A Bayesian Markov Model for Station-Level Origin-Destination Matrix Reconstruction

Victor Amblard, Amir Dib, Noëlie Cherrier ✉, and Guillaume Barthe

CITiO, 22 rue René Boulanger, 75010 Paris, France
{firstname}.{lastname}@cit.io

Abstract. This paper tackles Origin-Destination (OD) matrix reconstruction at a station level, which consists in computing the volume of passengers traveling between two different stations on a public transportation network. This information is critical for the transport operator to compute various indicators concerning the network's state and performance such as vehicle occupancy and travelers' behavior. Trip reconstruction for smart card holders, whose history of validations is available, has been thoroughly investigated in prior work. Conversely, trip reconstruction for non smart card holders has received less attention, mainly due to the difficulty of obtaining ground truth data. Among recent work in this domain, very few contributions have tackled large networks in their entirety, with millions of validations over a month and the computational challenges that come with it.

In this work, we present a new *Bayesian Markov Model* for OD matrix reconstruction. The novelty of our model lies in its scalability and the fact that it uses all available data, including Automated Fare Collection (*i.e.* smart card holders) data and Automatic Passenger Counting data (*i.e.* data from counting sensors), to accurately infer the trips' distribution. Moreover, the proposed approach produces proper OD matrices while taking into account sensor noise and fraud.

We empirically establish the relevance, robustness, and accuracy of the proposed method compared to the popular trip chaining algorithm and a previous Markov based approach on real-world, large-scale industrial datasets for two transportation networks in major cities.

Keywords: Origin Destination matrix · Bayesian · Markov model · Real world data · Automatic Passenger Counting data.

1 Introduction

Origin-Destination (OD) matrix reconstruction is a key element of public transport management. It provides insights regarding the network's performances and state, which drive strategic decisions regarding the network configuration, such as determining the line routing or evaluating the optimal level of service. OD reconstruction consists in reconstructing the flow of passengers who traveled from one station (origin) to another (destination) during a given period. The OD matrix is defined as the flows for all possible pairs of stations in the network. Since the origin stations are known in most cases (through user ticket validations when boarding the vehicle), accurately reconstructing these flows boils down to reconstructing passengers' alighting stations. For smartcard holders, most current approaches rely on a procedure called *trip chaining* that leverages consecutive validations within a predefined time frame. Each validation is tracked thanks to the related smart card unique identifier, and the associated alighting station is deduced from consecutive boarding stations.

Although very effective, this approach cannot be applied to single-use ticket holders or even smartcard holders whose behavior is not compatible with trip chaining rules based on expert knowledge. These drawbacks motivate the exploration of alternative approaches that use external sensors as additional data to reconstruct passengers' trips. Akin to traffic counts that provide information about vehicles entering and exiting a network of highways, counting cells are sensors installed at the vehicles' doors to count the number of passengers boarding and alighting the vehicle at each station. The availability of data from these detectors can often counterbalance the lack of information about individual passengers. However, the uncertainty associated with these sensors' measurements is significant due to intrinsic sensor noise and high false detection rates (passengers may trigger multiple detections). Hence, filtering and denoising raw sensor data is mandatory for these countings to be used. Finally, these sensors can be costly to install and maintain for transport operators leading to a partial equipment rate of the vehicle fleet. Altogether, these issues make OD reconstruction challenging and call for end-to-end approaches that consider sensor quality, scarcity, and scalability.

In this work, we propose a novel full Bayesian Markov-based model for OD reconstruction that considers all commonly available data sources. Our approach is based on the finding that sampling OD matrices based on Markov chain modeling of agents' behavior amounts to drawing from a multivariate hypergeometric distribution. Moreover, we overcome the short trip problem, which is the main drawback of such an approach, by considering a biased version of the hypergeometric sampling. Subsequently, we tackle two problems that commonly arise when dealing with real-world data: noise and scarcity. We propose a new denoising method for counting sensors that preserve the OD matrix structure and use a time series similarity metric to deal with unequipped vehicles. Finally, we show that this approach can be applied to large-scale networks with real-world data to better reconstruct the flow of passengers.

Section 2 introduces the basic concepts of OD matrix and trip reconstitution along with related works. Then, Section 3 presents the various aspects of our approach toward station-level OD reconstruction. Finally, Section 4 is devoted to the practical evaluation of our approach on real-world industrial use cases. Detailed proofs and derivations are deferred to the supplementary material.

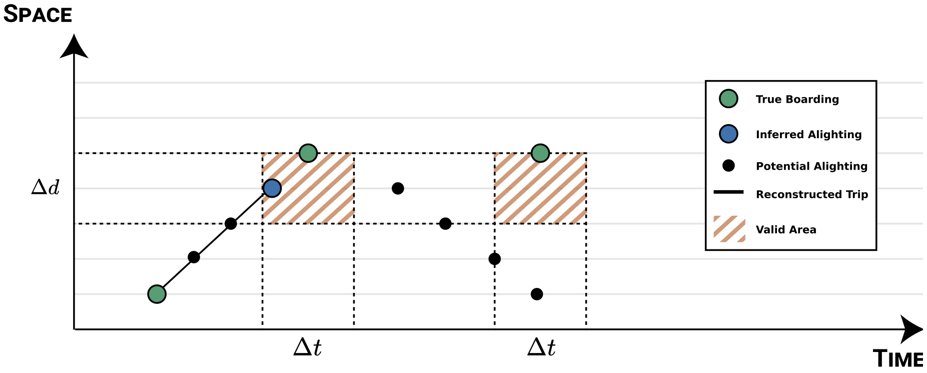

Fig. 1: Illustration of trip chaining with deterministic rules: the first trip is chained since a candidate alighting station lies within the time and distance thresholds Δt and Δd, while the second trip cannot be chained since no candidate alighting station abides by the thresholds.

2 Related work

Historically, OD matrices were obtained as part of the four-step model [25] for demand modeling using fully deterministic models inspired by physics such as the gravity [32] and the entropy models [31] are the best-known examples.

This work focuses on OD matrix reconstruction in public transport, a subfield of OD reconstruction that presents a few peculiarities, notably considering the amount and quality of available data. Thanks to the recent advancements in technologies, many transportation agencies are now using Automatic Data Collection (ADC) systems, that include Automated Fare Collection (AFC) systems, *i.e.* smart cards most of the time; Automatic Vehicle Location (AVL) systems, giving access to real arrival time of vehicles to stations; and Automatic Passenger Counting (APC) systems, with sensors installed on board the vehicles.

Although these increasingly abundant sources of data have been used for various applications in the last two decades (mining travel patterns, trip purpose detection among others) [5], this work tackles another major application which is station-level OD reconstruction (a review can be found in [13]). More specifically, it focuses on estimating alighting locations from known boarding locations (thanks to smart card validation data).

Until now, the area of OD reconstruction has been dominated by rule-based approaches using smart card data. Notably, *trip chaining* is a method that infers alighting locations from successive boarding locations, supposing the user has not traveled more than a distance threshold during a time threshold between the sought alighting and the next boarding [29,21,13] (see Fig. 1). Other advanced methods are probabilistic [11,17,12] or based on the full user's history [29,11,18,19].

Recently, increasing attention has been drawn to machine learning approaches [34], notably with neural networks [15,1]. Machine learning is expected to bridge the gap between different data sources, e.g. smartphone location data [33,9] or land use data [23,28]. More recent works use graph convolutional networks to infer OD flows [24,35,23] but require labeled training data.

Finally, while the vast majority of the literature focuses on exploiting AVL and AFC (*i.e.* vehicle location and smart card) data, only a few studies make use of other sources of data and especially APC data. APC data is mostly used as a scaling factor to the OD matrix extracted from previous methods, using methods such as Iterative Proportional Fitting (IPF) [2,26,14,7]. However, IPF as well as other optimization methods [16,22] do not enable any uncertainty estimation. On the opposite, statistical frameworks have been proposed and notably Bayesian approaches [20,10,36], with the recurrent drawback of being hardly scalable to larger and more complex networks. Also, the work from [3] derived a statistical approach that is inspired by the maximum entropy method, and the study from [14] proposed a Markov-chain Monte Carlo method to infer route OD with large amounts of APC data only.

However, few of these works consider the imprecision associated to APC data: they are usually considered 100% reliable while studies estimated the accuracy of standard infrared sensors to be around 80% [8]. Evaluating the quality and accuracy of the counting instruments is hard, while APC data can cover the entire network and make indicators easy to calculate [4]. In addition, the existing methods often lack validation through real transportation data, and when a validation procedure is proposed it is often on a very small perimeter, missing demonstration of scalability [13]. For instance, [27] validate their approach with an OD survey and a group of volunteers. This work proposes a denoising module for the counting cells data to be used more reliably by a Bayesian Markov model for OD reconstruction. The experiments are conducted on real data collected from Casablanca (Morocco) and Orléans Métropole (France) public transport networks.

3 Origin destination matrix reconstruction using ticketing and count data

This section describes the different steps of the proposed OD matrix reconstruction procedure. This method is based on a Bayesian Markov model inspired by [20] that takes into account the validations (AFC) and counting cells (APC) data per course. The latter is first denoised to get valid boarding and alighting counts. Then, a biased hypergeometric sampling integrating priors on trip lengths is proposed to simulate trips for each course based on the denoised counts. Finally, the posterior parameters of the Markov model are inferred and extrapolated to courses without counting cells.

In what follows, we consider a network with different routes (*i.e.* lines with specified directions). A course corresponds to a vehicle following a given route with a predefined schedule. For clarity, unless otherwise specified, we focus on a single course that occurs on a specific route. Let us denote by n the number of stations on the route and Y_i, Z_i respectively the number of passengers boarding and alighting at station i and p_{ij} the probability of alighting at station j conditionally to the fact that a passenger boarded at station i. The passengers are assumed to *behave the same way and independently of one another*. The passengers' behavior is described by a non-homogeneous Markov chain valued on a binary state space.

The inference relies on alighting counts, which in this case stem from counting cells measures and are typically tainted with imprecision. The following aims at correcting the counting cells noise before any inference of the model.

3.1 Count data preprocessing

This part presents a preprocessing method for counting cells measures. Due to multiple factors, all referred to as noise in what follows, the actual observed boarding counts $\tilde{Y_i}$ (resp. alighting counts $\tilde{Z_i}$) differ from the real ones by a noise $\eta_i^{+,IN} - \eta_i^{-,IN}$ (resp. $\eta_i^{+,OUT} - \eta_i^{-,OUT}$):

$$\tilde{Y_i} = Y_i + \eta_i^{+,IN} - \eta_i^{-,IN}, \quad \eta_i^{+,IN} \sim \text{Bin}(Y_i, p^+), \quad \eta_i^{-,IN} \sim \text{Bin}(Y_i, p^-). \quad (1)$$

The same applies for $(\tilde{Z_i})_i$ with $\eta_i^{+,OUT}$ and $\eta_i^{-,OUT}$ also following binomial distributions $\text{Bin}(Z_i, p^+)$ and $\text{Bin}(Z_i, p^-)$. Note that the noise is not required to be symmetric, as counting cells may over-count more than they under-count or conversely.

Fraud removal In this work, only trips corresponding to passengers who validated their tickets are reconstructed. Counting cells, however, record all passengers entering and exiting the vehicle, regardless of whether they did validate. Therefore, the total passenger count Z_i alighting at station i of a given course must be disaggregated between Z_i^F fraudsters and Z_i^V persons who validated their ticket. To estimate the number of fraudsters on board, a two-step approach first determines the total number of fraudsters during the course and then allocates them to different stations.

Let F be the total number of fraudsters on a given course and S the total number of passengers on the course. $S = F + \mathbb{1}^T Y^V$ is unobserved since neither the true boarding nor alighting counts are known. Nevertheless, two noisy versions of it are observed: $\tilde{S}_Y = \mathbb{1}^T \tilde{Y}$ and $\tilde{S}_Z = \mathbb{1}^T \tilde{Z}$. Therefore, S is the sum of the observed count $\tilde{S_Y}$ plus the sum of the noises for each station measure. Since the $(\eta_i^{+,IN})_i$ and $(\eta_i^{-,IN})_i$ are i.i.d variables, their sum also follows a binomial distribution of parameters $(\sum_i Y_i = S, p^+)$ and (S, p^-) respectively. Formally,

$$\tilde{S_Y} = \sum_{i=1}^{n} \tilde{Y_i} = S + \eta_\Sigma^{+,IN} - \eta_\Sigma^{-,IN}, \ \eta_\Sigma^{+,IN} \sim \text{Bin}(S, p^+), \ \eta_\Sigma^{-,IN} \sim \text{Bin}(S, p^-). \quad (2)$$

Thanks to the conditional independence between $\tilde{S_Y}$ an $\tilde{S_Z}$ conditionally to S, the posterior distribution for S is derived and therefore the number of frauding passengers F is sampled from this distribution:

$$\begin{aligned} p(S|\tilde{S_Y}, \tilde{S_Z}) &\propto p(\tilde{S_Y}|S)p(\tilde{S_Z}|S)p(S) \\ &\propto p(\eta_{\Sigma}^{+,IN} - \eta_{\Sigma}^{-,IN}|S)p(\eta_{\Sigma}^{+,OUT} - \eta_{\Sigma}^{-,OUT}|S)p(S). \end{aligned} \tag{3}$$

In addition, each station i is assigned a predetermined fraud rate f_i. The f_i can either be provided as prior expert knowledge or computed as the average fraud rate from boarding counting and ticketing data over all courses passing by station i in the opposite direction, making the hypothesis that the alighting fraudsters rate in one direction, aggregated over a sufficient number of courses, is approximated by the boarding fraudsters rate in the opposite direction. From there, the F fraudsters of a given course are disaggregated into F_i fraudsters alighting at station i, by sampling them from a Fisher's noncentral hypergeometric distribution with weights $\tilde{f_i}$ and initial number of objects $\tilde{Z_i}$. The F_i are removed from the $\tilde{Z_i}$ to yield adjusted alighting counts denoted as $\tilde{Z_i}^{ad} = \tilde{Z_i} - F_i$.

Alighting counts denoising with Gibbs sampling The following aims at refining the adjusted alighting counts $\tilde{Z_i}^{ad}$ to obtain a denoised alighting sequence that matches the validations boarding counts Y_i^V. Such an alighting sequence is further referred to as a *feasible alighting sequence*.

Definition 1. *A* feasible alighting sequence *with respect to a boarding sequence* $Y = (Y_1, ..., Y_{n-1}, 0) \in \mathbb{N}^n$ *is a sequence* $Z = (0, Z_2, ..., Z_n) \in \mathbb{N}^n$ *such that*

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} Z_i, \tag{4a}$$

$$\forall i \in [\![1, n-1]\!], \quad \sum_{k=1}^{i} Y_k \geq \sum_{k=1}^{i} Z_k. \tag{4b}$$

The feasible alighting set *is the set* $\mathcal{S}(Y)$ *of feasible alighting sequences w.r.t.* Y.

Conditions (4a) and (4b) simply enforce the following two physical constraints: the number of boarding passengers must be equal to those alighting during the course, and occupancy must always be nonnegative. The goal is thus to select a feasible alighting sequence close to the observed one ($\tilde{Z_i}$). Although the noise model presented in Eq. (1) is quite simple, the dependencies between the Z_i stemming from constraints (4a) and (4b) make it impossible to sample each count independently and call for a more sophisticated sampling algorithm. Hence, a Gibbs sampler approach is adopted to iteratively sample one of the alighting counts Z_i conditionally to all others sampled so far, so that constraint (4b) is satisfied all the times. Note that to abide by condition (4a), one of the values of the alighting sequence must act as a pivot. Z_1 is arbitrarily chosen to balance the sum of the remaining Z_i. From the noise model defined in Eq. (1), the conditional posterior probability of Z_i given $Z_{-i} = (Z_2, ..., Z_{i-1}, Z_{i+1}, ..., Z_n)$, Y and $\tilde{Z}$ writes

$$\begin{aligned} \forall k \in \mathbb{N}, \quad p(Z_i = k|Y, \tilde{Z}^{ad}, Z_{-i}) &\propto p(\tilde{Z_i}^{ad}|Z_i = k)\, p(Z_i = k) \\ p\left(\tilde{Z_1}^{ad}|Z_1 = S - \sum_{k \neq i} Z_k - k\right) &p\left(Z_1 = S - \sum_{k \neq i} Z_k - k\right). \end{aligned} \tag{5}$$

Algorithm 1 Gibbs sampler

```
Require: $Y, \tilde{Z}^{ad}, N_{IT}$ the number of sampling iterations, $n$ the number of stops

  $z^0 = (0, Y_1, ..., Y_{n-1})$
  for $t \in [\![1, N_{IT}]\!]$ do
    for $j \in [\![2, n]\!]$ do
      $m_j = \max_{i \in [\![1,j]\!]} \sum_{k=i+1}^{n} Y_k - \left( \sum_{k=i+1}^{j-1} z_k^t - \sum_{k=j+1}^{n} z_k^{t-1} \right)$
      $M_j = \mathbb{1}^T Y - \left( \sum_{k=2}^{j-1} z_k^t + \sum_{k=j+1}^{n} z_k^{t-1} \right)$
      For $k \in [\![m_j, M_j]\!]$, $p_k = p(Z_j = k | \tilde{Z}^{ad}, Z_1 = z_1^t, ..., Z_{j-1} = z_{j-1}^t, Z_{j+1} = z_{j+1}^{t-1}, ..., Z_n = z_n^{t-1})$
      Sample $z_j^t \sim$ Discrete($[p_{m_i}, ..., p_{M_i}]$)
    end for
    $z^t = \left( 0, \mathbb{1}^T Y - \sum_{k=2}^{n} z_k^t, z_2^t, ..., z_n^t \right)$
  end for
  return $Z = z^{N_{IT}}$
```

One can show that conditions (4a) and (4b) imply that Z_i has a finite support, *i.e.* that there exist two non-negative integers m_i and M_i such that $p(Z_i = k) = 0$ for all k not in $[\![m_i, M_i]\!]$. The full conditional probability is finally derived in closed form, provided that a prior is chosen for the true alighting counts. If no information is available, one could choose a uniform prior over the interval $[\![m_i, M_i]\!]$ for the alighting count Z_i. Finally, Gibbs sampling requires a valid initialization, *i.e.* an initial alighting sequence that belongs to the feasible alighting set. For instance, $z^0 = (0, Y_1^V, ..., Y_{n-1}^V)$ is a feasible alighting sequence w.r.t. Y^V. The full algorithm is presented in Algorithm 1. An improved initialization to reduce the number of iterations is proposed in the supplementary material.

3.2 Trips sampling and posterior estimation

The model presented in this section uses the denoised alighting counts for posterior parameter estimation and trip sampling. A first-order Markov model is first described as a basis for the proposed approach.

Definition 2. *The first-order Markov model is defined by a set of $n-1$ parameters $(\theta_2, ..., \theta_n)$ such that for all $i \in [\![2, n]\!]$,*

$$p(\xi_i = 0 | \xi_{i-1} = 1) = \theta_i, \quad p(\xi_i = 1 | \xi_{i-1} = 1) = 1 - \theta_i, \tag{6}$$

with ξ_i the variable indicating if the passenger is on board as the vehicle departs from station i ($\xi_i = 1$) or not ($\xi_i = 0$).

The above statement conveys that the model is without memory and "forgets" about the passengers' boarding stations, only focusing on whether they were on board the vehicle when it arrived at a station i.

The Markov property allows for simple derivation of the probability p_{ij} for a passenger to alight at a station j provided that they boarded at station i [20]

$$p_{ij} = \theta_j \prod_{k=i+1}^{j-1} (1 - \theta_k). \tag{7}$$

The parameters' likelihood is directly derived from the boarding and alighting counts [20] and writes

$$Z_i | \theta_i \sim Bin \left(\sum_{k=1}^{i-1} Y_k - Z_k, \theta_i \right). \tag{8}$$

Finally, to obtain a full Bayesian model, it is needed to choose a prior distribution over the set of parameters $(\theta_i)_{i=1}^n$. For clarity and simplicity of derivations, we set $\theta_j \sim \text{Beta}\,(\alpha_j, \beta_j)$ with hyperparameters α, β inferred from the chained trips. Once sampled from the posterior distribution, the model's parameters θ are used in Section 3.3 to extrapolate to courses without counting cells.

However, the first-order Markov model's shortcoming lies in that all passengers are considered equal: they all share the same probability to alight at a given station regardless of their boarding station, as long as they are in the vehicle. As a direct consequence, the longer the trip, the less likely it is since $p_{ij} = \theta_j \prod_{k=i+1}^{j-1}(1 - \theta_k) \sim \mathcal{O}(\theta^{j-i})$. The probability of staying in the vehicle for $j - i$ stations decays exponentially. This is far from being realistic and clashes with empirical observations. Indeed, over various networks and cities, it is frequent for the mode of the trip length distribution to be located around a length of 5 with a slow decay followed by a more rapid decay. Therefore, the following proposes a sampling procedure to overcome the short trips issue.

Let us denote by X_{ij} the number of passengers that boarded the vehicle at station i and alighted at station j. Formally we are looking for $(X_{ij})_{i,j}$ given $(Y_k)_k, (Z_k)_k$. Here, $X = (X_{ij})_{i,j}$ is the OD matrix. Although in general the underlying true trip length distribution is unknown, chained trips provide insights into this distribution and prior information. Once estimated, these priors are used to bias our sampling procedure using a Fisher's noncentral hypergeometric distribution for $(X_{ij})_j | Y, Z, L$ where L are the trip length priors. The practical details are deferred to the supplementary material.

This result extends the work of [20] in case priors are available and explores how to leverage biased multivariate hypergeometric distributions to sample directly from the model. As shown in the experiments section, it also alleviates the so-called short trips issue. Algorithm 2 summarizes the different steps to reconstruct the OD matrix from counting cells observations.

Algorithm 2 OD matrix reconstruction for courses with counting cells on a given route

```
Require: $p(\theta)$ prior for the $\theta$ parameters of the first-order model, $L$ prior for
    trips lengths for the considered route, $\mathcal{C}$ all courses, $(f_i)_i$ fraud rates by
    station $i$, $N_{sim}$ number of simulations
for course $c_t \in \mathcal{C}$ do
    for $k \in [\![1, N_{sim}]\!]$ do
        $\tilde{Z}^{ad}_{k,t}$ = Remove fraud from alighting counts($Y_t^V, \tilde{Y}_t, \tilde{Z}_t, f_i$) (Section 3.1)
        $Z_{k,t}$ = Sample feasible alighting counts($\tilde{Y}^{ad}_{k,t}, \tilde{Z}^{ad}_{k,t}, Y_t^V$) (Section 3.1)
        $\theta_{k,t}$ = Infer posterior parameters($Y_t^V, Z_{k,t}, p(\theta)$) (Section 3.2)
        $X_{k,t}$ = Sample feasible OD matrix($Y_t^V, Z_{k,t}, L$) (Section 3.2)
    end for
    $X_t$ = Mode$\big((X_{k,t})_k\big)$
end for
return $(X_t)_t, (\theta_{k,t})_{k,t}$
```

3.3 Extrapolation to courses without counting cells

Most of the times, due to the high cost of equipping vehicles with counting cells, only a fraction of the fleet operates with them. This is problematic since the proposed approach relies on alighting counts to simulate alighting stations. However, other courses associated to the same route can be used to extrapolate the first-order model's parameters on non-equipped courses.

More specifically, consider a target course $c_{T,r}$ that is associated with route $r \in \mathcal{R}$. The idea is to match the target course to some of the courses with count data on the same route $(c_{t,r})_t$ which are available in the data history.

The proposed method considers a course as a temporal series based on its station-wise validations: $Y_t^V = (Y_{1,t}^V, ...Y_{n-1,t}^V)$. Two courses are said to be similar if their validations are similar, for some time-series similarity metric. Here, Dynamic Time Warping (DTW) [30] is used as the similarity metric. Similar courses are the k-nearest neighbors for the DTW metric with k set experimentally.

$\mathcal{C}(T, r)$ is then the subset of courses $\{c_{t,r} | c_{t,r} \text{is similar to } c_{T,r}\}$ that contains all courses matched to $c_{T,r}$ and $\Theta(T, r)$ is the set of parameters of the first-order Markov model for the matched courses: $\Theta(T, r) = \{(\theta_{1,t}, ..., \theta_{n,t})_t | t \in \mathcal{C}(T, r)\}$. Then, the parameters $\theta_{T,i}$ for the target course are sampled as follows for all stations i in $[\![1, n]\!]$:

$$\theta_{T,i} \sim \mathcal{N}(\overline{\theta}, \sigma_\theta), \tag{9}$$

with $\overline{\theta} = (\overline{\theta_1}, ..., \overline{\theta_n})$ the experimental average and $\sigma_\theta = (\sigma_{\theta_1}, ...\sigma_{\theta_n})$ the standard deviation of these parameters over all courses belonging to $\mathcal{C}(T, r)$.

4 Experiments

This section aims at testing the proposed improved Bayesian Markov model. Some considerations on time and space complexity are developed, and the accuracy and robustness of the proposed method are discussed.

4.1 Experimental setup

The experiments are performed on two different networks. The first one is the Casablanca network with two streetcar lines of 30-40 stations each, totaling more than 100,000 boarding validations per day on average. This network is of high interest since passengers validate when they board but also when they alight, therefore providing ground truth data. However, none of the streetcars are equipped with counting cells, which have been simulated for the experiments (see supplementary material). The Orléans Métropole network is used for scalability assessment. It has a more complex topology than Casablanca's, with more than 50 bus and streetcar lines, numerous connections, and around 70,000 validations and 2,000 courses per day. Counting cells data is available, but this network does not give access to ground truth data since passengers validate only when they board. Both networks are illustrated in the supplementary material.

Five simulations are performed for each course in the dataset to come up with five candidate alighting stations for each passenger. The mode (*i.e.* the most probable station) is designated as the assigned alighting station. In the simulations, p^+ and p^- the counting cells noise parameters are both set to 0.4. The predefined fraud rates f_i are the same for all stations (the absolute value is not important since they only serve as bias for the hypergeometric sampling). All algorithms are implemented in Python and can run on multiple cores. The *BiasedUrn* library [6] is used for hypergeometric sampling.

4.2 Scalability

Table 1 summarizes and compares the time and space complexity of trip chaining and the proposed model. For trip chaining, passengers without an alighting station are aggregated by boarding station and their alighting stations are inferred simultaneously for the whole batch, which is done in $\mathcal{O}(n^2)$. The proposed model utilizes counting cells at the course level and therefore has a time complexity that is growing linearly with the number of courses $|\mathcal{C}|$. Moreover, the time complexity is directly proportional to the number of simulations N_s. Regarding the space complexity, since all of the passengers' candidate alighting stations are stored, the space complexity is proportional to the number of passengers P and the number of simulations.

However, the implementation still runs comfortably on a laptop: for instance, running 10 simulations on a month of data for the Orléans Métropole network (more than two million validations) takes up to 2 hours on an Apple M1 processor. Moreover, if multiples cores are available, courses can be inferred independently on different cores, speeding up the simulation process.

Table 1: Time and space complexity comparison.

Model	Time complexity	Spatial complexity
Trip chaining	$\mathcal{O}(n^2)$	$\mathcal{O}(P)$
Proposed model	$\mathcal{O}(n^2\|\mathcal{C}\|N_s)$	$\mathcal{O}(PN_s)$

Table 2: Comparison of the proposed method with three baselines along three metrics compared to ground truth in Casablanca network.

	KL divergence	Accuracy	Avg. max. occupancy error
Random model	0.45	6%	5.5%
Trip chaining [29]	0.15	10%	3.8%
Markov model [20] (5 simulations)	0.075	15%	**0%**
Proposed model (5 simulations)	**0.07**	**17%**	**0%**

4.3 Accuracy of trips reconstitution

Three baselines are considered: a random model that assigns to each passenger an alighting station randomly, the popular trip chaining algorithm [29], and the Markov model from [20], to be compared with the proposed improved Bayesian implementation. For trip chaining, passengers whose validation could not be chained are assigned an alighting station following the distribution of chained trips. The Casablanca network is used here with simulated noise-free counting cells and alighting validations removed: the models are run on boarding validations only, and the resulting OD matrices are compared to the true OD matrix obtained from both boarding and alighting validations.

Table 2 compares the proposed model to the baselines according to three metrics: the Kullback–Leibler (KL) divergence between the predicted and the true OD matrices, the accuracy of individual trips (whether the predicted alighting station is correct w.r.t. the ground truth) and the maximum relative error on the occupancy, averaged over all courses. The proposed model outperforms the baselines considering any metric. Both Markov model based approaches obtain a perfect occupancy estimation: indeed, the models are designed to comply with the provided boarding and alighting counts per course. Here, perfect counts are simulated, resulting in errorless occupancy estimation.

Trip length distribution The following experiment evaluates the impact of adding priors and biasing the Hypergeometric distribution to obtain more realistic trip lengths. The same Casablanca dataset as above is used. Fig. 2 compares the trip length distribution obtained by the vanilla Markov model (top figure) from [20] to the proposed one with priors over trip lengths (bottom figure). Incorporating priors results in a trip length distribution much closer to the true distribution: the sum of the absolute errors was reduced by over 50%.

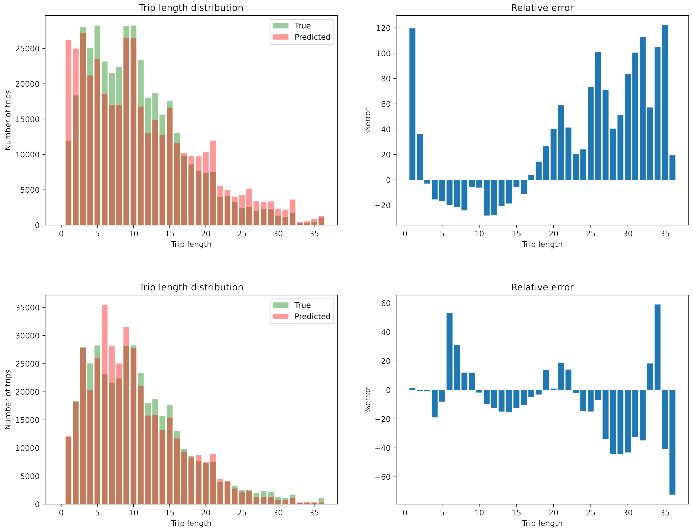

Fig. 2: **Top**: Markov model without priors. **Bottom**: Markov model with priors on trip length. The left plots display the distribution of trip lengths over all trips, and the right plots show the relative difference between the predicted distribution and the ground truth distribution.

4.4 Robustness

In this part, results are shown for the proposed model only since neither the random model nor trip chaining makes use of counting cells. Moreover, the Markov model from [20] does not deal with situations where count data is not perfect.

Influence of the noise level and fraud This experiment evaluates how the noise in counting cells data affects the different metrics when considered with and without fraud. The dataset from Casablanca is still used, but counting cells are simulated with a noise level p.

Table 3 presents the same three metrics with respect to Casablanca ground truth with different noise levels and with the presence or absence of fraud. The proposed model is shown resilient to noise: even with significant sensor noise levels, the KL divergence and the accuracy remain almost as high as when there is no noise. However, it is less robust to fraud, even in the absence of noise. This is explained by the fact that the fraud disaggregation algorithm assigns the inferred number of fraudsters to the course stations based on station fraud rates which are given as prior data and may be quite inaccurate. Future work may explore alternate approaches to station-level fraud rate estimation.

Table 3: Comparison of the performance metrics as a function of the noise level and of the presence of fraud. The first four lines do not include fraud, while the two last do.

		KL divergence	Accuracy	Avg. max. occupancy error
No fraud	No noise ($p = 0$)	0.069	16.6%(28.9%)	0%
	Low noise ($p = 0.1$)	0.071	16.5% (28.8%)	1.2%
	Moderate noise ($p = 0.2$)	0.074	16.3% (28.7%)	1.4%
	High noise ($p = 0.4$)	0.077	16.2% (28.4%)	1.6%
Fraud	No noise ($p = 0$)	0.101	14.9% (27.3%)	1.5%
	High noise ($p = 0.4$)	0.105	14.2% (26.6%)	1.8%

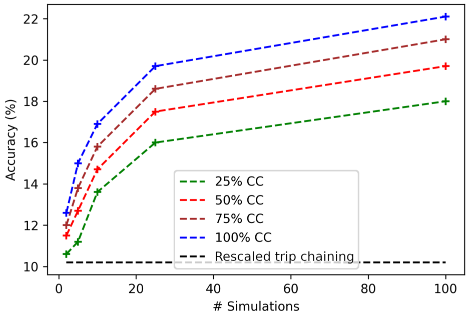

Fig. 3: Alighting station estimation accuracy with respect to the number of simulations of the proposed model, for varying equipment rates (25% CC means 25% of courses are equipped with counting cells), compared to trip chaining.

Influence of equipment rate Here, the sensitivity of the proposed model to lower coverage in counting cells is examined. To this end, counting cells are simulated only for a portion of all vehicles in the Casablanca network.

Fig. 3 depicts the accuracy of passenger trip reconstitution (*i.e.* the percentage of correctly inferred alighting stations) as a function of the number of simulations of the proposed model, for different scenarios depending on the percentage of courses equipped with counting cells. One can see that even with low equipment rates, the proposed model consistently outperforms trip chaining (black dotted line). This is particularly important as, for most networks, equipment rates do not exceed 50%. In addition, the accuracy loss resulting from having incomplete data can be compensated by an increased number of simulations at the cost of a linear increase in run time.

5 Conclusion and perspectives

5.1 Conclusion

This paper aims to reconstruct Origin-Destination matrices to better understand flows in public transport networks. The idea is to infer each passenger's alighting station with the data collected from the operators, the counting cells and the geolocalised stations. While recent statistical approaches may use sophisticated probabilistic models that are not scalable, we started from the model introduced by [20] which allows trips to be directly simulated and the parameters' distribution expressed in a closed form. Our implementation improved this model by using prior knowledge about the OD matrix that a trip chaining algorithm can provide. More importantly, several additions were built on top of this model, allowing us to tackle various phenomena that frequently occur when dealing with large-scale and real data and significantly affect the quality of the resulting OD matrix. Specifically, the objectives of these additions are to denoise count data and take fraudsters into account using a Bayesian approach. Dealing with both at once is challenging because their effects tend to mix and potentially cancel out. In the end, we demonstrated the robustness and accuracy of this approach on two real-world transportation networks. To the best of our knowledge, this approach is a novelty and as of today, extensive tests are performed on multiple networks in cities of different sizes.

5.2 Future work

Better understanding of the sensors Although the simulation environment enabled us to test different models with real-life phenomena, the lack of true counts to compare on the Orléans' network use case makes it challenging to estimate the correct value of the noise hyperparameters p^+ and p^- or the fraud rates at each station. It could be interesting to collect ground truth data for these sensors by manually counting passengers in vehicles. The value for these hyperparameters could then be estimated using an Expectation-Maximization algorithm.

Multi-source Although counting cells is very beneficial to OD matrix reconstruction, the problem remains highly uncertain. Indeed, many stations lead to high uncertainty in the resulting OD matrix. Nevertheless, adding other sensors, such as Bluetooth scanners, could reduce the system's underdetermination and increase the reconstruction's reliability.

Denoising method The actual statistical denoising method proposed in Subsection 3.1 is incomplete. Indeed, only observed alighting counts are denoised with respect to the boarding validations, which requires removing fraud beforehand. Thereby, denoising both boarding and alighting counts would give access to the total count of boarding and alighting passengers per station without needing to remove fraudsters, which is useful notably for occupancy estimation.

References

1. Assemi, B., Alsger, A., Moghaddam, M., Hickman, M., Mesbah, M.: Improving alighting stop inference accuracy in the trip chaining method using neural networks. Public Transport **12**(1), 89–121 (2020)
2. Ben-Akiva, M., Macke, P.P., Hsu, P.S.: Alternative methods to estimate route-level trip tables and expand on-board surveys. No. 1037, Transportation Research Board (1985)
3. Carvalho, L.: A Bayesian statistical approach for inference on static origin–destination matrices in transportation studies. Technometrics **56**(2), 225–237 (2014)
4. Egu, O., Bonnel, P.: Can we estimate accurately fare evasion without a survey? results from a data comparison approach in Lyon using fare collection data, fare inspection data and counting data. Public Transport **12**(1), 1–26 (2020)
5. Faroqi, H., Mesbah, M., Kim, J.: Applications of transit smart cards beyond a fare collection tool: a literature review. Advances in Transportation Studies **45** (2018)
6. Fog, A.: Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. Communications in Statistics—Simulation and Computation® **37**(2), 241–257 (2008)
7. Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H.: Estimation of population origin–interchange–destination flows on multimodal transit networks. Transportation Research Part C: Emerging Technologies **90**, 350–365 (2018)
8. Grgurević, I., Juršić, K., Rajič, V.: Review of automatic passenger counting systems in public urban transport. In: 5th EAI International Conference on Management of Manufacturing Systems. pp. 1–15. Springer (2022)
9. Harrison, G., Grant-Muller, S.M., Hodgson, F.C.: New and emerging data forms in transportation planning and policy: Opportunities and challenges for "track and trace" data. Transportation Research Part C: Emerging Technologies **117**, 102672 (2020)
10. Hazelton, M.L.: Network tomography for integer-valued traffic. The Annals of Applied Statistics **9**(1), 474–506 (2015)
11. He, L., Trépanier, M.: Estimating the destination of unlinked trips in transit smart card fare data. Transportation Research Record **2535**(1), 97–104 (2015)
12. Huang, D., Yu, J., Shen, S., Li, Z., Zhao, L., Gong, C.: A method for bus OD matrix estimation using multisource data. Journal of Advanced Transportation **2020** (2020)
13. Hussain, E., Bhaskar, A., Chung, E.: Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. Transportation Research Part C: Emerging Technologies **125**, 103044 (2021)
14. Ji, Y., You, Q., Jiang, S., Zhang, H.M.: Statistical inference on transit route-level origin–destination flows using automatic passenger counter data. Journal of advanced transportation **49**(6), 724–737 (2015)
15. Jung, J., Sohn, K.: Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. IET Intelligent Transport Systems **11**(6), 334–339 (2017)
16. Kumar, P., Khani, A., Davis, G.A.: Transit route origin–destination matrix estimation using compressed sensing. Transportation Research Record **2673**(10), 164–174 (2019)
17. Kumar, P., Khani, A., He, Q.: A robust method for estimating transit passenger trajectories using automated data. Transportation Research Part C: Emerging Technologies **95**, 731–747 (2018)

18. Lee, S., Lee, J., Bae, B., Nam, D., Cheon, S.: Estimating destination of bus trips considering trip type characteristics. Applied Sciences **11**(21), 10415 (2021)
19. Lei, D., Chen, X., Cheng, L., Zhang, L., Wang, P., Wang, K.: Minimum entropy rate-improved trip-chain method for origin–destination estimation using smart card data. Transportation Research Part C: Emerging Technologies **130**, 103307 (2021)
20. Li, B.: Markov models for Bayesian analysis about transit route origin–destination matrices. Transportation Research Part B: Methodological **43**(3), 301–310 (2009)
21. Li, T., Sun, D., Jing, P., Yang, K.: Smart card data mining of public transport destination: A literature review. Information **9**(1), 18 (2018)
22. Liu, X., Van Hentenryck, P., Zhao, X.: Optimization models for estimating transit network origin–destination flows with big transit data. Journal of Big Data Analytics in Transportation **3**(3), 247–262 (2021)
23. Liu, Z., Miranda, F., Xiong, W., Yang, J., Wang, Q., Silva, C.: Learning geo-contextual embeddings for commuting flow prediction. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 808–816 (2020)
24. Luca, M., Barlacchi, G., Lepri, B., Pappalardo, L.: A survey on deep learning for human mobility. ACM Computing Surveys (CSUR) **55**(1), 1–44 (2021)
25. McNally, M.G.: The four-step model. Emerald Group Publishing Limited (2007)
26. Mishalani, R.G., Ji, Y., McCord, M.R.: Effect of onboard survey sample size on estimation of transit bus route passenger origin–destination flow matrix using automatic passenger counter data. Transportation research record **2246**(1), 64–73 (2011)
27. Munizaga, M., Devillaine, F., Navarrete, C., Silva, D.: Validating travel behavior estimated from smartcard data. Transportation Research Part C: Emerging Technologies **44**, 70–79 (2014)
28. Simini, F., Barlacchi, G., Luca, M., Pappalardo, L.: A deep gravity model for mobility flows generation. Nature communications **12**(1), 1–13 (2021)
29. Trépanier, M., Tranchant, N., Chapleau, R.: Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems **11**(1), 1–14 (2007)
30. Vintsyuk, T.K.: Speech discrimination by dynamic programming. Cybernetics **4**(1), 52–57 (1968)
31. Wilson, A.G.: The use of entropy maximising models, in the theory of trip distribution, mode split and route split. Journal of transport economics and policy pp. 108–126 (1969)
32. Wilson, A.G.: A family of spatial interaction models, and associated developments. Environment and Planning A **3**(1), 1–32 (1971)
33. Wu, X., Guo, J., Xian, K., Zhou, X.: Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. Transportation Research Part C: Emerging Technologies **96**, 321–346 (2018)
34. Yan, F., Yang, C., Ukkusuri, S.V.: Alighting stop determination using two-step algorithms in bus transit systems. Transportmetrica A Transport Science **15**(2), 1522–1542 (2019)
35. Yao, X., Gao, Y., Zhu, D., Manley, E., Wang, J., Liu, Y.: Spatial origin-destination flow imputation using graph convolutional networks. IEEE Transactions on Intelligent Transportation Systems **22**(12), 7474–7484 (2020)
36. Zapata, L.P., Flores, M., Larios, V., Maciel, R., Antunez, E.A.: Estimation of people flow in public transportation network through the origin-destination problem for the South-Eastern corridor of Quito city in the smart cities context. In: 2019 IEEE International Smart Cities Conference (ISC2). pp. 181–186. IEEE (2019)