Bayesian origin-destination estimation in networked transit systems using nodal in- and outflow counts

Journal Article

Author(s):
Blume, Steffen O.P.; Corman, Francesco; Sansavini, Giovanni

Publication date:
2022-07

Permanent link:
https://doi.org/10.3929/ethz-b-000548494

Rights / license:
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

Originally published in:
Bayesian origin-destination estimation in networked transit systems using nodal in- and outflow counts

Steffen O.P. Blume\textsuperscript{a,b}, Francesco Corman\textsuperscript{c}, Giovanni Sansavini\textsuperscript{a,*}

\textsuperscript{a} Reliability and Risk Engineering Laboratory, Institute of Energy and Process Engineering, ETH Zürich, Leonhardstrasse 21, Zürich 8092, Switzerland
\textsuperscript{b} Future Resilient Systems, Singapore-ETH Centre, 1 Create Way, CREATE Tower #06-01, 138602, Singapore,
\textsuperscript{c} Institute for Transport Planning and Systems, ETH Zürich, Stefano-Franscini-Platz 5, Zürich 8093, Switzerland

ARTICLE INFO

Keywords:
OD-estimation
Bayesian inference
Transit network
Nodal passenger count data

ABSTRACT

We propose a Bayesian inference approach for static Origin-Destination (OD)-estimation in large-scale networked transit systems. The approach finds posterior distribution estimates of the OD-coefficients, which describe the relative proportions of passengers travelling between origin and destination locations, via a Hamiltonian Monte Carlo sampling procedure. We suggest two different inference model formulations, the instantaneous-balance and average-delay model. The average-delay model is generally more robust in determining accurate and precise coefficient posteriors across various combinations of observation properties. The instantaneous-balance model, however, requires lower resolution count observations and produces estimates comparable to the average-delay model, pending that certain count observation properties are met. We demonstrate that the Bayesian posterior distribution estimates provide quantifiable measures of the estimation uncertainty and prediction quality of the model. Moreover, the Bayesian approach is at least as accurate as existing optimisation approaches and proves robust in scaling to high-dimensional underdetermined problems without suffering from the curse of dimensionality. The Bayesian instantaneous-balance model is applied to the New York City subway network, with several years of entry and exit count observations recorded at several hundred station turnstiles across the network. The posterior distribution estimates provide intuitive demand patterns and are projected to be more valuable than point estimates, since they allow for robust transport network designs that account for the uncertainty of network parameters.

1. Introduction

The assessment, planning, and operation of transport systems relies heavily on measurements and estimates of travel demand rates between different locations of an investigated area or network. Research on Origin-Destination (OD)-estimation aims to reliably predict these travel demand rates. Due to its complexity regarding the scarcity or resolution of observable data, the number of unknown parameters, and continual technology changes, it has remained a several decade-long active field of research.
## Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>auto-regressive integrated moving average</td>
</tr>
<tr>
<td>CPU</td>
<td>central processing unit</td>
</tr>
<tr>
<td>DDR3</td>
<td>double data rate type 3</td>
</tr>
<tr>
<td>GB</td>
<td>Giga byte</td>
</tr>
<tr>
<td>GTFS</td>
<td>general transit feed specification</td>
</tr>
<tr>
<td>HMC</td>
<td>Hamiltonian Monte Carlo</td>
</tr>
<tr>
<td>HPD</td>
<td>highest posterior density</td>
</tr>
<tr>
<td>INGARCH</td>
<td>integer-valued generalised autoregressive conditional heteroscedasticity</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
</tr>
<tr>
<td>MNL</td>
<td>multinomial logit</td>
</tr>
<tr>
<td>MSE</td>
<td>mean squared error</td>
</tr>
<tr>
<td>MTA</td>
<td>metropolitan transport authority</td>
</tr>
<tr>
<td>NYC</td>
<td>New York City</td>
</tr>
<tr>
<td>NUTS</td>
<td>No U-Turn Sampler</td>
</tr>
<tr>
<td>OD</td>
<td>origin-destination</td>
</tr>
<tr>
<td>QP</td>
<td>quadratic programme</td>
</tr>
<tr>
<td>RAM</td>
<td>random access memory</td>
</tr>
</tbody>
</table>

## Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>OD-matrix</td>
</tr>
<tr>
<td>B</td>
<td>batch size of autoencoder neural network</td>
</tr>
<tr>
<td>C</td>
<td>normalizing constant</td>
</tr>
<tr>
<td>δ</td>
<td>observational evidence</td>
</tr>
<tr>
<td>F ̅</td>
<td>INGARCH model trend process</td>
</tr>
<tr>
<td>H</td>
<td>OD-assignment matrix of size $K \times S$</td>
</tr>
<tr>
<td>H_k</td>
<td>Sub-matrix of $H$ of size $N \times T_a \times S$</td>
</tr>
<tr>
<td>M</td>
<td>total number of rows of the OD-assignment matrix equal to $S \times N \times T_a$</td>
</tr>
<tr>
<td>K</td>
<td>non-negative ARIMA process</td>
</tr>
<tr>
<td>N</td>
<td>total number of observations</td>
</tr>
<tr>
<td>N_MC</td>
<td>number of (Markov Chain) Monte Carlo samples</td>
</tr>
<tr>
<td>N_nn</td>
<td>number of records within one batch of an autoencoder neural network training sample</td>
</tr>
<tr>
<td>N^+(μ,σ)</td>
<td>lower zero-truncated normal distribution, given location parameter $μ$ and scale parameter $σ$</td>
</tr>
<tr>
<td>̃R</td>
<td>potential scale reduction statistic</td>
</tr>
<tr>
<td>S</td>
<td>total number of stations</td>
</tr>
<tr>
<td>T</td>
<td>total number of time bins per observation window</td>
</tr>
<tr>
<td>T_a</td>
<td>total number of arrival time bins per arrival window</td>
</tr>
<tr>
<td>V_k</td>
<td>deterministic utility component of path alternative $k$</td>
</tr>
<tr>
<td>X</td>
<td>entry count matrix with size $N \times S$</td>
</tr>
<tr>
<td>Y</td>
<td>exit count matrix with size $N \times S$</td>
</tr>
<tr>
<td>Z_t</td>
<td>INGARCH model covariate vector</td>
</tr>
<tr>
<td>c</td>
<td>symmetric dirichlet concentration parameter</td>
</tr>
<tr>
<td>e_j</td>
<td>residual between the predicted and the observed exit count mean at station $j$</td>
</tr>
<tr>
<td>h_k_j</td>
<td>element of the OD-assignment matrix $H$ at index $(k, i)$</td>
</tr>
<tr>
<td>i</td>
<td>origin station index</td>
</tr>
<tr>
<td>j</td>
<td>destination station index</td>
</tr>
<tr>
<td>n</td>
<td>observation index</td>
</tr>
<tr>
<td>p(•</td>
<td>•)</td>
</tr>
<tr>
<td>r_j</td>
<td>intercept parameter for station $j$</td>
</tr>
<tr>
<td>t_a</td>
<td>arrival time bin index</td>
</tr>
<tr>
<td>t_d</td>
<td>departure time bin index</td>
</tr>
<tr>
<td>t_0</td>
<td>start of the departure window</td>
</tr>
<tr>
<td>t_1</td>
<td>start of the arrival window</td>
</tr>
<tr>
<td>Δt_{ij}</td>
<td>average travel delay (in number of time bins) between origin $i$ and destination $j$</td>
</tr>
<tr>
<td>q</td>
<td>placeholder variable</td>
</tr>
<tr>
<td>w</td>
<td>observation window width</td>
</tr>
<tr>
<td>x_i</td>
<td>observed mean entry counts at station $i$</td>
</tr>
<tr>
<td>x_{i}^{[n]}</td>
<td>observed entry counts at station $i$ during observation instance $n$</td>
</tr>
</tbody>
</table>
1.1. Brief overview of OD-estimation approaches

Considerable efforts have been devoted to OD-estimation, with notable works starting with the early beginnings of gravity models (Casey Jr, 1955; Reilly, 1931) and eventually followed by travel surveys (Ben-Akiva et al., 1985; Kuwahara and Sullivan, 1987), or inference and system identification approaches based on flow observations combined with assignment models (Bell, 1991; Cascetta, 2007). Generally, OD-estimation approaches are distinguished into road traffic and transit system OD-estimation, and further distilled into static and dynamic estimation methods. We refer to Antoniou et al. (2016), Cascetta et al. (2013), and Nuzzolo and Crisalli (2001) for thorough reviews of OD-estimation approaches. This paper focuses on static, inference-based OD-estimation in transit systems.

Many modern transit systems’ smart-card readers have set a new paradigm for OD-estimation, because every trip record generated from a smart-card “tap-in and tap-out” sequence includes OD-information and timestamps. This consequently explains that many of the recent works on transit network OD-estimation, transit assignment, and travel behaviour rely on smart-card data (Espinoza et al., 2013; Van Zuylen and Willumsen, 1980; Yang, 1995). Recent approaches include the use of mobile phone data (Bonnel et al., 2018; Hazelton, 2000; Maher, 1983; Nguyen et al., 1988; Robillard, 1975; Tesselkin and Khabarov, 2017; Toledo and Kolechkina, 2017), or inference and system identification approaches based on flow observations combined with assignment models (Bell, 1991; Cascetta, 2007; Tolouei et al., 2017) or automated fare collection data (i.e., smart-cards) in transit networks (Alsger, 2017; Ji et al., 2015; Zhao et al., 2018).
Most OD-matrix inference and closely related network tomography rely on measurements of the network flows between nodes, such as traffic volumes on links or the number of travellers on-board transit vehicles (Hazelton, 2010; Lawrence et al., 2006; Li and Cassidy, 2007; Nuzzolo and Crisalli, 2001; Tanaka et al., 2016; Vardi, 1996). These approaches can be distinguished into OD-reconstruction and OD-estimation, where OD-reconstruction aims to determine the actual (i.e., realized) OD trips, for instance on a specific day, and OD-estimation aims to determine mean OD-trip rates. Both streams typically rely on an assignment models taking into account either single or multiple route alternatives between each OD-pair (Hazelton, 2001a; Lo et al., 1996). In many high-capacity metro networks, however, link flows between nodes are unknown and only the flow entering and leaving the network are measured at the stations in terms of the number of travellers who access the network at origin stations (i.e., inbound flow) and the number travellers who egress from the network at destination stations (i.e., outbound flow). Recent related work by Jeong and Park (2021) considers an OD-reconstruction problem given only node counts (among other examples that also consider link and turn counts). However, they assume prior means and standard deviations of route flows as known. Conversely, the class of problems studied in this paper does not rely on the knowledge of network-internal link flows and path flows, or prior details about the latent OD-matrix. Estimates of the expected travel times, however, are typically available, with the transit system’s timetable information as the most approximate estimate of actual travel times.

Crucial to our problem is that we aim to estimate the OD-matrix, consisting of trip proportions between stations, from time series (i.e., sequenced observations) of the in- and outflow counts at every station of a transit network (e.g., daily measurements of in- and outflow within a defined observation window over a period of multiple years). In addition, we consider uninformative priors to the degree that a first estimate of the OD-matrix (i.e., a seed or target matrix) is not available. This stands in contrast to approaches that consider a Bayesian update of a prior OD-seed matrix (Carvalho, 2014; Hazelton, 2001b; Li, 2005; Maher, 1983; Tebaldi and West, 1998). However, thanks to the sequenced observations, identifiability of the OD-matrix parameters is theoretically feasible, albeit potentially requiring an excessive number of measurements (Hazelton, 2015; Vardi, 1996). The issue of identifiability can be further addressed by including various types of sensor measurements and retaining structural information contained in the observations, as studied by Yang et al. (2018), who consider link and path counts or additionally also prior OD demand data. They also point out that identifiability differs from observability, to the extent that identifiability refers to finding a unique solution for the parameters that describe travel demand distributions, whereas observability refers to the determinacy of OD-demand estimates given a single observation.

### 1.2. OD-estimation approaches relevant to this work

Our problem is concerned with OD-matrix estimation using time series of passenger counts at origin and destination stations. We therefore selected reference literature that concentrates on sequenced observations of in- and outflows at nodal locations in a network, without knowledge of network-internal flow measurements or prior knowledge of a seed matrix. We determined that such a problem is related to the estimation of highway OD-matrices by Nihan and Davis (1987), Gajewski et al. (2002), or Park et al. (2008), the estimation of turning movements at road intersections by Cremer and Keller (1987), and static OD-estimation with prior seed matrices by Carvalho (2014).

Nihan and Davis (1987) rely on a recursive method to solve the estimation problem’s system of equations. However, their method requires domain-specific and experienced fine-tuning of algorithm parameters to efficiently incorporate critical sum-to-one and non-negativity constraints.

Cremer and Keller (1987) propose a cross-correlation-based estimator, a least-squares estimator, or a Kalman filtering approach, that each partly relax these constraints and require additional normalization steps or approximations to enforce the constraints after a preliminary solution is found. However, these approximations and normalizations interfere with the reliability of the estimates (next to convergence issues of the Kalman filter), especially for networks with more than several tens of nodes. Cremer and Keller (1987) also propose a constrained optimisation formulation that explicitly includes the sum-to-one and non-negativity constraints. Even so, optimisation methods typically only perform well in low-dimensional, convex, well-defined problems, whereas realistic transit networks consist of hundreds of nodes and thus hundreds of thousands of unknown parameters.

Moreover, Gajewski et al. (2002) propose an $\ell_2$ estimation ($\ell_2E$) formulation that minimises the integrated squared error between the target and estimated densities of traffic counts as the goodness-of-fit criterion. Their formulation accounts for the non-negative sum-to-one constraints and is designed to be robust to outliers in the measured traffic volumes. However, it requires estimates of error variances that may not be trivially obtained (Park et al., 2008).

The three so far discussed works, produce point estimates, in spite of inherently uncertain processes regarding modelling error and the real-world variability of the modelled systems. Hence, related parameters should not be considered deterministic, especially in high-dimensional problems. Park et al. (2008) thus propose a Bayesian inference formulation solved via Markov Chain Monte Carlo (MCMC) to account for estimation uncertainty, while also accounting missing values and being insensitive to outliers in measured traffic volumes. However, Park et al. (2008) do not account for prior sum-to-one constraints (only non-negativity), which is found to be critical in the applications considered here, with high-dimensional parameter spaces and strongly underdetermined problems. In addition, the four aforementioned works consider a single simplified model that neglects travel time between locations of the network.

Carvalho (2014) propose a Bayesian model and MCMC Gibbs sampling solution for static OD-estimation given count measurements at nodal locations in a network. The proposed model respects critical parameter constraints and includes travel time information. The
author demonstrates that the traditional gravity, growth factor, and maximum entropy models can all be derived from their hierarchical Bayesian model formulation. However, their approach focuses on updating of a prior seed matrix, given a single observation of new measurements, and therefore does not address the problem of an uninformative prior.

In all five of the aforementioned works, the models and methods are demonstrated on a limited number of up to 10 count locations. Moreover, the works by Nihan and Davis (1987), Cremer and Keller (1987), Gajewski et al. (2002), and Park et al. (2008) exclusively deal with overdetermined problems (i.e., the number of observations exceeds the number of unknown model parameters). This work addresses OD-estimation problems that are underdetermined, with examples ranging in the order of $10^5$ parameters and $10^3$ observations. It is therefore critical to explicitly include non-negative sum-to-one parameter constraints that drastically and favourably reduce the solution domain. In addition, we treat problems that are without a seed matrix and therefore need to resort to non-informative improper priors for the OD-demand. Moreover, works so far have not studied the sensitivity of estimation accuracy and precision to the properties of count observations and measurement parameters.

1.3. This work’s contributions

In light of the above overview, most approaches to OD-estimation require network-internal link flow measurements or seed matrices as prior inputs. Neither of these are readily available for the cases studied here. Moreover, the reference approaches considered in Section 1.2 suffer from the curse of dimensionality and/or do not explicitly treat estimation uncertainty related to model assumptions and observation variability. In this paper, we therefore:

- demonstrate the identifiability of an OD-matrix estimate for underdetermined problems given sequenced count measurements and uninformative priors using a Hamiltonian Monte Carlo (HMC) sampling approach;
- study the sensitivity of estimation accuracy and precision to the characteristics that govern the observation sequence with regard to the width of the observation window, trends in the disaggregate count data contained within the window, and dispersion of aggregated count data;
- compare a simplified model against a more detailed model in order to assess how strongly neglected travel time assumptions affect the estimation results;
- showcase a real-world case study to exemplify the applicability of the approach for large networks that constitute in the order of $10^5$ OD-pairs.

We propose two Bayesian inference models with different levels of complexity regarding the underlying network flow assumptions. Both models identify distribution estimates for the OD-coefficients in large-scale urban transit networks based on time series measurements of the in- and outflow counts at stations. We term these two models the instantaneous-balance model and the average-delay model, where the instantaneous-balance model is subject to more approximate assumptions and constraints. They both formulate joint posterior distributions of the OD-coefficients that constitute the OD-matrix. These OD-coefficients determine the relative flow proportions bound for every destination, given an origin. Since the model solutions are analytically intractable, posterior distribution estimates are determined from an HMC sampling approach. We demonstrate and compare the performance of the two models on test networks and highlight the required preconditions for these models to generate reliable (i.e., accurate and precise) OD-coefficient estimates. Moreover, this comparison serves to evaluate the degree to which the simplified instantaneous-balance model is able to...
match the more detailed average-delay model results. In addition, we benchmark the HMC sampling approach against a quadratic programming (QP) approach originally established by Cremer and Keller (1987) as a way of comparing the distribution estimates to point estimates and point out their advantages in terms of uncertainty assessment and applicability to high-dimensional problems. At last, we showcase the methods on the New York City subway system, with several years of publicly available real-world turnstile count data. The proposed approach offers a quantification of the parameter uncertainty and delivers estimates that are at least as accurate as existing optimisation-based results for small-scale test networks, while not suffering from the curse of dimensionality as the size of the network increases. Test data and code files are available at Blume (2022).

2. Approach

Fig. 1 shows the workflow of this analysis. The starting point are time series of the in- and outflow counts of travellers at each station in a transit network. Since these data do not explicitly express the number of passengers travelling between the origin and destination locations, we develop two models to infer the latent OD-matrix. These models exploit that multiple measurements (i.e., time series) of nodal in- and outflow counts are recorded and, if possible, incorporate additional network information, namely, expected travel times that can for instance be obtained from service timetables. The following sections delineate these two model formulations that are developed to identify the elements of the OD-matrix, i.e., the “instantaneous-balance model” and the “average-delay model”.

For both model formulations, the aim is to identify the parameter values of the OD-matrix from entry and exit count observations at\textnumero S number of stations (also called nodes) in a fully connected network. The OD-matrix \( A_{S \times S} \) contains the OD-coefficients \( a_{ij} \) for each origin node \( i \in \{1, 2, \ldots, S\} \) and destination node \( j \in \{1, 2, \ldots, S\} \),

\[
A = \begin{bmatrix}
0 & a_{12} & \cdots & a_{1S} \\
a_{21} & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots \\
a_{S1} & \cdots & a_{S(S-1)} & 0
\end{bmatrix}_{S \times S},
\]

with

\[
\sum_{j=1}^{S} a_{ij} = 1; \ a_{ij} \geq 0; \ a_{ij} = 0 \ \forall \ i = j.
\]

The diagonal in Eq. (1) is zero, as no self-loops are permitted, i.e., none of the inflow leaves the network at its origin node. The OD-coefficients may be interpreted from two viewpoints: we can view OD-coefficients as the relative trip proportions of travellers entering at specific origins and bound for specific destinations; as such, they are also referred to as split coefficients (Cremer and Keller, 1987; Nihan and Davis, 1987) or split proportions (Gajewski et al., 2002; Park et al., 2008). Alternatively, OD-coefficients can be viewed as the conditional choice probabilities of selecting destination \( j \), given that a traveller accesses origin \( i \). We will use both viewpoints interchangeably, minding that they will converge in the limit of a large number of passengers.

2.1. Network assumptions

We rest our analysis on several key assumptions regarding the structure of the network and the observation sequence, that are in line with the assumptions by Gajewski et al. (2002) and Park et al. (2008):

(1) Observability: The network consists of links and nodes (i.e., edges and vertices). All points of the network where access and egress is possible are observable and passenger counts can be measured over steady state and periodic observation windows.

(2) Steady state: The OD-matrix and system operations stay constant during a single observation window (e.g., the OD-matrix remains constant during the 8:00 to 9:00 AM observation window).

(3) Mass conservation: The total inflow has to equal the total outflow across the network, i.e., every passenger who accesses the network has to eventually egress from the network – both the access event and the egress event need to be measurable.

Because of the observability assumption, any outliers or missing data will distort the inferred estimates to be less accurate and/or more uncertain. Therefore, such assumption violations need to either be corrected during data pre-processing or explicitly accounted for in the inference model. In this paper, inconsistencies due to outliers or missing sensor measurements are corrected in a pre-processing step through the autoencoder filtering approach in Appendix F, therefore neglecting severe measurement errors in the model formulation. For an alternative approach, Park et al. (2008) propose a model that inherently accounts for erroneous or missing data.

The steady state assumption requires that travel or traffic patterns are stable over the period of the observation window. As a result, this paper considers morning peak, midday off-peak, and evening peak travel times for the real-world case study. To assess the implications of partly violating this assumption, we parameterize test data using a trend scale which allows varying the degree of fluctuations in the entry and exit counts. The test case is described in Section 3.1 and Appendix E, the results analysis is presented in Section 4.2 which assesses the sensitivity of model accuracy and precision to the trend scale.

Mass conservation stipulates that none of the inflow remains in the network. If this assumption is violated, depending on the degree
of “leaking” inflow, models could be adapted by for instance including a slack node in the network to absorb any excess or missing inflow that does not match with the outflow. Alternatively, in- and outflow could be matched by proportionally allocating any excess across the network nodes, according the approach used in Appendix C.

Both the instantaneous-balance and the average delay model formulation adhere to the above assumptions. However, the instantaneous-balance model relaxes that passengers need time to travel between stations in the network, that is, mass conservation is achieved instantaneously such that over any chosen observation time window the total number of entering passengers balances with the total number of exiting passengers. The average-delay model therefore extends the list of assumptions, by including a fourth item,

(1) Expected travel time: Between every OD-pair exists an expected (measurable) travel time. When multiple route options exist between an OD-pair, the expected travel time is the mean travel time over all route options falling within a defined time margin from the earliest arrival (shortest) path.

As a result, the average-delay model requires fine grained, time-resolved observational data to construct a mapping between the entry and exit count observations. We present a detailed comparison of the instantaneous-balance and average-delay model.

This paper determines posterior distribution estimates of the model parameters of interest. Therefore, violations of the above assumptions do not imply that the models are fully misspecified or that estimates will strongly diverge from the true values. Instead, it is foreseen that distributions and associated interval estimates would be more dispersed or larger. An exhaustive assessment of the degree to which model assumption violations hinder reliable estimation of model parameters will be subject to future work.

2.2. Instantaneous-balance model

A simple regression model linking entry and exit count observations is the linear relation

\[
Y_{N \times S} = X_{N \times S} \cdot A_{S \times S} + \epsilon_{N \times S},
\]

where \( Y \) is a matrix of size \( N \times S \) and contains the number of exiting passengers at all \( S \) stations during \( N \) observations. The matrix \( X \) contains the corresponding entry count observations for all stations during the same observations. The second right-hand side term in Eq. (2) captures the errors \( \epsilon \) due to model deficiencies and measurement noise. These errors are assumed independent across stations and homoscedastic across observations. The inverse problem associated with Eq. (2) aims to determine the OD-coefficients contained in the matrix \( A \). Cremer and Keller (1987) solve a similar problem for the turning movements at intersections through a point-estimate-based constrained optimisation technique. By way of comparing the proposed Bayesian formulations and solution approaches against the point estimate formulation and a benchmark solution approach, Appendix A similarly expresses this problem as a constrained quadratic optimisation programme. The optimisation objective function includes an intercept variable and additional regularization terms that improve the OD-coefficient estimates, and are not included into the original constrained optimisation by Cremer and Keller (1987). These regularization terms are detailed in the context of the proposed model in Section 2.2.2. Since the coefficient estimates declare the relative flow proportions only, an estimate of the absolute flow volumes can be obtained according to \( \text{diag}(\pi)A \). The diagonal matrix \( \text{diag}(\pi) \) contains the sample means of the entry counts \( \pi \) of length \( S \).

2.2.1. The instantaneous-balance Bayesian model

When solving the quadratic optimisation programme in Eq. (A.1) of Appendix A, the returned solution gives point estimates of the OD-coefficients. These point estimates do not represent the inherent estimation uncertainty and are not robust for high-dimensional instances. In order to address these limitations, we express the model in Eq. (2) as a Bayesian inference model. The OD-split coefficient parameters \( a_{ij} \) thus are treated as random variables. Moreover, the model includes auxiliary parameters, namely, the scale parameter \( \sigma_{ij} \) and intercept \( r_{ij} \) to account for measurement and model error. Following Bayes’ rule, the posterior density for the unknown OD-split coefficients \( a_{ij} \) (and parameters \( \sigma_{ij} \) and \( r_{ij} \), denoted by \( p(\alpha, \sigma, r | \mathcal{D}) \), is expressed in terms of the likelihood \( p(\mathcal{D}|\alpha, \sigma, r) \) and prior \( p(\alpha, \sigma, r) \) according to

\[
\pi(\alpha, \sigma, r | \mathcal{D}) \propto p(\mathcal{D}|\alpha, \sigma, r) \pi(\alpha, \sigma, r) \]

The evidence \( D \) are the exit count observations \( y_{ij}^{(n)} \) at station \( j \in \{1, 2, \ldots, S\} \) during observation \( n \in \{1, 2, \ldots, N\} \). The corresponding entry count observations \( x_{ij}^{(n)} \) at station \( i \in \{1, 2, \ldots, S\} \) are part of the likelihood model. For each observation, entry and exit counts are recorded over the same fixed observation window, which assumes that system conditions are unchanged for each observation except for the inherent in- and outflow variability. The likelihood is specified according to

\[
p(\mathcal{D}|\alpha, \sigma, r) = \prod_{j=1}^{S} \prod_{n=1}^{N} p\left(y_{ij}^{(n)} | \mu_{ij}^{(n)}, \sigma_{ij}\right),
\]

where every observation of the number of exiting passengers \( y_{ij}^{(n)} \) is a realization of a truncated normal distribution (lower truncated at zero), \( N^{+} \), with location parameter \( \mu_{ij}^{(n)} \) and observation-invariant scale parameter \( \sigma_{ij} \), such that
and random intercept \( r_j \) can be parameterized as

\[
\mu^{(n)}_{ij}(\alpha, x, r) = \sum_{i=1}^{S} \alpha_{ij} x_{ij}^{(n)} + r_j .
\]

This likelihood model follows typical regression formulations, whereby the error term from Eq. (2) is decomposed into a residual \( \epsilon_j \) and random intercept \( r_j \), such that \( y_j^{(n)} = \sum_{i=1}^{S} \alpha_{ij} x_{ij}^{(n)} + r_j + \epsilon_j \) with \( \epsilon_j \sim N(0, \sigma_{\epsilon j}) \). This is equivalent to

\[
y_j^{(n)} - \left( \sum_{i=1}^{S} \alpha_{ij} x_{ij}^{(n)} + r_j \right) \sim N(0, \sigma_{\epsilon j})
\]

and reduces to Eq. (5) with the additional constraint that the normal distribution is truncated to reflect the non-negativity of exit counts \( y_j^{(n)} \). The scale parameter \( \sigma_{\epsilon j} \) is assumed constant across observations (i.e., across time), since it is assumed that the relative variation of \( \mu^{(n)}_{ij} \) across these periods is sufficiently modest. While the scale parameter accounts for uncertainties that are related to random errors, the intercept accounts for any biasing errors imposed by the model. The model does not explicitly account for any measurement errors in \( x_{ij}^{(n)} \), and therefore takes a measurement error-agnostic naïve regression modelling approach (Sorensen et al., 2015). This assumption is in line with other works, such as those by Gajewski et al. (2002) or Park et al. (2008). Nonetheless, this does not completely disqualify the model from cautiously using it with error-prone measurements. In fact, error-prone measurements will not only cause estimation bias, but also reflect negatively on the dispersion of the Bayesian coefficient posterior estimates. As a result, credibility intervals will become wider and help to detect spurious and potentially unreliable, highly uncertain estimates.

Whereas the likelihood captures the assumed model and weighs it against the evidence given by the observed data, the prior introduces any domain knowledge or information about the model parameters. The prior is specified according to

\[
\pi(\alpha, \sigma, r) = \prod_{i=1}^{S} p(\alpha_{ij}) \prod_{j=1}^{S} p(\sigma_{\epsilon j}) \prod_{j=1}^{S} p(r_j) .
\]

The notation indicates the row-wise simplex parameterization for \( \alpha_{ij} \), i.e., the non-negative, sum-to-one OD-coefficient constraint according to Eq. (1). As comparison, Park et al. (2008) define a truncated normal distribution for the OD-coefficients. In addition, we place a symmetric Dirichlet hyperprior on each row of the OD-matrix, according to

\[
a_{ij} \sim \text{Dir}(c1),
\]

where 1 is a vector of ones with size \( 1 \times (S - 1) \), and \( c \) is a concentration parameter. The symmetric Dirichlet hyperprior in Eq. (8) maintains the simplex structure and includes additional knowledge about the clustering of OD-coefficient values. By adjusting the concentration parameter in Eq. (8) away from 1, the prior biases towards either sparse (\( c < 1 \)) or dense (\( c > 1 \)) OD-coefficient matrix. Appendix B shows an example of a two-dimensional simplex and the influence of the concentration parameter. Unless domain expertise strongly suggests either a more sparse or dense OD-matrix, \( c \) should be set to 1. Alternatively, the Dirichlet hyperprior in Eq. (8) can be parameterized as \( a_{\bullet} \sim \text{Dir}(c_{\bullet}) \), where \( c_{\bullet} \) is a vector of concentration parameters \( c_{\bar{ij}} \) for each destination \( j \) and can be used to include any precursory information on each value of the OD-coefficients, similar to a seed matrix as implemented by Carvalho (2014). Moreover, enforcing zero constraints for infeasible OD-coefficients (\( a_{\bar{ij}} = 0 \)), similar to related works by Gajewski et al. (2002) and Park et al. (2008), can be achieved by setting the corresponding concentration parameters to zero (\( c_{\bar{ij}} = 0 \)). Part of the goals of this paper, however, is to demonstrate that OD-coefficient estimates can still be obtained even with uninformative priors, given sequenced records of count measurements, despite the general notion that proper choice of the seed matrix is crucial in obtaining accurate results particularly for dynamic OD-estimation (Cantelmo et al., 2014).

The scale parameters \( \sigma_{\epsilon j} \) as well as the intercepts \( r_j \) are assumed (partially) unconstrained and therefore given improper flat (uniform) priors according to

\[
\sigma_{\epsilon j} \sim U(0, \infty), \ r_j \sim U(-\infty, \infty) .
\]

In practice, the improper priors in Eq. (9) may be replaced by the proper uniform priors on the intervals \((0, M)\) and \((-M, M)\), respectively, for some sufficiently large value \( M \). Moreover, we assume that the likelihood model is informative enough to prevent \( \sigma_{\epsilon j} \) and \( r_j \) from diverging towards infinity.

### 2.2.2. Model regularisation

The intercept \( r_j \) in Eq. (6) corrects for the average biasing error between the linear model and the observable range of in- and outflow counts. However, an inadvertent consequence may be that the intercept model overfits the observed data, capturing the
observable range overly close and performing poorly when predicting on unseen data. Consequently, we extend the instantaneous-balance model by two regularization terms. Firstly, we include a minimal-bias penalty that aims to minimise the squared sum of the intercepts $r_j$, by defining an additional probability density according to

$$p\left(\sum r_j\right)^2 = \frac{1}{C} \exp\left(-\lambda \sum r_j\right) \propto \exp\left(-\lambda \sum r_j\right).$$

We can omit the normalizing factor $C^{-1}$, by assuming that it is constant and does not depend on $r_j$. The scaling parameter $\lambda$ controls the magnitude of the penalization. Throughout this paper, $\lambda = 1$, which accounts for a reasonable weighting between the posterior log-

![Fig. 2. Top panel: Schematic of the assumed average-delay observation sequence for a network consisting of S stations. The stations are connected through a fully connected network with known timetable information. Entry and exit counts are collected at every station within $N$ periodical observation windows. Every observation window is subdivided into time bins, starting with time bin $t = t_0$ and ending at $t = T$. System conditions and the underlying latent origin-destination matrix are assumed constant across all observations. One entry count or exit count corresponds to one passenger entering or exiting at a station, respectively. Bottom panel: The assumed access, travel time, and egress process for one observation window of the average-delay model with $T = 9$ time bins. The departure window starts at $t = t_0 = 1$. The arrival window starts at $t = t_1$, ($t_1 - t_0$) time bins after the departure window to account for passenger trips departing prior to and arriving during the arrival window. Legend: The arrows indicate example trips of passengers; *The passenger departed and arrives prior to the start of the arrival window; †The passenger departed prior to the arrival window and arrives during the arrival window; ‡The passenger departed and arrives during the arrival window; §The passenger arrival time at the destination station occurs after the observation window.](image-url)
probability and the penalization, given that count observations are typically in the order of $10^2$ to $10^3$ in the case studies considered. As an alternative to the penalty in Eq. (10), the improper prior for $\tau_j$ in Eq. (9) could be replaced by a normal prior. Here, the penalty is introduced in this way for consistency with the benchmark quadratic programming formulation in Appendix A.

Secondly, we include an expected value component, that tests the instantaneous-balance assumption against the mean entry and exit count observations. This component enters the likelihood model, and is defined according to

$$\tilde{y}_j | \mu_{y_j}, \sigma_{y_j} \sim N^y(\mu_{y_j}, \sigma_{y_j}),$$

(11)

with

$$\mu_{y_j}(a) = \sum_{i=1}^S a_{ij} \tilde{x}_i,$$

(12)

where $E[X_i] \approx \tilde{x}_i = \frac{1}{N} \sum_{n=1}^N x_n^{(n)}$ and $E[Y_j] \approx \tilde{y}_j = \frac{1}{N} \sum_{n=1}^N y_n^{(n)}$ are the sample means of the entry and exit count observations. Eq. (12) does not include an intercept. The scale parameter for the sample mean $\tilde{y}_j$ in Eq. (11) is the same as for the observations $y_n^{(n)}$ in Eq. (5), although one may expect it to be smaller for the sample mean. The scale parameter in Eq. (11) is therefore a generous estimate of the variance for the sample mean, in part owed to the nature of this ad hoc regularization but also to account for the missing intercept in Eq. (12).

By including the regularisation terms, we aim to control the generalisability (i.e., extrapolation outside of the observable range) of model predictions and avoid overfitting of the observed data. The minimal-bias and expected value expressions in Eqs. (10) and (11) propose that the missing-intercept model (i.e., the uncorrected instantaneous-balance assumption) is a generalisable approximation of the true underlying processes and thus valid both within and beyond the observable range, whereas the likelihood model in Eqs. (5) and (6) weighs in the proposal that the intercept model (i.e., the corrected instantaneous-balance assumption) captures more closely the observable range. The result is a trade-off between these two proposals that reflects positively in finding generalisable parameter estimates.

2.3. The average-delay model

The instantaneous-balance model neglects the time lags involved in travelling between origins and destinations. This assumption can give rise to significant estimation errors of the OD-matrix. In order to overcome this limitation, we propose the average-delay model, where delay refers to the travel time or lag between an origin and destination.

2.3.1. Formal model description

Fig. 2 shows diagrams of the passenger flow process in the average-delay model. Observations are recorded periodically, e.g., daily (Fig. 2, top panel). Every observation $n \in \{1, 2, \ldots, N\}$ considers a departure window and an arrival window (Fig. 2, bottom panel). The departure window starts at $t = t_0$ and comprises earlier times than the arrival window, which starts at $t = t_1$. Both end at the same time $t = T$. The time gap $t_1 - t_0$ between the start of the departure window and the start of the arrival window is chosen as the longest travel time between any two locations in the network. Alternatively, the time gap can be selected based on domain knowledge. For instance, the maximum planned travel time of any traveller may be assumed shorter than 1 h and thus the time gap would be set to 1 h. The departure and arrival window are discretized into time bins. The definitions of the arrival window, departure window, and the time gap between them entail that shorter trips that commence during the departure window and arrive before the arrival window are not accounted for. However, this choice ensures that all exit counts observed within the arrival window have a matching entry count within the departure window.

We determine the average travel time between every OD-pair (e.g., from timetable information) and construct an OD-assignment matrix $H_k \times S$, whose elements are the number of entering passengers at origin $i \in \{1, 2, \ldots, S\}$ during observation $n$ and time bin $t \in \{1, 2, \ldots, T\}$. A share of these entering passengers exits at destination $j$ with a delay of $t + \Delta t_j$, and same observation $n$. This share is given by the OD-coefficient $a_{ij}$ that is, the proportion of entering passengers at $i$ that are bound for $j$. The average travel time $\Delta t_j$ is given in terms of the number of lagged time bins required to travel between the two stations. It is assumed as given, either because it is measurable or can be estimated external to the OD-estimation problem. The OD-assignment matrix has $K = S \times N \times T_0$ rows, where $T_0$ is the number of arrival time bins.

Formally, the average-delay model describes the exit count observations at destination $j$, according to

$$y_{j n} = H_k \cdot a_j + \frac{\epsilon}{(N \times T_0) \times 1},$$

(13)

where $y_{j n}$ is a column vector of length $N \times T_0$ and contains the number of passengers exiting at station $j$ during time bin $t_0 \in \{1, 2, \ldots, T_0\}$ and observation $n \in \{1, 2, \ldots, N\}$. The sub-matrix $H_k$ contains the entry count observations at origin $i \in \{1, 2, \ldots, S\}$ during observation $n \in \{1, 2, \ldots, N\}$ at lagged time bin $(t_0 - \Delta t_j)$ with respect to destination $j$. The column vector $a_j$ corresponds to the $j$-th column of the OD-matrix $A$ in Eq. (1). The second right-hand side term in Eq. (13), $\epsilon$, accounts for the modelling and measurement errors. The element $h_{kj}$ at index $(k, j)$ of the OD-assignment matrix is given by
\[ h_{ij} = \begin{cases} x_{ij}^{(n, t_d)}, & \text{if } t_d \in \{1, 2, \ldots, T_s\}, \\ 0, & \text{otherwise,} \end{cases} \]  

(14)

where \( x_{ij}^{(n, t_d)} \) corresponds to the entry counts at station \( i \) during observation \( n \) and departure time bin \( t_d \in \{1, 2, \ldots, T_s\} \), with \( T \) being the number of departure time windows. The row index \( k \) is given in terms of the origin station \( i \), destination station \( j \), the observation \( n \), and the arrival time bin \( t_a \) according to

\[ k = (j - 1) \times N \times T_a + (n - 1) \times T_a + t_a. \]  

(15)

The arrival time bin depends on the specific OD-pair and is given by

\[ t_a = t_d + \Delta t - t_i + 1. \]  

(16)

2.3.2. The average-delay Bayesian model

We express the average-delay model into a Bayesian formulation. The posterior density for the unknown parameters \( \alpha \) is given by

\[ \pi(\alpha, \sigma | \mathcal{D}) \propto \exp(\mathcal{D} | \alpha, \sigma) \pi(\alpha, \sigma). \]  

(17)

where the evidence \( \mathcal{D} \) are the exit count observations \( y_{j}^{(n, t_a)} \) at station \( j \), during observation \( n \) and time bin \( t_a \). Moreover, the model includes only the auxiliary scale parameters \( \sigma_{ij} \) to account for measurement and modelling errors. Given that the average-delay model accounts more closely for travel time and truncates any passenger arrivals outside the arrival window, we omit the intercepts \( r_{ij} \) that capture the average bias errors in the instantaneous-balance model. We define the likelihood as

\[ y_{j}^{(n, t_a)} | \mu_{ij}^{(n, t_a)}, \sigma_{ij} \sim N^{+} \left( \mu_{ij}^{(n, t_a)}, \sigma_{ij} \right), \]  

(18)

with

\[ \mu_{ij}^{(n, t_a)} = \sum_{i=1}^{S} \alpha_{ij} h_{ij}. \]  

(19)

\( N^{+} \) in Eq. (18) denotes the truncated normal distribution. The prior is specified according to

\[ \pi(\alpha, \sigma) = \prod_{i=1}^{S} \rho(\alpha_{i}) \prod_{j} \rho(\sigma_{ij}). \]  

(20)

The OD-split coefficients \( \alpha_{ij} \) are constrained as simplexes for each origin station \( i \) and we place the symmetric Dirichlet hyperprior in Eq. (8). The observation- and time bin-invariant scale parameters \( \sigma_{ij} \) are constrained by improper uniform priors,

\[ \sigma_{ij} \sim U(0, \infty). \]  

(21)

2.4. OD-matrix estimate via Markov Chain Monte Carlo

The point estimates, that are for instance obtained from the optimisation model in Appendix A, do not express the prevalent parameter uncertainty due to modelling and measurement errors. In contrast, the Bayesian approach propagates the uncertainty regarding model assumptions and measurements into the posterior distribution estimates of the OD-coefficients. These posterior distribution estimates can hence be used to quantify each parameter’s expectation, that is, “the weighted average” over its (marginal) posterior distribution.

The OD-estimation problem proposed in this paper aims to find estimates for \( S \times (S - 1) \) OD-coefficients and \( S \) scale parameters, with additional \( S \) intercepts for the instantaneous balance model. Given the multidimensional parameter space and constraints, the models in Eqs. (3) and (17) are analytically intractable, and prevent from obtaining a closed form expression of the OD-coefficients’ posterior distribution. Moreover, the high-dimensional parameter space makes numerical approximation of the normalization integral associated with Bayes’ rule practically infeasible. Markov Chain Monte Carlo (MCMC) approaches, however, are able to find sampling estimates of the posterior distribution (Geyer et al., 2011). In the context of MCMC, the targeted posterior distribution is also known as the equilibrium distribution.

We thus resort to MCMC sampling to estimate the multidimensional posterior distribution \( \pi(\alpha, \sigma, r | \mathcal{D}) \) of the OD-matrix coefficients and auxiliary parameters. Traditional MCMC samplers like the Metropolis-Hastings algorithm struggle to converge and sample from the equilibrium distribution of high-dimensional parameter spaces, since they are based on random walk processes. The particularly large number of OD-coefficients thus poses a problem. Hamiltonian Monte Carlo (HMC), however, has proven to overcome this limitation by taking targeted and tuned steps through the posterior distribution based on derivative information of the target distribution and numerical integration in a transformed parameter space (Duane et al., 1987; Neal, 2011, 1996). The result is a significant improvement in sampling efficiency. We apply the HMC variant known as NUTS (No-U-Turn Sampler) (Hoffman and Gelman, 2014), which includes tuning of the algorithm parameters prior to sampling from the target distribution. At last, the MCMC sampling...
approach is also the reason why the normalizing constant in Eq. (10) can be omitted, as it cancels out during the derivative and Metropolis acceptance step.

3. Generative processes and test networks

The capability of the instantaneous-balance and average-delay model in estimating the latent OD-matrix coefficients of transit systems is demonstrated on two synthetic datasets and on one real-world dataset.

3.1. Synthetic datasets

Test network A encompasses the synthetic dataset to test whether the inverse problem tackled by the instantaneous-balance inference model is solvable. It consists of 15 stations, provides daily balanced entry and exit counts, and possesses a minimum level of network information, i.e., the link structure is unknown and the travel time between origins and destinations in the network is assumed to be zero. However, the true OD-matrix is known. The generative process for test network A is detailed in Appendix D, whereby entry counts are generated from a negative binomial distribution. Exit counts are determined according to a probabilistic choice model given the true OD-coefficients. For both test network A and test network B, neither the entry counts, nor the exit counts are additionally distorted by synthetic measurement error.

Real-world data do not entirely fulfill the assumptions of the instantaneous-balance model, which is, therefore, subject to modelling error. For instance, we expect discrepancies between the observed and predicted exit counts, since the instantaneous-balance model omits travel time information between stations. Therefore, test network B enforces realistic assumptions regarding the generative processes associated with the observational data. Foremost, several path alternatives with varying travel times exist between each OD-pair. In addition, the number of entering and exiting passengers within the fixed and truncated observation windows are unbalanced. Test network B assumes a known true OD-coefficient matrix, generates time-variant entry count series, and simulates exit counts via a path choice and travel delay model. In contrast to the instantaneous-balance model, the resolution of the passenger count measurements is increased – observations are again recorded daily; in addition, every daily observation window is segmented into five-minute time bins. The detailed generative process for test network B is given in Appendix E. It is based on a subcomponent and timetable data of the New York City (NYC) subway (Metropolitan Transport Authority, 2019a), consisting of 35 stations, 8 lines, and 39 transfer linkways (i.e., footpaths) shown in Fig. E.1. Test network B encompasses the synthetic dataset to compare the instantaneous-balance model to the higher resolution average-delay inference model. It is therefore used to assess the ramifications of the instantaneous-balance model assumptions, when tested on realistic data and to benchmark the gap with respect to the higher resolution average-delay model.

Next to generic model testing, the synthetic datasets are used to demonstrate model robustness against the properties of the underlying generative processes and chosen observation horizons. For test network A, the variance of the entry and exit count

---

Fig. 3. The New York City subway network and city boroughs (Metropolitan Transport Authority, 2019b). The borough of Manhattan is divided into Lower, Midtown, and Upper Manhattan. The fifth borough, Staten Island, is not shown on this map.
3.3. Accuracy and precision of OD-coefficient estimates

Observations is modulated through an adjustable dispersion parameter $\varphi$. Test network B additionally includes a trend scale $\eta$ to control within-day entry count fluctuations, and an observation window width $w$ to adjust the daily observation horizon.

3.2. Real-world dataset

In addition to synthetic datasets, we aim to take a step towards validating the proposed models on real-world datasets, i.e., the NYC subway network shown in Fig. 3. Starting in April 2010, the MTA publishes turnstile entry and exit counts for every station (Metropolitan Transport Authority, 2019a). The count data are originally collected in 4 h intervals and recorded together with a turnstile identifier. Every station has multiple turnstiles. We conservatively up-sampled the counts to a higher-resolution time granularity. The up-sampling method is based on a monotonic rational quadratic spline interpolation (Delbourgo and Gregory, 1983) of the cumulative, and thus monotonically increasing, 4 h interval count values. The up-sampled cumulative count values are differentiated to obtain the synthetic higher-resolution count data. The $C^2$-continuity of the interpolation ensures that the differentiated interpolated values remain smooth. We chose a 5-minute upsampling interval. These upsampled counts are conservative in that they sum to the original 4 h interval count data. In addition, we apply an auto-encoder data filtering approach to correct outliers and missing data that are interspersed throughout the real-world dataset. Details about the up-sampling approach and subsequent filtering procedure for outlier detection and missing data imputation are given in Appendix F.

Furthermore, the observed total exit counts of the NYC subway system are consistently lower than the observed total entry counts. For instance, the average total number of entering passengers is 5.9 million versus 4.3 million exiting passengers, when summed over a full daily cycle. This is presumably due to passengers using emergency exits instead of the official turnstile gates. In order to comply with the mass conservation assumption, we consequently balanced the daily in- and outflow, by distributing the excess inflow proportionally across stations as given by the procedure in Appendix C. Assuming that the entry counts follow Negative Binomial distributions for every station, we determine the equivalent dispersion parameter values, given the sample mean and variance of the entry count data. The dispersion parameter concentrates near values of order $10^2$, indicating mostly overdispersed count data.

These pre-processing steps result in time series of the number of travellers accessing and egressing at each station in the NYC subway system in 5-minute intervals over a time span of approximately 8 years (April 2010 to August 2018). In this analysis, however, we only consider weekday counts and the months from April until October. Moreover, the counts are aggregated over pre-defined station identifier. Every station has multiple turnstiles. We conservatively up-sampled the counts to a higher-resolution time granularity. The up-sampling method is based on a monotonic rational quadratic spline interpolation (Delbourgo and Gregory, 1983) of the cumulative, and thus monotonically increasing, 4 h interval count values. The up-sampled cumulative count values are differentiated to obtain the synthetic higher-resolution count data. The $C^2$-continuity of the interpolation ensures that the differentiated interpolated values remain smooth. We chose a 5-minute upsampling interval. These upsampled counts are conservative in that they sum to the original 4 h interval count data. In addition, we apply an auto-encoder data filtering approach to correct outliers and missing data that are interspersed throughout the real-world dataset. Details about the up-sampling approach and subsequent filtering procedure for outlier detection and missing data imputation are given in Appendix F.

The observed total exit counts of the NYC subway system are consistently lower than the observed total entry counts. For instance, the average total number of entering passengers is 5.9 million versus 4.3 million exiting passengers, when summed over a full daily cycle. This is presumably due to passengers using emergency exits instead of the official turnstile gates. In order to comply with the mass conservation assumption, we consequently balanced the daily in- and outflow, by distributing the excess inflow proportionally across stations as given by the procedure in Appendix C. Assuming that the entry counts follow Negative Binomial distributions for every station, we determine the equivalent dispersion parameter values, given the sample mean and variance of the entry count data. The dispersion parameter concentrates near values of order $10^2$, indicating mostly overdispersed count data. These pre-processing steps result in time series of the number of travellers accessing and egressing at each station in the NYC subway system in 5-minute intervals over a time span of approximately 8 years (April 2010 to August 2018). In this analysis, however, we only consider weekday counts and the months from April until October. Moreover, the counts are aggregated over pre-defined observation windows, as detailed in Section 4.3. The purpose of constraining the data records in this way is to comply with the models’ assumptions regarding recurrent system conditions and unchanged OD-demand. The resulting number of observations is equal to $N = 1315$ records of the in- and outflow counts of passengers at each of the $S = 471$ stations of the NYC subway system.\(^1\)

3.3. Accuracy and precision of OD-coefficient estimates

Since the true OD-matrices of test network A and B are known, we can quantify the accuracy of the OD-coefficient estimates. We define the Bayesian posterior accuracy in terms of the mean-squared-error (MSE) between the posterior means $\hat{\mu}_{ij}$ and the OD-coefficients’ true values $\bar{a}_{ij}$. In addition, we define the accuracy of the QP optimisation solution in terms of the MSE between the point estimates $\hat{\alpha}_y$ and the true values $\bar{a}_y$, such that

\[
\text{MSE - MCMC} = \frac{1}{S^2} \sum_{s=1}^{S} \sum_{j=1}^{S} (\hat{\mu}_{ij} - \bar{a}_{ij})^2,
\]

\[
\text{MSE - QP} = \frac{1}{S^2} \sum_{s=1}^{S} \sum_{j=1}^{S} (\hat{\alpha}_y - \bar{a}_y)^2. \tag{22}
\]

The highest-posterior-density (HPD) interval of the Bayesian OD-coefficient posterior is a measure of the estimates’ uncertainty. We define the aggregate uncertainty statistic, which we will refer to as the estimates’ precision, in terms of the mean 95% HPD interval across all coefficient posterior estimates,

\[
\bar{\mu}_{\text{HPD}} = \frac{1}{S^2} \sum_{s=1}^{S} \sum_{j=1}^{S} \text{HPD}(\bar{a}_{ij})_{95\%}. \tag{23}
\]

We do not define a precision statistic for the optimisation results because they are point estimates. The accuracy statistic measures the estimation error with respect to the true values, whereas the precision statistic measures the estimates’ posterior resolution irrespective of how far the estimates are from their true values.

4. Results and discussion

We encode and run the proposed models in Stan (Carpenter et al., 2017), which implements the NUTS algorithm. Data processing is implemented in Python 3.6. The model code and compilation of the model are executed on the Python package pystan. In order to satisfy the OD-coefficients’ non-negative sum-to-one constraint, the coefficients are parameterised as simplexes. The definition of a

\(^1\) In fact, there are currently 472 stations in the NYC subway system. This study is based on data prior to the re-opening of the station Cortlandt Street in 2018, after being rebuilt in the aftermath of the 9/11 attacks.
simplex poses an inherent sampling difficulty, especially in high dimensions, given that any new sample in an arbitrary direction is hardly guaranteed to satisfy the simplex constraint. Stan implements a stick-breaking transformation to map from the \((S - 1)\)-dimensional simplex to an unconstrained, more sampling efficient \((S - 2)\)-dimensional space.

Throughout the analysis, we assume unbiased prior knowledge about the concentration of the OD-coefficients and set \(c\) in Eq. (8) to 1. This results in a flat Dirichlet distribution of the OD-coefficients, i.e., the probability mass is uniformly spread over the entire domain of the multidimensional OD-coefficient simplex support. To estimate the average travel time of path alternatives between origins and destinations for the average-delay model, the travel times of all path alternatives falling within a 10-minute arrival time margin from the earliest arrival path are averaged.

4.1. Instantaneous-balance model verification

We consider test network A, estimate the OD-coefficients based on the instantaneous-balance model using the proposed Bayesian

---

Fig. 4. OD-coefficient estimates for a 15-node network, based on \(N = 30\) entry and exit count observations at each station node. The entry count dispersion parameter of the generative process model is \(\phi = 10\). The red stars mark the true OD-coefficient values, the green dots indicate the quadratic programme solution, and the horizontal dashes mark the sample means of the MCMC posterior estimate. Vertical lines indicate the 95% probability highest posterior density (HPD) intervals of the MCMC posterior estimate.
Table 1
Mean entry count levels for the 15-station instantaneous-balance test network A.

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>μ_{x,i}</td>
<td>600</td>
<td>900</td>
<td>1400</td>
<td>400</td>
<td>550</td>
<td>650</td>
<td>900</td>
<td>1000</td>
<td>750</td>
<td>450</td>
<td>200</td>
<td>650</td>
<td>750</td>
<td>1000</td>
<td>1300</td>
</tr>
</tbody>
</table>

Fig. 5. Predictive test on a single count observation under instantaneous-balance model assumptions. The red vertical line indicates the observed exit counts; The dashed green line indicates the prediction based on the QP coefficient point estimates; The greyed histogram indicates the posterior predictive distribution based on the MCMC coefficient posterior estimates; The dotted black line indicates the mean of the posterior predictive distributions.
inference approach, and compare these estimates against the QP estimates. The instantaneous-balance model results are based on the regularized formulations of Section 2.2.2 (Appendix G provides additional detail on the comparison between the regularized and unregularized formulation results).

4.1.1. Bayesian and QP model comparison

Fig. 4 illustrates the OD-coefficient estimates according to both approaches,\(^2\) using observations from the generative process in Appendix D with \(\phi = 10\) and \(N = 30\). Both the QP optimisation and the Bayesian MCMC approach reliably reconstruct the latent OD-matrix from only the entry and exit count observations of test network A, with an MSE of approximately 0.001 for both the QP point estimates and MCMC posterior means. However, the QP estimates do not provide a measure of the parameters’ uncertainty, whereas MCMC posterior samples provide an interval estimate. The vertical bars in Fig. 4 indicate the 95% probability HPD interval, which provides uncertainty bounds (i.e., credibility intervals) of the MCMC-derived parameter estimates. Remarkably, several origin stations consistently exhibit coefficient estimates with large uncertainties (i.e., wide HPD intervals), which are correlated with small mean entry counts (as reported in Appendix D, Table D.1). This is because the exit counts at destination stations are the result of compounded OD-flows from all origin stations, and, consequently, OD-flows that originate from stations with small inflows are more difficult to discern against the more dominant contributions of high-inflow stations.

The posterior predictive distribution of the exit counts is found by marginalizing the likelihood model over the posterior distribution. We determine the posterior predictive distribution during the MCMC sampling routines, by using the sampled OD-coefficient values and substituting them into the model description in Eq. (6) and sampling from the truncated normal distribution in Eq. (5). Fig. 5 shows the posterior predictive distributions for a single observation of the exit counts at the 15 stations of test network A. The observations fall within the predictive distribution’s support, indicating that the model can recover the structure of the observed data. In contrast, the QP-derived point estimates only provide a crisp value, and their accuracy can only be judged by their residual error, which in the example of Fig. 5 deviate similarly from the observations as the MCMC derived predictive distribution means. When predicting the exit counts on 10,000 unseen observations that were not used for parameter estimation, the MSE between the QP predictions and observations is 1006.8, versus an MSE of 934.1 between the MCMC posterior predictive means and the observations. These deviations from the observed counts are generally due to model inadequacies and limited sample size – here, the deviations are merely due to the limited sample size, since both the inference model and the generative process are based on the same network assumptions. Appendix G provides additional discussions on the influence of the number of observations on accuracy.

---

\(^2\) The MCMC sampler is configured to the default control and sampling parameters. The number of warm-up iterations is set to 1500; the number of sampling iterations is 1000. The sampler is randomly initialised on four chains, resulting in a total of 4000 posterior samples. The MCMC sampler does not exhibit any divergences and none of the tree searches saturates at the maximum tree-depth of 10. The potential scale reduction statistic, \(\hat{R}\), is close to 1.0 for all parameters (\(\hat{R}_{\text{max}} = 1.002\)). The effective sample size ratio \(\text{r}_{\text{eff}} = N_{\text{eff}}/N_{\text{MC}}\) is generally close to 1.0 across all estimated parameters with (\(\text{r}_{\text{eff}}\text{min} = 0.25\), \(\mu_{\text{r}_{\text{eff}}} = 1.1\), and \(\sigma_{\text{r}_{\text{eff}}} = 0.23\)) – The NUTS algorithm used in Stan can produce samples with \(N_{\text{eff}} > N_{\text{MC}}\). The QP is solved on the commercial solver CPLEX.
The optimisation solver and the MCMC sampler (4 chains with 2500 draws each) complete in 0.1 s and 60 s, respectively, on an Intel 2.9 GHz i5 CPU, 16 GB DDR3 RAM laptop computer. Therefore, the optimisation approach outperforms the Bayesian approach when applied to very small transit systems like test network A, if estimates’ uncertainty is not required. In contrast, the Bayesian approach provides a measure of the estimates’ uncertainty and considers the concentration of probability mass of the parameter posterior distribution, which reveals essential in high dimensional problems. In large-scale transit networks, we therefore focus on the Bayesian approach. Moreover, we prompt to Section 4.3, where we revisit the QP approach on a high-dimensional parameter estimation problem for a large-scale network and find that the QP approach produces unreliable OD-estimates, regarding the coherence with the model assumptions.

Fig. 7. The effect of (a) model type (instantaneous-balance or average-delay), (b) observation window width \( w \in \{5, 15, 30, 60\} \) min, (c) trend scale \( \eta \in \{0.0, 0.5, 1.0\} \), and (d) count dispersion \( \phi \in \{10, 100, 1000\} \) on the accuracy and precision of the Bayesian OD-estimates with \( N = 100 \) observations, resulting in \( 2 \times 4 \times 3 \times 3 = 72 \) test sites. The arrows indicate the average direction of the change in accuracy and precision of the two model types as the observation window width increases, the trend scale increases, and the count dispersion increases (i.e., as \( \phi \) decreases), in (b), (c), and (d), respectively. The boxed markers in (a) indicate the worst estimate of the instantaneous-balance and average-delay model. The triangular-framed and circled markers in (a) indicate the best estimate of the instantaneous-balance model and of the average-delay model, respectively; these are further discussed in 0.
4.1.2. Effect of count data dispersion on accuracy and precision

The dispersion parameter $\phi$ controls the variance of the entry count observations. We vary the value of $\phi$ considering $N = 30$ observations and constant entry count means given in Appendix D, Table D.1. Fig. 6 demonstrates that an increase in the variance of the entry count observations (i.e., the distribution of entry counts becomes more overdispersed) improves the accuracy of the Bayesian and QP estimates and the precision of the Bayesian estimates.

The OD-coefficient estimates are solely distilled from the entry and exit count observations and rely on identifying how changes in the entry counts at one station influence the changes in the exit counts at another station. Therefore, if the $N$ observations of entry counts at all stations were constant, there would be infinitely many solutions to the OD-estimation problem. Conversely, as the variability of the entry counts becomes more pronounced, the signals that entry count fluctuations impose on the exit count observations are more discernible and enable the OD-estimation method to more accurately and precisely infer the OD-coefficients.

4.2. Instantaneous-balance and average-delay model sensitivity

We apply the average-delay model and the instantaneous-balance model to test network B and compare their estimation results. For test network B, the cumulative entry and exit counts may be imbalanced over the observation window. To comply with the instantaneous-balance assumption, we correct for this imbalance using the approach in Appendix C prior to applying the instantaneous-balance model. For this discussion, we solely focus on the Bayesian estimates. We analyse the sensitivity of the OD-estimation accuracy and precision with respect to (a) the model type, i.e., the instantaneous-balance model and the average-delay model, (b) the observation window width $w$, (c) the trend scale $\eta$, and (d) the count data dispersion $\phi$. Fig. 7 shows the accuracy and precision of the OD-coefficient estimates for the instantaneous-balance and the average-delay model; markers closer to the origin indicate more accurate and precise estimates.

Fig. 7a shows that the average-delay model estimates are overall more accurate and precise. Therefore, the average travel delay model is more reliable in retrieving the true OD-matrix from the time-binned count data than the time-aggregated instantaneous-balance model and is more robust to changes in count characteristics. Additionally, Fig. 7b–d illustrate the effect of observation window width, trend scale, and Negative Binomial dispersion, respectively. The arrows in Fig. 7 indicate the stepped average change of the accuracy and precision given a change of the respective test parameter values. The observation window width (in the case of the average-delay model this corresponds to the arrival window width) significantly influences the accuracy and precision of both the instantaneous-balance and average-delay model estimates, since it prescribes the information horizon that the models are able to observe and the number of binned data points considered in the average-delay model. The trend scale unnoticeably influences the results of the average-delay model, while the instantaneous-balance struggles to determine reliable parameter estimates as the trend scale increases and interferes with the instantaneous-balance assumptions. Moreover, the estimates of the models improve as the count dispersion increases, because stronger fluctuations in the count observations provide clearer evidence of the level of coupling between origin and destinations as shown in Section 0. Appendix H further supports these findings based on the correlation and metamodel-based sensitivity analysis between the test parameters and the accuracy and precision statistics.

The estimation ability of the average-delay model, however, demands specific data requirements and is subject to higher computational complexity. In particular, the average-delay model relies on granular time series and extensive information regarding the transit network timetable and transfer times between stations. Furthermore, while both the average-delay and instantaneous-balance model require matrix evaluations associated with a time complexity of $\mathcal{O}(n^2 \pi)$ to $\mathcal{O}(n^3)$ depending on the specific algorithm, the average-delay model requires 5 looped matrix evaluations for the subsets of the OD-assignment matrix for each destination station. For 100 observations on test network B, the execution time of the MCMC sampler varies between 1–30 min for the average-delay model and between 30 s to 3 min for the instantaneous-balance model on an Intel 2.9 GHz i5 CPU, 16 GB DDR3 RAM laptop.
computer. The variability of the execution time is subject to the combinations of different observation window widths, trend scales, and Negative Binomial dispersions, creating irregularly complex posterior distributions to sample from.

In case the computations for the average-delay model are prohibitively costly (e.g., in the case of the NYC subway network), or travel time information is not readily available, the instantaneous-balance model shows to perform well for (a) a sufficiently long observation window to span the travel times between origins and destinations, (b) flat or small trends in the periodically repeating (e.g., daily) count series, (c) a suitable degree of dispersion in the count series and aggregated observed data, and (d) an ample number of observations across which system conditions (travel demand and supply) can be assumed to be invariant.

### 4.3. OD-estimation for the NYC subway network

We apply the proposed instantaneous-balance Bayesian OD-estimation model to the NYC subway network. With $S = 471$ stations, the total number of OD-coefficient estimates is equal to $S(S - 1) = 221,370$. We consider all weekday counts in three observation windows, namely, 8 AM to 9 AM, 1 PM to 2 PM, and 6 PM to 7 PM between April and October with records starting in 2010 and ending in 2018, resulting in $N = 1315$ observations. The three windows characterize the morning peak, the midday off-peak, and the evening peak window.
peak, respectively, and are shown in Fig. 8 along with the average total number of entering and exiting passengers throughout the day. According to the timetable data in (Metropolitan Transport Authority, 2019b), the average travel time across the network is roughly 55 min and therefore justifies the width of the one hour observation window for the morning peak, midday off-peak, and evening peak window. Passenger count data for NYC are only available in an aggregated 4 h resolution basis and, therefore, are solely applicable to the instantaneous-balance OD-estimation model. Moreover, in preliminary testing it was found that using approximately ten days of count observations together with the average-delay formulation results in approximately the same computation time as using the full 1315 daily observations with the instantaneous-balance formulation.

The posterior distribution estimates of the OD-coefficients are determined from the MCMC sampling algorithm described in Section 2.4. The MCMC sampler completes in roughly 4 days for 2000 iterations (1000 warm-up and 1000 sampling iterations) on each of 10 parallel chains, when executed on a computing cluster with 64 Intel® Xeon® IE5-2699 v3 2.30 GHz CPU’s and 200 GB of DDR3 RAM. This results in a sampling with a total of 10,000 MCMC draws from the (marginal) posterior distribution of each of the 221,370 OD-split coefficients for each observation window. Regarding the NUTS sampler, no divergences are recorded during the sampling runs of all three observation windows. The tree search depth of the NUTS algorithm is set to 10. The sampler saturates this tree depth for all draws in all three sampling runs. While the absence of divergences indicates that the model is well specified and the associated MCMC draws are valid, tree-depth saturations indicate that the algorithm is not sampling efficiently. Nonetheless, the drawn samples are valid MCMC draws, even if the search tree of the NUTS algorithm saturates at the maximum defined depth.

### 4.3.1. OD-coefficient posterior means and absolute OD-demand

The OD-coefficient posterior means vary distinctly across origins and destinations, with noticeable patterns of smaller and larger valued estimates in the range from $10^{-6}$ to 0.7. In comparison, the Dirichlet OD-coefficient hyperprior in Eq. (8) defines a mean equal to $(S - 1)^{-1} \approx 0.002$ for all coefficients. Since the concentration parameter in Eq. (8) is set to $c = 1$, the simplex-constrained OD-coefficient support is subject to uniform probability mass throughout. The distinct OD-coefficient patterns therefore indicate that the OD-coefficient posterior sampling trajectory is mostly guided by the likelihood model and not the hyperprior.

Indeed, this becomes more evident in the absolute OD-flow. Fig. 9 shows the absolute passenger flow between origin and destination locations following the estimate of the average absolute OD-flow detailed in 0. Fig. 9a demonstrates that during the morning peak, most passenger trips are bound for stations along the subway line corridors that stretch across the business and commercial centres of Upper East and West Side, Midtown, and Lower Manhattan. During the midday off-peak in Fig. 9b, the passenger ridership reduces, while most passenger trips continue to be bound towards stations in Manhattan from stations in the surrounding boroughs. Moreover, stations in Queens and the Bronx start to more strongly attract passenger trips. During the evening peak in Fig. 9c, the number of passenger trips bound towards stations in Lower and Midtown Manhattan reduces, while a large number of passenger trips end at stations located in residential areas in Brooklyn, Queens, and the Bronx. Only few stations in Lower and Midtown Manhattan continue to attract many passenger trips, due to their vicinity to regional train services and activity centres.

### 4.3.2. Posterior predictive checks

The instantaneous-balance model predictions and the real-world observations for the NYC subway are compared using posterior predictive checks, in order to establish whether the model can reproduce the original observations.

To this aim, the means of the observed exit counts $\bar{y}_{ij}$ are compared to predictions generated from the instantaneous-balance model formulation in Eq. (6). The intercept $r_j$ is omitted, so that the predictive model satisfies the uncorrected instantaneous-balance assumption. In this way, it is assumed that passenger exit counts are solely the result of passengers entering the network and distributing across the destinations according to the relative proportions given by the OD-coefficients. Moreover, the predictive model determines the predictive distribution of the uncorrected latent location parameter $\tilde{\mu}_{yj}$, as opposed to the predictive distribution of the

![Fig. 10.](image-url) Instantaneous-balance model posterior predictive exit count distributions versus the observed exit counts for the 471 stations of the NYC subway network during the (a) morning peak, (b) midday off-peak, and (c) evening peak window. For each station the distribution consists of 1000 samples. The grey annotated lines indicate the (relative) percentage difference contours.
exit count observations. The average entry counts of the 1315 observations are substituted, as well as a thinned sample of 1000 draws from the OD-coefficient posterior.

Fig. 10 compares the mean of the observed exit counts $\bar{y}_j$ with the posterior predictive distributions of the exit count location parameter $\hat{\mu}_{y_j}$ (for large exit counts this converges to the mean). The coefficient of variation is less than 0.04 across all predictive distributions. Consequently, the predictive distribution samples (i.e., 1000 samples for each station) that are plotted in Fig. 10 are narrowly bundled for every station, visually almost converging to a single point. The model closely captures the observed data across all three observation windows, particularly for stations with large exit counts. Indeed, for 382, 419, and 452 of the 388, 421, 452 stations with mean exit counts of ≥100 passengers per hour for the morning, midday, and evening window, respectively, the relative absolute prediction errors are below 10%. However, one particular station exhibits anomalous prediction errors; for Wall Street station (lines 4 and 5) the relative prediction errors are approximately equal to 25%, despite generally large exit count observations, i.e., the observed mean exit counts during the morning, midday, and evening observation window are in the order of 9500, 2500, and 1500 passengers per hour. We attribute these discrepancies to inconsistent data records of the entry and exit count observations at Wall Street station, where in 2014 the observations significantly dropped (c.f., Fig. F.1d in Appendix F), and thus violate the assumptions regarding the periodicity of recurring system conditions. Finally, at stations with exit counts roughly below 100 passengers per hour, the model over-predicts by up to almost four-fold (i.e., 300% relative error). These errors are negligible, considering that the total exit counts sum to approximately $3 \times 10^5$ to $5 \times 10^5$ passengers per hour.

### 4.3.3. MCMC convergence and autocorrelation

Convergence monitoring is essential in qualifying whether the MCMC samples are drawn from the equilibrium distribution, i.e., each randomly initialised chain has converged to the equilibrium distribution. This is quantified by the potential scale reduction statistic, $\hat{R}$, as the ratio between the sum of the average within-chain variance plus cross-chain sample variances, and the average within-chain variance (Gelman et al., 1992). A value of $\hat{R}$ close to unity indicates that the chains of the MCMC algorithm sample from the same equilibrium posterior distribution. Fig. 11 shows the distribution of $\hat{R}$ for the 221,370 OD-coefficients and the three observation windows. Fig. 11a and b show that most $\hat{R}$ statistics are smaller than 1.2 and predominantly concentrate close to 1.0 for the morning and midday observation window. Hence, the sampling chains converge for 99.96% and 99.89% of the 221,370 OD-coefficients for the morning and midday observation windows, respectively, considering $\hat{R} = 1.2$ as the convergence threshold. However, Fig. 11c shows that the evening observation window only exhibits convergence for few OD-coefficients for the same $\hat{R}$ threshold (constituting 5.18% of all coefficient posteriors), where the non-converged samples account for approximately 98% of the overall OD-flow between stations. Therefore, the posterior distribution for the evening observation window is more difficult to explore and sample from. Data records for the evening observation window are slightly less dispersed than for the morning and midday observation windows, which could be the cause for the sampling deficiencies. Despite this, the posterior estimates for the evening observation window are considered sufficiently reliable, because the posterior predictive check in Section 4.3.2 indicates that the model adequately captures the observations. Finally, autocorrelation analysis (see Appendix K) shows that 80% of all coefficient estimates exhibit an effective sample size ratio $N_{\text{eff}}/N_{\text{MC}} > 0.8$ for the morning peak, $N_{\text{eff}}/N_{\text{MC}} > 0.25$ for the midday off-peak, and $N_{\text{eff}}/N_{\text{MC}} > 0.02$ for the evening peak. Therefore, the estimates’ variance can mostly be attributed to the inherent observation variability and modelling uncertainties rather than to autocorrelated samples, in particular for the morning peak. For the evening peak, the effective sample size ratio reveals that samples are more autocorrelated, which underscores the difficulties in sampling efficiently from the evening peak posterior distribution. We
therefore presume that sufficient dispersion in the count observations is a necessary prerequisite for determining more precise OD-coefficient posteriors with acceptable convergence.

4.3.4. Inadequacy of QP estimates

Fig. 12 illustrates the OD-coefficient matrix resulting from the QP optimisation model in Appendix A and applied to the NYC subway system with observations from the morning observation window. The QP model is solved in CPLEX resulting in an optimal solution after approximately 6 h on the same computing cluster as for the MCMC sampler. Despite the generally favourable non-sparse output of QP optimisation models, the QP optimisation produces a highly sparse OD-matrix, with a matrix sparsity of 99% (for a zero round-off for values below $10^{-9}$). In contrast, the MCMC-estimates result in a matrix sparsity of 0.23%, where 64% of the coefficients are larger than $10^{-3}$. Moreover, the maximum coefficient for the QP-derived OD-matrix is equal to 1.0, suggesting that all inflow at an origin station is assigned to a single destination station. However, the passenger inflow at an origin station is likely to distribute to multiple destination stations – even those that do not dominantly attract many passenger trips. Most of the OD-flow according to the QP derived OD-matrix is restricted to only a few OD-pairs, while the MCMC-derived matrix distributes origin inflow considerably more evenly. A possible cause of this discrepancy can be that QP optimisation is sensitive to outliers in the observed data. However, outliers were rectified prior to solving the optimisation problem (see Appendix F) and tests with a linear programme formulation, that is less sensitive to outliers, similarly result in a highly sparse result. Another cause is that the high-dimensional problem, combined with the limited number of count measurements (i.e., the underdetermined problem of 1315 observations versus $471 \times 471 = 221,841$ OD-coefficients), may have multiple local optima, such that the OD-matrix estimate shown in Fig. 12 is only one of numerous possible solutions. Moreover, the high-dimensional underdetermined problem, which leads to overfitting of the observed data and difficulty in retrieving a possible global optimum, is additionally hampered by model assumptions, that may partly mis-specify the real-world processes. For this reason, the QP optimisation routine presumably fails to adequately determine physically interpretable estimates.

Fig. 12. The inferred OD-matrix for the NYC subway system for the morning peak window with $N = 1315$ observations, based on the quadratic programme optimisation approach. OD-coefficient estimates with $\hat{\alpha}_{ij} < 10^{-9}$ are plotted as $10^{-9}$.

Fig. 13. Samples from a Dirichlet distribution with Dir($\mathbf{c} \mid 1, 1, 1$). Variates are either more or less concentrated towards the vertices of the $(k-1)$-dimensional simplex (with $k = 3$), depending on the value of the concentration parameter $c$. For $c = 1$ the samples are uniformly distributed over the simplex.
Fig. 14. Test network B – a subcomponent of the NYC subway system; Colours indicate the different subway lines. Dotted black lines indicate transfer footpaths. The annotated stations are assumed to be the only access and egress locations of the network.

Fig. 15. The mean and 25th to 75th percentile range of the entry count time series at a single station for 24 time bins and evaluated over 100 observations, based on the hybrid ARIMA-INGARCH generative entry count model. Dots connected by solid lines indicate the mean, and grey-shaded areas indicate the percentile range. The graphs correspond to different values of the scale parameter $\eta$ and Negative Binomial dispersion parameter $\phi$; (a) $\phi = 10$; (b) $\phi = 100$; (c) $\phi = 1000$. 
5. Conclusion

This work proposes a Bayesian inference approach for static Origin-Destination (OD)-estimation and system identification from time series measurements of nodal in- and outflows in a networked transit system. The approach is particularly suited for high-dimensional and underdetermined problems with several hundred network nodes and uninformative priors about OD-demand. We suggest two different model formulations: the instantaneous-balance and average-delay model. Overall, the average travel delay model is more robust to changes in the observation characteristics and is able to more accurately and precisely determine the OD-coefficient posterior estimates, under the condition that high-resolution count data are available. However, given the appropriate observation characteristics, the instantaneous-balance model recovers the OD-matrix sufficiently well at smaller computational cost and coarser
data resolution. Moreover, we demonstrate that the Bayesian instantaneous-balance model can be applied to large-scale real-world instances, such as the NYC subway system. In fact, OD-coefficient estimates for the NYC subway system overall reflect intuitive demand patterns alongside satisfying convergence, autocorrelation, and posterior predictive tests.

We deem sufficient dispersion (i.e., variability) in the count observations as being key in determining accurate and precise OD-estimates. Indeed, the real-world dynamics of passenger access and egress rates in transit systems provide sufficient variability to discern most of the relevant OD-flows. Moreover, the variability of observed in- and outflows is sufficient to derive reasonable credibility intervals for most of the OD pairs.

This paper shows that despite using uninformative priors, the Bayesian OD-estimation approach is able to determine mean parameter estimates that are at least as good as point estimates obtained from an optimisation approach, given sufficiently large data sets. In addition, the Bayesian approach provides distribution estimates that allow for additional information compared to the traditional point estimates. For instance, the distribution estimates are useful for determining reliable transport network designs, where a system could be dimensioned more conservatively according to the upper quantiles of OD-demand.

Table 2

The sensitivity of the models’ estimation accuracy and precision to test parameters. The table lists the correlations between the test parameter values and the models’ estimation accuracy and precision, and the coefficients of the fitted model in Eq. (H.1) along with their p-values as well as the test parameters’ first-order and total Sobol’ indices (\(S\) and \(S_T\), respectively). Columns with header I correspond to the instantaneous-balance model; Columns with header II correspond to the average-delay model. Positive correlation with respect to the window width, trend scale, or dispersion parameter means that a decrease in either test parameter will improve the model’s accuracy or precision, and vice versa. The variance/dispersion of the samples drawn from the negative binomial distribution is inversely proportional to the dispersion parameter \(\phi\).

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Window width, (w)</th>
<th>Trend scale, (\eta)</th>
<th>Dispersion, (\phi)</th>
<th>Intercept, (\beta_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Corr.</td>
<td>-0.8986</td>
<td>-0.6378</td>
<td>0.2195</td>
<td>-0.0645</td>
</tr>
<tr>
<td>Coeff.</td>
<td>-0.0050</td>
<td>-0.0022</td>
<td>0.0011</td>
<td>-0.0002</td>
</tr>
<tr>
<td>p-val.</td>
<td>4.65 (\times) 10^{-6}</td>
<td>4.68 (\times) 10^{-6}</td>
<td>8.70 (\times) 10^{-4}</td>
<td>5.82 (\times) 10^{-1}</td>
</tr>
<tr>
<td>(S)</td>
<td>0.9273</td>
<td>0.7832</td>
<td>0.0450</td>
<td>0.0105</td>
</tr>
<tr>
<td>(S_T)</td>
<td>0.9235</td>
<td>0.7748</td>
<td>0.0471</td>
<td>0.0067</td>
</tr>
</tbody>
</table>

Table 3

The worst and best-case estimation results for the instantaneous-balance and average-delay model. The red stars are the true OD coefficient values, the vertical and horizontal bars visualise the mean and HPD interval of the estimates. Coefficients are sorted from lowest to greatest along the vertical axis. (a) Both models’ worst-case; \(w = 5\) min, \(\eta = 1\), \(\phi = 1000\). (b) The instantaneous-balance model best-case; \(w = 1\) h, \(\eta = 0\), \(\phi = 10\). (c) The average-delay model best-case; \(w = 1\) h, \(\eta = 1\), \(\phi = 10\). The left- and right-hand depictions in each sub-figure juxtapose the instantaneous-balance next to the average-delay model results.

Fig. 18. The worst and best-case estimation results for the instantaneous-balance and average-delay model. The red stars are the true OD coefficient values, the vertical and horizontal bars visualise the mean and HPD interval of the estimates. Coefficients are sorted from lowest to greatest along the vertical axis. (a) Both models’ worst-case; \(w = 5\) min, \(\eta = 1\), \(\phi = 1000\). (b) The instantaneous-balance model best-case; \(w = 1\) h, \(\eta = 0\), \(\phi = 10\). (c) The average-delay model best-case; \(w = 1\) h, \(\eta = 1\), \(\phi = 10\). The left- and right-hand depictions in each sub-figure juxtapose the instantaneous-balance next to the average-delay model results.
However, the choice of the observation time window has impact on the quality and performance of the approach and particularly so for the instantaneous-balance model. In networks with few routes or with a clear lower bound on the travel time between origin and destination, the observation window can be kept narrow and include most of the passenger trips that both commence and end within the window. In networks with unreliable travel times or particularly long routes, the estimation accuracy and precision will be hampered by the truncation of observation window and missed passenger trips.

With identifiability being a key concern in determining a reliable OD-estimate, further research will be required to establish the sensitivity to the problem size and properties of the observational data. For instance, we do not establish a generalizable theoretical condition under which the Bayesian inference model or the benchmark QP approach cease to provide accurate estimates. In this regard, this work merely identifies an example for which the QP formulation is unable to provide dependable estimates. Future work should also investigate how the proposed solution, using ad-hoc regularizations and gradient-based HMC sampling routines, can be integrated with established approaches that make use of seed matrices or assignment models in case of measurable route or link flows. It would be moreover interesting to understand the possible benefits of an even more complex model where errors can be modelled independently based on several factors (for instance, measurement noise in X; different measurement noise characteristics; measurement noise versus model deficiencies).

With the advent of novel fare collection systems, transit OD-estimation is becoming considerably more refined. The presented work will be complementary to smart-card-enabled methods that can extract OD-information from trip records. Moreover, we propose that this work is not exclusive to transit network OD-estimation and is presumably equally amenable to road traffic OD-estimation or network tomography.

**Funding**

This work was supported by the National Research Foundation (NRF) of Singapore (FI 370074011), as part of NRF’s Campus for Research Excellence and Technological Enterprise (CREATE) programme.

**CRediT authorship contribution statement**

**Steffen O.P. Blume:** Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Writing – original draft, Visualization. **Francesco Corman:** Writing – review & editing. **Giovanni Sansavini:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing – review & editing.

**Declaration of Competing Interest**

The authors declare no conflict of interest.

**Acknowledgments**

We would like to thank the anonymous reviewers for their valuable inputs to considerably improve the quality of this work.
Appendix A. The instantaneous-balance optimisation model

The proposed Bayesian inference model is compared with an equivalent point estimate formulation. The point estimate formulation finds OD-coefficients estimates, such that the total sum of the squared residuals between the model predictions and observations is minimised. It is formulated as a constrained quadratic optimisation programme that minimises the $L_2$-norm (i.e., sum-squared-error objective function) between the predicted and observed exit counts, where the formal expressions are given by

\[
\begin{align*}
\text{minimize} & \quad \sum_{j \in S} \left( \epsilon_j^{(n)} \right)^2 + \lambda \sum_{j \in S} \left( r_j \right)^2 + \sum_{j \in S} \left( e_j \right)^2 \\
\text{subject to} & \quad \sum_{i = 1}^S a_{ij} \hat{x}_i^{(n)} + r_j - y_j^{(n)} = e_j \forall j \in \{1, 2, \ldots, S\}, n \in \{1, 2, \ldots, N\}, \\
& \quad \sum_{i = 1}^S a_{ij} = 1 \forall i \in \{1, 2, \ldots, S\}, \\
& \quad 0 \leq a_{ij} \leq 1 \forall i \neq j.
\end{align*}
\]  

(A.1)

Eq. (A.1) is similar to the constrained optimisation model by Cremer and Keller (1987), who used it for OD-estimation of the turning movements at intersections. We extend the formulation by adding the intercept $r_j$ and including the minimal-bias and expected value regularization terms into the objective function (i.e., $\lambda \sum_{j \in S} (r_j)^2$ and $\sum_{j \in S} (e_j)^2$, where $\lambda$ is the scaling parameter from Eq. (10)). These terms are further explained in Section 2.2.2, which also describes their implementation into the Bayesian model formulation.

The optimisation objective function could similarly be expressed in terms of the sum-absolute-error (i.e., the $L_1$-norm), resulting in a linear programme formulation. The linear and quadratic formulation each come with canonical properties that make them favourable depending on the problem setting and desired outcome. The least absolute deviations problem associated with the linear programme is robust to outliers in the data and produces sparse outputs, whereas the least squares problem associated with the quadratic programme is sensitive to outliers and tends to produce non-sparse outputs. While it is reasonable to assume that the inflow at an origin will be dominantly bound for only a few destinations, it is implausible that none of the inflow will ever be bound for the remaining destination nodes. Hence, a sparse solution that results in many zero-valued OD-coefficients is neither desired nor plausible. We therefore choose the quadratic programme formulation in Eq. (A.1) and rectify data outliers prior to solving it.

Appendix B. The Dirichlet hyperprior

Section 2.2.1 describes that we place a Dirichlet hyperprior on the rows of the OD-matrix. Here, we illustrate the influence of the concentration parameter $c$ included in the hyperprior. Fig. B.1 shows samples from a Dirichlet distribution according to $\text{Dir}(x; c[1, 1, 1])$. This particular distribution spans a two-dimensional simplex equivalent to the triangle with vertices (1, 0, 0), (0, 1, 0), and (0, 0, 1). By reducing the concentration parameter $c$ below 1, the variates are biased to be sparser, whereas $c > 1$ creates denser, hence concentrated, variates. For the MCMC-based OD-estimation problem, values of $c$ considerably below 1 may make the simplex prohibitively difficult to sample over, given the concentration of probability mass near the vertices of the simplex. A value larger than 1, may put too much bias towards equal OD-coefficients.

Appendix C. Count imbalance correction

As a consequence of the cumulative count aggregation for the instantaneous-balance model, the entry count observations are not guaranteed to balance with the exit count observations that are accumulated over the same observation window. A possible solution to re-balance these counts is to distribute the excess total counts across the stations by rule of proportion, accounting for the stations’ relative contribution to the total entry and exit counts, such that
chosen destination are incremented by 1; This process is repeated

\begin{equation}
\begin{cases}
\tilde{x}^{(n)}_i = x^{(n)}_i - \frac{x^{(n)}_i \bar{x}^{(n)}_i}{x^{(n)}_i} = x^{(n)}_i y^{(n)}_i, & \text{if } x^{(n)}_i < 0 \\
\tilde{y}^{(n)}_j = y^{(n)}_j - \frac{y^{(n)}_j \bar{y}^{(n)}_j}{y^{(n)}_j} = y^{(n)}_j x^{(n)}_i, & \text{if } x^{(n)}_i > 0 
\end{cases}
\end{equation}

(C.1)

where \( \tilde{x}^{(n)}_i \) and \( \tilde{y}^{(n)}_j \) are the balanced entry and exit counts at stations \( i \) and \( j \), respectively. The excess entry counts \( x^{(n)}_i \) are determined from the total summed entry and exit counts over all stations for the particular observation, according to

\begin{equation}
x^{(n)}_i = x^{(n)}_i - y^{(n)}_i,
\end{equation}

(C.2)

where

\begin{equation}
x^{(n)}_i = \sum_{j=1}^{S} x^{(n)}_i \cdot y^{(n)}_i = \sum_{j=1}^{S} y^{(n)}_j.
\end{equation}

Appendix D. Test network A

The following assumptions hold for test network A, which consist of \( S \) stations: (i) The network is fully connected and all in- and outflow at every station node can be fully observed; (ii) Travel time between nodes is ignored, such that any changes in inflow instantaneously affect changes in outflow; (iii) The total number of entering passengers is equal to the total number of exiting passengers (i.e., flows are balanced) in every observation window.

D.1 OD-matrix and count data generative process. The generative process creates the test data from an assumed known (i.e., hypothetical) OD-matrix. The non-negative OD-coefficients are generated for each station \( i \) from a symmetric Dirichlet distribution with concentration parameter \( c_i \), uniformly sampled between 0 and 2, to create an \((S - 1)\)-simplex and enforce the sum-to-one constraint,

\begin{equation}
c_i \sim U(0, 2),
\end{equation}

\begin{equation}
\bar{a}_{i:1:(S-1)} \sim \text{Dir}(c_i, 1),
\end{equation}

(D.1)

where \( \bar{a}_{i:1:(S-1)} \) is a row vector of size \((S - 1)\). By letting the concentration parameter range up to 2, a stronger degree of similarity between coefficients associated with the same origin is possible, that is, for \( c_i > 1 \) coefficients concentrate more densely towards the centre of the simplex. The sampled OD-coefficients are mapped to the \( i \)-th row of the OD-matrix in Eq. (1) according to

\begin{equation}
a_{ij} = \begin{cases}
\bar{a}_{ij}, & \text{if } i > j, \\
0, & \text{if } i = j, \\
\bar{a}_{i,j-1}, & \text{if } i < j .
\end{cases}
\end{equation}

(D.2)

At every station node \( i \) and for every observation \( n \), cumulative passenger entry counts \( x^{(n)}_i \) are generated from a Negative Binomial distribution with mean parameter \( \mu_{x,i} \) and dispersion parameter \( \varphi \in (0, \infty) \),

\begin{equation}
x^{(n)}_i \sim \text{NegBin}(\mu_{x,i}, \varphi)
\end{equation}

(D.3)

The Negative Binomial distribution is parameterized in the following form

\begin{equation}
P(Z=z|\mu, \varphi) = \frac{\Gamma(\phi + z)}{\Gamma(z + 1)\Gamma(\phi)} \left( \frac{\phi}{\phi + \mu} \right)^{\phi} \left( \frac{\mu}{\phi + \mu} \right)^{\mu}, \quad z = 0, 1, \ldots .
\end{equation}

(D.4)

where \( \Gamma \) is the gamma function. In this way, the variance of the Negative Binomial distribution is defined by

\begin{equation}
\sigma^2 = \text{Var}[z|\mu, \varphi] = \mu + \frac{\mu^2}{\phi}.
\end{equation}

(D.5)

As the dispersion parameter \( \varphi \) becomes smaller, the variance increases and the distribution becomes more overdispersed; as \( \varphi \) increases, the more \( \sigma^2 \to \mu \) and hence the more the distribution becomes equidispersed.

The generative process views the OD-split coefficients as conditional probabilities, where every split coefficient is the probability of choosing destination \( j \) given that a passenger accesses the network at origin \( i \). Given the choice probabilities, the generative process creates the cumulative exit count observations in the following way: For every passenger count out of the total entry counts \( x^{(n)}_i \) at origin \( i \) during observation interval \( n \), choose a destination \( j \) according to a categorical distribution with probabilities given by \( \bar{a}_{i:1:(S-1)} \), such that for origin \( i \) and destination \( j \),

\begin{equation}
P(j|\bar{a}_{i:}) = a_{ij}, \quad \text{with } j \in \{1, 2, \ldots, S\},
\end{equation}

(D.6)

where \( \bar{a}_{i:} \) is the choice probability vector given by the \( i \)-th row of the OD-matrix in Eq. (E.1); Next, the total number of exit counts at the chosen destination are incremented by 1; This process is repeated \( x^{(n)}_i \) times for every origin \( i \) and observation \( n \). The resulting total exit
counts at destination \( n \) during observation are denoted as \( \gamma_j^{(n)} \).

D.2 Test data generation. Test network A consists of \( S = 15 \) stations. The mean entry count levels across all observation windows and for each station are given in Table D.1. The dispersion parameter of the Negative Binomial model in Eq. (D.3) is initially fixed to \( \phi = 10 \). Exit counts are subsequently determined according to the aforementioned generative process, and thus by sequentially applying the categorical choice model in Eq. (D.6) for every observation instance and entry count at each station. The number of observation windows is initially set to \( N = 30 \).

Appendix E. Test network B

The steps used to generate the count data for test network B are explained in the upcoming sections, starting with the generation of the OD-coefficient matrix in Section E.1, whereas Sections E.2 and E.3 delve into the entry count and route choice model.

E.1 OD-matrix generative process.

The unknown true OD-matrix is assumed constant across all time bins and observations. It is generated by sampling from a Dirichlet distribution for every origin \( i \). The Dirichlet distribution takes a vector \( c_i \) of size \((S-1)\), such that

\[
\alpha_i \sim \text{Dir}(\epsilon_i).
\]  

(E.1)

The vector \( c_i \) consists of the concentration parameters \( c_j \), with \( j \in \{1, 2, \ldots, S\} \), where \( j \neq i \), such that \( c_i = (c_1, c_2, \ldots, c_{i-1}, c_{i+1}, \ldots, c_S - 1) \). The concentration parameters are sampled once from a uniform distribution, according to

\[
c_j \sim U(0, 1).
\]  

(E.2)

Here, we constrain the concentration parameters between 0 and 1. Since Eq. (E.1) defines a generic Dirichlet distribution, the sparsity of the coefficient samples is controlled by the relative weight of each of the concentration parameters with respect to all other concentration parameters (the magnitude of the concentration parameters only controls the variance of the Dirichlet distribution). In addition, we point out that the value for \( c_j \) is only sampled once for every station \( j \); That is, for every sample \( \alpha_i \), the concentration parameter vector \( c_i \) consists of the same source values \( c_j \in\{c_1, c_2, \ldots, c_{S-1}\} \) with \( j \neq i \). Each concentration parameter represents the attractiveness of a specific destination \( j \). This is to replicate that certain stations in a network will attract more trips than others; For instance, during the morning commute hours when most passengers will travel from residential areas to business and commercial districts. The vector \( \alpha_i \), is mapped to the \( i \)th row of the OD-matrix in Eq. (1) according to Eq. (D.2).

E.2 The entry count hybrid ARIMA-INGARCH model. Passenger entry counts are assumed as a superposition of a time-dependent trend and a stochastic process distortion. Entry counts are therefore simulated from an auto-regressive integrated moving average (ARIMA) model paired with an integer-valued generalized autoregressive conditional heteroscedasticity (INGARCH) model (Ferland et al., 2006), where we use the forecast R-library to simulate from the ARIMA model and the tcount R-library (Loboschik et al., 2017) to simulate from the INGARCH model.

The integer-valued generalized autoregressive conditional heteroscedasticity (INGARCH) model (Ferland et al., 2006) describes a count time series \( \{Y_t: t \in \mathbb{N}\} \) by modelling its conditional mean as

\[
E(Y_t|F_{t-1}) = \lambda_t,
\]

where \( \lambda_t \) is a process with \( \{\lambda_t: t \in \mathbb{N}\} \) the history \( \mathcal{F}_{t-1} \) is the history of the joint process \( \{Y_{t}, \lambda_{t}, \lambda_{t-1} + 1: t \in \mathbb{N}\} \) up to time \( t \) for \( Y_t \) and \( \lambda_t \), and \( t + 1 \) for the time-varying \( r \)-dimensional covariate vector \( Z_t = (Z_{t,1}, \ldots, Z_{t,r})^T \). The general form of the joint process is

\[
g(\lambda_t) = \lambda_0 + \sum_{k=1}^{n} \gamma_k \hat{g}(Y_{t-k}) + \sum_{i=1}^{q} \xi_i \hat{g}(\lambda_{t-i}) + \eta Z_t,
\]  

(E.3)

where \( g: R^r \to R \) is a link function and \( \hat{g} : \mathbb{N}_0 \to R \) is a transformation function. The parameter \( \beta_0 \) is the intercept. The second right-hand side term in Eq. (E.3) is a regression on past observations \( Y_{t-i}, Y_{t-i-1}, \ldots, Y_{t-1} \) with integers \( 0 < i_1 < i_2 < \ldots < i_p < \infty \) and \( p \in \mathbb{N}_0 \). The third term defines a regression on the lagged conditional means \( \lambda_{t-i}, \lambda_{t-i-1}, \ldots, \lambda_{t-i}, \) with \( 0 < j_1 < j_2 < \ldots < j_p < \infty \) and \( q \in \mathbb{N}_0 \). The parameters \( \gamma_k \) and \( \xi_i \) are the respective weights on regressed lags. The parameters \( \eta = (\eta_1, \ldots, \eta_r)^T \) weigh the effects of the covariates.

For the purpose of simulating the entry counts at stations we assume that \( g \) and \( \hat{g} \) are equal to the identity, replace the observations \( Y_t \) with our notation for the entry counts \( x_t^{(n)} \) (the entry counts at time \( t \) during observation \( n \); for the sake of clarity, we omit the station sub-index \( i \);} However, entry count time series are simulated for every station separately). Moreover, we assume a Negative Binomial (c. f., Eq. (D.4)) entry count likelihood model,

\[
x_t^{(n)}|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t^{(n)}, \phi),
\]  

(E.4)

where \( \phi \in (0, \infty) \) is the dispersion parameter, such that \( \text{Var}[x_t^{(n)}|\mathcal{F}_{t-1}] = \lambda_t + \lambda_t^2/\phi \). In addition, the covariates are a compound of a trend and a so-called intervention. Interventions model idiosyncrasies in the count time series such as sudden spikes or shifts. In our case, the intervention is solely used to simulate the initial value of the time series. The resulting model for the mean function of the count time series process is

\[
x_t^{(n)} = \lambda_0 + \sum_{k=1}^{n} \gamma_k x_{t-k}^{(n)} + \sum_{i=1}^{q} \xi_i x_{t-i}^{(n)} + \omega_0 \delta_{t-1}^{(n)}(t = \tau) + \eta Z_t,
\]  

(E.5)
where \( \delta_0^{-1}(t) \) denotes the deterministic intervention covariate, with \( t \) being the time occurrence and \( \delta_0^{-1} \) being the constant decay rate starting from time \( t \). Since the intervention models the times series’ initial value, we fix \( t = 0, \delta_0^{-1} = 0 \) (corresponds to a spike), and

\[
\alpha_0 = \mu/(1 - \eta).
\] (E.6)

The parameter \( \mu \) is the mean of a trend process \( F_t \) and \( \eta \in (0, 1) \) is a scale parameter that captures the effect strength of the trend on the time series. The covariate weights are given by \( \gamma = (\alpha_0, 1)^T \), with covariate vector \( Z_t = (\delta_0(t - 0), F_t)^T \).

The remaining component to be defined is the trend process. We model the trend according to a seasonal auto-regressive integrated moving average (ARIMA) process. The ARIMA model will be denoted with the common notation as \( \text{ARIMA}(p, d, q)(P, D, Q)_m \), where the lower- and uppercase “p, d, q” parameters are non-negative integers defining the order of the autoregressive \( p \), differencing \( d \), and moving average \( q \) components of the non-seasonal part (lowercase) and seasonal part (uppercase), respectively. The span of the period is given by \( m \).

Since ARIMA models are defined over all reals and we require that the trend is non-negative, we subtract the simulated trend \( \delta_t \) from \( \delta_t \) to get \( \hat{\delta}_t \) and define \( \eta \) constrained to \( \hat{\delta}_t \)

\[
\hat{\delta}_t = \delta_t - \delta_t.
\]

Moreover, schedule information is available to generate multiple path alternatives for every OD-pair. Path choice is based on the multinomial logit model and considers the utility (i.e., generalized travel cost) of each path alternatives and generates the probabilities within a set of alternatives. Here, the alternatives are the path alternatives between the origins and destinations of the passengers. The choice between the path alternatives is generated from a categorical choice model, according to

\[
P(k|p) = p_k, \quad k \in \{1, 2, \ldots, K\}, \quad p = (p_1, p_2, \ldots, p_K).
\] (E.10)

**E.3 The multinomial logit path choice model.** Passengers choose between different path alternatives when travelling between an origin and a destination. The travel time to reach their destination varies and thus the time bin when the corresponding exit counts are incremented varies, depending on the selected path alternative. The path choice model is based on a network with a known link structure. Moreover, schedule information is available to generate multiple path alternatives for each OD-pair. Path choice is based on a multinomial logit model and considers the utility (i.e., generalized travel cost) of each path alternatives and generates the probabilities of choosing each alternative in the choice set.

The multinomial logit model (McFadden, 1974) is based on random utility choice models, and aims to determine the choice probabilities within a set of alternatives. Here, the alternatives are the path alternatives between the origins and destinations of the transit network. The utility of each path alternative is equal to the weighted sum of attributes \( z_k \) with fixed attribute coefficients \( \xi \). The attributes \( z_k = (z_{k,1}, z_{k,2}, z_{k,3}, z_{k,4}) \) of path alternative \( k \) are the total in-vehicle time \( z_{k,1} \), total wait time \( z_{k,2} \), total transfer time \( z_{k,3} \), and number of transfers \( z_{k,4} \). The probability of choosing a particular path \( k \) with utility \( V_k \) from a set of \( K \) possible path alternatives is given by \(^3\)

\[
p_k = \frac{e^{V_k}}{\sum_{i=1}^{K} e^{V_i}},
\] (E.8)

with

\[
V_i = \xi^T z_i.
\] (E.9)

The choice between the \( K \) alternatives is generated from a categorical choice model, according to

\[
P(k|p) = p_k, \quad k \in \{1, 2, \ldots, K\}, \quad p = (p_1, p_2, \ldots, p_K).
\] (E.10)

**E.4 Test data generation.** The average-delay test case is based on a subcomponent of the NYC subway network. The chosen subcomponent is shown in Fig. E.1. The subcomponent network consists of 35 stations, 8 lines, and 39 transfer linkways (i.e., footpaths). Out of the 35 stations, we pick \( S = 15 \) stations for which we generate \( N \) entry count time series, following the generative process described earlier. The departure window starts at \( t_0 = 7 \) AM, ends at 9 AM, and is discretized into 5-minute time bins (i.e., \( T = 24 \)).

After generating the OD-matrix according to Eq. (E.1), any coefficients corresponding to station pairs connected by a direct transfer path are set to zero. The corresponding OD-matrix row is re-normalized to ensure the simplex constraint is satisfied.

Entry count series are simulated for each of the 15 stations over a fixed-length window and at pre-defined time intervals, according to the hybrid ARIMA-INGARCH process in Section E.2. The trend covariate \( F_t \) (c.f., Eq. (E.7)) is generated once for each station; Next, we generate \( N \) count series for each station, according to the INGARCH model in Eq. (E.5). The intercept \( \nu_0 \) is fixed at \( 1 \). The observation and mean regression parameter values are \( \nu_1 = 0.1 \) and \( \nu_2 = 0.05 \), and \( \xi_1 = 0.1 \) and \( \xi_2 = 0.05 \), respectively. The trend is simulated from an \( \text{ARIMA}(1, 1, 1)(1, 1, 1)_T \) model, where \( T \) is the number of simulated time bins and corresponds to the number of time bins contained in the departure window defined in Section 2.3. The coefficients of the autoregressive and moving average components are set to \( 0.8 \) and \( 0.9 \) for the non-seasonal part, and \( 0.5 \) and \(-0.5 \) for the seasonal part.

The adjustable parameters are the trend scale \( \eta \) and the count dispersion \( \varphi \). Fig. E.2 shows examples of the entry count time series mean and percentile range for different \( \eta \) and \( \varphi \) parameter values. In this example, entry count series are simulated \( N = 100 \) times over

---

\(^3\) In order to avoid numerical overflow due to the exponential terms in Eq. (E.8), we use an arithmetic reformulation using the log-sum of exponential terms. By first determining the maximum utility \( V_{\text{max}} \) the expression in Eq. (E.8) can be rewritten to \( p_k = \exp(V_k - (V_{\text{max}} + \log(\sum \exp(V_i - V_{\text{max}}]))) \).
an assumed 2 h window, segmented into 5-minute intervals, resulting in \( T = 24 \) time bins. We simulate \( T + 1 \) time bins and truncate the first sample to avoid a fixed starting value.

Path alternatives are generated based on the schedule information published by the Metropolitan Transport Authority (MTA) of NYC (Metropolitan Transport Authority, 2019b), whereby all path alternatives that fall within a 10-minute margin from the earliest arrival path are considered. The schedule information is available in the General Transit Feed Specification (GTFS) format and parsed into queryable data structures. We fix the proxy departure time to 8 AM — i.e., any passenger entering the system during the 7 AM to 9 AM window will see the same path alternatives corresponding to a departure time of 8 AM. The path alternatives provide all necessary route attribute information required for the multinomial logit model in Eq. (E.8), with attribute coefficients fixed to \( \xi = (-0.002, -0.006, -0.002, -1.0)^T \).

The generative process simulation proceeds to increment the exit counts according to the entry counts at each origin, the departure time bins, given destination, and path choice of every passenger entering the system. The arrival window starts at \( t_1 = 8\) AM. The maximum travel time through the test network is about 40 min. The start time of the departure time window at \( t_0 = 7\) AM thus provides sufficient delay to capture passengers who are arriving at their destination station at exactly 8 AM and later.

Once \( N \) entry count time series at each of the \( S \) stations are simulated, the generative model iterates over every station, observation, time bin, and entry count. For every entry count increment, a passenger chooses a destination according to the categorical choice model in Eq. (D.6), where the choice probabilities are given by the OD-matrix coefficients generated from Eq. (E.1). Next, the passenger chooses a path alternative according to the choice model in Eq. (E.10). The departure time bin, chosen path, and corresponding travel time decide when the traveller arrives at their destination. If the arrival time bin is within the arrival window, the exit counts during that respective arrival time bin are incremented. This process repeats until all entry counts are processed.

Appendix F. Raw data processing and filtering

The turnstile count records of the NYC subway system contain entry and exit counts in 4 h intervals and erroneous measurements. Therefore, the raw count records of the number of passengers accessing and egressing at stations must be cleaned and processed before they can be used as input data in the OD-inference model. First, they are up-sampled and aggregated over the desired observation windows. Despite this pre-processing of the data, the count records are still distorted by noise, missing data, and outliers. Thus, in a second step, the data are filtered to impute new estimates for missing and corrupted data points.

F.1 Count up-sampling. The first step in treating the data records is to up-sample the entry and exit counts to the desired recording intervals. A critical property is to up-sample the counts conservatively in order to preserve the total number of counts over the original 4 h intervals. We do this by monotonically interpolating the cumulative counts at points of the desired sampling frequency and differencing the interpolated values to obtain the up-sampled counts. This way, when integrating over the up-sampled counts, the original cumulative counts can be recovered. The interpolated values are, however, not integer count values, but positive reals. The interpolator of the cumulative counts thus needs to be monotonically increasing. Additionally, the interpolator needs to be \( C^2 \)-continuous to ensure smoothness of the differentiated function values. Rational Quadratic Spline interpolation (Delbourgo and Gregory, 1983) is such an interpolator. We implement the interpolation routines into a custom up-sampling method, that finds contiguous segments of count records, monotonically interpolates their cumulative counts, and up-samples the segments to the desired frequency. Here, we up-sample the counts to 5 min intervals and aggregate them over the respective observation windows before applying the instantaneous-balance inference model – up-sampling to 1 h intervals would give the same resulting counts; the 5 min interval count data, however, allow for a more flexible use in testing the average-delay inference model.

F.2 Filtering and missing data imputation. Once the up-sampled data are aggregated over the desired observation window, they are still interspersed with missing and corrupted data, requiring a filtering procedure. The filtered data are obtained from a de-noising autoencoder neural network implemented using the tensorflow package for Python. The unsupervised learning algorithm aims to minimise the target loss, defined as the sum of the reconstruction loss and a regularization loss. The reconstruction loss measures the mean squared error between the (normalized) count observations and the neural network predictions. The regularization loss is defined by an \( L_2 \)-regularisation term over the hidden layers’ weights to prevent overfitting. The regularisation parameter is set to 0.0001.

The input layer includes observation neurons that feed in count observations, as well as masking neurons that feed in a Boolean mask to identify missing data and dynamically changing outliers. The number of observation and masking neurons depends on the number of stations and a specified sub-frame width. The neural network has three hidden layers. The input layer is connected to the first hidden layer via a drop-out layer with a 30% de-noising drop-out rate. The first hidden layer uses a rectified linear activation function. The output layer has the same number of neurons as the input layer. We choose 150 neurons for the first and third hidden layer, and 75 neurons for the second hidden layer.

Training the autoencoder neural network requires a defined number of epochs. Every epoch consists of multiple iterations to train the neural network. Every iteration draws a new batch of multiple training samples of contiguous records with specified sub-frame width \( M \). The sub-frame width defines the number of contiguous records in every sample. The samples are drawn at random from the full set of recorded data. The batches of training samples are fed into the neural network and are processed forward and backward to train the neural network while minimizing the loss due to reconstruction error and weighing in the regularization error. Every processed batch completes one iteration.

Once all iterations have been completed, the neural network is tested by evaluating its predictions versus the observed count records. Prior to testing, the predictions are filtered through a moving average window from the first to the last record. The width of the moving average window is chosen to be the same as the sub-frame width \( M \). These filtered predictions are then used to determine the
squared error for non-missing data. We use the squared error to determine the upper percentile of observations that qualify to be outliers, based on an estimated minimum fraction of outliers in the data. The estimated fraction of outliers is set to 5%. To correct for potentially having misidentified outliers, an outlier drop-out rate is fixed at 0.1 to randomly re-qualify data points as valid. According to the identified outliers, the outliers mask is changed to reflect whether a data point is flagged as an outlier or not. Next, outliers are corrected by a weighted average of the observation value and the prediction. The weights are set to 0.5, effectively determining the average between the observation and prediction. In addition, missing data points are imputed with a weighted average of the previous corrected values for the outliers and missing data.

The batch size $B$ is specified according to user input. The total number of records is denoted as $N_{\text{tr}}$. The number of iterations per epoch is defined by the sub-frame width and batch size, according to $(N_{\text{tr}} - M)/B$, rounded to the nearest integer. Thus, during every epoch, the neural network is trained over $(N_{\text{tr}} - M)/B$ iterations. Each iteration draws one batch consisting of $B$ samples of count records. Every sample consists of $M$ contiguous records. In this case, we choose $M = 50$ records and a batch size of $B = 15$ samples for every batch. The number of input neurons is equal to $2S \times M$, where $S$ is the number of stations. The number of epochs is set to 100.

Fig. F.1 shows examples of the processed time series of the number of passengers who entered and exited at selected stations within the 8:00-to-9:00 AM window from 2010 to 2018. The grey curves are the pre-processed time series. The black curves are the filtered time series, determined from the autoencoder neural network.

**Appendix G. Effect of model regularization**

The instantaneous-balance Bayesian inference model and QP optimisation model are run on test samples with either 5, 10, 30, or 100 observations derived according to the generative process for test network A in Appendix D. Next, the inferred OD-coefficient estimates are used to predict the exit counts for out-of-sample validation data according to the uncorrected instantaneous-balance model formulation $Y = XA$. The validation data consisting of 10,000 out-of-sample observations are denoted as $y$ and derived from the same generative process as the sampled observations for test network A. However, they are not used in the estimation of the OD-coefficients. The Bayesian model predictions $\hat{\mu}_y$ are determined by inserting the OD-coefficient posterior means into $A$; The QP model predictions $\hat{y}$ are determined by inserting the OD-coefficient point estimates. This procedure is carried out twice; Once, by determining OD-coefficient estimates according to the regularized formulations in Eq. (3) for the Bayesian inference model and Eq. (A.1) for the QP inference model; And again, by using the otherwise same formulations except that the regularization terms are omitted. Fig. G.1 shows that the MSE of the MCMC and of the QP estimate decreases as the number of observations used for the OD-estimation increases, given that more observations provide more evidence of the underlying, latent OD-matrix. Furthermore, Fig. G.1 confirms that the regularized solution outperforms the non-regularized solution when tested against validation data.

**Appendix H. Sensitivity analysis**

We assess the correlation between accuracy and precision of the instantaneous-balance and average-delay estimates, and their sensitivity to the observation window width $w$, the trend scale $\eta$ and the count dispersion $\varphi$ using a metamodel/emulator, which describes the relation between each statistic (i.e., the response variable) and these parameters (i.e., the regressors) according to the linear model

$$q = \beta_0 + \beta_1 w + \beta_2 \eta + \beta_3 \varphi + \epsilon,$$  

(H.1)

where $q$ is a placeholder for either the accuracy or precision statistic (i.e., either $MSE(\mu_0, \alpha_0)$ or $\hat{\mu}_{\text{HPD}}$), $w$ denotes the observation window width, and $\eta$ and $\varphi$ are the trend scale and Negative Binomial dispersion. We fit the model via the MinMax-normalized regressors and by determining the least-squares estimate of the coefficients $\hat{\beta}$. In this way, we construct four metamodels for the accuracy and precision of the instantaneous-balance and average-delay model. For each metamodel, we test for the statistical significance of its coefficients and determine the first-order and total Sobol’ sensitivity indices. The Sobol’ indices are based on a Monte Carlo estimate with 10,000 samples from a uniform distribution for each parameter and 1000 bootstrap replicates. All corresponding statistics regarding the correlation metrics, metamodel parameters, and Sobol’ indices are collected in Table H.1.

Table H.1 confirms that increasing the window width improves significantly both models’ accuracy and precision. Indeed, accuracy and precision are negatively correlated with the window width, i.e., the larger the window width, the smaller the estimates’ MSE and mean HPD. P-values indicate high significance of the window width regressor and Sobol’ indices indicate that the accuracy exhibits large first-order sensitivities, i.e., large variance of the conditional expectation of the accuracy with respect to the window width and relative to the total variance, and little indication of any higher-order interactions, i.e., total indices are similar to first-order indices. The underlying assumption for both the generative and inferential model is that the OD-matrix stays constant throughout the observation window. Consequently, increasing the window width will increase the number of time bins contained within the observation window and, as a result, will add more data points and thus evidence for the average-delay model to infer the OD-coefficients. Moreover, as the observation window becomes wider, a larger proportion of passenger trips occur within the window itself, which complies with the instantaneous-balance assumption and, thus, improves the instantaneous-balance model’s accuracy and precision.

Table H.1 also confirms that the average-delay model does not exhibit significant sensitivity to the trend scale. Because the average-delay model explicitly includes the travel time between stations, any behaviour in the trend of the entry counts at one station is mapped to the exit counts by a lagged time bin at another station. Since the trend stays unchanged for all observations, this mapping will be
repetitive and not contribute any new information to improve the inference, whether or not the trend strongly persists. In contrast, the instantaneous-balance model is influenced by the strength of the time series trend, and both the accuracy and precision deteriorate as $\eta$ increases. The instantaneous-balance model ignores the travel time between stations and expectedly will perform best if steady conditions persevere over the observation window. However, the stronger the trend, the less the steady-state balance of the entry flow and exit flow is maintained. As a result, time aggregation over the limited window will induce larger errors to the instantaneous-balance assumption, which consequently degrades the accuracy and precision of the instantaneous-balance model.

Finally, the accuracy and precision of the OD-coefficient estimates improve as the count observations become more dispersed. This is in line with the analysis of the instantaneous-balance model in Section 4.1. Indeed, both the average-delay and the instantaneous-balance model rely on the inherent variability of the entry and exit count observations. The more pronounced the exit-entry count patterns change, the stronger they correlate between the exit and entry count observations, the stronger the evidence is of a possible linkage between particular OD-pairs. Under the assumption that the latent OD-matrix does not change, a stronger dispersion in the entry counts will propagate into a stronger dispersion of the exit counts. As a result, inferring the associated OD-coefficients will become easier.

Appendix I. Critical parameter combinations

The best- and worst-cases reported in Fig. 7, for $N = 100$ number of observations, identify the critical combinations of test parameters that produce the smallest and largest estimation errors and uncertainties. Both the instantaneous-balance and average-delay model exhibit the largest error and uncertainty for an observation window with $w = 5$ min, trend scale $\eta = 1$, and dispersion $\varphi = 1000$. For this worst-case condition, both models equally struggle to identify the true OD-coefficients. The corresponding estimates are visualized in Fig. I.1a. In contrast, the instantaneous-balance model returns the best estimates for $w = 1$ h, $\eta = 0$, and $\varphi = 10$, shown in Fig. I.1b. Fig. I.1c shows the results subject to the average-delay model’s best-case conditions with $w = 1$ h, $\eta = 1$, and $\varphi = 10$. This confirms previous analyses in Appendix H, that establishes that longer observation windows and larger count dispersion improve the accuracy and precision of the estimates. Moreover, it supports the reasoning that the instantaneous-balance model favours flat trends, whereas the average-delay model is weakly impacted by the trend scale.

Appendix J. OD-flow in the NYC subway system

Fig. J.1 shows an estimate of the absolute passenger demand between all OD-pairs, given the OD-coefficient posterior means and the average weekday passenger access counts between May 2nd and May 13th, 2016. The plotted values in Fig. J.1 are computed from \( \text{diag}(\mathbf{X}) \mathbf{A} \), where \( \mathbf{X} \) is a column vector of length \( S \), containing the average weekday passenger access counts for either the morning peak, midday off-peak, or evening peak window; \( \text{diag}(\mathbf{X}) \) is a diagonal matrix with the elements of \( \mathbf{X} \) on the diagonal; and \( \mathbf{A}_{S \times S} \) is the OD-matrix consisting of the OD-coefficient posterior means.

The three OD-matrices in Fig. J.1 display vertical bands of OD-coefficients. These bands are marked by the vertical arrows at the bottom of each plot and correspond to the destination stations that consistently observe more than 3000 exiting passengers per hour across all three observation windows and, thus, persistently attract passenger trips throughout the day and from throughout the network. However, the patterns of dominant bands are visually more dissolved in Fig. J.1c as compared to Fig. J.1a. This is because a group of only few stations attract a large number of passenger trips from stations all across the network during morning peak demand in Fig. J.1a. By the afternoon off-peak in Fig. J.1b, this dedicated OD-demand gradually reduces and spreads to more stations in the network, that is, the vertical bands become more diffused. During the evening peak in Fig. J.1c, the dedicated OD-demand vanishes and numerous stations throughout the network attract passenger trips from across the network.

Appendix K. Effective sample size

As much as Markov chains provide desirable properties to be able to sample from the posterior distribution, their inherent stochastic dependence causes their samples to be autocorrelated. An artefact of this correlation is that the standard errors and uncertainties of parameter estimates increase. The effective sample size \( N_{\text{eff}} \) can be used to measure the degree to which the estimates are influenced by the samples’ autocorrelation (Geyer et al., 2011). Here, we report the ratio with respect to the autocorrelated samples \( N_{\text{MC}} \) – the closer this ratio is to 1.0, the more the estimates’ uncertainty can be expected to not stem from the samples’ inherent autocorrelation. Since the autocorrelation for a chain with multiple parameters and joint probability function cannot be directly calculated (the joint probability function is the posterior density of interest), Stan (Carpenter et al., 2017) estimates the autocorrelation and thus effective sample size from a Fast Fourier Transform (FFT) procedure and combines the estimates from multiple chains. In the case of the NYC OD-estimation results, we use a thinned sample of \( N_{\text{MC}} = 1000 \) MCMC draws for the estimation of the effective sample size.

---

4 These stations are: Grand Central - 42 St (lines 4, 5, and 6), 34 St - Herald Sq (lines B and D), 34 St – Penn Station (lines A, C, and E), 34 St – Penn Station (lines 1, 2, and 3), and 14 St - Union Square (lines N, Q, R, and W)


