IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 9, 2022, accepted February 27, 2022, date of publication March 2, 2022, date of current version March 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156078

Point-Process Modeling and Divergence Measures Applied to the Characterization of Passenger Flow Patterns of a Metro System

GABRIEL VIDAL[1], JUAN I. YUZ[1,2], (Member, IEEE), RONNY VALLEJOS[1,3], AND FELIPE OSORIO[3]

[1]Advanced Center for Electrical and Electronic Engineering, AC3E, Valparaíso 2390123, Chile
[2]Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso 2390123, Chile
[3]Department of Mathematics, Universidad Técnica Federico Santa María, Valparaíso 2390123, Chile

Corresponding author: Juan I. Yuz (juan.yuz@usm.cl)

This work was supported in part by the Agencia Nacional de Investigación y Desarrollo (ANID) (Chile) through the Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) under Grant 1181090 and Grant AC3E FB0008, and in part by MATH-AMSUD under Grant 20-MATH-03.

ABSTRACT The problem of characterizing the passengers' movement in a public transport system has been considered in the literature for analysis, simulation and optimization purposes. In particular, origin-destination matrices are commonly used to describe the total number of passengers that travel between two points during a given time interval. In this paper, we propose to model the instantaneous rate of arrival of passengers for the origin-destination pairs of a metro system using point processes. More specifically, we apply the Expectation-Maximization algorithm to estimate the parameters of a Gaussian mixture intensity function for the daily flow of passengers using data from multiple days provided by EFE Valparaíso. The uncertainty in the parameter estimates is quantified computing standard errors and confidence intervals. Secondly, we quantitatively analyze the similarity of the obtained intensity functions among the different origin-destination pairs. In particular, we propose a dissimilarity index based on the Kullback-Leibler divergence and we apply this index in hierarchical agglomerative and partitioning methods to cluster origin-destination pairs with similar daily flow of passengers. The obtained numerical results confirm expert knowledge about the passengers' behavior in EFE Valparaíso metro system and, more interestingly, provide additional insights on the passengers' behaviour for specific origin-destination pairs.

INDEX TERMS Expectation-maximization, cluster analysis, Kullback-Leibler divergence, public transport.

I. INTRODUCTION

Efficient and reliable transport systems have become increasingly important to support the large urban and suburban daily flow of passengers in large cities all over the world. In particular, metro systems are one of the most commonly used means of transport due to their cost for passengers, effectiveness and speed. At the end of 2017, there were metros in 182 cities in 56 countries carrying on average a total of 168 million passengers per day [37].

The modeling of the passengers flow in a metro system may be of interest to reduce operating costs, to improve service quality and for long-term investment in the metro system infrastructure [28]. The passengers flow is usually modeled using origin-destination (OD) matrices that gather the information of the flow of passengers between different pairs of stations of the line. Transit agencies in the past used to rely on survey to estimate these matrices, however, nowadays database management and geographic information systems provide large amounts of data that can be exploited [42]. In fact, payment systems are the main source of information for passengers movement, where some transport lines only validate payment cards at the passenger entry, and in others, payment cards are validated at both ends of the trip. Boarding and alighting data have been used with, for example, the iterative proportional fitting method, which could be considered the state-of-the-art practical OD estimation method [14]. However, this kind of methods generally describe the total number of passengers traveling in a given period (hours, days, weeks or years) for a given OD pair, and not the variation within that time interval.

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

Poisson processes have also been used to model transport systems. For example, in [22] short-term demand for bus arrival was predicted by describing an intensity using neural networks, and in [23], the demand for taxis was modeled using a discrete intensity. Moreover, Poisson processes have been used to model patient arrival to emergency departments using a piecewise constant approximation of intensity [9]. The modeling of the intensity of the passengers flow is not always possible and, in such cases, researchers have proposed methods such as the short-term forecasting of real-time OD matrices [5], [43]. For public transport, Munizaga and Palma [24] estimated an OD matrix combining bus and metro system data.

In this paper, we propose a point process model for the instantaneous rate of arrival of passengers for every OD station pairs of a metro network during one day of operation. The modeling of passengers arrival was previously considered in Allende *et al.* [1], where three different point process models were considered: Hawkes-Phan process, Poisson process with Gaussian mixtures, and a novel Hawkes-Gauss process (that combined the previous two). In that work, the intensity function was obtained applying maximum likelihood and the EM algorithm, however, using only single-day real data provided by EFE Valparaíso. Additionally, Bayesian information criterion was used in [1] to determine the (best) model order and, thus, the number of parameters for the different point processes intensity functions.

In this paper, our first contribution is to continue and extend the work in [1] by estimating the intensity function of the passengers arrival using data from multiple days using the superposition principle [17]. We consider the nonhomogeneous Poisson processes (also considered in [1]) by defining a time-dependent intensity function $\lambda(t)$ as a Gaussian mixture model. To estimate the parameters of the intensity Gaussian mixture function for each OD pair, we use the EM algorithm [8], [21]. This algorithm has been applied to truncated Hawkes point process modeling in [11].

For estimation from real data, the accuracy analysis is an important issue to detect high uncertainty in the parameter estimates. In this work, the accuracy of the parameter estimates is analyzed via confidence intervals and standard errors. The Oakes method [25] is applied to obtain the derivatives of the conditional expected function of the EM algorithm to obtain the observed information matrix that leads to estimating confidence intervals. In addition, parametric bootstrap [10] based on a Monte Carlo simulation is used to approximate standard errors. Moreover, we analyze the impact of combining data from multiple days to reduce the associated uncertainty in the parameter estimates of the intensity functions. It is worth noticing that the derivation of the observed information matrix using the Oakes method is a novel side result, which, to the best of our knowledge, has not been reported previously for Gaussian mixture models.

In the literature, different works have considered the characterization of passengers flow pattern in transport systems, focusing on the study on clustering days with similar behavior of user arrivals. For example, Weijermars and Berkum [36] classified daily traffic profiles using hierarchical Ward's clustering for a Dutch highway. Yang *et al.* [40] applied dimensionality reduction to OD matrices to classify a set of days of a metro system using affinity propagation. For daily traffic data, Yu and Hellendoorn [41] proposed a clustering algorithm for mixture models. Caceres *et al.* [3] applied hierarchical clustering to Euclidean distances of road section features to estimate traffic flow.

In this paper, our second contribution is a methodology to characterize OD pairs with similar passengers flow patterns using a quantitative index to compare the associated intensity functions. In particular, we propose a dissimilarity index between OD pairs using a symmetrized Kullback-Leibler divergence [6]. In this way, a dissimilarity matrix between all OD pairs of Metro Valparaíso is obtained. To characterize similar passengers flow patterns, we use two clustering algorithms: agglomerative hierarchical methods using single, complete and average linkage [15] and partitional methods using the k-medoids algorithm [16]. A key problem of clustering algorithms in practice is the validation of the obtained set of clusters [27], that is dependent on the context of the application. For the metro system, the choice of the number of clusters may be considered as a design parameter depending on the number of groups required for analysis, simulation or optimization purposes.

The remainder of the paper is structured as follows: in Section II, the parameter estimation of the intensity function is presented, and the associated confidence intervals and standard errors are obtained. Section III provides background on the Kullback-Leibler divergence and how it can be used to define a symmetric distance between OD intensity functions. Then, in Section IV, clustering methods and algorithms are presented. In Section V we present the numerical results obtained by the proposed analysis using real data provided by EFE Valparaíso. Finally, in Section VI, we discuss the obtained results and draw conclusions.

II. ESTIMATION OF INTENSITY FUNCTIONS

Origin-destination (OD) matrices are used to model the movement of passengers in a public transport system. They usually summarize the total number of passengers that travel between given origin and destination points (or stations) during a given period of time. For instance, for a set of t stations, the matrix

$$\begin{pmatrix} 0 & N_{12} & \cdots & N_{1t} \\ N_{21} & 0 & \cdots & N_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ N_{t1} & N_{t2} & \cdots & 0 \end{pmatrix},$$

is an OD matrix where N_{ij} represents the number of trips made from station i to station j over a given period of time. Note that it is assumed that no trips are allowed having the same origin and destination station.

In this section, we introduce point processes intensity functions to model the instantaneous rate of arrival of passengers

for each OD pair of a metro system during one day of operation. We assume that the intensity associated with an OD pair is denoted by $\lambda(t)$ and corresponds to the intensity of a Poisson point process as described in [1], [31].

Definition 1: A mixture of one-dimensional Gaussian densities is a finite linear combination of g Gaussian densities of the form

$$f(y;\boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \cdot \phi(y;\mu_i,\sigma_i^2), \tag{1}$$

where

$$\phi(y;\mu_i,\sigma_i^2) = (2\pi\sigma_i^2)^{-1/2}\exp\{-\tfrac{1}{2}(y-\mu_i)^2/\sigma_i^2\}, \tag{2}$$

denotes the Gaussian density function with mean μ_i and variance σ_i^2, and π_i is the weight associated with each Gaussian component. If a random variable Y has a density function (2), we shall denote $Y \sim \mathsf{N}(\mu_i,\sigma_i^2)$.

In (1), the parameter vector associated to the Gaussian mixture model is $\boldsymbol{\theta} = (\boldsymbol{\pi}^\top,\boldsymbol{\mu}^\top,\boldsymbol{\sigma}^\top)^\top$ with $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_{g-1})^\top$, $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_g)^\top$ and $\boldsymbol{\sigma} = (\sigma_1^2,\ldots,\sigma_g^2)^\top$. Accordingly, the parameter space is defined as

$$\begin{aligned}\Theta_g = \Big\{&(\boldsymbol{\pi}^\top,\boldsymbol{\mu}^\top,\boldsymbol{\sigma}^\top)^\top \in \mathbb{R}^p : \\ &\pi_g = 1 - \sum_{j=1}^{g-1}\pi_j,\ \pi_i > 0,\ \mu_i \in \mathbb{R},\ \sigma_i^2 > 0, \\ &\forall\, i = 1,\ldots,g\Big\},\end{aligned} \tag{3}$$

where $p = 3g-1$ is the dimension of Θ_g.

Following [1], for a fixed day d, we consider (1) to model the intensity of a point process for an OD pair such that the Gaussian mixture of densities is weighted by the number of arrivals n_d to the destination. In this framework, we define the intensity through

$$\lambda_d(t;\boldsymbol{\theta}) = n_d \sum_{i=1}^{g}\pi_i\,\phi(t;\mu_i,\sigma_i^2). \tag{4}$$

The above model is an extension of the problem addressed in [1] in which the parameter vector $\boldsymbol{\theta}$ was estimated using data from one day only. Here, we consider data from a set of days where, for each day, we assume that the intensity function of the Poisson point process depends on the same fixed parameter vector $\boldsymbol{\theta}$. Furthermore, if we assume that the Poisson processes for different days are independent of each other, then from the superposition principle [17],

$$\lambda(t;\boldsymbol{\theta}) = \frac{1}{N_d}\sum_{d}\lambda_d(t;\boldsymbol{\theta}) = \bar{n}_d \sum_{i=1}^{g}\pi_i\,\phi(t;\mu_i,\sigma_i^2) \tag{5}$$

is the intensity of the realizations of the Poisson point processes of the selected N_d days and $\bar{n}_d$ is the average number of daily arrivals.

In the framework of discrete mixtures of Gaussian distributions, the EM algorithm [8] is often used to obtain the maximum likelihood (ML) estimates of the parameters of interest. The simplicity and stability of this estimation approach has allowed it to become a popular algorithm (see, for instance, [39]).

The power of the EM algorithm lies in considering a data augmentation scheme $\boldsymbol{Y}_{\mathsf{aug}}$ by including latent variables or missing data to the observed data $\boldsymbol{Y}_{\mathsf{obs}} = \mathcal{M}(\boldsymbol{Y}_{\mathsf{aug}})$ for some many-to-one mapping $\mathcal{M}$.

Then, we proceed to maximize the observed log-likelihood function $\ell(\boldsymbol{\theta})$ iteratively based on a surrogate function known as the log-likelihood function of augmented data $\boldsymbol{Y}_{\mathsf{aug}} = (\boldsymbol{Y}_{\mathsf{obs}}^\top,\boldsymbol{Y}_{\mathsf{mis}}^\top)^\top$ denoted by $\ell_c(\boldsymbol{\theta})$.

The EM procedure iteratively computes the ML estimator by alternating between the following steps:

- **E-step:** For a current estimation $\boldsymbol{\theta}^{(k)}$, compute the conditional expectation
$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = \mathsf{E}\{\ell_c(\boldsymbol{\theta})|\boldsymbol{y}_{\mathsf{obs}},\boldsymbol{\theta}^{(k)}\}.$$
- **M-step:** Update $\boldsymbol{\theta}^{(k+1)}$ by maximizing $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ as a function of $\boldsymbol{\theta}$.

Under mild general conditions [38], the EM algorithm increases the observed data log-likelihood function after each iteration, and the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 1}$ converges to a stationary point of $\ell(\boldsymbol{\theta})$.

To obtain the maximum likelihood estimates in the context of discrete mixtures of normal, we augment the observed data $\boldsymbol{Y}_{\mathsf{obs}} = (Y_1,\ldots,Y_n)^\top$ by incorporating latent variables to obtain $\boldsymbol{Y}_{\mathsf{aug}} = (Y_1,\ldots,Y_n,\boldsymbol{Z}_1^\top,\ldots,\boldsymbol{Z}_n^\top)^\top$, where $Z_{ij} = (\boldsymbol{Z}_j)_i$ is an indicator variable that identifies if the observation Y_j belongs to the ith component of the mixture. This leads to the following hierarchical model:

$$Y_j|z_{ij} = 1 \sim \mathsf{N}(\mu_i,\sigma_i^2), \quad \boldsymbol{Z}_j \sim \mathsf{Mult}_g(1,\boldsymbol{\pi}), \tag{6}$$

where $\mathsf{Mult}_g(1,\boldsymbol{\pi})$ denotes the multinomial distribution with parameters 1 and probabilities $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)^\top$. Thus, the augmented-data log-likelihood function is given by

$$\begin{aligned}\ell_c(\boldsymbol{\theta}) &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\log\left(\pi_i\phi(y_j;\mu_i,\sigma_i^2)\right) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\log\pi_i \\ &\quad + \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\Big\{-\frac{1}{2}\log 2\pi\sigma_i^2 - \frac{1}{2\sigma_i^2}(y_j-\mu_i)^2\Big\}.\end{aligned}$$

It is straightforward to show that the conditional expectation required to evaluate the E-step of the EM algorithm is given by

$$z_{ij}^{(k)} = \mathsf{E}(Z_{ij}|\boldsymbol{Y};\boldsymbol{\theta}^{(k)}) = \frac{\pi_i^{(k)}\phi(y_j;\mu_i^{(k)},\sigma_i^{2(k)})}{\sum_{r=1}^{g}\pi_r^{(k)}\phi(y_j;\mu_r^{(k)},\sigma_r^{2(k)})}, \tag{7}$$

for $j = 1,\ldots,n$; $i = 1,\ldots,g$. This allows us to obtain the conditional expectation of the log-likelihood of the augmented data. Moreover, $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ is given by

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = Q_1(\boldsymbol{\pi};\boldsymbol{\theta}^{(k)}) + Q_2(\boldsymbol{\mu},\boldsymbol{\sigma};\boldsymbol{\theta}^{(k)}),$$

where

$$Q_1(\boldsymbol{\pi};\boldsymbol{\theta}^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}^{(k)} \log \pi_i,$$

$$Q_2(\boldsymbol{\mu},\boldsymbol{\sigma};\boldsymbol{\theta}^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}^{(k)} \Big\{ -\frac{1}{2}\log 2\pi\sigma_i^2 - \frac{1}{2\sigma_i^2}(y_j-\mu_i)^2 \Big\}.$$

Thus, the M-step related to $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are given by

$$\pi_i^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n} z_{ij}^{(k)}, \tag{8}$$

$$\mu_i^{(k+1)} = \frac{1}{\sum_{j=1}^{n} z_{ij}^{(k)}} \sum_{j=1}^{n} z_{ij}^{(k)} y_j, \tag{9}$$

$$\sigma_i^{2(k+1)} = \frac{1}{\sum_{j=1}^{n} z_{ij}^{(k)}} \sum_{i=1}^{n} z_{ij}^{(k)} (y_j-\mu_i^{(k)})^2, \tag{10}$$

for $i = 1, \ldots, g$.

In the algorithm, the E- and M-steps in (7) and (8)-(10), respectively, are iterated until a convergence threshold is achieved.

Note that the updates of $\mu_i^{(k+1)}$ and $\sigma_i^{2(k+1)}$ correspond to weighted averages whose weights are given by $z_{ij}^{(k)}$. This is computationally inexpensive and guarantees non-negative variance estimates.

Under mild regularity conditions, the ML estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is asymptotically normally distributed, that is,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} \mathsf{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})),$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathsf{E}\{-\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top\}$ corresponds to the Fisher information matrix. Note that the Fisher information matrix is necessary not only to obtain the standard error of ML estimators but also to evaluate confidence intervals and hypothesis test statistics such as Wald and score statistics [2].

The following alternatives for estimating the Fisher information matrix have been frequently suggested:

$$\boldsymbol{I}_{\mathsf{o}}(\boldsymbol{\theta}) = -\frac{1}{n}\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}, \quad \bar{\boldsymbol{I}}_{\mathsf{o}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\theta})\boldsymbol{U}_i^\top(\boldsymbol{\theta}),$$

which correspond to observed and empirical versions of the information matrix, respectively. Here $\boldsymbol{U}_i(\boldsymbol{\theta})$ represents the score function for a single observation, which is common when the likelihood function is additive.

Another alternative to derive the observed information matrix uses the missing information principle [19].

In this paper, we obtain the observed information matrix in the Gaussian mixture model using the method proposed by Oakes [25], i.e., we compute:

$$-\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = -\Big\{\frac{\partial^2 Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} + \frac{\partial^2 Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\widehat{\boldsymbol{\theta}}^\top}\Big\}\Big|_{\theta=\widehat{\theta}}.$$

where matrices $\ddot{\boldsymbol{Q}}_{\theta\theta}(\boldsymbol{\theta}) = \partial^2 Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ and $\ddot{\boldsymbol{Q}}_{\theta\widehat{\theta}}(\boldsymbol{\theta}) = \partial^2 Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ have the following structure:

$$\ddot{\boldsymbol{Q}}_{\theta\theta}(\boldsymbol{\theta}) = \begin{pmatrix} \ddot{\boldsymbol{Q}}_{\mu\mu}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\mu\sigma}(\boldsymbol{\theta}) & \mathbf{0} \\ \ddot{\boldsymbol{Q}}_{\sigma\mu}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\sigma\sigma}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddot{\boldsymbol{Q}}_{\pi\pi}(\boldsymbol{\theta}) \end{pmatrix}, \tag{11}$$

$$\ddot{\boldsymbol{Q}}_{\theta\widehat{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \ddot{\boldsymbol{Q}}_{\mu\widehat{\mu}}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\mu\widehat{\sigma}}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\mu\widehat{\pi}}(\boldsymbol{\theta}) \\ \ddot{\boldsymbol{Q}}_{\sigma\widehat{\mu}}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\sigma\widehat{\sigma}}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\sigma\widehat{\pi}}(\boldsymbol{\theta}) \\ \ddot{\boldsymbol{Q}}_{\pi\widehat{\mu}}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\pi\widehat{\sigma}}(\boldsymbol{\theta}) & \ddot{\boldsymbol{Q}}_{\pi\widehat{\pi}}(\boldsymbol{\theta}) \end{pmatrix}. \tag{12}$$

and the entries of matrices $\ddot{\boldsymbol{Q}}_{\theta\theta}(\boldsymbol{\theta})$ and $\ddot{\boldsymbol{Q}}_{\theta\widehat{\theta}}(\boldsymbol{\theta})$ are presented in Appendix A.

To the best of our knowledge, the use of the Oakes method to obtain the observed information matrix for Gaussian mixture models has not been reported previously in the literature.

III. DIVERGENCE MEASURES

In this section, we introduce a dissimilarity index based on the Kullback-Leibler divergence to quantify the similarity among intensity functions of the different OD pairs of the metro line.

Let $Z \in \mathbb{R}$ be a random variable with density function $f_Z(z)$. The Shannon entropy, or expected information, is given by [6]

$$H(Z) = -\mathsf{E}[\log f_Z(z)] = -\int_{\mathbb{R}} \{\log f_Z(z)\} f_Z(z)\, \mathsf{d}\, z.$$

Now, assume two random variables X and Y with probability density functions $f(x)$ and $g(y)$, respectively, having the same support. Based on the entropy notion, we can define divergence measures between the distributions of X and Y. One of the most common measures to determine the divergence between two distributions is the Kullback-Leibler (KL) divergence given by [6]

$$D(f\,\|\,g) = \int_{\mathbb{R}} \log\Big(\frac{f(x)}{g(y)}\Big) f(x)\, \mathsf{d}\, x = \mathsf{E}_f\Big[\log\Big(\frac{f(x)}{g(y)}\Big)\Big],$$

where the notation emphasizes that the expectation is defined with respect to the probability density function $f(x)$.

The KL divergence measures the distance between two densities, it is only a pseudodistance measure, since, in general, the KL from f to g is not the same as the KL from g to f. However, from the statistical point of view, it is relevant that $D(f\,\|\,g) \geq 0$ and $D(f\,\|\,g) = 0$ if and only if $f = g$ almost everywhere. A familiar symmetric variant of the KL divergence is the J-divergence (see, for instance, [6]), which takes the following definition:

$$J(f\,\|\,g) = D(f\,\|\,g) + D(g\|f). \tag{13}$$

These divergence measures have several useful applications including telecommunications, image analysis and econometrics. An excellent description of the properties and extensions of these procedures is given in [34].

Although it is possible to obtain explicit expressions for the KL divergence between normal, uniform or gamma variables, closed forms for the KL divergence are not available for the class of discrete mixture of normal distributions. This has motivated considerable effort in proposing procedures to

approximate the KL divergence in the context of discrete mixtures of normals. For a review of a wider variety of methods, see [7] and [12]. In the numerical experiments in Section V, we use a Monte Carlo approach. In this method, assuming that f and g are density functions following a discrete mixture of normal densities (1), a Monte Carlo estimate of $D(f\|g)$ is based on a random sample $\{x_1, \ldots, x_M\}$ from $f(x)$, given by

$$D_M(f\|g) = \frac{1}{M}\sum_{r=1}^{M} \log\left(\frac{f(x_r)}{g(x_r)}\right).$$

By the Law of Large Numbers, we have that $D_M(f\|g) \to D(f\|g)$ as $M \to \infty$. Moreover, this procedure allows us to obtain a Monte Carlo estimate for the variance of $D_M(f\|g)$, given by

$$\mathrm{var}\left(D_M(f\|g)\right) = \frac{1}{M^2}\sum_{r=1}^{M} f(x_r)\left[\log\left(\frac{f(x_r)}{g(x_r)}\right) - D_M(f\|g)\right]^2.$$

In addition,

$$\frac{D_M(f\|g) - \mathsf{E}_f[\log(f(x)/g(x))]}{\sqrt{\mathrm{var}(D_M(f\|g))}} \xrightarrow{D} \mathsf{N}(0,1).$$

Notice that the accuracy of this method can be improved by simply increasing the number of generated random variables.

IV. CLUSTER ANALYSIS

In this section, we briefly describe the most common methods used in clustering analysis in order to apply them to group similar OD pairs. We seek to determine whether clustering techniques with a symmetrized KL distance, i.e., the J-divergence in (13), is a suitable technique to characterize different types of passenger flow between different pairs of stations in the OD matrix. In particular, we expect a high value of the divergence between an OD pair with a higher morning passengers flow and an OD pair with a peak of passengers flow in the afternoon. Thus, obtaining clearly separated clusters. Other profiles of daily passengers flow are also expected to appear in the analysis.

A. CLUSTERING METHODS

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the finite set of all possible elements to be grouped. In this case, $\mathcal{X} \subset \mathcal{F}_{\Theta_g}$ is the set of all mixed Gaussian densities associated with an OD pair. Now, the network concept can be introduced [4].

Definition 2: A network N is a pair $(\mathcal{X}, D)$ where $\mathcal{X} = \{x_1, \ldots, x_n\}$ is a finite set of points to be grouped, and $D(\cdot\|\cdot) : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a divergence measure. The set of all networks is denoted as $\mathcal{N}$.

A partition $\mathcal{P}$ of $\mathcal{X}$ is a collection of sets $\mathcal{P} = \{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$ such that $\mathcal{G}_i \subset \mathcal{X}$ for all $i = 1, \ldots, k$, $\bigcup_{i=1}^{k} \mathcal{G}_i = \mathcal{X}$, and $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ for all $i \neq j$. We will call the partition $\mathcal{P}$ a grouping or clustering of $\mathcal{X}$. A partition $\mathcal{P}$ depends on the network we used; thus, two different divergences measures will not necessarily be associated with the same partition.

Among all possible clustering techniques, we consider two methods: partitional and hierarchical clustering. Within the first approach, there are two well-known techniques: k-means and k-medoids. These methods require to choose the number of groups k and the use of a symmetric divergence measure $J(\cdot\|\cdot)$ between pairs of elements in $\mathcal{X}$. Because the k-means method is well known in the literature (see, for instance, [15]), in the Appendix we briefly describe the k-medoids technique.

On the other hand, hierarchical clustering techniques proceed by either successive mergers or successive divisions. The way in which these techniques use the nearest neighbor between items is known as linkage methods. Hierarchical clustering generates a series of nested partitions [32], the description of which has been relegated to the Appendix.

B. SELECTION OF NUMBER OF CLUSTERS

One of the issues in clustering is choosing the number of groups for an available data set. Depending on the method used or the type of data, there are several techniques for the validation of the groups, of which heuristic-based indexes are usually used for the selection of the number of groups. The main difficulty is to find the right number of clusters so that there are not too few clusters with very dissimilar data but not too many clusters with very dissimilar data. Among all existing coefficients, analysts prefer to use techniques that satisfy an optimal condition. We mention three coefficients of this type: the elbow method, the Silhouette method, and the Gap statistic [30], [33]. These methods are widely known, and implementations in `R` are available on several websites. See, for instance, https://uc-r.github.io/kmeans_clustering#gap.

For the numerical experiments in Section V, the silhouette method is used to select the number of clusters. However, for the case of a metro system, the choice of the number of clusters may be considered as a design parameter depending on the number of groups required for analysis, simulation or optimization purposes.

V. CASE STUDY: EFE VALPARAISO

In this section, we show numerical results to gain insights into passenger behavior in the EFE Valparaíso metro line, analyzing a real data set consisting of passenger trips of August 2019. EFE is the chilean state railway company and EFE Valparaíso is the largest metro line outside Santiago, capital city of Chile, having 20 stations over a 43 kilometers line and moving around 20 million passengers per year.

A. PARAMETER ESTIMATES

Our first goal is to estimate the parameters of a Gaussian mixture model (1) for one of the OD pairs. In particular, we estimate the intensity function for the pair (17,1) using the data of all Tuesdays during August 2019. We also provide confidence intervals for all parameters of model (1).

Later, in Section V-B, we numerically analyze the flow passenger patterns of the metro system considering all OD pairs of EFE Valparaíso. Based on the experience of line operators, we expect clear patterns in different OD pairs. Trips from the last to the first stations correspond mainly to the work population, so a greater flow of passengers is expected

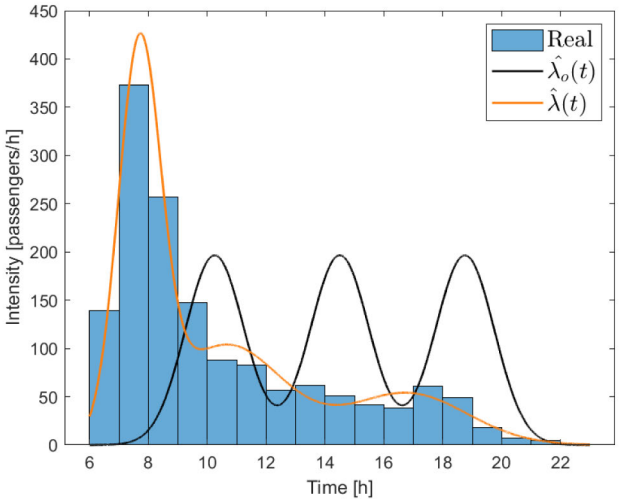


FIGURE 1. The blue bars represent the histogram of passengers arrivals for Tuesdays during August 2019 associated with pair (17,1). The yellow line show the intensity function with the initial parameters, whereas the orange line show the estimated intensity obtained with the EM algorithm.

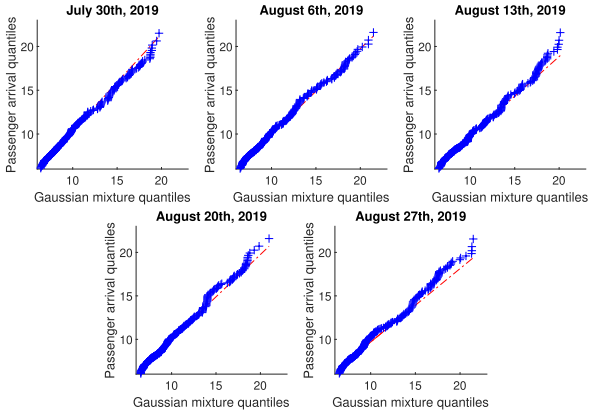


FIGURE 2. QQ plots of daily passenger arrival vs. estimated intensity quantiles generated by Monte Carlo simulation of 10^6 samples.

in the morning. For the opposite direction, when mainly workers return to their homes, a higher flow is expected in the afternoon.

In [1], Bayesian Information Criterion (BIC) was used to determine the (best) model order and, thus, the number of parameters for the different point processes intensity functions. That analysis showed that the choice of $g = 3$ Gaussian components for the different OD pairs of the EFE Valparaíso metro line was the optimal trade-off between fitting of the real data and model complexity. Thus, $g = 3$ is used for all the experiments carried out in this section.

To define the initial parameters for model (1) that are required for the EM estimation algorithm, the time instants were defined based on the available operation schedule as $T_i = 6$ and $T_f = 23$. The initial parameter vector was $\boldsymbol{\theta}^{(0)} = (0.33, 0.33, 10.25, 14.5, 18.75, 1, 1, 1)^\top$. The bootstrap confidence intervals for the components of $\boldsymbol{\theta}$ were generated using $B = 1000$ bootstrap samples [10].

For OD pair (17,1), the total number of passenger arrivals for every Tuesday of August 2019 were 275, 309, 302, 305 and 287, i.e., giving a daily average $\bar{n}_d = 295.6$. Figure 1 shows the initial intensity and the intensity obtained with the parameter estimates given by the proposed EM algorithm.

In addition, Table 1 shows the obtained parameter estimates, its standard errors, and the corresponding Oakes and bootstrap 95% confidence intervals for all parameters in model (1). Notice that, from (3) the weight of the third Gaussian density is given by $\pi_3 = 1 - \pi_1 - \pi_2$.

The Oakes method obtained the lower standard errors in 7 of the 8 parameters of the model, and the bootstrap method gives the best results for π_1, the Gaussian component with the highest passenger arrivals (see Table 1). The Oakes method also achieves the shortest confidence intervals for the second and third Gaussian components, which have lower passenger flows. The confidence intervals for π_1, μ_1 and σ_1^2 exhibit similar behavior for all methods.

The estimated intensity in Figure 2 shows a good fit, in particular, for every day included in the study. The passenger arrival quantiles vs. the Gaussian mixture quantiles show small deviations from the straight line, supporting the goodness of fit in each case.

Figure 3 shows the 95% confidence intervals of the parameter estimates obtained by the Oakes methods, when the number of days used as data for estimation is increased. It can be noticed that the uncertainty around the point estimate of the parameter is clearly reduced as the number of days increases. In fact, all the parameter estimates (except σ_1) show a 50% reduction in the associated confidence intervals. In the figure, it can also be noticed that the intervals obtained for 3 or more days of data are similar for each parameter.

In the next subsection we are interested in quantifying the similarity between the intensity functions for different OD pairs. In order to estimate these densities, the same procedure in this subsection was followed: a mixture of $g = 3$ Gaussian densities in (1), and we use the data for all Tuesdays of August 2019.

B. CLUSTERING STATIONS PAIRS

In this subsection, we use the OD pair densities of the line model to estimate a dissimilarity matrix by the Monte Carlo method using 10^6 random samples. Later, the dissimilarity matrix is used as input to hierarchical and partitional clustering algorithms.

We compute the symmetrized KL divergence (13) for all 380×380 OD pairs in the metro line. The values of the dissimilarity index are approximately in the range from 0 to 8, where low values of divergence are mostly observed. Figure 4 shows the KL based dissimilarity index where the maximum of the color scale has been set to 2.25, which includes 90% of distances between OD pairs. In the figure, it can be noticed that the pairs containing 11 as an entry station divide the matrix into quadrants. The most significant contrast of the matrix appears in the lower left quadrant that compares OD pairs with entry stations between station 1 and 11, with respect to pairs with entry stations between 12 and 20.

TABLE 1. Point estimates and confidence intervals for parameters of model (1) using (17,1) pair. Standard errors of estimates obtained using Oakes and bootstrap methods. Lower and upper limits are shown for 95% asymptotic confidence intervals.

		Standard errors		95% Confidence intervals	
Parameter	Estimate	Oakes	Bootstrap	Oakes	Bootstrap
π_1	0.4822	0.0276	0.0314	(0.4282, 0.5363)	(0.4206, 0.5438)
π_2	0.3334	0.0256	0.0373	(0.2831, 0.3836)	(0.2602, 0.4065)
μ_1	7.6973	0.0459	0.0455	(7.6073, 7.7872)	(7.6080, 7.7865)
μ_2	10.6518	0.2117	0.2664	(10.2369, 11.0667)	(10.1296, 11.1739)
μ_3	16.7923	0.2218	0.3293	(16.3575, 17.2271)	(16.1469, 17.4378)
σ_1^2	0.5173	0.0512	0.0592	(0.4170, 0.6176)	(0.4013, 0.6333)
σ_2^2	3.6116	0.4063	0.7836	(2.8152, 4.4079)	(2.0758, 5.1474)
σ_3^2	4.1250	0.5656	0.8609	(3.0165, 5.2336)	(2.4377, 5.8124)

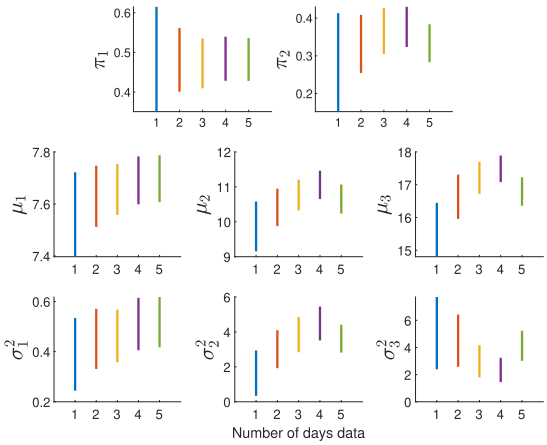


FIGURE 3. Confidence intervals (95%) obtained by the Oakes method as the number of the days of data used for the intensity function estimation is increased.

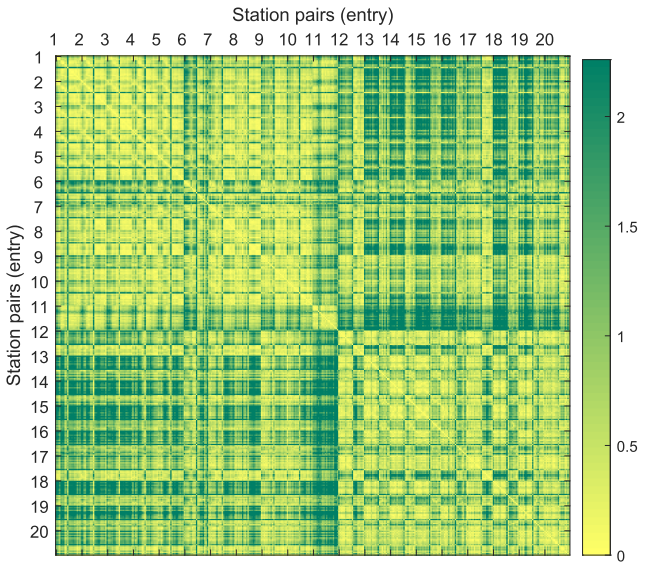


FIGURE 4. Symmetric Kullback-Leibler divergence for all OD pairs (color scale adjusted to the interval [0,2.25] that includes 90% of the pairs).

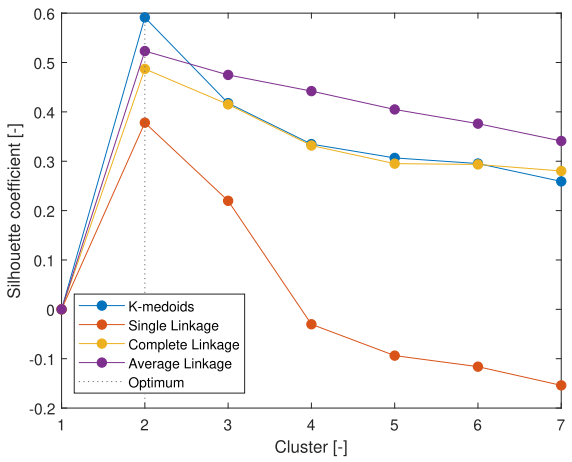


FIGURE 5. Silhouette coefficient for different clustering methods of all OD pairs with partitions between 2 to 7 clusters.

Regarding the choice of the numbers of clusters, the silhouette coefficient suggests that best fit for all algorithms is a partition into 2 clusters (Figure 5).

Although complete linkage has the second worst silhouette coefficient, Figure 6 shows that this method produces more balanced hierarchies with respect to a single linkage (that tends to link OD pairs one by one). For instance, when using complete linkage, a partition into 2 clusters produces clusters with 155 and 225 elements, and a partition into 7 clusters still does not produce a singleton. Average linkage is an intermediate case between single and complete linkage because outliers are detected for higher values of dissimilarity. However, average linkage is also capable of recognizing clusters with many densities but for lower dissimilarity values (see Figure 7).

The choice of the number of clusters for the OD pairs may be considered as a design parameter depending on the number of groups required for analysis, simulation or optimization purposes. Although the silhouette coefficient suggests a partition into 2 clusters, we analyze the partition of 5 medoids to recognize more traffic profiles in the metro system. The number of grouped densities for each cluster from 1 to 5 are 131, 74, 42, 63 and 70. Figure 8 shows that most OD pairs above the main diagonal of the OD matrix (cluster 1 in blue) exhibit an afternoon peak traffic, when a large number of passengers return home after work. As a way of contrast, OD pairs corresponding to clusters 3 and 5 show a higher flow in the morning, when passengers most probably travel to their workplaces. Notice that these two clusters are different

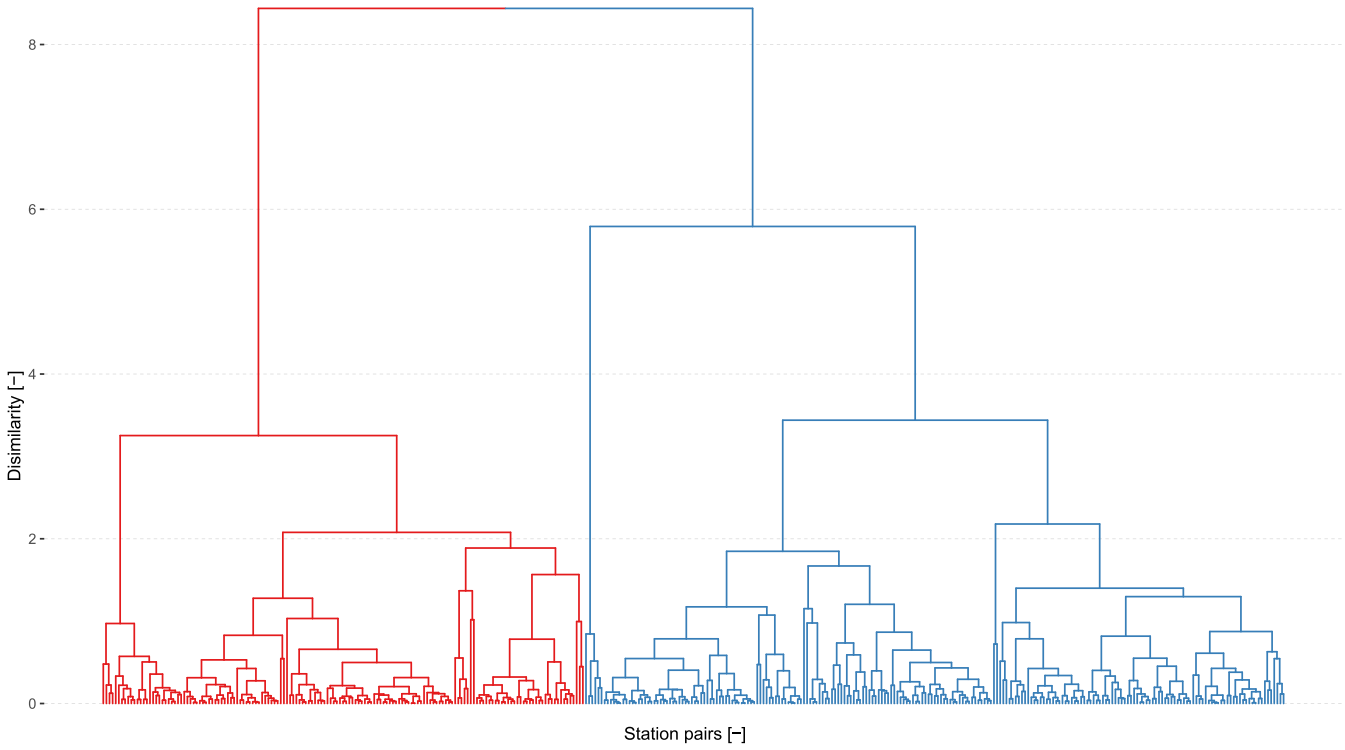


FIGURE 6. Dendrogram using complete linkage with partition of 2 clusters for all OD pairs.

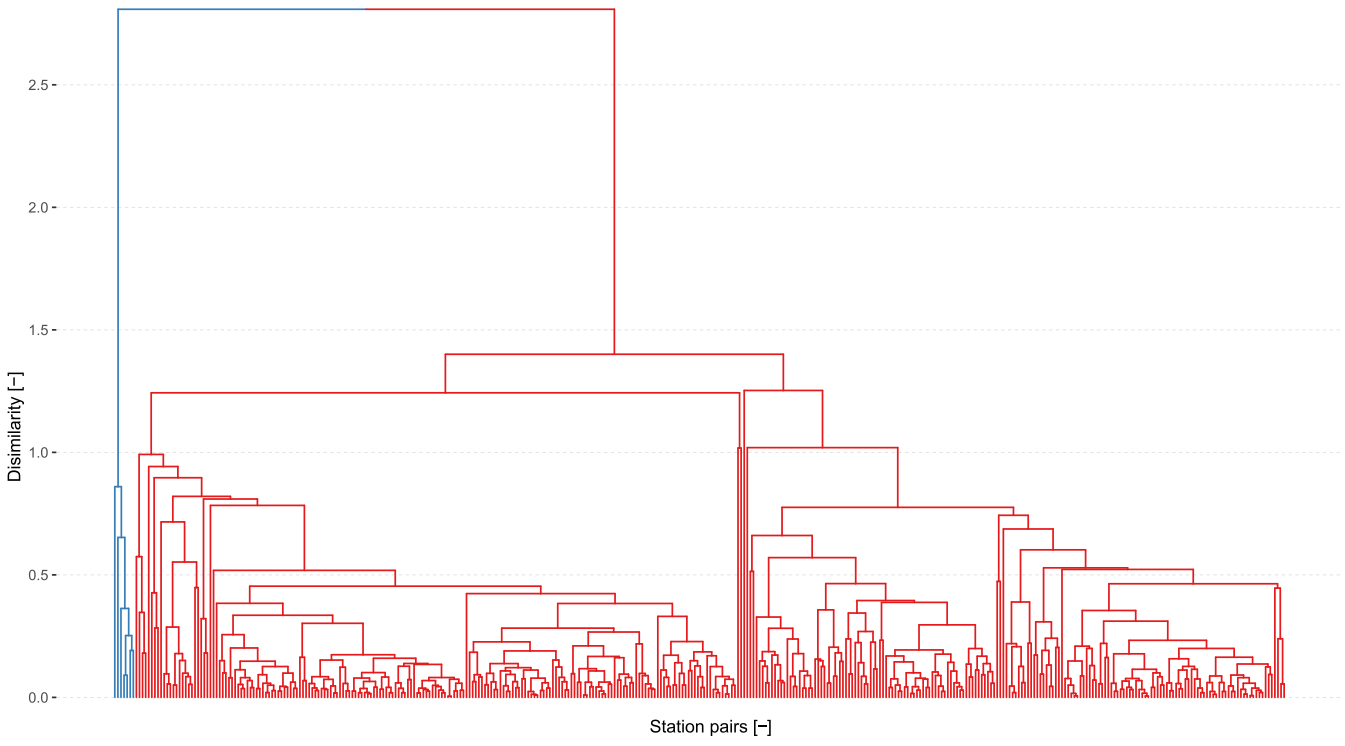
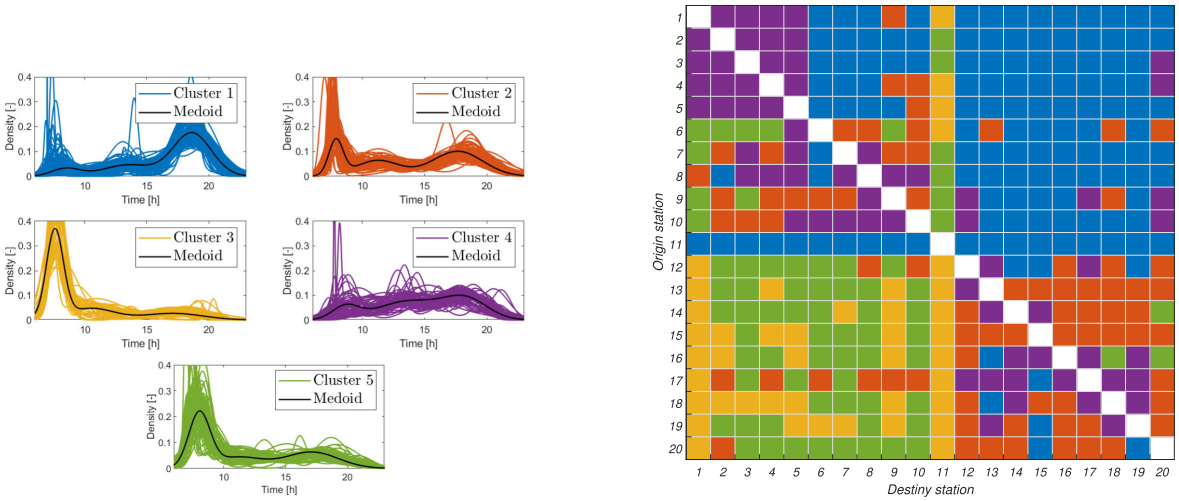


FIGURE 7. Dendrogram using average linkage with partition of 2 clusters for all OD pairs.

because 3 shows a more pronounced peak. On the other hand, cluster 2 shows both a morning and afternoon hour peak. Finally, a key characteristic appears in cluster 4 where a higher passenger flow at midday can be noticed.



(a) (b)

FIGURE 8. **Clustering using 5-medoids method for all OD pairs. (a) Clustered densities. (b) Distribution of clusters in OD matrix.**

Figure 8 also shows the singular behavior of station 11. When considered as an origin station (row 11 of the matrix), we notice an afternoon peak (cluster 1 in blue). On the other hand, when considered a destination station (column 11), peak traffic appears in the morning. This behaviour might be expected since this station corresponds to El Salto industrial area, where people travel to in the morning to work (i.e., is a destination station), and leave in the afternoon to go home (i.e., is an origin station).

An interesting insight that arises from the clusters and matrix shown in Figure 8 is related to the OD pairs close to the diagonal. Most of these pairs are in clusters 2 and 4, which correspond to a more even flow of passengers during the day. In fact, this kind of passengers' behaviour has not been previously recognized by EFE Valparaíso operators, showing an advantage of the proposed methodology.

VI. CONCLUSION

In this paper, we have proposed a methodology to model passengers movement in a metro line and to characterize the flow patterns that appear among the different origin-destination pairs. The paper makes two contributions. First, the instantaneous rate of arrival of passengers for each origin-destination pair is modeled using a Gaussian mixture intensity function, exploiting the EM algorithm to obtain parameter estimates and the associated confidence intervals. In fact, the use of data from multiple days leads to an improvement in the uncertainty in the parameter estimates showing, for example, a 50% reduction of the confidence intervals obtained by the Oakes method when going from 1 day to 3 or more days of data.

The case study presented in the paper confirms that the Gaussian mixture model provides good approximations of the (observed) intensity functions throughout the day. Moreover, the confidence intervals of the estimates obtained by different approaches are similar and allow us to detect estimates with poor accuracy, for example, for stations with fewer passengers.

The second contribution presented in the paper is to quantitatively compare the intensity functions of different origin-destination station pairs of the line. The proposed dissimilarity index to measure the distance between the intensity functions was derived from the KL divergence. This approach provides asymptotic properties that enhance its use in practice since approximate confidence intervals and hypothesis tests could be derived as a direct consequence.

In addition, the hierarchical and medoid clustering methods proposed in the paper to characterize similar patterns among the origin-destination intensity functions, provide both visual and quantitative information about different passenger behavior in the line. The results obtained show patterns that are consistent with knowledge about passenger behavior in morning and evening peak hours and, more significantly, provide additional insights for specific origin-destination pairs.

We believe that the general methodology proposed in the paper can be applied to other similar transport networks, providing useful information for analysis, simulation and optimization.

APPENDIX A
OBSERVED INFORMATION MATRIX

The blocks of the Hessian matrix (11) are given by

$$\ddot{Q}_{\mu\mu}(\boldsymbol{\theta}) = -\operatorname{diag}\left(\widehat{z}_{1\bullet}/\sigma_1^2, \ldots, \widehat{z}_{g\bullet}/\sigma_g^2\right),$$

$$\ddot{Q}_{\mu\sigma}(\boldsymbol{\theta}) = -\operatorname{diag}\left(\frac{1}{\sigma_1^4}\sum_{j=1}^{n}\widehat{z}_{1j}(y_j-\mu_1), \ldots \right.$$

$$\left. \ldots, \frac{1}{\sigma_g^4}\sum_{j=1}^{n}\widehat{z}_{gj}(y_j-\mu_g)\right),$$

$$\ddot{Q}_{\sigma\sigma}(\boldsymbol{\theta}) = \tfrac{1}{2}\operatorname{diag}\left(\widehat{z}_{1\bullet}/\sigma_1^4, \ldots, \widehat{z}_{g\bullet}/\sigma_g^4\right) - \operatorname{diag}\left(\frac{1}{\sigma_1^4}\sum_{j=1}^{n}\widehat{z}_{1j}\Big(\frac{y_j-\mu_1}{\sigma_1}\Big)^2, \ldots, \frac{1}{\sigma_g^4}\sum_{j=1}^{n}\widehat{z}_{gj}\Big(\frac{y_j-\mu_g}{\sigma_g}\Big)^2\right),$$

$$\ddot{Q}_{\pi\pi}(\boldsymbol{\theta}) = -\operatorname{diag}\left(\widehat{z}_{1\bullet}/\pi_1^2, \ldots, \widehat{z}_{g-1,\bullet}/\sigma_{g-1}^2\right),$$

where $\widehat{z}_{i\bullet} = \sum_{j=1}^{n}\widehat{z}_{ij}$, for $i = 1, \ldots, g$, while the entries of the matrix $\ddot{Q}_{\theta\widehat{\theta}}(\boldsymbol{\theta}) = (\ddot{Q}_{\theta_r\widehat{\theta}_s})$ defined in (12) are

$$\ddot{Q}_{\theta_r\widehat{\theta}_s} = \frac{\partial^2 Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}})}{\partial\theta_r\partial\widehat{\theta}_s},$$

$$\ddot{Q}_{\mu_r\widehat{\mu}_s} = \begin{cases} -\widehat{\pi}_r\widehat{\pi}_s\displaystyle\sum_{j=1}^{n} p_{rs}(y_j;\widehat{\boldsymbol{\theta}})\Big(\frac{y_j-\mu_r}{\sigma_r^2}\Big)\Big(\frac{y_j-\widehat{\mu}_s}{\widehat{\sigma}_s^2}\Big), & r \neq s, \\ \widehat{\pi}_r\displaystyle\sum_{j=1}^{n} q_r(y_j;\widehat{\boldsymbol{\theta}})\Big(\frac{y_j-\mu_r}{\sigma_r^2}\Big)\Big(\frac{y_j-\widehat{\mu}_r}{\widehat{\sigma}_r^2}\Big), & r = s, \end{cases}$$

$$\ddot{Q}_{\mu_r\widehat{\sigma}_s} = \begin{cases} \dfrac{\pi_r\widehat{\pi}_s}{2\widehat{\sigma}_s^2}\displaystyle\sum_{j=1}^{n} p_{rs}(y_j;\widehat{\boldsymbol{\theta}})\Big(\frac{y_j-\mu_r}{\sigma_r^2}\Big) \times \Big(1-\Big(\frac{y_j-\widehat{\mu}_s}{\widehat{\sigma}_s}\Big)^2\Big), & r \neq s, \\ -\dfrac{\pi_r}{2\widehat{\sigma}_r^2}\displaystyle\sum_{j=1}^{n} q_r(y_j;\widehat{\boldsymbol{\theta}})\Big(\frac{y_j-\mu_r}{\sigma_r^2}\Big) \times \Big(1-\Big(\frac{y_j-\widehat{\mu}_r}{\widehat{\sigma}_r}\Big)^2\Big), & r = s, \end{cases}$$

$$\ddot{Q}_{\mu_r\widehat{\pi}_s} = \begin{cases} -\dfrac{\widehat{\pi}_r\widehat{\pi}_s}{\pi_r}\displaystyle\sum_{j=1}^{n} p_{rs} \times (y_j;\widehat{\boldsymbol{\theta}})\Big(\frac{y_j-\mu_r}{\sigma_r^2}\Big), & r \neq s, \\ \dfrac{\widehat{\pi}_r}{\pi_r}\displaystyle\sum_{j=1}^{n} q_r(y_j;\widehat{\boldsymbol{\theta}})\Big(\frac{y_j-\mu_r}{\sigma_r^2}\Big), & r = s, \end{cases}$$

$$\ddot{Q}_{\sigma_r\widehat{\sigma}_s} = \begin{cases} -\dfrac{\widehat{\pi}_r\widehat{\pi}_s}{4\sigma_r^2\widehat{\sigma}_s^2}\displaystyle\sum_{j=1}^{n} p_{rs}(y_j;\widehat{\boldsymbol{\theta}}) \times \Big(1-\Big(\frac{y_j-\mu_r}{\sigma_r}\Big)^2\Big)\Big(1-\Big(\frac{y_j-\widehat{\mu}_s}{\widehat{\sigma}_s}\Big)^2\Big), & r \neq s, \\ \dfrac{\widehat{\pi}_r}{4\sigma_r^2\widehat{\sigma}_r^2}\displaystyle\sum_{j=1}^{n} q_r(y_j;\widehat{\boldsymbol{\theta}}) \times \Big(1-\Big(\frac{y_j-\mu_r}{\sigma_r}\Big)^2\Big)\Big(1-\Big(\frac{y_j-\widehat{\mu}_r}{\widehat{\sigma}_r}\Big)^2\Big), & r = s, \end{cases}$$

$$\ddot{Q}_{\sigma_r\widehat{\pi}_s} = \begin{cases} \dfrac{\widehat{\pi}_r\widehat{\pi}_s}{2\widehat{\sigma}_r^2\pi_s}\displaystyle\sum_{j=1}^{n} p_{rs}(y_j;\widehat{\boldsymbol{\theta}})\Big(1-\Big(\frac{y_j-\widehat{\mu}_s}{\widehat{\sigma}_s}\Big)^2\Big), & r \neq s, \\ -\dfrac{\widehat{\pi}_r}{2\widehat{\sigma}_r^2\pi_r}\displaystyle\sum_{j=1}^{n} q_r(y_j;\widehat{\boldsymbol{\theta}})\Big(1-\Big(\frac{y_j-\widehat{\mu}_r}{\widehat{\sigma}_r}\Big)^2\Big), & r = s, \end{cases}$$

$$\ddot{Q}_{\pi_r\widehat{\pi}_s} = \begin{cases} -\dfrac{\widehat{\pi}_r}{\pi_r}\displaystyle\sum_{j=1}^{n} p_{rs}(y_j;\widehat{\boldsymbol{\theta}}), & r \neq s, \\ \dfrac{1}{\pi_r}\displaystyle\sum_{j=1}^{n} q_r(y_j;\widehat{\boldsymbol{\theta}}), & r = s, \end{cases}$$

with

$$p_{rs}(y;\widehat{\boldsymbol{\theta}}) = \frac{\phi(y;\widehat{\mu}_r,\widehat{\sigma}_r^2)\phi(y;\widehat{\mu}_s,\widehat{\sigma}_s^2)}{\{f(y;\widehat{\boldsymbol{\theta}})\}^2},$$

$$q_r(y;\widehat{\boldsymbol{\theta}}) = \frac{\phi(y;\widehat{\mu}_r,\widehat{\sigma}_r^2)}{\{f(y;\widehat{\boldsymbol{\theta}})\}^2}\big[f(y;\widehat{\boldsymbol{\theta}}) - \widehat{\pi}_r\phi(y;\widehat{\mu}_r,\widehat{\sigma}_r^2)\big].$$

APPENDIX B
CLUSTERING ALGORITHMS

Let $m_i \in \mathcal{G}_i$ be the medoid of the group $\mathcal{G}_i$; then, for all $i = 1, \ldots, k$, m_i is obtained as

$$m_i = \arg\min_{x\in\mathcal{G}_i}\sum_{x'\in\mathcal{G}_i} J(x\|x'). \tag{14}$$

Let $\mathcal{S} \subset \mathcal{X}$ be the set of all selected medoids. The objective of the method is to determine the set $\mathcal{S} = \{m_1, \ldots, m_k\}$ that minimizes the sum of all dissimilarities between each medoid and the elements of the respective group. i.e.,

$$\arg\min_{\mathcal{S}}\sum_{i=1}^{k}\sum_{x'\in\mathcal{G}_i} J(m_i\|x'). \tag{15}$$

Algorithm 1 summarizes the necessary steps to yield the final clusters.

Algorithm 1 k-Medoids Method

1. Define k initial medoids $m_1, \ldots, m_k$.
2. Assigns to each $x \in \mathcal{X}$ the nearest medoids cluster:
$$x \in \mathcal{G}_i \Leftrightarrow m_i = \arg\min_{m\in\mathcal{S}} J(x\|x').$$
3. Update $\mathcal{S}$ through Equation (14).
4. Repeat steps *2* and *3* until there are no new updates in $\mathcal{S}$.

One of the most commonly used methods to obtain a set of medoids is partitioning around medoids (PAM) introduced in [18]. This methods consists of two steps BUILT and SWAP. More details can be found in [18, Section 2.4].

In order to describe the hierarchical agglomerative methods we state the following definition.

Definition 3: A partition $\mathcal{P}_1$ of k clusters is said nested within $\mathcal{P}_2$ that has $r < k$ clusters if for each $\mathcal{G} \in \mathcal{P}_1$ there exists $\mathcal{G}' \in \mathcal{P}_2$ such that $\mathcal{G} \subseteq \mathcal{G}'$. The nested partitions are denoted as $\mathcal{P}_1 \sqsubset \mathcal{P}_2$.

Hierarchical clustering generates a hierarchy of nested partitions. Let $\mathcal{P}_0$ be the partition of $\mathcal{X}$ into n groups, i.e., singletons, and let $\mathcal{P}_{n-1}$ be the partition of a single group containing all objects. Agglomerative hierarchies generate the following nested sequence of partitions:

$$\mathcal{P}_0 \sqsubset \mathcal{P}_1 \sqsubset \ldots \sqsubset \mathcal{P}_{n-1},$$

in which if two objects are joined together in a group, they will never be separated again. On the other hand, divisive hierarchies generate the reverse sequence

$$\mathcal{P}_{n-1} \sqsubset \mathcal{P}_{n-2} \sqsubset \ldots \sqsubset \mathcal{P}_0.$$

Merging of the clusters is carried out using one of the three linkage criteria described as follows. Single linkage: Groups are formed from the individual entities by merging nearest neighbors. i.e.,

$$d_{SL}(\mathcal{G}_i, \mathcal{G}_j) = \min_{x \in \mathcal{G}_i, x' \in \mathcal{G}_j} J(x \| x'). \tag{16}$$

Complete linkage: At each stage, the distance between clusters is determined by the distance between the two elements, one from each cluster, that are most distant. i.e.,

$$d_{CL}(\mathcal{G}_i, \mathcal{G}_j) = \max_{x \in \mathcal{G}_i, x' \in \mathcal{G}_j} J(x \| x'). \tag{17}$$

Average linkage: Computes the average between dissimilarities between objects from different clusters through

$$d_{AL}(\mathcal{G}_i, \mathcal{G}_j) = \frac{1}{|\mathcal{G}_i| \cdot |\mathcal{G}_j|} \sum_{x \in \mathcal{G}_i} \sum_{x' \in \mathcal{G}_j} J(x \| x'), \tag{18}$$

with $|\mathcal{G}_i|$ being the number of objects belonging to $\mathcal{G}_i$.

The hierarchical agglomerative method is summarized in Algorithm 2.

Algorithm 2 Hierarchical Agglomerative Method

1. For the set $\mathcal{X}$, compute the divergence matrix $\boldsymbol{D} = (d_{ik})$.
2. Choose the linkage method. Denote the distance between clusters as $d_{\cdot L}$.
3. Search the pair $\mathcal{G}_i$ and $\mathcal{G}_j$ having the smallest dissimilarity according to the method selected in step *2*. Store the value $d_{\cdot L}(\mathcal{G}_i, \mathcal{G}_j)$.
4. Merge clusters $\mathcal{G}_i$ and $\mathcal{G}_j$ to form the new cluster $(\mathcal{G}_i\mathcal{G}_j) = \mathcal{G}_i \cup \mathcal{G}_j$. Update the entries in the distance matrix by eliminating the rows and columns corresponding to clusters $\mathcal{G}_i$ and $\mathcal{G}_j$ and adding a row and column giving the distances between cluster $(\mathcal{G}_i\mathcal{G}_j)$ and the remaining clusters.
5. Repeat steps *3* and *4* a total of $n - 1$ times. Record identity of clusters are merged and the levels at which the merges take place.

ACKNOWLEDGMENT

The authors would like to thank the comments and suggestions made by the associate editor and anonymous reviewers that have allowed them to improve the paper, and also would like to thank EFE Valparaíso for providing the data set used in the numerical examples.

REFERENCES

[1] J. Allende-Bustamante, J. I. Yuz, and J. Rodo, "Application of point processes estimation to a metro system," in *Proc. Austral. Control Conf. (AuCC)*, Newcastle, NSW, Australia, pp. 3–4, Nov. 2016.

[2] D. D. Boos and L. A. Stefanski, *Essential Statistical Inference: Theory and Methods*. New York, NY, USA: Springer, 2013.

[3] N. Caceres, L. M. Romero, and F. G. Benitez, "Estimating traffic flow profiles according to a relative attractiveness factor," *Proc. Social Behav. Sci.*, vol. 54, pp. 1115–1124, Oct. 2012.

[4] G. Carlsson, F. Memoli, A. Ribeiro, and S. Segarra, "Hierarchical clustering methods and algorithms for asymmetric networks," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Beijing, China, Nov. 2013, pp. 352–360.

[5] Z. Cheng, M. Trépanier, and L. Sun, "Real-time forecasting of metro origin-destination matrices with high-order weighted dynamic mode decomposition," *Transp. Sci.*, pp. 1–15, Feb. 2022, doi: 10.1287/trsc.2022.1128.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[7] S. Cui, "Comparison of approximation methods to Kullback–Leibler divergence between Gaussian mixture models for satellite image retrieval," *Remote Sens. Lett.*, vol. 7, pp. 651–660, Jul. 2016.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Roy. Stat. Soc., B*, vol. 39, pp. 1–38, Sep. 1977.

[9] A. De Santis, T. Giovannelli, S. Lucidi, M. Messedaglia, and M. Roma, "Determining the optimal piecewise constant approximation for the nonhomogeneous Poisson process rate of emergency department patient arrivals," *Flexible Services Manuf. J.*, pp. 1–34, Mar. 2021, doi: 10.1007/s10696-021-09408-9.

[10] B. Efron and R. Tibshirani, *An Introduction to Bootstrap*. London, U.K.: Chapman and Hall, 1993.

[11] B. I. Godoy, V. Solo, and S. A. Pasha, "Truncated Hawkes point process modeling: System theory and system identification," *Automatica*, vol. 113, Mar. 2020, Art. no. 108733.

[12] J. R. Hershey and P. A. Olsen, "Approximating the Kullback–Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Honolulu, HI, USA, Apr. 2007, pp. 317–320.

[13] A. Janßen and P. Wan, "K-means clustering of extremes," *Electron. J. Statist.*, vol. 14, pp. 1211–1233, Jan. 2020.

[14] Y. Ji, Q. You, S. Jiang, and H. M. Zhang, "Statistical inference on transit route-level origin-destination flows using automatic passenger counter data," *J. Adv. Transp.*, vol. 49, no. 6, pp. 724–737, Oct. 2015.

[15] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. New York, NY, USA: Prentice-Hall, 2002.

[16] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2005.

[17] J. F. C. Kingman, *Poisson Processes*. Oxford, U.K.: Oxford Univ. Press, 1993.

[18] K. Leonard and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: Wiley, 1990.

[19] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. Roy. Stat. Soc., B*, vol. 44, pp. 226–233, Jan. 1982.

[20] G. McLachlan and D. Peel, *Finite Mixture Models*. New York, NY, USA: Wiley, 2000.

[21] G. McLachlan and T. Krishnan, *The EM Algorithm Extensions*, 2nd ed. Hoboken, NJ, USA: Wiley, 2008.

[22] A. K. Menon and Y. Lee, "Predicting short-term public transport demand via inhomogeneous Poisson processes," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 2207–2210.

[23] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi-passenger demand using streaming data," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1393–1402, Sep. 2013.

[24] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, Chile," *Transp. Res. C, Emerg. Technol.*, vol. 24, pp. 9–18, Oct. 2012.

[25] D. Oakes, "Direct calculation of the information matrix via the EM," *J. Roy. Stat. Soc., B Stat. Methodol.*, vol. 61, no. 2, pp. 479–482, Apr. 1999.

[26] C. Patil and I. Baidari, "Estimating the optimal number of clusters K in a dataset using data depth," *Data Sci. Eng.*, vol. 4, no. 2, pp. 132–140, Jun. 2019.

[27] J. Prades, G. Safont, A. Salazar, and L. Vergara, "Estimation of the number of endmembers in hyperspectral images using agglomerative clustering," *Remote Sens.*, vol. 12, no. 21, p. 3585, Nov. 2020.

[28] Y. Qu, H. Wang, J. Wu, X. Yang, H. Yin, and L. Zhou, "Robust optimization of train timetable and energy efficiency in urban rail transit: A two-stage approach," *Comput. Ind. Eng.*, vol. 146, Aug. 2020, Art. no. 106594.

[29] S. Ross, *Stochastic Processes*. New York, NY, USA: Wiley, 1996.

[30] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[31] D. L. Snyder and M. I. Miller, *Random Point Processes Time Space*. New York, NY, USA: Springer, 2012.

[32] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. London, U.K.: Academic, 2008.

[33] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Statist. Methodol.*, vol. 63, no. 2, pp. 411–423, 2004.
[34] A. Ullah, "Entropy, divergence and distance measures with econometric applications," *J. Stat. Planning Inference*, vol. 49, no. 1, pp. 137–162, Jan. 1996.
[35] L. Wasserman, *All of Nonparametric Statistics*. New York, NY, USA: Springer, 2006.
[36] W. Weijermars and E. van Berkum, "Analyzing highway flow patterns using cluster analysis," in *Proc. IEEE Intell. Transp. Syst.*, Sep. 2005, pp. 308–313.
[37] (Sep. 2018). *World Metro Figure 2018*. Statistics Brief, International Association of Public Transport. [Online]. Available: https://www.uitp.org/publications/world-metro-figures/URL
[38] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.
[39] X. Wu, V. Kumar, J. S. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, pp. 1–37, Jan. 2008.
[40] C. Yang, F. Yan, and X. Xu, "Daily metro origin-destination pattern recognition using dimensionality reduction and clustering methods," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 548–553.
[41] Y. Hu and H. Hellendoorn, "Mixture-model-based clustering for daily traffic volumes," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 2757–2762.
[42] J. Zhao, A. Rahbee, and N. H. M. Wilson, "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 22, pp. 376–387, Jul. 2007.
[43] F. Zúñiga, J. C. Muñoz, and R. Giesen, "Estimation and prediction of dynamic matrix travel on a public transport corridor using historical data and real-time information," *Public Transp.*, vol. 13, no. 1, pp. 59–80, Mar. 2021.

GABRIEL VIDAL received the Ingeniero Civil Matemático professional title from Universidad Técnica Federico Santa María (UTFSM), in 2021. He was worked as a Researcher with the Advanced Center in Electrical and Electronic Engineering AC3E and the Cetaqua.

JUAN I. YUZ (Member, IEEE) received the Ingeniero Civil Electrónico professional title and M.Sc. degrees in electronic engineering from UTFSM, Chile, in 2001, and the Ph.D. degree in electrical engineering from The University of Newcastle, Australia, in 2006. From 2015 to 2019, he was the Director of the Advanced Center for Electrical and Electronic Engineering AC3E—UTFSM, where he is currently a Full Professor with the Departamento de Electrónica. He is the coauthor of the book *Sampled-Data Models for Linear and Nonlinear Systems* (Springer, in 2014). His research interests include control and identification of sampled-data systems and their applications to power electronics and biomedical systems.

RONNY VALLEJOS received the B.Sc. and M.Sc. degrees in mathematics from UTFSM, Chile, in 1995 and 1998, respectively, the master's degree in statistics from the University of Connecticut, USA, in 2002, and the Ph.D. degree in statistics from the University of Maryland, Baltimore County, in 2006. He is currently an Associate Professor with the Department of Mathematics, UTFSM. His research interests include spatial statistics, statistical image processing, time series, and agreement measures.

FELIPE OSORIO received the degree in statistical engineering from the Universidad de Valparaíso, Chile, in 2001, and the D.Sc. degree in statistics from the Universidade de São Paulo, Brazil, in 2006. He is currently an Assistant Professor with the Department of Mathematics, UTFSM, Chile. His research interests include models for data with longitudinal structure (mixed-effects models, GEE) and diagnostic methods and the computational implementation of such techniques. He is the creator and the maintainer of several packages for the R statistical environment.