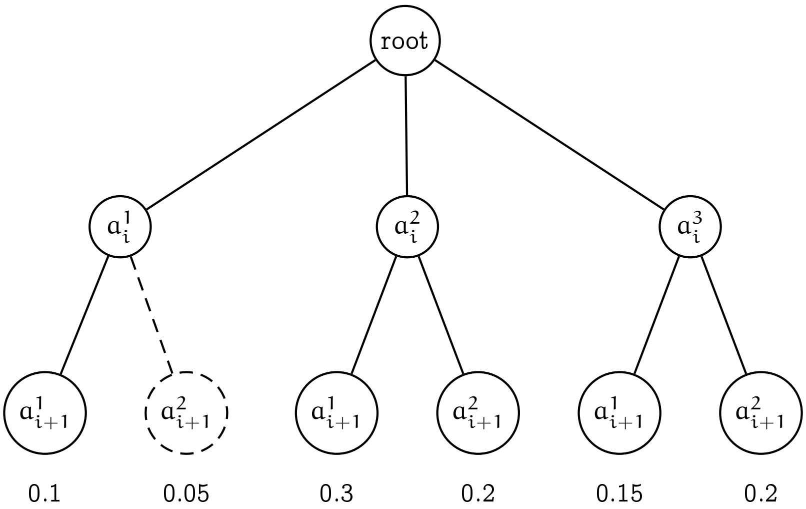

Figure B.21: The tree of Figure B.20 with the probability for each leaf to be the correct one. If $L = 5$, we eliminate the candidate represented by the dotted line, as it is associated with the lowest probability.

Appendix C. Illustration of the methodology

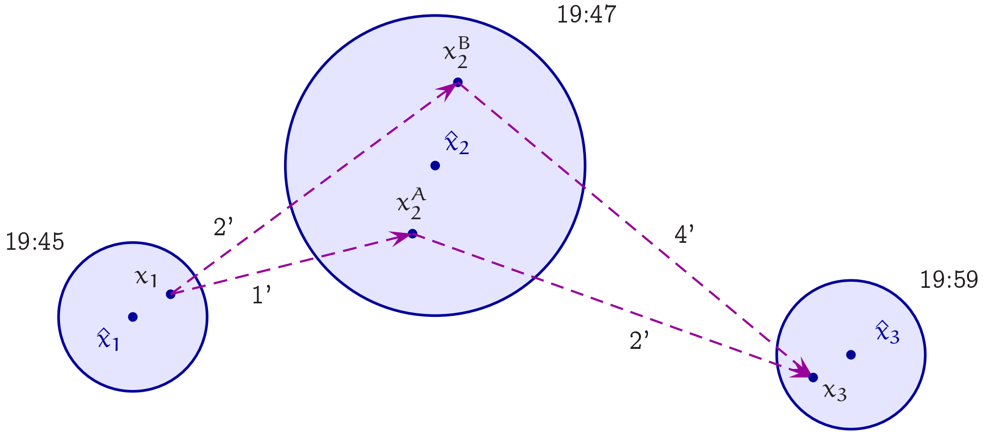

Figure C.22: The first measurement $\hat{s}_1 = (\hat{x}_1, \hat{t}_1)$ takes place at $\hat{t}_1 = 19{:}45$. Its DDR contains only one POI, x_1. It is connected with the two elements x_2^A and x_2^B of the DDR of measurement $\hat{s}_2 = (\hat{x}_2, \hat{t}_2 = 19{:}47)$. At 19:47, there are two candidates, (a_1, a_2^A) and (a_1, a_2^B). Let's assume a_2^A is twice more attractive than a_2^B, and x_2^A and x_2^B are in the same distance of the measurement location $\hat{x}_2$. Thus, (a_1, a_2^A) is twice more likely than (a_1, a_2^B). If $L = 1$, only (a_1, a_2^A) is kept, and then associated with x_3. With travel times from the picture, start and end times for a_2^A are generated: $t_2^- \sim U(19{:}45 + 1', 19{:}47)$, $t_2^+ \sim U(19{:}47, 19{:}59 - 2')$. Estimated time spent at a_2^A is 5'30. If $L = 2$, both (a_1, a_2^A) and (a_1, a_2^B) are kept and associated with x_3. Then, start and end times for a_2^B are generated: $t_2^- \sim U(19{:}45 + 2', 19{:}47)$, $t_2^+ \sim U(19{:}47, 19{:}59 - 4')$, for an expected time spent at a_2^B of 4'. As it is less than 5', a_2^B is eliminated and the two candidates are now (a_1, a_2^A, a_3) and (a_1, a_3). (a_1, a_3) is the most likely sequence since the measurement likelihoods are the same but the priors are $P(a_1, a_3) > P(a_1, a_2^A, a_3)$.

Appendix D. Weighted shortest path algorithm

Algorithm 2: Weight definition procedure for each edge in the pedestrian network

```
if door = closed then
    weight = ∞;
else
    if Major Route then
        hierarchical factor = 1;
    else if Inter-building Route then
        hierarchical factor = 1.2;
    else if Intra-building Route then
        hierarchical factor = 1.5;
    else if Access to Offices then
        hierarchical factor = 2.0;

    floor factor = 1;
    if Up then
        if Ramp then
            floor factor = 3;
        if Stairs then
            floor factor = 15;
    if Down then
        if Ramp then
            floor factor = 2;
        if Stairs then
            floor factor = 12;

    lift factor = 0;
    if Elevator then
        elevator factor = 40;

    weight = length · hierarchical factor · floor factor + elevator factor;
```

The length between floors is defined as being 0.5 meters in our pedestrian network. It explains why the floor factor is quite high compared to the hierarchical factor. In the case of elevators, the length is 0 and this is why we add 40 in the final weight formula.