Multi-class speed-density relationship for pedestrian traffic

Marija Nikolić * Michel Bierlaire *
Matthieu de Lapparent * Riccardo Scarinci *

January 15, 2017

Report TRANSP-OR 170115
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

*Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Switzerland,
{marija.nikolic, michel.bierlaire, matthieu.delapparent, riccardo.scarinci}@epfl.ch

Abstract

We introduce a modeling approach for pedestrian speed-density relationship. It is motivated by a high scatter in real data that precludes the use of traditional equilibrium relationships. To characterize the observed pattern we relax the homogeneity assumption of equilibrium relations and propose a multi-class model. In addition to the general modeling framework, we also present some concrete model specifications. Real data is utilized to test the performance of the approach. The approach is able to reveal fundamental properties causing the heterogeneity in population and describe their impact on pedestrian movement. We also show the advantages of the proposed approach compared to approaches from the literature. The proposed model is flexible, and it provides better fit and richer information than traditional models.

Keywords: pedestrian traffic, speed-density relationship, heterogeneity, latent class model, individual trajectories

1 Introduction

The analysis of pedestrian facilities has recently gained a lot of interest due to intense urbanization. The examples range from transportation hubs, such as airports (Kalakou et al., 2014) and train stations (Daamen, 2004; Van den Heuvel and Hoogenraad, 2014; Hänseler et al., 2014), to museums (Yoshimura et al., 2014; Kanda et al., 2007), music festivals (Naini et al., 2011; Duives et al., 2014), commercial centers (Lam and Cheung, 2000; Yaeli et al., 2014), university campuses (Danalet et al., 2014), crosswalks (Lam et al., 2002; Rastogi et al., 2013) or even religious infrastructures (Algadhi and Mahmassani, 1990; Helbing et al., 2007). A sophisticated understanding and modeling of complex pedestrian movement patterns (Bierlaire and Robin, 2009) is necessary for (i) an efficient planning and management of future pedestrian facilities and (ii) the optimization of current infrastructure and operations.

Data on pedestrian movement is traditionally collected by revealed and stated preferences surveys, manual counting, mechanical counting and infrared beans (Bauer et al., 2009). Thanks to advances in modern technologies (e.g. WiFi, Bluetooth, visual, depth and infrared sensors) a number of comprehensive empirical analysis (Hoogendoorn and Daamen, 2004, Helbing et al., 2005) and pedestrian traffic models (Duives et al., 2013, Hänseler et al., 2014, Hoogendoorn et al., 2014, Hänseler et al., 2017) have been reported in the literature in the last decades.

The fundamental variables characterizing the traffic of pedestrians are density (k), flow (q) and velocity (v). *Density* is expressed as the number of pedestrians

per unit of space at a given moment in time; *flow* is interpreted as the number of pedestrians per unit of time and per unit of length; *velocity* is expressed in unit of length per unit of time. The relationships between density and flow, density and speed, and flow and speed are referred to as the fundamental relationships. They play an important role in planning and designing pedestrian facilities. They are also central in the specification of models for pedestrian dynamics. These relationships are derived from the assumption that the traffic system is at equilibrium, that is stationary and homogenous.

The empirical analysis of real data reveals a high scatter that questions the assumption of an equilibrium flow. To describe the observed nature of the data, we assume heterogeneity in the pedestrian population, that results from the existence of multiple pedestrian classes that are characterized by different behavior. As these classes cannot in general be characterized or even observed, we adopt a latent class approach to derive the relationship between speed and density.

The structure of the paper is as follows. A review of related research from the literature is provided in Section 2. Section 3 describes the proposed methodological framework for the derivation of the multi-class speed-density relationship. Section 4 presents a case study and corresponding empirical analysis. In Section 5 we empirically illustrate the performance of the approach. Finally, Section 6 summarizes the outcomes of the proposed methodology and determines future research directions.

2 Literature review

The first fundamental traffic relationship was established in the field of vehicular traffic in the thirties of the last century (Greenshields et al., 1935). Over the years, various studies have been reported on fundamental relationships (see Zhang, 2012 for a complete review). In the context of pedestrian traffic, speed-density relationships are predominantly considered. The main stream of the literature establishes the relationships empirically, by fitting curves to empirical observations. The relationships that are usually used in the literature are listed in Table 1, where v_f is the free flow speed, k_j the jam density, and θ and γ are parameters. Other studies focus on the derivation of the relationships via simulation-based models (Blue and Adler, 1998) or from microscopic pedestrian traffic principles (Flötteröd and Lämmel, 2015; Hoogendoorn et al., 2014). The physical features that they share are (i) the decreasing trend of the speed with increase in density, and (ii) the deterministic nature, due to assumed equilibrium conditions (conditions when dynamics of the system are ignored and the population is homogenous).

The findings from several empirical analysis (Cheung and Lam, 1998; Daamen et al., 2005; Steffen and Seyfried, 2010) question the deterministic approach

of the listed studies. They report heavy scatter in the empirical speed-density relationship, which cannot be predicted by the proposed deterministic models. The observed scattering is explained in the literature by violation of equilibrium assumptions (Kim and Zhang, 2008; Wang et al., 2013; Jabari et al., 2014; Cheung and Lam, 1998; Weidmann, 1993).

Source	Specification	Parameters
Older (1968) Navin and Wheeler (1969) Fruin (1971) Tanaboriboon et al. (1986) Lam et al. (1995)	$v(k) = v_f - \theta k$	v_f, θ
DiNenno (2002)	$v(k) = v_f - v_f \theta k$	v_f, θ
Tregenza (1976)	$v(k) = v_f \exp\left(-(\frac{k}{\theta})^{\gamma}\right)$	v_f, γ, θ
Weidmann (1993)	$v(k) = v_f \left\{1 - \exp\left(-\gamma\left(\frac{1}{k} - \frac{1}{k_j}\right)\right)\right\}$	v_f, k_j, γ
Rastogi et al. (2013)	$v(k) = v_f \exp(-\frac{k}{\theta})$	v_f, θ
Units: $k[\mathrm{ped}/\mathrm{m}^2], v[\mathrm{m}/\mathrm{s}]$		

Table 1: Deterministic fundamental relationships - pedestrian traffic

Everyday experience suggests that the pedestrian population is heterogeneous. Various studies have shown empirically that the differences among pedestrians, such as trip purpose, age, gender, health conditions, etc., influence walking speeds of pedestrians (Weidmann, 1993; Bowman and Vecellio, 1994; Campanella et al., 2009). Microscopic approaches capture this complex phenomena by modeling the exact underlying walking process and interactions at the level of individuals (Johansson et al., 2007; Hoogendoorn and Bovy, 2004). Although being highly precise, these approaches suffer from high computational time and require a great deal of dissagregate data.

At the macroscopic level (speed-density relationship) there are several ways to account for heterogeneity. In a two-stage approach, the data is first segmented based on some observed characteristics (e.g. socio-economic or demographic variables), using automated clustering schemes (Ge et al., 2012; Lee et al., 2007) or manually. The assignment of the individual observations to different segments is deterministic in this approach. In the second stage, a separate model is estimated for each predefined segment in the population (Weidmann, 1993). The issue of imprecise parameter estimates may arise due to potentially small sample sizes in some segments. Also, segmentation is usually performed based on a sin-

gle characteristic and assumed to be error-free. In reality, the heterogeneity may come from multiple factors, which may introduce errors in the second stage.

An alternative to the two-stage approach for dealing with heterogeneity is the probabilistic model-based method (Fraley and Raftery, 2002). Probabilistic models in the field of vehicular traffic are usually derived by adding Gaussian noise to the existing deterministic relationships (Wang et al., 2013). This can potentially lead to unrealistic outcomes (e.g. negative speed values). Jabari et al. (2014) proposed a probabilistic speed-density relationship based on a microscopic car-following model. Probabilistic features are incorporated by introducing random parameters capturing population heterogeneity. However, limited behavioral basis exists to help in the specification of the distribution of these parameters. Also, such approach has the disadvantage of potentially allowing for behaviorally and physically implausible parameter values, depending on the choice of distributions. In Nikolić et al. (2016), a probabilistic speed-density relationship for pedestrians (*PedProb-vk*) is derived by bringing together first principles and a data-driven approach. The model specification ensures the physical correctness of the results, but lacks behavior-oriented explanatory power.

We propose in this paper an alternative approach to account for the heterogeneity of speed, as observed in the data. The suggested model is a latent class model (LCM). The LCM approach has been proven to be valuable in capturing unobserved heterogeneity and characterization of the latent classes (Walker and Li, 2007; Jintanakul et al., 2009).

3 Methodology

This section first describes the general modeling framework and presents concrete suggestions for the model specification. Then, data requirements and model estimation method are discussed.

3.1 Modeling framework

Let (v_i, k_i, X_i) be a triplet representing the speed v_i, the density k_i and the vector of observable characteristics X_i (such as age, trip purpose, etc.) associated with individual i. We assume that the population is partitioned into J classes (sub-populations) of pedestrians. In this framework, the individual speeds v_i are random variables. It is assumed that the speed is influenced by the prevailing density, and that this relationship varies across classes. Therefore, the distribution of v_i is characterized by its probability density function conditional on the density k_i experienced by individual i and the class j

$$f_j(v_i|k_i, j; \theta_j(k_i)), \tag{1}$$

where $\theta_j(k)$ are parameters. We refer to this distribution as a class-specific model (CSM). Each class may be characterized by a different probability distribution of the individual speeds. However, in most applications it is assumed that all class densities arise from the same parametric distribution family (Frühwirth-Schnatter, 2006). We assume that this distribution is continuous with positive support. This property is in accordance with the physical characteristic of the speed, being that the speed is a continuous variable whose values cannot be negative. Note that the density k_i is not class dependent, as all pedestrians contribute to the density, irrespectively of the class they belong to. The specification of the parameters $\theta_j(k_i)$, which characterize the class-specific speed distribution, is assumed to vary with the density of pedestrians. The assumption is motivated by empirical observations (Cheung and Lam, 1998) that suggest different trend of speed distribution for different density levels (e.g. the mean and the spread of the speed distribution decrease with increase in density of pedestrians). For instance, this dependency can be represented using deterministic speed-density relationships, such as those presented in Table 1. Some concrete examples are shown in Section 3.2.

The class of pedestrian i cannot be directly observed. Therefore, we propose a class membership model (CMM) that provides the probability that a pedestrian i, characterized by her socio-economic characteristics X_i, belongs to class j

$$\Pr(j|X_i;\beta_j), \tag{2}$$

where β_j are parameters. The CMM can take a number of forms. A typical assumption is based on a fitness function, that is a continuous variable measuring how much individual i fits into class j. For example, a linear formulation would consist in

$$U_{i,j} = V_{i,j} + \varepsilon_{i,j} = CSC_j + \beta_j X_i + \varepsilon_{i,j}, \tag{3}$$

where CSC_j and β_j are unknown parameters to be estimated from data, and $\varepsilon_{i,j}$ is a random term. The assumption is that the individual belongs to the class with the highest value of the fitness function. A specific distribution assumption for $\varepsilon_{i,j}$ leads to a specific probability model. The exact specification of $V_{i,j}$, and in particular the exact list of characteristics involved in X_i, is application dependent. We show some examples in Section 3.2.

The *multi-class speed-density model* (MC-vk) is obtained by combining the CSM and the CMM as follows

$$f_{MC-vk}(v_i|k_i, X_i; \theta_j(k_i), \beta_j) = \sum_{j=1}^{J} \underbrace{f_j(v_i|k_i, j; \theta_j(k_i))}_{\text{CSM}} \underbrace{\Pr(j|X_i;\beta_j)}_{\text{CMM}}. \tag{4}$$

The latent classes can be assumed a priori or interpreted a posteriori. The a priori specification of classes can be based on the information from the literature

about different pedestrian sub-populations (e.g business, leisure travelers or children, adults, seniors), and their preferred walking speed and the attitude towards congestion (Daamen, 2004, Weidmann, 1993). The a posteriori interpretation of latent classes should be supported by the estimation results.

3.2 Exemplary specification

For the illustration of the model, we need to assume the number of classes (J), and to specify the exact form of the CMM and the CSM. We assume the existence of two classes ($J = 2$), denoted as class C_1 and class C_2, in order to keep the model parsimonious and to avoid potential over-fitting. Note that other models with higher number of classes are estimated for the train station context and compared using statistical tests in Section 5.1.3. The results of this comparison show that the two-class model is superior for this case study.

We suggest the Rayleigh model for the distribution of the speed in each class

$$f_j(v_i|k_i, j, \theta_j(k)) = \frac{v_i}{\theta_j^2(k_i)} \exp(-\frac{v_i^2}{2\theta_j^2(k_i)}), \tag{5}$$

which is driven by only one parameter, the scale $\theta_j(k_i)$. The choice of the Rayleigh distribution is motivated by its properties (continuous distribution that is defined on the positive support) that are in accordance with the physical properties of the speed. The mean of a Rayleigh random variable is expressed as $\mu_j(k_i) = \theta_j(k_i)\sqrt{\pi/2}$. We model the mean as the linear class-specific equilibrium speed-density relationship for all classes

$$\mu_j(k_i) = v_{f,j} - \gamma_j k_i, \tag{6}$$

where $v_{f,j}$ and γ_j are class-specific parameters, referring to the free-flow speed, respectively the sensitivity to congestion. These parameters are expected to vary across classes such that they reflect the class-specific behavior. The equilibrium relationship in (6) is derived in Nikolić et al. (2016) from the social force model proposed by Helbing and Molnar (1995), for isotropic, homogenous and stationary traffic conditions. It also corresponds to the relationships proposed by Older (1968), Navin and Wheeler (1969), Fruin (1971), Tanaboriboon et al. (1986) and Lam et al. (1995) (Table 1).

Each pedestrian is assumed to belong to one class only (pedestrians do not switch among classes over time). This is consistent with the literature on vehicular traffic (van Wageningen-Kessels, 2013). We assume that the error term $\varepsilon_{i,j}$ in (3) is i.i.d. type 1 Extreme Value (EV1(0,1)) across classes and individuals. This assumption yields the binary logit CMM, defined as

$$Pr(j|X_i; \beta_j) = \frac{e^{V_{i,j}}}{\sum_{j=1}^{2} e^{V_{i,j}}}. \tag{7}$$

The quality of the CMM depends on the available information about the behavioral profiles of pedestrians, that constitutes the deterministic part of the model ($V_{i,j}$). For instance, various studies have shown that, in general, the age (children, adults, seniors), the gender (female, male) and the trip purpose (leisure, commuters, shoppers, business) of pedestrians have an effect on their movement behavior (Section 2). If this information is available, the deterministic parts of the fitness function for each class and pedestrian can be defined as

$$\begin{aligned} V_{i,1} &= CSC_1 + \beta_{CHILD,1}CHILD_i + \beta_{ADULT,1}ADULT_i + \beta_{FEMALE,1}FEMALE_i \\ V_{i,2} &= \beta_{LEISURE,2}LEISURE_i + \beta_{COMMUTERS,2}COMMUTERS_i + \\ &\quad \beta_{SHOPPERS,2}SHOPPERS_i, \end{aligned} \tag{8}$$

where SENIOR, MALE and BUSINESS are considered as the reference levels of the corresponding discrete variables. Depending on the context, some additional information may be also useful. For instance, in transportation hubs, such as train stations, metro stations or airports, the type of passenger (arriving, departing, transferring), the time to departure, the distance that pedestrians need to traverse, the presence of luggage and walking in groups appear as relevant factors. Similarly, the opening hours of shops in commercial centers, or highlights in museums, interacted with the cultural background of pedestrians, can be valuable in explaining the class membership.

Note that the framework described in Section 3.1 is general and allows for different specifications to be tested. To assess the performance of our approach, we present the analysis on a real case study in Section 5.

3.3 Data requirements

In order for the presented framework to be applied, the following data must be available. The key type of data, like for any such model, is the traffic condition data. It includes individual speed and density observations, necessary for the estimation of the model. They can be extracted from the individual trajectory data, that is the data provided in the form of individual-specific pairs of consecutive time and location observations. Ideally, trajectories are collected using precise pedestrian tracking systems with high temporal resolution. For instance, the systems based on optical, thermal and depth sensors (Alahi et al., 2011; Alahi et al., 2014) or digital cameras (Daamen and Hoogendoorn, 2003). Pedestrian trajectories can be also obtained using wireless technologies such as WiFi or Bluetooth. The issue with these technologies is the low temporal resolution and strong sample bias (Danalet, 2015). In this case, the combination of WiFi or Bluetooth traces with count data may provide better understanding about prevailing traffic conditions (Hänseler, 2016).

Although the class membership model could be specified without any explanatory variable, the quality of the model would benefit from characteristics of individuals. This means that the more information is available about pedestrian characteristics, the easier it will be to obtain a good class membership model. This information includes typical socio-economic and demographic data, such as age, gender, health conditions, culture, trip purpose, etc. As mentioned in Section 3.2, depending on the type and the purpose of pedestrian facilities, additional types of data would be useful. For instance, the apparent types of data are (i) the type of passenger, the time to departure, the presence of luggage and walking in groups in public transport facilities, (ii) the points of pedestrians' interest in museums and the attractivity of these points, (iii) opening hours of shops and restaurants in commercial centers, (iv) concert schedules and toilets' locations in music festivals, (v) pollution and noise levels experienced by pedestrians in urban streets, etc. To collect the mentioned characteristics various recall methods may be used, including paper-based surveys distributed to individuals (Bachu et al., 2001; Kalakou et al., 2014), smartphone-based applications (Ohmori et al., 2005; Cottrill et al., 2013; Ball et al., 2014; Zhao et al., 2015), and web-based methods (Bohte and Maat, 2009).

In case data collection using recall methods is not performed, it is possible to compensate, to some extent, by relying on the available information from pedestrian trajectories and some other sources (e.g. attributes of a facility, timetables and schedules in transportation hubs, etc.). This is demonstrated in Section 4.2.2.

3.4 Estimation procedure

With respect to traffic condition data, we assume the availability of individual observations collected over multiple time instants. This means that traffic condition data has panel nature, and the speed observations of a single individual are likely to be correlated across time. This is usually referred to as serial correlation. The issue arises due to unobserved individual factors that persist over time. If ignored it leads to consistent but inefficient estimators (Gourieroux et al., 1984). To address the issue of serial correlation we introduce an agent effect, α_i, in the mean of the speed distribution that drives the CSM (Wooldridge, 2005). We assume that it is exponentially distributed

$$f(\alpha_i; \mu_{\alpha_i}) = \frac{1}{\mu_{\alpha_i}} \exp(-\frac{\alpha_i}{\mu_{\alpha_i}}), \tag{9}$$

where μ_{α_i} is the mean. The agent effect is assumed to be independently and identically distributed across pedestrians, but remains constant within the observations of a given pedestrian. The choice of the exponential specification is motivated by its positive support. Also, it is well suited to avoid any arbitrariness in imposing

the upper bound, while preventing arbitrary high values at the same time. The likelihood conditional on class j for the observations of pedestrian i in the described panel setting is as follows

$$f(v_{i1}, ..., v_{iT_i}|k_{i1}, ..., k_{iT_i}, j; \theta_j(k)) = \int_{\alpha_i} \prod_{t=1}^{T_i} f_j(v_{it}|k_{it}, j, \alpha_i; \theta_j(k)) f(\alpha_i; \mu_{\alpha_i}) d\alpha_i, \tag{10}$$

where T_i is the number of observations of the individual i, and $v_{i1}, ..., v_{iT_i}$ and $k_{i1}, ..., k_{iT_i}$ are the sequences of speed, respectively density observations associated with this individual.

Contrary to traffic condition data, pedestrian characteristics (e.g. age, gender, etc.) are assumed to be time independent during the observation period. Each pedestrian is associated with one observation of the considered characteristics. For instance, the age category of a pedestrian is *adult* and her gender is *female*. This means that the class membership probability is calculated only once, and it remains constant across all speed-density observations for a given pedestrian (the error terms $\varepsilon_{i,j}$ (3) are not time dependent). Therefore, the issue of serial correlation does not appear at the CMM level. We also assume that the error terms of the CSM and the CMM are not correlated, as in Walker and Li (2007).

Combining the CSM and the CMM, the contribution of individual i to the likelihood is given as

$$\begin{aligned} &f(v_{i1}, ..., v_{iT_i}|k_{i1}, ..., k_{iT_i}, X_i; \theta_j(k), \beta_j) = \\ &\sum_{j=1}^{2} \left\{ \int_{\alpha_i} \prod_{t=1}^{T_i} f_j(v_{it}|k_{it}, j, \alpha_i; \theta_j(k)) f(\alpha_i; \mu_{\alpha_i}) d\alpha_i \right\} Pr(j|X_i; \beta_j). \end{aligned} \tag{11}$$

The integral in (11) is approximated via simulation as

$$\begin{aligned} &f(v_{i1}, ..., v_{iT_i}|k_{i1}, ..., k_{iT_i}, X_i; \theta_j(k), \beta_j) = \\ &\sum_{j=1}^{2} \left\{ \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} f_j(v_{it}|k_{it}, j, \alpha_{i,r}; \theta_j(k)) \right\} Pr(j|X_i; \beta_j), \end{aligned} \tag{12}$$

where R refers to the number of draws from $f(\alpha_i; \mu_{\alpha_i})$. The likelihood function for the sample of N individuals ($i = 1, ..., N$) is given by

$$\begin{aligned} \mathcal{L} = &\prod_{i=1}^{N} f(v_{i1}, ..., v_{iT_i}|k_{i1}, ..., k_{iT_i}, X_i; \theta_j(k), \beta_j) = \\ &\prod_{i=1}^{N} \left\{ \sum_{j=1}^{2} \left\{ \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} f_j(v_{it}|k_{it}, j, \alpha_{i,r}; \theta_j(k)) \right\} Pr(j|X_i; \beta_j) \right\}, \end{aligned} \tag{13}$$

that is to be maximized.

4 Case study and empirical analysis

We illustrate the methodology and analyze its performance on a dataset collected in a pedestrian underpass of the train station of Lausanne, Switzerland, described below.

4.1 Lausanne train station

Figure 1 shows the layout of the studied area. It covers approximately 685 m^2. The underpass is frequently used especially during the morning and afternoon peak hours since it connects the exterior of the train station to the main platforms. It also acts as a connection between a residential area in the south and the center of the city in the north.

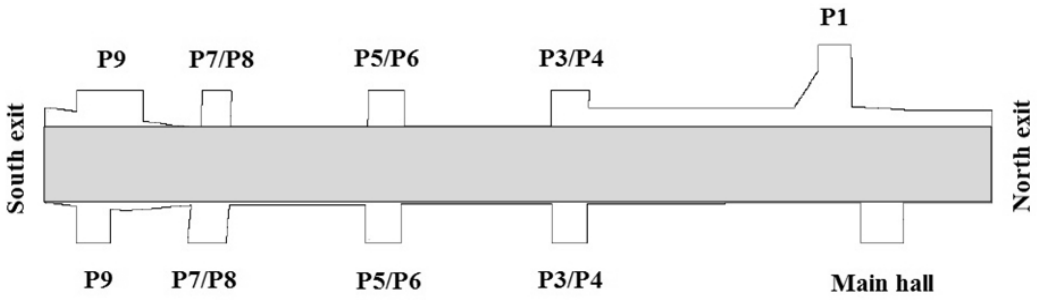


Figure 1: Lausanne train station - pedestrian underpass West

To collect the raw data, a large-scale network of smart sensors has been deployed in the station. The underlying technology is based on optical, thermal and depth sensors that detect silhouettes and track each pedestrian in the area covered by the network. The tracking engine uses a sparsity driven framework (Alahi et al., 2011; Alahi et al., 2014) to link detected pedestrians over the network of sensors.

In the underpass (Figure 1), 33 depth sensors are installed. The placement of the sensors is such that the major part of the underpass is monitored (Alahi et al., 2013). However, blind areas exist (the areas that are not covered by the sensors), where missing data are completed using an inter-sensor tracking algorithm (Alahi et al., 2011). We call the observations from the covered areas the “detected" observations, and those from the uncovered areas the “imputed" observations. We need this distinction in order to reduce the effect of errors due to measurement noise and technological issues in the calculation of the density indicator, as elaborated in Section 4.2.1.

The tracking results in a dataset of 25,603 trajectories, collected during a time period between 07:00 and 08:00 on February 12, 13, 14, 15 and 18, 2013. The temporal resolution of every trajectory ranges from 10 to 25 points per second and it has been processed to obtain the position of every pedestrian in the scene at every second. The average length of the trajectories is 78 meters and the duration of a pedestrians' stay in the underpass ranges from 15 seconds to 2.2 minutes.

Note that we have selected only trajectories collected in the main corridor of the underpass, represented by the shaded area in Figure 1. The trajectories from the ramps and stairs (denoted as P1-P9) are not considered in this study. Indeed, as explained by Daamen (2004) and Weidmann (1993), the walking behavior and, therefore, the speed-density relationship, varies with the type of infrastructure.

In addition to detailed pedestrian trajectories, the infrastructure data, that is detailed plans containing the locations and dimensions of all relevant parts of the monitored system, is also available. We also have access to the train timetable for the period under study. The arrival and departure times and the assigned tracks are thus known for all trains.

In the rest of the paper, we refer to this case study as the Lausanne case study.

4.2 Empirical analysis

We first present the analysis of traffic condition data, that is speed and density observations. We then discuss the factors used to explain the class membership of individuals.

4.2.1 Speed and density observations

The speed-density profile corresponding to the Lausanne case study (Figure 2) is obtained from the Voronoi-based measurement method presented in Nikolić et al. (2016). The Voronoi space decomposition assigns a personal region to each pedestrian i, based on the positions of pedestrians. This is done in such a way that each point in the personal region is closer to i than to any other pedestrian, with respect of the Euclidean distance. The position of "detected" observations is considered to be accurate, while the position of "imputed" observations might be subject to inter-sensor tracking algorithm errors (Alahi et al., 2011). Therefore, the Voronoi spatial discretization is performed based on "detected" observation only. We, however, need to account for the existance of the pedestrians whose observations are marked as "imputed". To do so, the density at each point $p = (x, y, t_s)$ is computed as

$$k(x, y, t_s) = \frac{n_{V_{is}}^{\text{detected}} + n_{V_{is}}^{\text{imputed}}}{|V_{is}|}, (x, y) \in V_{is}, \tag{14}$$

where $t_s = (t_0, ..., t_f)$ corresponds to the available sample, $|V_{is}|$ is the area of Voronoi cell V_{is} assigned to "detected" pedestrian i at time t_s, and $n_{V_{is}}^{detected}$ and $n_{V_{is}}^{imputed}$ refer to the number of "detected", respectively "imputed" observations within the cell V_{is}. The speed is approximated using finite differences, based on the "detected" observations (Nikolić et al., 2016).

In Figure 2, each circle corresponds to one observation, that is, one pedestrian at one specific time. The x coordinate of the circle corresponds to the density, and its y coordinate corresponds to the speed. The figure plots 154,417 observations corresponding to the peak hour of February 13, 2013. The same pattern was observed on any weekday.

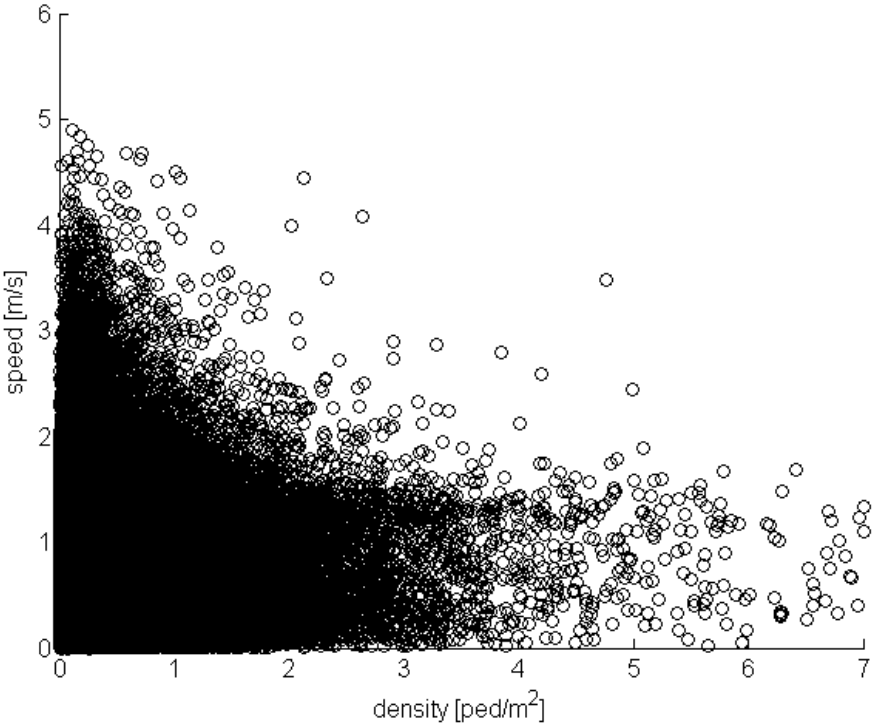


Figure 2: Speed-density profile

A high scattering is observed in Figure 2. The density ranges from 0 to approximately 7 pedestrians per square meter, and 99% of the observations are below 1.4 pedestrians per square meter. The speed ranges from 0 to approximately 6 meters per second, and 99% of the observations are below 2.42 meters per second.

The deterministic models for the speed-density relationship proposed in the literature (Section 2) appear to be inadequate for representing the observed pattern.

4.2.2 Characteristics associated with pedestrians

In addition to the speeds and densities, we extract additional variables that are used in the model: the pedestrian type, the OD distance, the peak and off-peak periods and the time to departure. In the following, we explain each factor separately.

The pedestrian type refers to the classification of pedestrians based on their OD pairs. The definition of OD areas depends on the layout of a train station, and it includes the stairs and ramps to the platforms (denoted as P1-P9 in Figure 1) and the entrance/exit areas. We consider four pedestrian types

1. arriving passenger (AP) - pedestrians originating from a platform and exiting the station,

2. departing passenger (DP) - pedestrians walking to a platform to embark on their trains,

3. transferring passenger (TP) - pedestrians whose origin and destination are different platforms,

4. non-passenger (NP) - pedestrians whose origin and destination are not a platform (e.g. pedestrians that go shopping in the station , or use the underpass to reach the other side of the city).

Between 8% and 9% of pedestrians each day are not classified, due to the mismatch of their initial and/or final observations with any of the predefined zones that indicate origins and destinations. These observations are considered for the calculation of density, but are not taken into account at the level of the CMM. Figure 3 shows the percentage of pedestrians belonging to each of the four types across days. The pattern is relatively stable over days, and it indicates that the majority of pedestrians are arriving and departing. Figure 4 illustrates the speed distributions for these two types of pedestrians, for a particular density level. It shows that arriving passengers tend to walk with lower speeds, compared to departing passengers. The shift of the distribution for departing passengers towards higher values can be explained by the fact that they are more likely to be under time pressure (to embark on the trains) than the arriving ones.

The OD distance is associated with each pedestrian based on the corresponding OD pair. The OD distances are calculated as the shortest Euclidean physical distance between each origin and destination. The distances range from approximately 3 meters to 80 meters, as shown in Figure 5.

The peak and off-peak periods are defined based on the number of people observed over time for each day (Figure 6). For the temporal aggregation we consider the intervals that are 5 minutes long. Peak periods (PP) refer to the

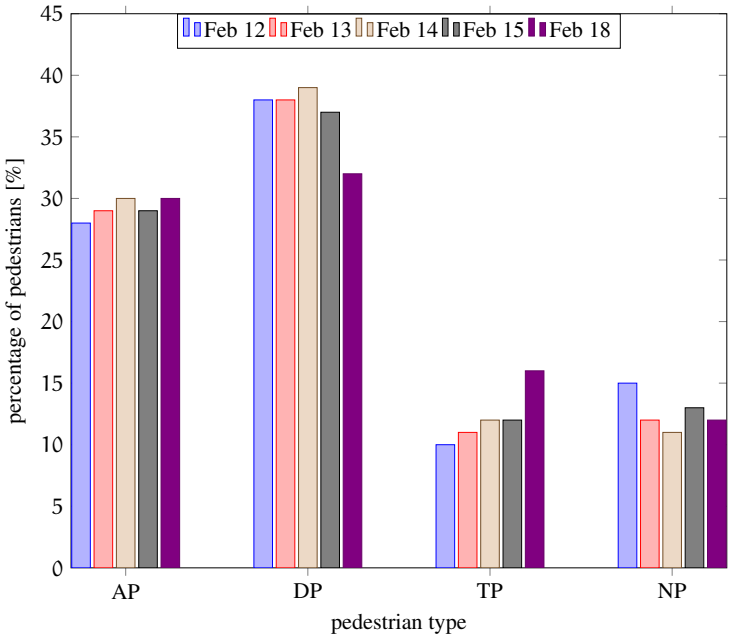


Figure 3: Number of pedestrians per pedestrian type

intervals where the local maxima of the number of people are observed, and off-peak periods (OPP) refer to all other intervals. For February 13, 2013, for instance, the peak periods are 07:15-07:20 and 07:40-07:45. We expect that pedestrians are characterized with different walking behavior, depending on the period when they are observed.

The time to departure is obtained by exploiting the information contained in the train timetable. We define the time to departure as the difference between the departure time of the next train from the platform that the pedestrian is going to and the time at which the pedestrian is first observed in the underpass. The distribution of time to departure is shown in Figure 7. It ranges from a few seconds to approximately 50 minutes. The distribution suggests that most of the people arrive to the train station approximately 3 minutes (the mode of the distribution) before the train departure. It is natural to assume that people that have more time to the departure of their trains behave differently from those rushing to catch their trains.

We also explored the impact of the group behavior on the speeds of pedestrians. We adopted spatial clustering together with temporal frequent patterns analysis to identify the pedestrians walking in groups. In the spatial clustering step, we used the values proposed by McPhail and Wohlstein (1982) for the features characterizing pedestrians that walk in groups (e.g. distances pedestrians keep between each other, differences in speeds and directions). However, less

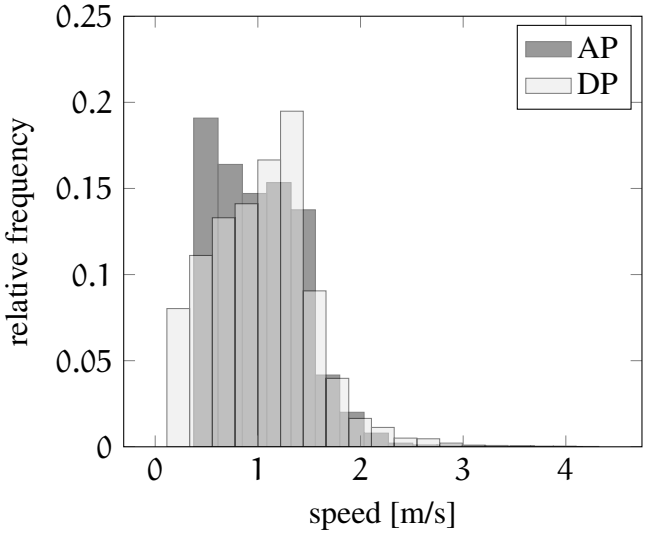


Figure 4: Speed distribution per pedestrian type (k=0.6 ped/m^2)

than one percent of the population was identified to walk in groups, and the analysis with respect to this factor showed no significant effect on the speed at which pedestrians move. The factor is therefore excluded from further analysis.

To investigate the case study data in more details, we consider the speed distributions with respect to the mentioned factors at various density levels. The speed observations are aggregated based on Levels of Service (LoS) standard for pedestrian facilities proposed by Fruin (1971). The levels are labeled from A to F, as shown in Table 2. We consider the levels A-E, given that the largest part of

LoS	Density range [ped/m^2]
A	$k \leq 0.31$
B	$k \in (0.31 - 0.43]$
C	$k \in (0.43 - 0.71]$
D	$k \in (0.71 - 1.11]$
E	$k \in (1.11 - 2.17]$
F	$k > 2.17$

Table 2: LoS (Fruin, 1971)

the observations falls below the LoS F (99%) of the observations are below 1.4 pedestrians per square meter.

The box-plots of the distributions are shown in Figure 8, Figure 9, Figure 10 and Figure 11 for different pedestrian types, respectively OD distances, periods and time to departure. For the purpose of the analysis we discretize the range of

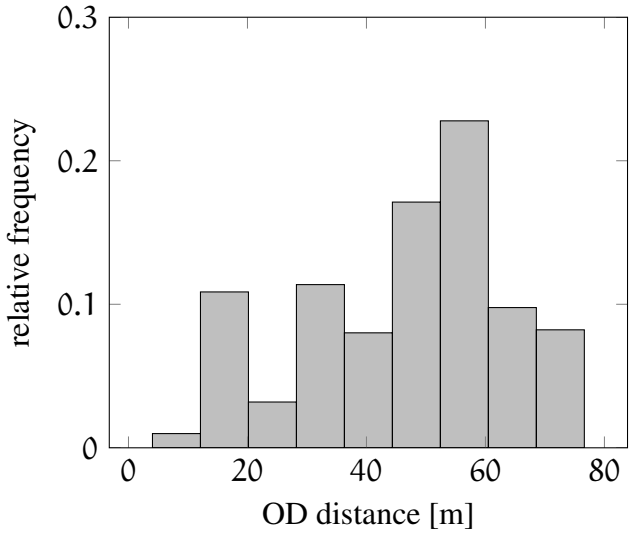


Figure 5: OD distance distribution

OD distances and time to departure into segments. We consider distances shorter than 10 meters (short distances - SD), distances between 10 meters and 30 meters (medium distances - MD), and distances longer than 30 meters (long distances - LD). Time to departure interval is segmented into time intervals shorter than 5 minutes (short intervals - SI), and time intervals longer than 5 minutes (long intervals - LI). The bottom and top of the boxes in Figure 8 - Figure 11 are the first and third quartiles, and the line inside the box is the median. The ends of the whiskers represent the 2^{nd} and the 98^{th} percentiles. The analysis indicates dissimilar trends, in particular with respect to speed distributions for different pedestrian types (Figure 8) and different ranges of OD distances (Figure 9).

For the quantitative analysis, we use Kolmogorov-Smirnov test (Massey, 1951). For each LoS and each factor we test the hypothesis that speed distributions for different factor values represent the same population. The results are shown in Table 3. The test rejects the hypothesis in all cases, at the 5% significance level.

The analysis presented in Figure 8 - Figure 11 and Table 3 are in agreement with our assumption about the pedestrian heterogeneity reflected through different walking speeds. The analysis above suggest that density is not the only factor influencing pedestrians' speed and that the observed heterogeneity in speed values might come from multiple factors.

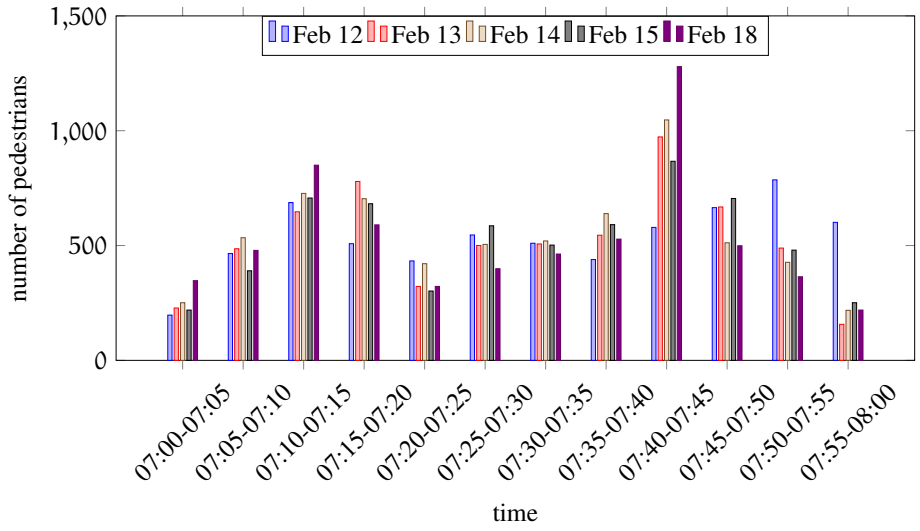


Figure 6: Number of pedestrians grouped into intervals of 5 minutes

5 Applying the framework

The framework proposed in Section 3 is illustrated on the Lausanne case study. The CSM is specified in the model (5). The CMM is specified as the logit model given in (7), with the deterministic parts of the fitness function defined as

$$\begin{aligned} V_{i,1} &= CSC_1 + \beta_{DP,1} DP_i + \beta_{TP,1} TP_i + \beta_{NP,1} NP_i \\ V_{i,2} &= \beta_{TTD,2} TTD_i + \beta_{PP,2} PP_i + \beta_{OD,2} OD_i, \end{aligned} \tag{15}$$

where the variables are described in Section 4.2.2. This CMM specification is motivated by the analysis presented in Section 4. We present below the model estimation results and detailed examination of each component of the model. Also, comparisons with the existing models and practical implications are discussed.

5.1 Model estimation

For the estimation of the model parameters, we use maximum likelihood procedure. The estimation results of the presented model are shown in Table 4. We use $R = 200$ draws from (9) for the simulation of the integral in (13). The standard errors are calculated using bootstrapping (Nikolić et al., 2016). All estimates have the expected sign and value. The results also show the low standard errors of the parameters and their statistical significance at a usual significance level (0.05).

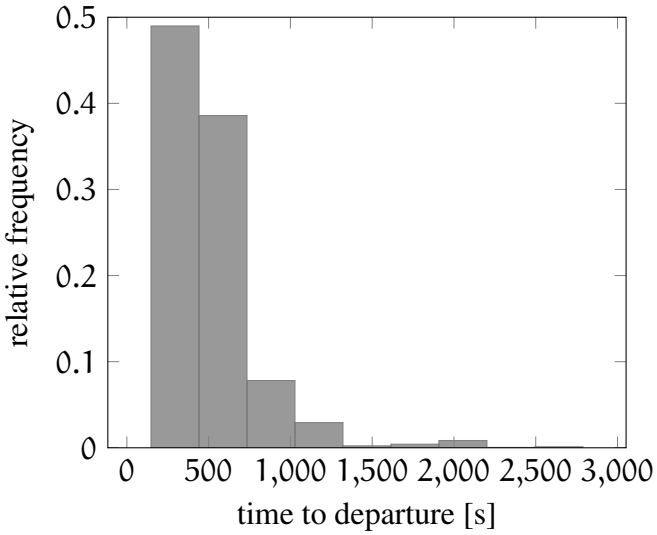


Figure 7: Distribution of time to departure

5.1.1 Class-specific model

The parameter estimates of the class-specific models show how movement behavior vary across classes.

The signs and the estimated values of the parameters $\nu_{f,1}$,γ_1,$\nu_{f,2}$ and γ_2 are consistent with the ones reported in the literature. The corresponding class-specific speed-density relationships are shown in Figure 12.

The parameters μ_{α_1} and μ_{α_2} show that agent effect distributions are characterized with similar mean in both classes.

We make inferences on the movement behavior of each class based on the estimated parameters of the CSM. The C_1 class pedestrians are characterized with a higher free-flow walking speed and significantly lower sensitivity to congestion, compared to pedestrians belonging to C_2 class. We thus call C_1 the class of “pedestrians less sensitive to congestion" and C_2 the class of “pedestrians more sensitive to congestion". The classes are relevant for the context of pedestrian movement in train stations.

5.1.2 Class membership model

The parameter estimates of the class membership model show what are the underlying factors leading to different movement behavior.

The class-specific constant CSC_1 has negative sign denoting that, the rest of fitness functions being equal, pedestrians in the train station are less likely to belong to C_1 class than to C_2.

The positive sign of the parameters $\beta_{DP,1}$, $\beta_{TP,1}$ and $\beta_{NP,1}$ indicates that departing, transferring and non-passengers are more likely to be in class C_1, compared to arriving passengers.

The positive sign of the parameters $\beta_{TTD,2}$ and $\beta_{OD,2}$ shows that as pedestrians have more time to departure and longer distances to traverse, they are more likely to belong to C_2. Similarly, the sign of the parameter $\beta_{PP,2}$ shows that pedestrians are more likely to be of C_2 during peak periods with respect to off-peak periods.

Figure 13(a) shows the average probability over the sample of belonging to each class

$$w_j = \frac{\sum_{i=1}^{N} \Pr(j|X_i;\beta_j)}{N}, \tag{16}$$

where N refers to the number of individuals. According to the predictions of the model, 39.77% of pedestrians are "pedestrians less sensitive to congestion" (C_1) and 60.23% are "pedestrians more sensitive to congestion" (C_2). Figure 13(b) indicates the higher share of C_1 class in the case of departing and transferring passengers, compared to arriving passengers. This is expected, given that departing and transferring passengers are usually associated with higher time pressure that drives their more "aggressive" behavior. It is however not clear whether the high share of C_1 class in the case of non-passengers can be related to some behavioral aspects, or to the fact that OD pairs of non-passengers correspond to the main flow direction in the underpass. It is documented in the literature that pedestrians constituting the main flow are characterized by higher walking speeds (Wong et al., 2010).

Figure 14 shows the predicted value of time to departure and OD distance for pedestrians in each class. The predicted values represent the weighted average of each variable, where the weight is the probability of being in a particular class. As expected, pedestrians in class 1 (less sensitive to congestion) on average have less time to departure and shorter distances to traverse, compared to pedestrians in class 2 (more sensitive to congestion).

The results indicate the existence of relatively strong profiling. However, the segmentation of a population is not deterministically determined. Even though the parameter estimates in Table 4 suggest that the characteristics associated with pedestrians significantly influence the segmentation of considered population, a probabilistic model is necessary.

5.1.3 Alternative specifications

We have also performed a sensitivity analysis to investigate the potential existence of one and three sub-populations. The suggested model represents a form of a clustering analysis of data. Compared to the standard clustering methods, it is a probabilistic model-based approach. The issue of determining the correct number

of classes is therefore reduced to model selection problem in the probabilistic framework. We asses the goodness of the model by the means of the Bayesian information criterion (Schwarz et al., 1978)

$$\mathrm{BIC} = -2 \cdot \log \mathcal{L} + \mathfrak{m} \cdot \log(\mathfrak{n}), \tag{17}$$

where $\mathfrak{n}$ is the number of observations, $\mathfrak{m}$ is the number of unknown parameters and $\mathcal{L}$ is the value of the likelihood function of the model. The statistics are reported in Table 5, indicating that models with multiple classes are preferred over a single-class model. Although the best value of the $\log \mathcal{L}$ is achieved for the three-class model, the BIC statistic suggests that the two-class model represents the best compromise between the model accuracy and simplicity among all evaluated specifications. It also provides the most satisfactory behavioral interpretation of the results.

Model	1 class	2 classes	3 classes
$\log \mathcal{L}$	-529000.76	-519050.63	-519007.51
# parameters	3	13	23
# observations	747385	747385	747385
BIC	1058042.09	1038277.07	1038326.07

Table 5: Goodness of fit - BIC

Various alternative specifications of the model with two classes have been investigated, and the results are listed in Table 6. They differ from the proposed model in terms of assumed class-specific speed-density relationships and agent effect distributions. Linear speed-density relationships in Table 6 correspond to those proposed by Older (1968), Navin and Wheeler (1969), Fruin (1971), Tanaboriboon et al. (1986) and Lam et al. (1995) (Table 1). Exponential speed-density relationships refer to the relationship proposed by Rastogi et al. (2013) (Table 1). All alternative specifications resulted in a poorer fit with respect to the proposed model (BIC=1038277.07, Table 5), according to the BIC.

Model	M_a	M_b	M_c	M_d	M_e
Speed-density rel. C_1	Linear	Linear	Exponential	Linear	Linear
Speed-density rel. C_2	Linear	Exponential	Exponential	Exponential	Exponential
Agent effect distribution C_1	Rayleigh	Rayleigh	Rayleigh	Log-normal	Exponential
Agent effect distribution C_2	Rayleigh	Rayleigh	Rayleigh	Log-normal	Exponential
$\log \mathcal{L}$	-519952.55	-525143.19	-525064.29	-529320.93	-519070.04
# parameters	13	13	13	13	13
# observations	747385	747385	747385	747385	747385
BIC	1040080.92	1050462.20	1050304.40	1058817.69	1038315.91

Table 6: Goodness of fit for alternative specifications - BIC

5.2 Comparison at the aggregate level

The performance of the proposed *MC-vk* model is compared with the models proposed in the literature (Section 2) at the aggregate level. We have estimated the parameters of the deterministic models from Table 1 using linear regression on our dataset. The parameters of *PedProb-vk* are estimated on the same dataset using maximum likelihood procedure.

For *MC-vk*, the average speed for each density level (k) is given by

$$\bar{v}_{MC-vk} = \sum_{j=1}^{2} \mu_j(k) \Pr(j|X_i;\beta_j), \tag{18}$$

with the parameters described in Table 4.

A comparison among four deterministic models, the aggregate speed calculated from *PedProb-vk* and *MC-vk*, and the observed values are shown in Figure 15. The analysis is performed for density levels ranging from 0 to 1.4 ped/m^2, corresponding to 99% of the observed values. Table 7 reports the goodness of fit, in terms of the adjusted coefficient of determination $\bar{R}^2$ (Srivastava et al., 1995), and the mean squared error. The adjusted coefficient of determination is defined as

$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-m-1}, \tag{19}$$

where n is the number of observations, m is the total number of explanatory variables in the model, and R^2 is the coefficient of determination. The mean squared error (MSE) is defined as

$$MSE = \frac{1}{p}\sum_{\ell=1}^{p} (\bar{v}_{model}(k_\ell) - \bar{v}_{data}(k_\ell))^2, \tag{20}$$

where $p = 15$ density levels and $k_\ell = 1.4(\ell-1)/(p-1)$. The value $\bar{v}_{data}(k_\ell)$ is the average of the observed speeds corresponding to densities ranging from $(k_\ell + k_{\ell-1})/2$ and $(k_\ell + k_{\ell+1})/2$ (Nikolić et al., 2016).

The *MC-vk* is more complex than both traditional speed-density relationships and *PedProb-vk*. Yet, it relies on relatively few assumptions and it is easy to analyze by using simulation. The results in Table 7 indicate that the model *MC-vk* achieves the best fit among the analyzed models. Even though it has been calibrated at the disaggregate level, it is more consistent with empirical observations at the aggregate level.

5.3 Posterior analysis

The estimated model allows for the calculation of posterior probabilities of class-membership for each observation

$$\Pr(j|\nu_i, k_i, X_i; \theta_j(k), \beta_j) = \frac{\Pr(j|X_i; \beta_j) f_j(\nu_i|k_i, j; \theta_j(k))}{\sum_{j=1}^{2} \Pr(j|X_i; \beta_j) f_j(\nu_i|k_i, j; \theta_j(k))}. \quad (21)$$

The assessment of posterior probabilities allows to assign an observation to multiple classes with different degrees of membership, or to perform the so-called “soft" clustering. This form of clustering is desirable when the information about probabilities is useful. It provides more information than classical methods, where each observation belongs to only one class (“hard" clustering).

Figure 16 shows the posterior class-membership probabilities of the observations in the speed-density plot for the class of “pedestrians less sensitive to congestion" (C_1). They indicate how strongly each observation belongs to this class. Darker gray color indicates higher posterior probability, and lighter gray color corresponds to lower posterior probability. The results indicate relatively strong separation of the observations. It can be observed that higher speed values belong to C_1 class more strongly than lower speed values, for all density levels. The results of the analysis are in line with our expectations. They show the capability of the model to capture the structure of the data better than the traditional models.

5.4 Practical implications

The proposed model can be used by the operators as an instrument for policy making (e.g. long run planning) and daily operations (e.g. to control the flow). For instance, using the estimated CMM it is possible to examine the influence that the modification of the explanatory variables values has on the split of pedestrians among the classes. We look at the effect of the reduction of time to departure for departing and transferring passengers, assuming all the rest remains unchanged. Such a reduction is typically the result of the modification of the train time table in the train station. The reduction of 10%, 20%, 30%, 40% and 50% is considered. Figure 17 indicates the increase in the share of “pedestrians less sensitive to congestion" class, and decrease in the share of “pedestrians more sensitive to congestion" class, with the decrease in time to departure. This observation is in accordance with our expectations.

The presented analysis suggests that the model can be utilized for the analysis of the effects of such scenarios on the movement behavior of pedestrians. This is important for a variety of applications, such as the impact of different scenarios on the resulting LoS within train stations. Also, the evaluation can be further augmented by posterior analysis of the class membership probabilities (Section 5.3).

Moreover, the proposed methodology is not case specific. It is general and can be simply adopted to other specific applications.

6 Conclusion and future work

In this paper, a latent class model for speed-density relationship for pedestrian traffic is proposed. Differently from approaches in the literature, it is a multi-class model designed to account for the heterogeneity of speed observed in the data. The model uses the latent classes to relax the homogeneity assumption of equilibrium speed-density relationships.

To illustrate the proposed methodology, we use the data collected in the train station in Lausanne, Switzerland. The analysis confirms the existence of multiple classes of pedestrians characterized by different movement behavior. The resulting behavior is relevant for the studied situation. In addition to density, the observed pattern is explained by factors related to pedestrian type, train timetable and infrastructure. The proposed model thus represents a flexible tool for (i) recognizing the main factors driving the heterogeneity in the population of pedestrians and (ii) describing the impact of heterogeneity on pedestrian movement. Being conceptually insightful, the model can be further used to support the derivation of pedestrian flow models from the first principles.

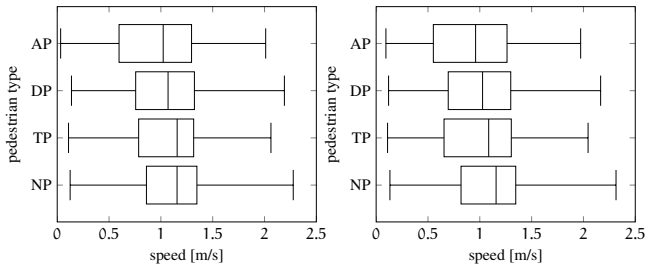
We present various tests using empirical data that validate the specification of the model. The model is also shown to outperform other models from the literature at the aggregate level. In contrast to the existing approaches, the suggested model features a behavior-oriented explanatory power. As such, it provides more realistic representation of the observed phenomena, and it is better suited for forecasting analysis.

In future work we plan to investigate the validity of stationary aspects of the model. We will analyze whether the speed of pedestrians observed at one time instant depends on the speed values observed in the previous time instants. This would lead to a dynamic model including the lagged speed variable (e.g. Markov dynamic model). To develop the solution for the initial conditions problem and related endogeneity bias in estimation, we could consider the conditional maximum likelihood estimation using a correction as proposed by Wooldridge (2005).

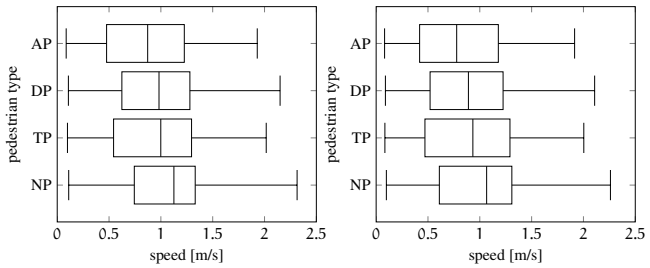
Our future research will also be directed towards the examination of additional explanatory factors (e.g. attractivity of certain zones), and their influence on the performance of the model. Finally, the performance of the approach in other situations, such as other transportation hubs, urban streets, museums or shopping malls, would also be an interesting research direction.

Acknowledgements

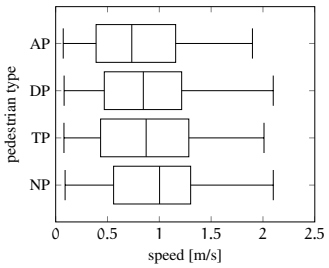
This research is supported by the Swiss National Science Foundation Grant 200021-141099 "Pedestrian dynamics: flows and behavior". The authors would like to thank SBB-CFF-FFS and Alexandre Alahi, who provided the dataset collected in the Lausanne train station. We are thankful for the constructive comments obtained from the audience during the fifth European Association of Research in Transportation (hEART 2016) and ninth Triennial Symposium on Transportation Analysis (TRISTAN IX), where different development phases of our methodological framework were presented.



(a) LoS A (b) LoS B

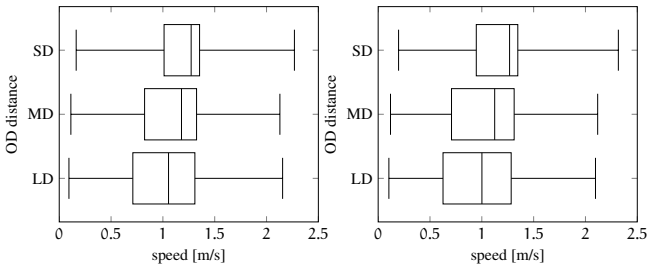


(c) LoS C (d) LoS D

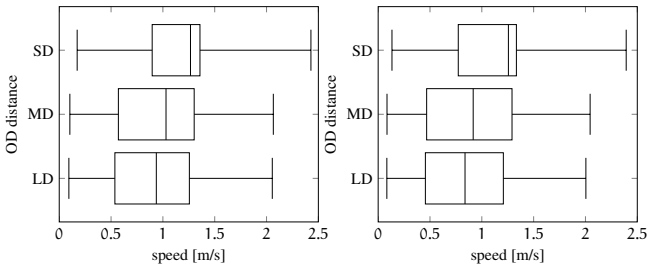


(e) LoS E

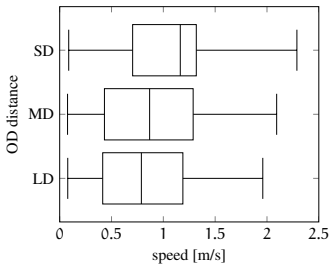
Figure 8: Speed distribution for different pedestrian types and different LoS



(a) LoS A (b) LoS B

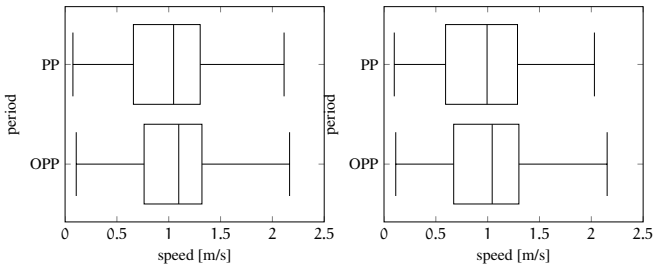


(c) LoS C (d) LoS D

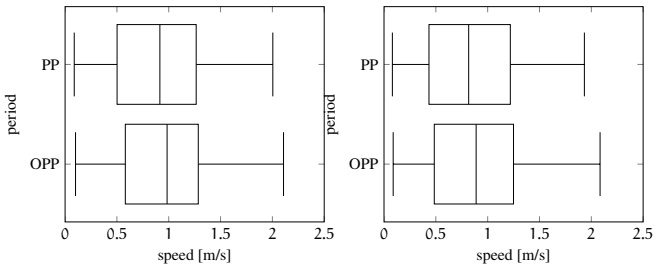


(e) LoS E

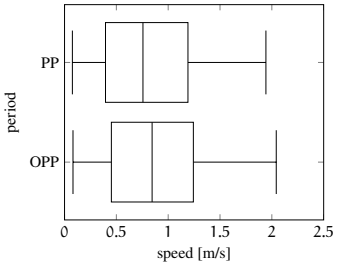
Figure 9: Speed distribution for different OD distances and different LoS



(a) LoS A (b) LoS B

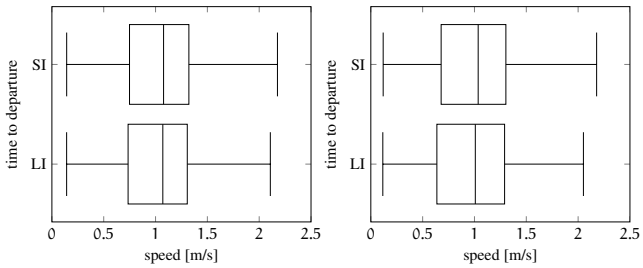


(c) LoS C (d) LoS D



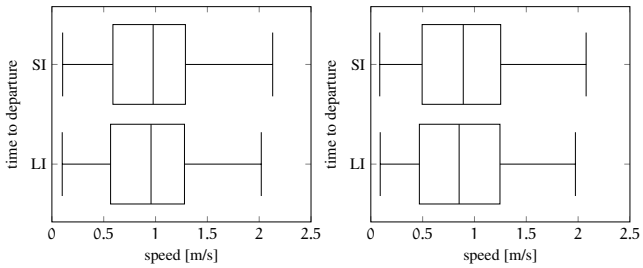
(e) LoS E

Figure 10: Speed distribution for peak and off-peak periods and different LoS



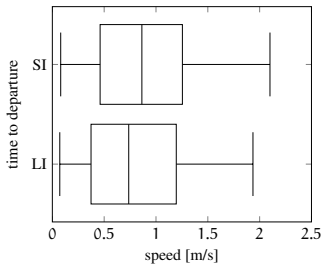
(a) LoS A

(b) LoS B



(c) LoS C

(d) LoS D



(e) LoS E

Figure 11: Speed distribution for different time to departure and different LoS

LoS	p-value					
	AP-DP	AP-TP	AP-NP	DP-TP	DP-NP	TP-NP
LoS A	0	0	0	$2.38e^{-213}$	0	$5.16e^{-65}$
LoS B	$3.21e^{-113}$	$9.02e^{-94}$	$3.73e^{-256}$	$1.04e^{-38}$	$6.75e^{-133}$	$1.99e^{-37}$
LoS C	$4.61e^{-188}$	$1.88e^{-106}$	0	$7.71e^{-24}$	$3.97e^{-147}$	$5.80e^{-96}$
LoS D	$3.16e^{-82}$	$1.28e^{-93}$	$4.22e^{-192}$	$1.53e^{-37}$	$1.40e^{-84}$	$2.29e^{-34}$
LoS E	$1.70e^{-24}$	$8.80e^{-29}$	$1.13e^{-51}$	$3.63e^{-09}$	$7.73e^{-19}$	$3.00e^{-11}$

(a) Pedestrian types

LoS	p-value		
	SD-MD	SD-LD	MD-LD
LoS A	$2.55e^{-187}$	0	0
LoS B	$3.25e^{-51}$	$2.14e^{-175}$	$4.08e^{-156}$
LoS C	$9.71e^{-89}$	$6.05e^{-215}$	$8.78e^{-131}$
LoS D	$1.55e^{-50}$	$1.81e^{-106}$	$1.22e^{-70}$
LoS E	$2.57e^{-14}$	$9.45e^{-30}$	$4.50e^{-24}$

(b) OD distances

LoS	p-value
LoS A	$1.89e^{-206}$
LoS B	$9.28e^{-48}$
LoS C	$1.24e^{-79}$
LoS D	$6.76e^{-36}$
LoS E	$1.24e^{-18}$

(c) Peak/off-peak periods

LoS	p-value
LoS A	$2.54e^{-12}$
LoS B	$1.11e^{-04}$
LoS C	$2.83e^{-02}$
LoS D	$4.35e^{-02}$
LoS E	$2.66e^{-06}$

(d) Time to departure

Table 3: p-value of Kolmogorov-Smirnov statistic

Parameter	Value	Std err	t-test
CSC_1	-0.682	$2.82e^{-02}$	$-2.35e^{01}$
$\beta_{DP,1}$	2.02	$7.50e^{-03}$	$2.69e^{02}$
$\beta_{TP,1}$	1.46	$2.90e^{-02}$	$4.99e^{01}$
$\beta_{NP,1}$	7.16	$2.63e^{-02}$	$2.72e^{02}$
$\beta_{TTD,2}$	0.00125	$2.61e^{-04}$	4.95^{00}
$\beta_{PP,2}$	1.31	$2.00e^{-02}$	$6.58e^{01}$
$\beta_{OD,2}$	0.0207	$2.65e^{-03}$	$8.24e^{00}$
$\nu_{f,1}$	1.08	$2.24e^{-03}$	$4.84e^{02}$
γ_1	0.0478	$3.05e^{-03}$	$1.52e^{01}$
$\nu_{f,2}$	0.984	$3.26e^{-02}$	$2.90e^{01}$
γ_2	0.186	$2.51e^{-03}$	$7.38e^{01}$
μ_{α_1}	0.0985	$4.09e^{-03}$	$2.32e^{01}$
μ_{α_2}	0.121	$1.87e^{-02}$	$5.98e^{00}$
$\log \mathcal{L}$	-519050.63		
Number of parameters	13		
Number of observations	747385		

Table 4: Estimation results

Model	Weidmann, 1993	Tregenza, 1976	Rastogi et al., 2013	Linear	*PedProb-vk*	*MC-vk*
MSE	$4.81e^{-03}$	$3.63e^{-03}$	$3.95e^{-03}$	$4.99e^{-03}$	$3.17e^{-03}$	$1.51e^{-03}$
$\bar{R}^2$	$2.64e^{-01}$	$4.45e^{-01}$	$3.96e^{-01}$	$2.37e^{-01}$	$5.16e^{-01}$	$7.69e^{-01}$

Table 7: Goodness of fit - MSE and $\bar{R}^2$

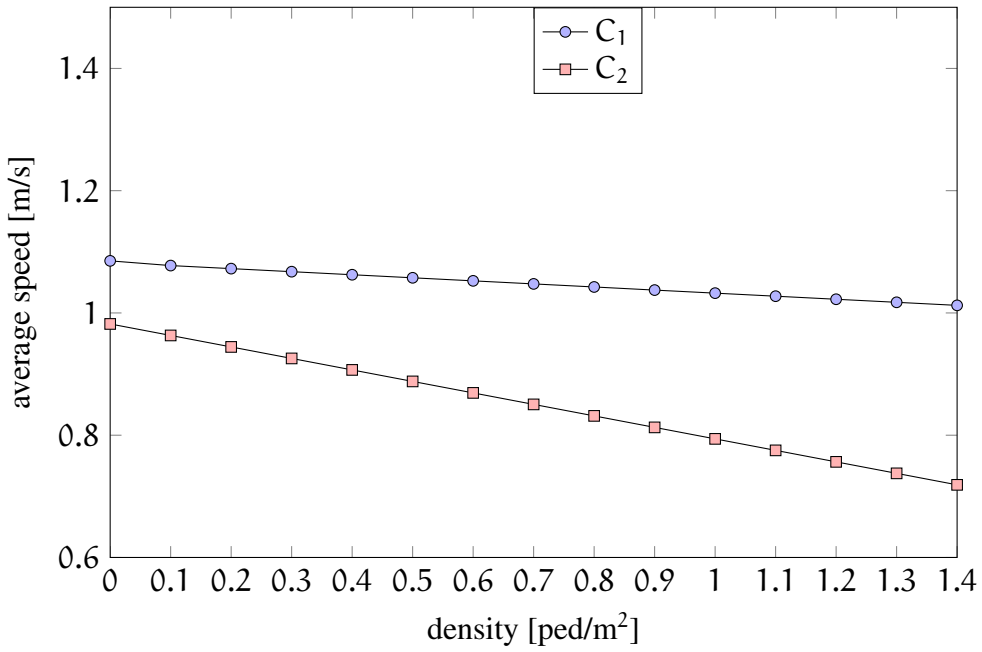
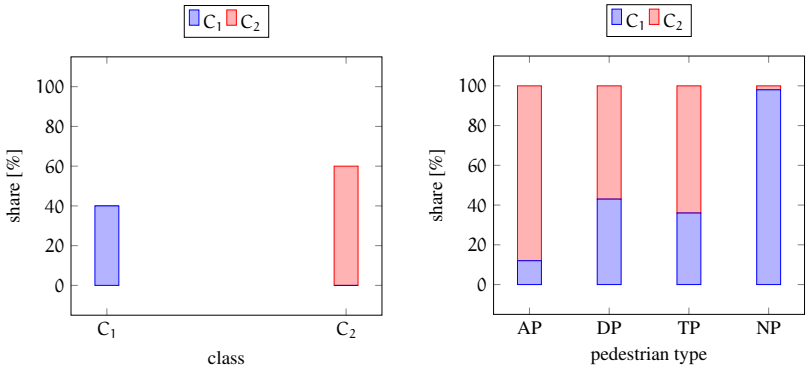
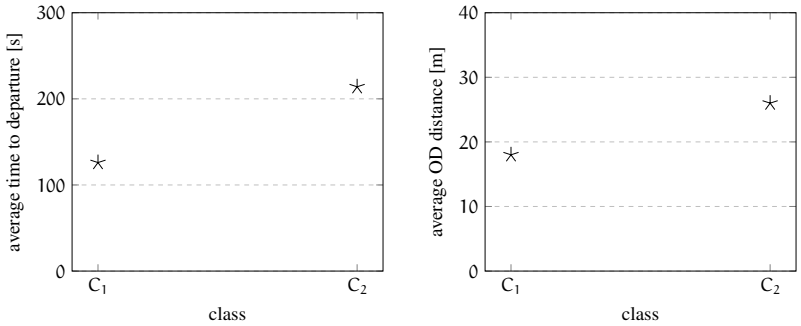


Figure 12: Class-specific speed-density relationships



(a) Overal (b) Per pedestrian type

Figure 13: Shares per class



(a) Avg. time to departure (b) Avg. OD distance

Figure 14: Class profiling

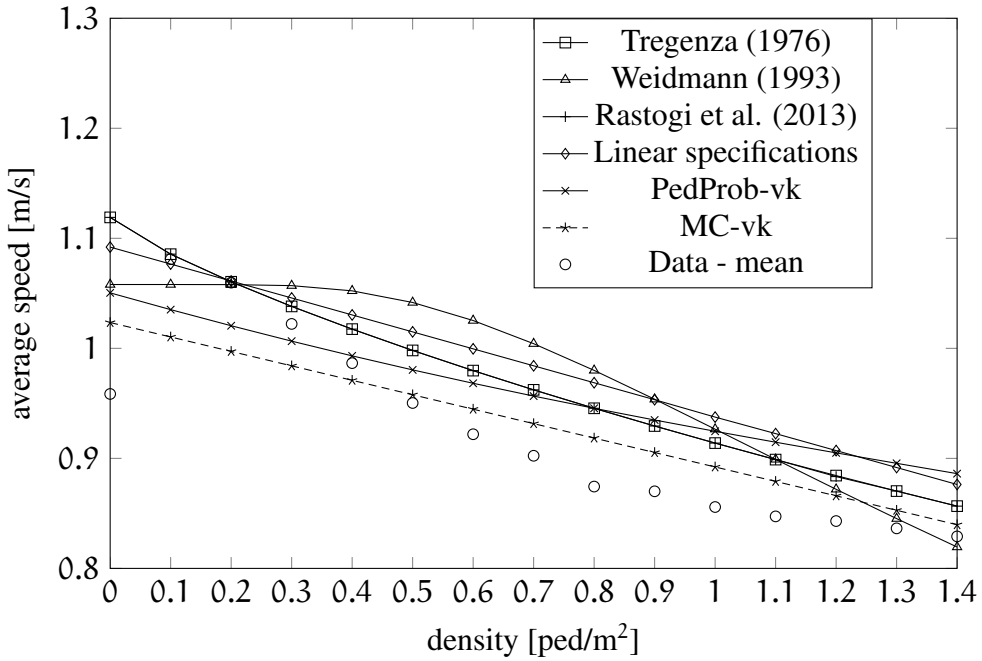


Figure 15: Comparison between deterministic models predictions and the aggregated probabilistic model predictions

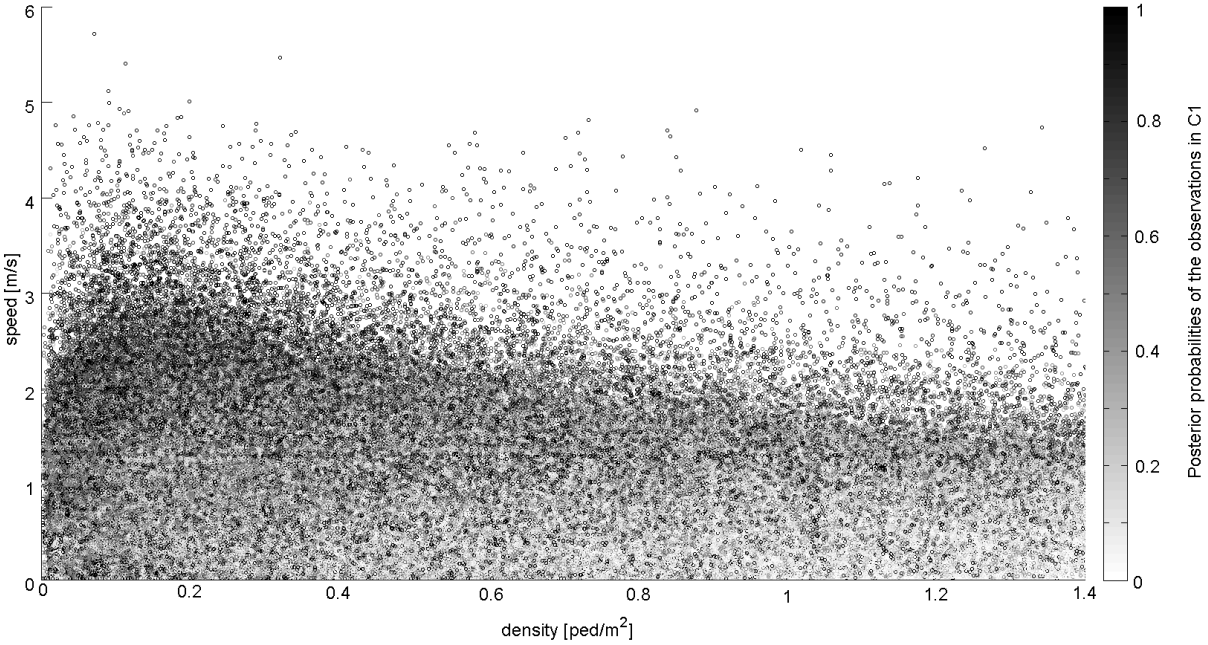


Figure 16: Posterior probabilities of the observations for C_1

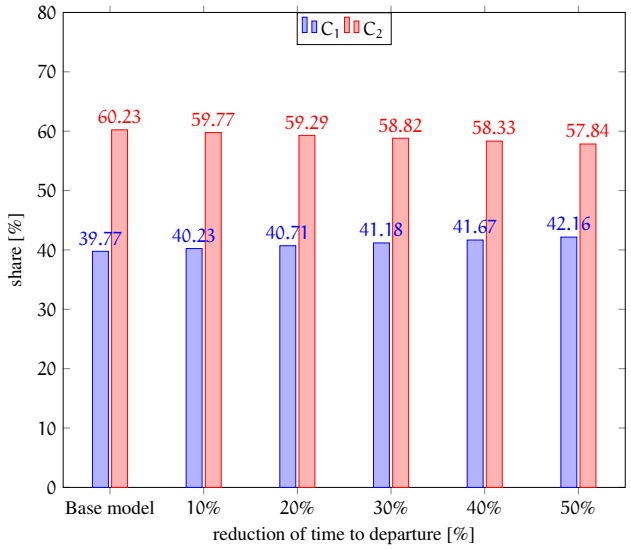


Figure 17: Scenario analysis - train timetable modification

References

Alahi, A., Bagnato, L., Chanel, D. and Alahi, A. (2013). Technical report for SBB network of sensors, *Technical report*, VisioSafe SA, Switzerland.

Alahi, A., Bierlaire, M. and Vandergheynst, P. (2014). Robust real-time pedestrians detection in urban environments with a network of low resolution cameras, *Transportation Research Part C: Emerging Technologies* **39**: 113–128.

Alahi, A., Jacques, L., Boursier, Y. and Vandergheynst, P. (2011). Sparsity driven people localization with a heterogeneous network of cameras, *Journal of Mathematical Imaging and Vision* **41**(1-2): 39–58.

Algadhi, S. A. and Mahmassani, H. S. (1990). Modelling crowd behavior and movement: application to makkah pilgrimage, *Transportation and traffic theory* **1990**: 59–78.

Bachu, P., Dudala, T. and Kothuri, S. (2001). Prompted recall in global positioning system survey: Proof-of-concept study, *Transportation Research Record: Journal of the Transportation Research Board* (1768): 106–113.

Ball, R., Ghorpade, A., Nawarathne, K., Baltazar, R., Pereira, F. C., Zegras, C. and Ben-Akiva, M. (2014). Battery patterns and forecasting in a large-scale smartphone-based travel survey, *10th International Conference on Transport SurveyMethods, Fairmont Resort, Jamison Valley,BlueMountains National Park, Australia*.

Bauer, D., Brandle, N., Seer, S., Ray, M. and Kitazawa, K. (2009). Measurement of pedestrian movements: A comparative study on various existing systems, *in* H. Timmermans (ed.), *Pedestrian Behavior: Models, Data Collection and Applications*, Emerald Group Publishing Limited, pp. 325–344.

Bierlaire, M. and Robin, T. (2009). Pedestrians choices, *in* H. Timmermans (ed.), *Pedestrian Behavior. Models, Data Collection and Applications*, Emerald Group Publishing Limited, pp. 1–26. ISBN:978-1-84855-750-5.

Blue, V. and Adler, J. (1998). Emergent fundamental pedestrian flows from cellular automata microsimulation, *Transportation Research Record: Journal of the Transportation Research Board* (1644): 29–36.

Bohte, W. and Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands, *Transportation Research Part C: Emerging Technologies* **17**(3): 285–297.

Bowman, B. L. and Vecellio, R. L. (1994). Pedestrian walking speeds and conflicts at urban median locations, *Transportation research record* (1438): 67–73.

Campanella, M., Hoogendoorn, S. and Daamen, W. (2009). Effects of heterogeneity on self-organized pedestrian flows, *Transportation Research Record: Journal of the Transportation Research Board* (2124): 148–156.

Cheung, C. and Lam, W. H. (1998). Pedestrian route choices between escalator and stairway in mtr stations, *Journal of transportation engineering* **124**(3): 277–285.

Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M. and Zegras, P. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in singapore, *Transportation Research Record: Journal of the Transportation Research Board* (2354): 59–67.

Daamen, W. (2004). *Modelling passenger flows in public transport facilities*, PhD thesis, Delft University of Technology, Delft.

Daamen, W. and Hoogendoorn, S. P. (2003). Controlled experiments to derive walking behaviour, *European Journal of Transport and Infrastructure Research* **3**(1): 39–59.

Daamen, W., Hoogendoorn, S. P. and Bovy, P. H. (2005). First-order pedestrian traffic flow theory, *Transportation Research Record: Journal of the Transportation Research Board* **1934**(1): 43–52.

Danalet, A. (2015). *Activity choice modeling for pedestrian facilities*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Danalet, A., Farooq, B. and Bierlaire, M. (2014). A bayesian approach to detect pedestrian destination-sequences from wifi signatures, *Transportation Research Part C: Emerging Technologies* **44**: 146–170.

DiNenno, P. J. (2002). *SFPE handbook of fire protection engineering*, National Fire Protection Association, Quincy, Massachusetts.

Duives, D. C., Daamen, W. and Hoogendoorn, S. (2014). Trajectory analysis of pedestrian crowd movements at a dutch music festival, *Pedestrian and Evacuation Dynamics 2012*, Springer, pp. 151–166.

Duives, D. C., Daamen, W. and Hoogendoorn, S. P. (2013). State-of-the-art crowd motion simulation models, *Transportation Research Part C: Emerging Technologies* **37**: 193–209.

Flötteröd, G. and Lämmel, G. (2015). Bidirectional pedestrian fundamental diagram, *Transportation research part B: methodological* **71**: 194–212.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American statistical Association* **97**(458): 611–631.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, Springer Series in Statistics, Springer Science & Business Media, LLC.

Fruin, J. J. (1971). Designing for pedestrians: A level-of-service concept, number 355, Highway Research Board, Washington, DC, pp. 1–15.

Ge, W., Collins, R. T. and Ruback, R. B. (2012). Vision-based analysis of small groups in pedestrian crowds, *IEEE transactions on pattern analysis and machine intelligence* **34**(5): 1003–1016.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory, *Econometrica: Journal of the Econometric Society* pp. 681–700.

Greenshields, B., Bibbins, J., Channing, W. and Miller, H. (1935). A study of traffic capacity, *Proceedings of the Highway Research Board*, Vol. 14, Highway Research Board, Washington, DC.

Hänseler, F. (2016). *Modeling and estimation of pedestrian flows in train stations*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Hänseler, F., Bierlaire, M., Farooq, B. and Mühlematter, T. (2014). A macroscopic loading model for time-varying pedestrian flows in public walking areas, *Transportation Research Part B: Methodological* **69**: 60 – 80.

Hänseler, F., Lam, W., Bierlaire, M., Lederrey, G. and Nikolic, M. (2017). A dynamic network loading model for anisotropic and congested pedestrian flows, *Transportation Research Part B: Methodological* **95**: 149–168.

Helbing, D., Buzna, L., Johansson, A. and Werner, T. (2005). Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions, *Transportation science* **39**(1): 1–24.

Helbing, D., Johansson, A. and Al-Abideen, H. Z. (2007). Dynamics of crowd disasters: An empirical study, *Physical review E - Statistical, Nonlinear, and SoftMatter Physics* **75**(4): 1–7.

Helbing, D. and Molnar, P. (1995). Social force model for pedestrian dynamics, *Physical review E* **51**(5): 4282–4286.

Hoogendoorn, S. and Daamen, W. (2004). Self-organization in walker experiments, *Traffic and Granular Flow*, Vol. 3, pp. 121–132.

Hoogendoorn, S. P. and Bovy, P. H. (2004). Pedestrian route-choice and activity scheduling theory and models, *Transportation Research Part B: Methodological* **38**(2): 169–190.

Hoogendoorn, S. P., van Wageningen-Kessels, F. L., Daamen, W. and Duives, D. C. (2014). Continuum modelling of pedestrian flows: From microscopic principles to self-organised macroscopic phenomena, *Physica A: Statistical Mechanics and its Applications* **416**: 684–694.

Jabari, S. E., Zheng, J. and Liu, H. X. (2014). A probabilistic stationary speed–density relation based on Newell's simplified car-following model, *Transportation Research Part B: Methodological* **68**: 205–223.

Jintanakul, K., Chu, L. and Jayakrishnan, R. (2009). Bayesian mixture model for estimating freeway travel time distributions from small probe samples from multiple days, *Transportation Research Record: Journal of the Transportation Research Board* **2136**(1): 37–44.

Johansson, A., Helbing, D. and Shukla, P. K. (2007). Specification of the social force pedestrian model by evolutionary adjustment to video tracking data, *Advances in complex systems* **10**(2): 271–288.

Kalakou, S., Bierlaire, M. and Moura, F. (2014). Effects of terminal planning on passenger choices, *14th Swiss Transport Research Conference (STRC), Monte Verità, Ascona, Switzerland*.

Kanda, T., Shiomi, M., Perrin, L., Nomura, T., Ishiguro, H. and Hagita, N. (2007). Analysis of people trajectories with ubiquitous sensors in a science museum, *Proceedings 2007 IEEE International Conference on Robotics and Automation*, IEEE, pp. 4846–4853.

Kim, T. and Zhang, H. (2008). A stochastic wave propagation model, *Transportation Research Part B: Methodological* **42**(7): 619–634.

Lam, W. H. and Cheung, C.-y. (2000). Pedestrian speed/flow relationships for walking facilities in hong kong, *Journal of transportation engineering* **126**(4): 343–349.

Lam, W. H., Lee, J. Y. and Cheung, C. (2002). A study of the bi-directional pedestrian flow characteristics at hong kong signalized crosswalk facilities, *Transportation* **29**(2): 169–192.

Lam, W. H., Morrall, J. F. and Ho, H. (1995). Pedestrian flow characteristics in Hong Kong, *Transportation Research Record* (1487): 56–62.

Lee, J.-G., Han, J. and Whang, K.-Y. (2007). Trajectory clustering: a partition-and-group framework, *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, ACM, pp. 593–604.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American statistical Association* **46**(253): 68–78.

McPhail, C. and Wohlstein, R. T. (1982). Using film to analyze pedestrian behavior, *Sociological Methods & Research* **10**(3): 347–375.

Naini, F. M., Dousse, O., Thiran, P. and Vetterli, M. (2011). Population size estimation using a few individuals as agents, *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, IEEE, pp. 2499–2503.

Navin, F. and Wheeler, R. (1969). Pedestrian flow characteristics, *Traffic Engineering, Inst Traffic Engr* **39**: 30–36.

Nikolić, M., Bierlaire, M., Farooq, B. and de Lapparent, M. (2016). Probabilistic speed–density relationship for pedestrian traffic, *Transportation Research Part B: Methodological* **89**: 58–81.

Ohmori, N., Nakazato, M. and Harata, N. (2005). Gps mobile phone-based activity diary survey, *Proceedings of the Eastern Asia Society for Transportation Studies*, Vol. 5, pp. 1104–1115.

Older, S. (1968). Movement of pedestrians on footways in shopping streets, *Traffic engineering and control* **10**: 160–163.

Rastogi, R., Ilango, T. and Chandra, S. (2013). Pedestrian flow characteristics for different pedestrian facilities and situations, *European Transport* **53**: 1–21.

Schwarz, G. et al. (1978). Estimating the dimension of a model, *The annals of statistics* **6**(2): 461–464.

Srivastava, A. K., Srivastava, V. K. and Ullah, A. (1995). The coefficient of determination and its adjusted version in linear regression models, *Econometric reviews* **14**(2): 229–240.

Steffen, B. and Seyfried, A. (2010). Methods for measuring pedestrian density, flow, speed and direction with minimal scatter, *Physica A: Statistical mechanics and its applications* **389**(9): 1902–1910.

Tanaboriboon, Y., Hwa, S. S. and Chor, C. H. (1986). Pedestrian characteristics study in singapore, *Journal of Transportation Engineering* **112**(3): 229–235.

Tregenza, P. (1976). *The design of interior circulation*, Van Nostrand Reinhold, New York, USA.

Van den Heuvel, J. and Hoogenraad, J. (2014). Monitoring the performance of the pedestrian transfer function of train stations using automatic fare collection data, *Transportation Research Procedia* **2**: 642–650.

van Wageningen-Kessels, F. (2013). *Multi-class continuum traffic flow models: analysis and simulation methods*, PhD thesis, Delft University of Technology/TRAIL Research school, Delft.

Walker, J. L. and Li, J. (2007). Latent lifestyle preferences and household location decisions, *Journal of Geographical Systems* **9**(1): 77–101.

Wang, H., Ni, D., Chen, Q.-Y. and Li, J. (2013). Stochastic modeling of the equilibrium speed–density relationship, *Journal of Advanced Transportation* **47**(1): 126–150.

Weidmann, U. (1993). Transporttechnik der fussgänger, *Technical Report Schriftenreihe des IVT Nr. 90*, Institut für Verkehrsplanung,Transporttechnik,Strassen- und Eisenbahnbau, ETH Zürich. (In German).

Wong, S., Leung, W., Chan, S., Lam, W. H., Yung, N. H., Liu, C. and Zhang, P. (2010). Bidirectional pedestrian stream model with oblique intersecting angle, *Journal of transportation Engineering* **136**(3): 234–242.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity, *Journal of applied econometrics* **20**(1): 39–54.

Yaeli, A., Bak, P., Feigenblat, G., Nadler, S., Roitman, H., Saadoun, G., Ship, H. J., Cohen, D., Fuchs, O., Ofek-Koifman, S. et al. (2014). Understanding customer behavior using indoor location analysis and visualization, *IBM Journal of Research and Development* **58**(5/6): 3:1–3:12.

Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardin, F., Carrascal, J. P., Blat, J. and Sinatra, R. (2014). An analysis of visitors' behavior in the louvre museum: A study using bluetooth data, *Environment and Planning B: Planning and Design* **41**(6): 1113–1131.

Zhang, J. (2012). *Pedestrian fundamental diagrams: Comparative analysis of experiments in different geometries*, PhD thesis, Forschungszentrum Jülich.

Zhao, F., Pereira, F. C., Ball, R., Kim, Y., Han, Y., Zegras, C. and Ben-Akiva, M. (2015). Exploratory analysis of a smartphone-based travel survey in singapore, *Transportation Research Record: Journal of the Transportation Research Board* **2**(2494): 45–56.