IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 June 2022, accepted 7 September 2022, date of publication 15 September 2022, date of current version 26 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206838

APPLIED RESEARCH

Using Data Mining to Estimate Patterns of Contagion-Risk Interactions in an Intercity Public Road Transport System

TERESA CRISTÓBAL, ALEXIS QUESADA-ARENCIBIA, GABRIELE S. DE BLASIO, GABINO PADRÓN, FRANCISCO ALAYÓN, AND CARMELO R. GARCÍA

Institute for Cybernetics, University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain

Corresponding author: Carmelo R. García (ruben.garcia@ulpgc.es)

This work was supported in part by the University of Las Palmas de Gran Canaria (ULPGC) through Project COVID19-03; and in part by the Spanish Ministry of Science, Innovation and Universities (MCIU), the State Research Agency (AEI), and the European Regional Development Fund (ERDF) under Project RTI2018-097263-B-I00.

ABSTRACT The COVID-19 pandemic has had very negative effects on public transport systems. These effects have compromised the role they should play as enablers of social equity and environmentally sustainable mobility and have caused serious economic losses for public transport operators. For this reason, in the context of pandemics, meaningful epidemiological information gathered in the specific framework of these systems is of great interest. This article presents the findings of an investigation into the risk of transmission of a respiratory infectious disease in an intercity road transport system that carries millions of passengers annually. To achieve this objective, a data mining methodology was used to generate the data required to ascertain the level of risk. Using this methodology, the occupancy of vehicle seats by passengers was simulated using two different strategies. The first is an empirical approach to the behaviour of passengers when occupying a free seat and the second attempts to minimise the risk of contagion. For each of these strategies, the interactions with risk of infection between passengers were estimated, the patterns of these interactions on the different routes of the transport system were obtained using k-means clustering technique, and the impact of the strategies was analysed.

INDEX TERMS Close contact patterns, clustering, COVID-19, data mining, epidemics, information management, intelligent transport systems, public health.

I. INTRODUCTION

As a result of the COVID-19 pandemic, the world has had to face multiple challenges in its efforts to mitigate the spread and consequences of the disease. In the context of public transport systems, operators and authorities have had to develop measures to prevent infections among their users. The implementation of these measures, together with the public perception of the risk of infection associated with the use of public transport, has on occasion caused a considerable loss of passengers on public transport systems, up to 90% in some cases [1], [2], [3], jeopardising their role as enablers of social equity [4]. Therefore, in the context of epidemics or pandemics, it is of interest to have information that allows for an objective assessment of infection risk on public transport, as this knowledge can be used to develop and assess effective measures to mitigate it.

This article presents the findings of a research study designed with the aim of determining the infection risk among passengers on the different routes of a public road transport system (PRTS). This information can be used to identify the routes with the highest risk of infection and to assess the impact of different measures to minimise the potential risk. Such measures include devoting more resources to those routes with the highest risk, or other measures that do not involve more resources, such as using certain seating strategies for passengers. The latter type of measures provide alternatives to those that have been routinely implemented

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

during the COVID-19 pandemic, such as suspending transport services or limiting the capacity of public transport vehicles. In this study, the concept of contact has been formalised to estimate the contacts that may have occurred during the scheduled services provided by a road transport network and to analyse how they vary according to different aspects. This research is in line with works that propose artificial intelligence [5] and big data [6] techniques for epidemiological control. To the best of our knowledge, the research presented in this paper is original in terms of the objectives pursued and the methodology used. The methodology takes as its starting point a formal framework based on epidemiological parameters, commonly used entities in the definition of transport networks and the operations of a PRTS; it was implemented using the data mining paradigm. In this methodology, the initial source of data is the data generated in the transit system, where the data provided by smart payment systems based on contactless cards are of special relevance. These potentially massive starting data are processed to obtain a dataset, with a graph structure, which is used to estimate the number of interactions that carry a risk of infection in the public transportation system.

In addition to this first introductory section, this article is organised into six sections. The second section presents a review of related work. The third section presents the methodology used in the research; this methodology was applied to a real use case of an intercity PRTS used by millions of passengers annually. The results and a discussion thereof are presented in the fourth and fifth section respectively. The sixth section presents the limitations of this study. Finally, in the seventh section, we present our conclusions.

II. RELATED WORKS

This review of related studies has been developed in the context of the epidemiology of respiratory infectious diseases and is organised in two parts. The first part deals with studies that aim to obtain information on the patterns of person-to-person contacts that can lead to the transmission of this type of disease and, based on these patterns, to model the dynamics of disease transmission and/or design epidemiological control measures. In this part, two types of publications have been reviewed: studies in which patterns are generated from inferred contacts and those in which patterns are obtained from sensor networks data. The second part reviews studies on the role of public transport systems in the transmission of this type of disease.

A. USE OF CONTACT PATTERNS IN THE EPIDEMIOLOGY OF RESPIRATORY INFECTIOUS DISEASES

In the 1880s, Carl Flügge observed that droplets expelled by an infected person when talking, coughing or sneezing could contain the pathogens that cause infectious diseases. Later, Wells [7], in his research on tuberculosis, made the distinction between ''large droplet'' and ''small droplet''. According to Wells, large droplets are deposited in the immediate vicinity of the infected person before they evaporate, in contrast to small droplets, which evaporate before they are deposited, forming residual particulates from the dried material, called aerosols or droplet nuclei. Building on this contribution, it was considered useful to collect data reflecting human-to-human contacts, as these data provide patterns of disease spread and enable effective disease control measures to be implemented [8], [9], [10]. Specifically, when studying the dynamics of disease spread, the social contact hypothesis is used. This assumes that the number of potentially infectious contacts between people is proportional to the number of social contacts, with this proportionality factor being an indicator of the infectivity of the disease [11]. A mathematical model used in these studies [12] uses the next-generation matrix (N) to estimate how many people in different age groups will become infected as a result of contact with an infected person in a given group. This matrix is of such relevance that a considerable number of studies have been carried out to obtain it using different methodologies.

1) STUDIES BASED ON INFERRED CONTACT NETWORKS

Modelling infectious interactions between people in large populations is a scientific challenge of interest. To do this, network theory is used, representing the interactions in a network called a contact network. As such, techniques that attempt to synthesise these networks have been developed. These techniques can be classified into two types: those that generate the interaction network using real or simulated egocentric data, and those that generate the network from the simulated behaviour of individuals. In works based on egocentric data, these data are provided by people (egos) whose identity is known and refer to interactions they have had with other people (alters) whose identity is unknown, but some data are provided, such as their approximate age, for example. The result is a set of interaction networks with a star topology, in which egos are connected to their alters, which provides valuable information about the heterogeneity of the contact network and the patterns of interactions between different population groups. Valuable information is provided by this type of study, such as the patterns that these interactions follow and the probability of interactions between different alters from egocentric data [13].

Ferguson *et al.* [14] and Longini *et al.* [15] describe how to estimate patterns of social contacts from census data, assuming that they reflect the distribution of groups in the population and household size. There is a comprehensive set of studies inferring these patterns from data obtained from surveys of the populations under study. The methodology used in these studies have often organised in three stages. The first stage consists of a survey of selected individuals, in which they are asked to provide information on their close contacts. The second stage consists of obtaining a representation of the network of contacts between different population groups, using a contact matrix (C). Finally, the third step consists of analysing the dynamics of the disease using the next-generation matrix (N), which is obtained from the C matrix and available epidemiological

data. Wallinga *et al.* [11] present a study on how to obtain transmission parameters by age group from a social contact survey on the conversational partners of the participants. The survey was conducted by face-to-face interviews in Utrecht, the Netherlands, in 1986, where 3084 were invited, 2106 completed a questionnaire and 1813 met the criteria for further analysis. A survey-based study conducted in the framework of the European POLYMOD project is presented in [16]. The study involved 7297 participants from 8 European countries. In [17] the BBC Pandemic project is presented. This study was developed in the UK and reported social contact information from 40 177 participants who completed the study, out of the 86 000 participants initially recruited. Other studies using a similar methodology have been carried out in France [18], Russia [19], Hong Kong [20], Japan [21], Taiwan [22], Vietnam [23], in rural areas in Kenya [24], in South Africa [25], Peru [26] and Senegal [27]. In the context of the COVID-19 pandemic, different studies have been conducted on contact patterns in pre-pandemic and pandemic periods in Luxemburg [28], China [29], UK [30], Netherlands [31] and USA [32]. These studies show that as a result of various measures to ensure social distancing, social contacts are reduced by between 40% and 85%, depending on the country. In [33] a study is carried out on the patterns of social contacts in lockdown and post-confinement periods, which are compared with patterns obtained in pre-pandemic periods.

The generation of contact networks from simulated population behaviour is another technique used in epidemiology. Stochastic simulation based on agents that emulate the behaviour of individuals in households and workplaces, resulting in a network of interactions between potentially contagious individuals, is used in [8] to study the effectiveness of a mass vaccination campaign versus a targeted vaccination campaign to control the spread of smallpox disease in structured communities of 2000 people. In [15], the same simulation technique is used to study the effectiveness of the use of antivirals, quarantines and vaccination against an avian influenza pandemic in a population of 500 000 people distributed over 5625 km^2 and structured according to the 2000 population census of Thailand. Agent-based stochastic simulation is also used in [14] to evaluate the effectiveness of using antivirals as a containment measure for an early-stage avian influenza pandemic. The simulation emulates the behaviour of a population of 85 million people located in a 100 km^2 area in Thailand incorporating households, workplaces and schools. In [34] the same technique is used to emulate the behaviour of the population in the UK and the USA to analyse different epidemiological control measures (household quarantines, perimeter confinements, school and workplace closures, travel restrictions and clinical treatments) in the context of an influenza pandemic. In [35], an agent-based simulation is used to model the movements of 1.5 million people in Portland, Oregon, USA, between 180 000 different locations. The goal of the simulation was to detect the presence of two people in the same place at the same instant in time, to generate a static network of contacts, and to predict the number of contacts that occur at each location. The researchers found that the network was highly heterogeneous, in terms of the number of contacts, and had properties analogous to ''small-world'' networks. In order to better predict outbreaks of SARS (Severe Acute Respiratory Syndrome), Meyers *et al.* [36] obtained the contact network of an urban population using different mathematical models and through a stochastic simulation of the behaviour of the people in the population, where contacts occur randomly, in homes, schools, workplaces, hospitals and other public places. The researchers drew on population data from the city of Vancouver, British Columbia. Stochastic simulation of the behaviour of individuals belonging to large populations was also used in [37]. The researchers found that the dynamics of influenza epidemics modelled using the contact network generated from the simulation was consistent with epidemiological data from the 1957–1958 and 2009 influenza pandemics.

2) STUDIES BASED ON CONTACT DATA COLLECTED VIA SENSOR NETWORKS

Technological advances in mobile communications and sensor networks have also been applied by researchers to epidemiological monitoring. In this context, and more specifically in the epidemiological monitoring of airborne diseases, close contact is defined as two persons spending a certain amount of time at a distance of less than a given threshold. The following is a review of literature on contact networks generated in different contexts of social relationships, using different types of sensors. The methodologies followed by all these studies have the same objective, which is to obtain data useful for modelling the dynamics of infectious disease, using a compartmental SIR (Susceptible-Infected-Recovered) model [38], or to evaluate the impact of epidemiological control measures. These data are: frequency of contacts, duration, location of contacts, contact network, and contact matrices between different clusters of participants. Because they do not coincide with the aims of the research presented in this article, we have not considered studies on the tracing of contacts for epidemiological control in health crisis situations.

Isella *et al.* [39], analysed contact data from a scientific conference and a museum exhibition using RFID technology. The number of contact records analysed was 10 000 for a scientific conference and 230 000 for an exhibition. Cattuto *et al.* [40] presented a scalable, high-resolution environment for the acquisition and analysis of person-to-person contacts using RFID technology. Three use cases are described in this work: an exhibition, in which 25 people participated, resulting in 8700 contacts, and two scientific conferences, in which 575 and 405 people participated in each event, resulting in 17 000 and 60 000 contacts, respectively. Salathé *et al.* [41] proposed a mobile sensor network, based on TelosB motes to obtain the network proximity interactions (up to 3 metres) in a high school. During a month, 788 people

participated in the study and 21 489 991 interactions were recorded. Isella *et al.* [42] presented a study conducted to obtain the close contact network (up to 1.5 metres) between patients, healthcare staff and caregivers in a hospital. In this work, they used RFID technology, recording 16 000 close contacts for 7 days, in the peak period of the 2009 A/H1N1 influenza pandemic. Stehlé at al. [43] presented a study to obtain the close contact network (up to 1.5 metres) in a primary school. They used RFID technology and the participants were 242 people. The number of close contacts recorded was 77 602 for two days. Around 500 students from the Technical University of Denmark participated in a study [44], which used Bluetooth technology to record the proximity between them. The proximity records were used to analyse how the proximity between people affects the spread of an infectious disease, using a SIR model. Génois and Barrat [45] analysed the properties of contact networks at different spatial resolutions, studying the differences between real face-to-face contact networks and surrogate face-to-face contact networks obtained from co-presence data.

In the context of the COVID-19 pandemic and in order to monitor crowded environments, the use of thermal sensors installed on Unmanned Aereal Vehicles (UAV) have been proposed in [46] and [47]. A study of disease spread is presented in [48], using a SEIR (Susceptible-Exposed-Infectious-Recovered) model [49] in which people's mobility was obtained from disaggregated mobility data from mobile phone services. The study was conducted in 10 of the largest cities in the US, recording the movements of 98 million people every hour from 1 March to 2 May 2020. A comprehensive review of the use of different technological advances to mitigate the impact of this pandemic, including the use of IoT, UAV and 5G for epidemiological control purposes, is presented in [50].

B. STUDIES ON THE ROLE OF PUBLIC TRANSPORT IN THE SPREAD OF RESPIRATORY INFECTIOUS DISEASES

Public transport systems are used daily by millions of people all over the world. For this reason, epidemiological knowledge regarding these systems is of interest, both from a scientific perspective and from the point of view of epidemiological control. The following studies focus on modelling how the spread of respiratory infectious diseases occurs. Merler and Ajelli [51] analysed the spread of a Europe-wide influenza epidemic by modelling long-distance travel using data from the European railway system and relating population heterogeneity to the mobility of people in the spread of the pandemic. A simulation tool for analysing the spread of an influenza epidemic in New York City was presented by Cooley [52]. This simulation was based on a compartmental SEIR influenza disease transmission dynamics model, and drew on epidemiological parameters obtained from the 1957–1958 pandemic and simulated subway ridership for a total of approximately 8 million passengers. A study on the relationship between crowded environments in public transport systems and the spread of airborne infections was presented by Goscé and Johansson [53]. The study used epidemiological parameters of influenza-like illnesses and mobility data from the London Underground obtained from travel records generated by its passengers using an automatic payment system based on a contactless card. Troko *et al.* [54] examined whether the use of public transport is a risk factor for acute respiratory infection. The authors used epidemiological data obtained in the 2008–2009 influenza season and related it to data on bus and tram usage using multiple regression techniques.

Recently, in the context of the COVID-19 pandemic, Luo *et al.* [55] described a contact-tracing study on an outbreak in Hunan Province, China, involving 10 passengers on two public transport buses. A case of community transmission among bus passengers was reported by Shen *et al.* [56]. The authors suggested that this outbreak was due to poor vehicle ventilation. A study on the risk of COVID-19 transmission among passengers on a high-speed train system in China was presented by Hu *et al.* [57]. In this study, the authors developed a model that quantifies the risk of transmission on the basis of travel time and distance between passengers. Severo *et al.* [58] analysed the role that the urban rail system in the city of Lisbon played in the transmission of COVID-19 in said city. The authors used confirmed SARS-CoV-2 data in this city for the period from 2 March to 5 July 2020 and, using geographical data, linked the cases to the train stations closest to the homes of the infected passengers. The authors concluded that there is no relationship between proximity to train stations and illness, suggesting that socioeconomic factors affect infection dynamics.

III. METHODOLOGY

The objective of this study is to acquire information on the risk of infection on the routes operated in a PRTS. The knowledge gained can be used to identify the routes with the highest risk and to evaluate the impact of different measures to minimise this risk. To achieve this objective, data are required which, on many occasions, are not available and therefore have to be estimated by processing a large volume of data. For this reason, a data mining methodology was used. The formal framework used in the methodology, and then the methodology itself, which consists of two stages, as illustrated in Fig. 1, are set out below.

This methodology differs significantly from the methodologies employed in the studies cited in the previous section on related works. With regard to the studies that use surveys to infer the network of contacts, this methodology makes it possible to obtain a large number of samples without first having to select the elements that form the sample to be analysed. Potentially, all passengers who use the public transport system under consideration contribute with their trips to the initial sample.

Compared to studies that infer the contact network by simulating the behaviour of the study population, this methodology uses data that reflect the real movements of people and does not simulate these movements. This avoids the high

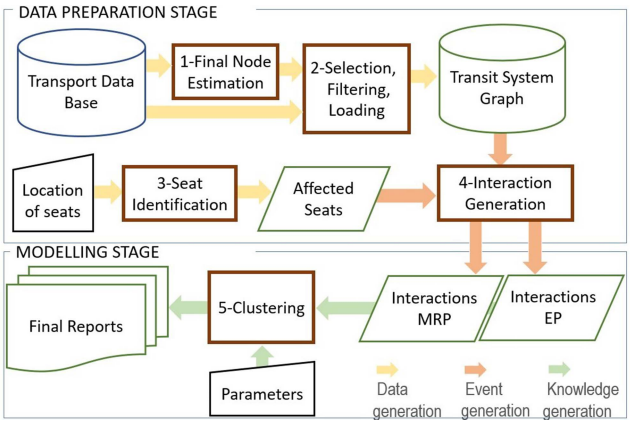

FIGURE 1. Scheme of the processes and data of the two stages of the methodology.

computational cost involved in such simulation techniques. As for the studies that use sensors to determine the contacts between people, the methodology presented herein estimates these contacts without the need for any technological implementation, as it uses data taken from the transport operations. Finally, in the specific context of public transport systems, this methodology differs from the methodologies described in the section on related works in terms of the objective pursued.

The objective of the methodology is to estimate the risk of infection among users of a public transport system based on their travel behaviour. To do this, the challenge of estimating data that are not known, but which are required to estimate the infection risk, must be addressed. Moreover, as will be seen, the methodology is complete and parametrisable, both from the point of view of the public road transport system and from the epidemiological point of view. To the best of our knowledge, these features make it an original methodology.

A. FORMALIZATION

In the context of respiratory infectious diseases, in general and at community level, a risk of infection is considered to exist when an uninfected person has been in close contact with an infected person. In the case of COVID-19, an uninfected person is considered to have been in close contact with an infected person when, within a 24-hour interval, these two persons have been within a distance of less than 2 metres for at least 15 minutes. These 15 minutes may be a single exposure or multiple exposures with a cumulative duration of 15 minutes or more [59]. In order to generalise the concept of close contact, in this methodology it is determined by three parameters: v_1, v_2 and v_3. A close contact occurs when an uninfected person has been with an infected person at a distance of less than v_1, for a time equal to or greater than v_2, in a period or time window of duration v_3. The values of v_1, v_2 and v_3 depend on the infectious disease in question. For instance, in the case of COVID-19, distance v_1 is 2 metres, cumulative time value v_2 is 15 minutes and time window v_3 is 24 hours. Based on this definition, the objective of the study is not to estimate the number of close contacts in the PRTS, since no personal passenger data are collected and therefore it is not known whether or not the passenger is infected, but to estimate the number of close interactions between passengers. In the formal framework used, an interaction is defined as the event in which two passengers physically remain in the same public transport vehicle for a period of time, at a distance of less than value v_1. When one or more interactions with a cumulative duration equal to or greater than v_2 occur between two passengers in period of duration v_3, then a close interaction event occurs between them.

For the purposes of this research, the entities of interest for the PRTS are: the transport network, the routes defined in this network operated by public transport vehicles, and the vehicle journeys made by these vehicles along these routes. The transport network is represented as a directed graph $G = G(N, A)$, where N represents the set of nodes of the network and each node of this set represents a point in the transport network where passengers can board or alight from the vehicles $N = \{n_i\}$, where subscript i is the point identifier, and A represents the set of simple arcs linking two nodes $A = \{a_i\}$, where subscript i is the arc identifier. The next entity to be defined is the route. A route is defined as the journey taken by vehicles carrying passengers. Considering graph G, a route is defined as an ordered sequence of arcs $(a_i, \ldots, a_n)$, where $a_i, \ldots, a_n \in A$. The set of routes defined in the transport network is represented by $R = \{r_i\}$, where subscript i is the route identifier. A segment of route r_i is defined as an ordered sequence of arcs $(a_p, \ldots, a_q)$ along route r_i. The entity associated with the planning of operations performed in the transport network is the vehicle journey. The set of completed vehicle journeys is represented by $J = \{J_i\}$, where J_i is the set of journeys completed on the route identified by subscript i. Alternatively, the set of vehicle journeys, irrespective of the route followed, that are completed in a time period T is represented by the notation J_T. The set of vehicle journeys that consist of carrying passengers on route i during time period T is represented by $J_{i,T}$. If instead of time period T, we have moment of time t, then $J_{i,t}$ represents the set of vehicle journeys on route i for which the start time is t. Finally, if v identifies a vehicle, then $J_{i,t,v}$ represents a vehicle journey on route i that begins at time t and is performed by vehicle v. The trip taken by a passenger on vehicle journey $J_{i,t,v}$ is defined as the route segment $(a_p, .., a_q)$ that the vehicle has travelled while the passenger is on the vehicle. The duration of the trip the passenger has made is the time elapsed since the passenger boards the vehicle at origin node a_p of the arc and alights at destination node a_q of the arc.

At this point, the concept of an interaction event between two passengers, p_1 and p_2, on the PRTS used in the methodology can be formalised. Specifically, an interaction event is said to occur if the following three conditions are met:

Condition 1. Both have travelled on the same vehicle journey, $j_{i,t,v}$.

Condition 2. The trips made by p_1 and p_2 on $J_{i,t,v}$ have at least one common arc.

TABLE 1. Notation of the formal model used by the methodology.

Notation	Meaning
n_i	Node on the transport network. Each node is associated with a stop. Subscript i is an integer value that uniquely identifies the node
N	Set of transport network nodes
a_i	Transport network arc. Each arc directly links two nodes of set N of the transport network
A	Set of arcs on the transport network directly linking two nodes
$G(N, A)$	Directed graph representing the transport network
r_i	Route on the transport network. Subscript i is an integer value that uniquely identifies the route
R	Set of defined routes on the transport network
J_i	Vehicle journey on route i.
J	Set of vehicle journeys on all defined routes on the transport network
T	Period of time
t	Moment of time in period T
J_T	Set of vehicle journeys that have been completed in time period T
$J_{i,T}$	Set of vehicle journeys on route i that have been completed in period T
$J_{i,t}$	Set of vehicle journeys on route i that started at moment t
v	Public transport vehicle
$J_{i,t,v}$	Vehicle journey on route i that started at moment t carried out by vehicle v
E_T	Set of interaction events occurring on vehicle journeys during period T
$E_{i,T}$	Set of interaction events produced on vehicle journeys on route i in period T
$E_{i,t,v}$	Set of interaction events on vehicle v on route i that started at time t

Condition 3. Passengers p_1 and p_2 have been less than v_1 metres apart during the common arcs of the trips made by p_1 and p_2 in $J_{i,t,v}$.

In addition, if during a time window of duration v_3, the cumulative duration of all interaction events is equal to or greater than v_2, then a close interaction event occurs. The interaction events that occur on all routes of the transport network during time period T are represented by E_T. The events that occur during time period T on route i are represented by $E_{i,T}$. Therefore, $E_T = \{E_{i,T}\}$. The set of interaction events occurring on vehicle journey $J_{i,t,v}$ is represented by $E_{i,t,v}$. Table 1 summarises the entities used in this formal framework.

To study the interaction events between passengers in the transport network, information is needed about the trip made by each passenger: the origin and destination nodes, the date and time of the start of the trip, and in the case of close interactions, the distance of separation from other passengers with whom he or she travelled during a vehicle journey. Most PRTSs do not use pre-assigned seating, so it is not known how far apart passengers were during the trip and in certain cases, depending on the payment system used by the passenger, their destination is not known either. Therefore, a challenge in this research was how to estimate this unknown data.

B. DATA PREPARATION STAGE

The objective of this stage is to generate the data records representing the interaction events that may occur on each of the routes of the PRTS during the selected study period T. The data structures and procedures are shown in Fig. 1. The main source of the data is the Transport Data Base (TDB), which contains all data relating to the definition of the transport network, the planning of operations and the provision of services. The Transport System Graph (TSG) is a graph database that contains, firstly, all the entities mentioned in the previous section, completed, consolidated and coherent in the study period — fundamental aspects when handling a large volume of data — to facilitate, secondly, the process of estimating interactions that are meaningful and persistent.

This stage comprises four processes. The first two processes — final node estimation and selection, filtering and loading — generate and complete the set of entities and relationships to be represented in the TSG. The first — final node estimation — estimates the destination node of the trips made by the users when necessary and will be explained in detail in Section III-B1. The second — selection, filtering and loading — encompasses all the tasks related to the generation and loading of the TSG from, on the one hand, the records contained in the TDB relating to the transport network, vehicles, users, cards, services and trips made, and on the other, the destination stops as estimated by the previous procedure, guaranteeing the reliability, accuracy, completeness and consistency of all the data. The third — seat identification — obtains, for each seat of each type of bodywork in the fleet of vehicles, the set of seats that are at a distance less than or equal to a parameter called the safety distance, based on a two-dimensional representation of the vehicle bodywork (location of seats). This safety distance may correspond both to the epidemiological parameter v_1 and to the distance threshold of the different seat allocation policies. Once the three processes described above have been executed, the data necessary for the estimation of the interaction events that take place in the vehicle journeys are generated. This estimate is obtained by means of the fourth process in this stage — interaction generation — which, based on parameter v_1 and the seat allocation simulation, which will be explained in Section III-B2, generates a record of the total estimated interactions for each of the completed vehicle journeys, composed of the fields shown in Table 2, where field NI_1 is the total number of interactions lasting 1 minute, NI_2 the total number lasting 2 minutes, and NI_m the total number of estimated interactions lasting longer in the vehicle journey. As this is an estimation process that under certain conditions performs a random allocation of vacant seats, repeated execution of this process will generate different sets of records, which are of interest in the modelling phase.

Fig. 2 shows the interaction event records of the vehicle journeys on two routes. The records represented in Fig. 2(a) correspond to those of a 22-stop route with an estimated

TABLE 2. Structure of records of estimated interactions on a vehicle journey.

Route ID	Vehicle Journey ID	Start Date time	NI_1	NI_2	NI_3	...	NI_m

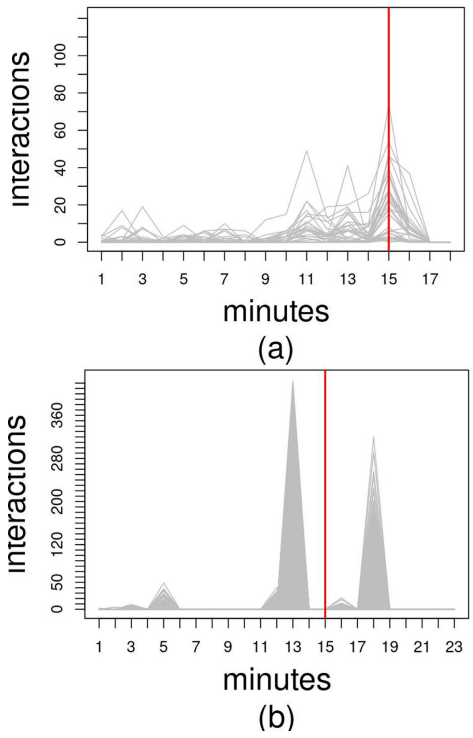

FIGURE 2. Representation of interaction events on vehicle journeys along two different routes: (a) correspond to those of a 22-stop route with an estimated journey time of 18 minutes and (b) correspond to those of a 5-stop route with an estimated journey time of 23 minutes.

journey time of 18 minutes. The records represented in Fig. 2(b) correspond to those of a 5-stop route with an estimated journey time of 23 minutes. The horizontal axis represents the duration, in minutes, of the interaction events and the vertical axis represents the estimated number of interaction events. Each grey curve represents the estimated interaction event records for a vehicle journey on the route. The red vertical line identifies the boundary of the number of events lasting 15 minutes or more, above which close interactions are considered.

1) DESTINATION STOP ESTIMATION PROCESS

The objective of this process of the methodology is to solve a problem that frequently arises in data mining projects. This problem consists of handling data sets with missing values. In a general context, this problem is addressed by Dinh *et al.* [60] by proposing a novel method, called Clustering Mixed Numerical and Categorical Data with Missing Values (k-CMM), to classify datasets with a high number of missing values. In the specific context of traffic accident data analysis, this challenge has been addressed by Deb and Liew [61], who proposed a method based on decision trees. Considering previous works that address how to estimate the destination stop [62], [63], a procedure was developed to infer the final destination of the trips made by passenger p — from one of the two categories above — when this information has not been recorded.

With the technologies commonly used by intercity road transport services, it is possible to obtain information about the trip made by passengers — at which node they started, which vehicle they used and at which moment in time they boarded the vehicle — but the end point and the duration of their trip are not always recorded. This problem can be overcome in the case of frequent travellers because they generally use specific personal payment systems, such as contactless cards, which automatically record payment transactions and identify the user. There are several types of frequent users, among which the most common are:

- Passengers that make multi-stage trips, such that the end node of one stage (transfer node) is close to the start node of the next stage.
- Passengers who make single-stage trips to their place of work, study, public service or leisure and who also return using the PRTS.

These types of trips exhibit a common pattern: on two consecutive trips made by the same passenger, the destination node of the first is located within a short distance of the origin node of the second. This proximity will be determined by a distance threshold depending on the type of transport network, smaller in the case of urban transport and larger in the case of intercity transport. This procedure is based on the known data for two consecutive trips made by p. For each trip made by p on vehicle journey $J_{i,t,v}$, node n at which p started the trip and time t' of the beginning of the trip are known, where node n is an origin node of one of the arcs forming the sequence of arcs $(a_p, \ldots, a_q)$ that form the segment of route i travelled on $J_{i,t,v}$. Moreover, $t \leq t'$, meaning that the start of the user's trip t' is equal to or later than the start of vehicle journey t. The purpose of the procedure is to ascertain the final stop of the trip made by p on $J_{i,t,v}$ and, therefore, the sequence of arcs that form the segment of route i travelled by p. To estimate final stop q of journey J_{i_1,t_1,v_1}, the procedure uses the known data for the next trip made by p. If J_{i_2,t_2,v_2} is the next trip made by p, then node n_2 and time t'' at which he or she started the journey are known. If nodes n_1 and n_2, the starting nodes of the two vehicle journeys, are not the same, and are not within a distance threshold that determines that they are similar (on both sides of a two-way road, at an intersection, or are close consecutive nodes on the same route), then final stop q of the trip made by p on J_{i_1,t_1,v_1} would be the stop on route i_1 closest to stop n_2 at which p started the trip on J_{i_2,t_2,v_2}, provided that this final node q is at a distance from n_2 not greater than the proximity threshold indicated above, that is, it is not too far away. Once the final stop has been deduced, the time of the trip made by p will be the sum of the time taken by v to traverse the sequence of arcs $(a_{n_1}, \ldots, a_{q_1})$.

Fig. 3 illustrates this procedure. It represents, by means of a graph, a generalisation of the procedure in the case of

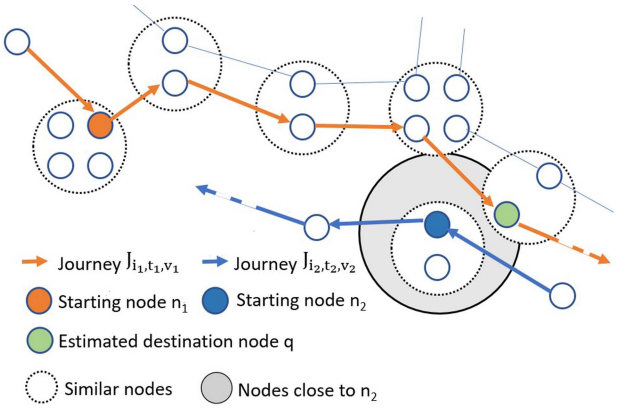

FIGURE 3. Procedure for estimating the destination stop of a trip.

two consecutive trips in time of passenger p, trip i and trip $i+1$. The orange arcs represent the route of the scheduled service used by the passenger on trip i, and the node at which the passenger starts this trip is highlighted. The blue arcs represent the route of the scheduled service used by the passenger on trip $i+1$, and the node at which the passenger starts this trip is highlighted.

The green node is the estimated end node of trip i (the objective of the procedure). The nodes encircled by a dotted line represent nodes considered similar due to their geographical proximity. The grey shaded area denotes the maximum distance for determining nodes close to the start node of trip $i+1$. The algorithmic description of the procedure is presented in Algorithm 1.

To validate the proposed method, it was applied to a set of trips where the destination stop is known and the result of the estimation for each trip was compared with the known destination. The test dataset was obtained from the transport system selected as a use case in Section IV, where the results are presented. The data contained in this test dataset correspond to the records generated from trips made using a contactless card as a means of payment and a fare option that requires the passenger to check in at the start of the trip and to check out upon arrival at the destination, where the destination stop is recorded. The number of trips in this test dataset was 278 694. For this set of trips, the proposed method estimated the destination stop in 205 183 cases (73.6 %) and failed to do so for 73 511 trips (26.3%).

Table 3 presents the numbers of trips for which the destination stop was estimated as a function of the Euclidean distance between the estimated stop and the known stop. The first column shows the distance between the estimated stop and the known stop (D). The second column shows the number of trips (NT) in which the destination stop was estimated as a function of the value of D, and the percentage of these trips in relation to the total number of trips in which the destination stop could be estimated.

In the parameterisation of the destination stop estimation algorithm, the value used for the DSmax parameter — which represents the distance threshold for considering two stops

Algorithm 1 Estimating the Destination of a Trip

Input data:

- Vehicle journey J_{i_1,t_1,v_1} taken by passenger p
- Node n_1 at which p started journey J_{i_1,t_1,v_1}
- Time t' at which p started journey J_{i_1,t_1,v_1}
- Next vehicle journey J_{i_2,t_2,v_2} made by the passenger
- Node n_2 at which p started journey J_{i_2,t_2,v_2}
- Maximum distance DPmax at which two nodes are considered to be close
- Maximum distance DSmax at which two nodes are considered to be similar

Goal:

- Node q, estimated destination of p on vehicle journey J_{i_1,t_1,v_1}

if Euclidean distance between n_1 and $n_2 >$ DSmax **then**
  Obtain sequence of arcs of route i_1 starting at node n_1. Output data for this step: sequence of arcs $(a_{n_1}, \ldots, a_{q1})$ that form the largest possible segment of route i_1 travelled by p
  for each route arc of the sequence $(a_{n_1}, \ldots, a_{q_1})$ **do**
    Obtain the Euclidean distance between the destination node of the route arc and node n_2. Output data for this step: sequence of distances $d_{a_{n_1}}, \ldots, d_{a_{q_1}}$
  end for
  Obtain the minimum value dmin of the sequence $d_{a_{n_1}}, \ldots, d_{a_{q_1}}$ and arc a_{j_1} in which this value has been obtained. Output data for this step: destination node q of arc a_{j_1}
  if (dmin < DPmax) and (Euclidean distance between n_1 and $q >$ DSmax) **then**
    The estimated destination stop of p on journey J_{i_1,t_1,v_1} is the final stop of arc a_{j_1}
  else
    The destination stop cannot be determined. There is no near stop to the starting stop of the next journey, or it is similar to the starting stop of the previous journey
  end if
else
  The destination stop cannot be determined. The starting stops are the same or similar
end if

to be similar — was 500 metres; the value of the DPmax parameter — which indicates when two stops are close to each other — was 1 km. Considering that in the case of intercity transport, stops are spaced along the length of a route, these distance thresholds are reasonable and conservative.

As can be seen in Table 3, for all trips for which the destination stop was estimated, in 71.6% the estimated destination stop was less than 1 km from the actual destination stop. Considering the results of this validation test, this parameterisation of the DSmax and DPmax values makes it possible to obtain an estimate of the destination stop for a

TABLE 3. Numbers of trips for which the destination stop was estimated. The first column shows the distance between the estimated stop and the actual stop (*D*). The second column shows the number of trips (NT) for which the destination stop was estimated as a function of the value of *D*.

D (km)	NT
0	98 926 (48.21%)
$0 < D < 0.5$	20 630 (10.05%)
$0.5 \leq D < 0.75$	21 939 (10.69%)
$0.75 \leq D < 1$	5452 (2.65%)
$1 \leq D < 2$	12 114 (5.90%)
$2 \leq D < 3$	6373 (3.10%)
$3 \leq D < 4$	9967 (4.85%)
$4 \leq D < 5$	3733 (1.81%)
$5 \leq D$	26 049 (12.69%)

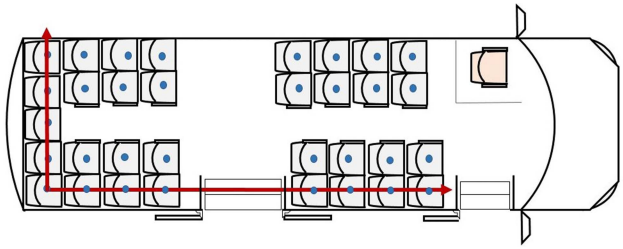

FIGURE 4. Representation of the passenger areas of a vehicle.

significant proportion of the trips for which this information is not available.

2) SIMULATION OF SEAT ASSIGNMENT ON A VEHICLE JOURNEY

In intercity PRTSs that are not long distance, it is not always possible to know the distance between passengers during a vehicle journey, since passengers are not assigned a seat when they travel and, therefore, when passengers board the vehicle, they can occupy the vacant seat of their choice. To address this challenge, it is assumed that all passengers are seated in one of the available vacant seats (when the vehicle capacity is exceeded, the process rules out new passengers until a seat becomes available) and that they remain in the same seat during their journey, and simulates the choice of seat based on various assumptions about passenger behaviour. The way the procedure is implemented is shown below.

In the methodology, the distance between two passengers travelling in a vehicle is defined as the distance between the centre points of the seats they occupy. In order to systematise the process of obtaining the seat centre points and thus automatically obtain the distances between seats, a two-dimensional representation model of the vehicle space for passengers has been developed that takes into account the wide variety of bodywork types used in intercity PRTSs. Fig. 4 shows a representation of the passenger zones of one of the vehicle types considered in this study in two of the vehicle types considered in this study. A reference system to obtain the coordinates of the centre of each seat is also shown in red in this figure.

In this research, two alternative seat allocation policies were considered. The first is the Empirical Policy (EP). This policy is based on observed behaviour whereby a passenger prefers not to sit next to another passenger, without any other consideration. The second policy aims to reduce the risk of infection and is called the Minimise Risk Policy (MRP). It consists of assigning the user to the free seat that is more than 2m away from the largest number of passengers, in order to avoid as many interactions as possible with passengers on board the vehicle when boarding. In both policies, if the occupancy of the vehicle does not permit strict application of the allocation criterion, then a seat is randomly allocated from the vacant seats that are in the best circumstances according to the allocation policy used.

The allocation procedure is based on three parameters the values of which vary according to the allocation policy. The first parameter is the *safety distance*, which is determined by the allocation strategy. The second parameter is the *affected seats list*, which is a list associated with each seat of each type of bodywork in the vehicle fleet that contains the list of seats that are affected by its occupancy, and which is directly dependent on the value of the safety distance parameter. The third parameter is the *risk potential*, which is a value assigned to each of the vacant seats in the vehicle during the course of a vehicle journey and which determines its potential risk: it increases as the seats in which it appears in the affected seats list are occupied and decreases when any of these seats are vacated.

The procedure simulates seat occupancy by passengers on each vehicle journey $J_{i,t,v}$, taking as input parameters the *affected seats list* pertaining to the vehicle bodywork type, and the *safety distance* of the policy to be applied and the origin and destination stops of each of the trips made by the passengers on that vehicle journey. Following the route order established for that vehicle journey, each stop is treated by the procedure in the following way: first, it vacates the seat of the passengers arriving at their destination and assigns it the corresponding *risk potential* according to the occupancy of the affected seats, lowering the *risk potential* of the seats that are vacant in the list of affected seats, and then it allocates the passengers starting their trip a seat with the lowest *risk potential* among those that are randomly vacant, increasing the *risk potential* of the vacant affected seats. The algorithmic description of the seating procedure is described in Algorithm 2.

C. MODELING STAGE

In general, in a data mining project, the modelling phase is designed to generate new knowledge, applying techniques of varying nature — both statistical and machine learning — depending on the type of problem posed. As has already been noted, the objective of the methodology is to obtain information by detecting the patterns followed by interaction events between passengers on the different routes of the PRTS over a given period of time. To obtain these patterns, a clustering process was implemented (see Fig. 1),

Algorithm 2 Assignment of Seats in a Vehicle During a Vehicle Journey

Input data:

- Safety distance. In the case of EP, this is the minimum distance between the centres of two adjacent seats, and in the case of MRP, it is 2 metres.
- Affected seats list. This is a list for each seat in each bodywork type in the fleet, showing the number of seats that are affected by occupancy of the seat. This list depends directly on the value of the safety distance parameter as determined by the allocation policy used.

Goal:

- Potential risk of a seat. This is a value that is assigned to each of the free seats in the vehicle during the course of a vehicle journey. The value increases as the seats that appear in the affected seats list are occupied and decreases when any of these seats are vacated.

When a vehicle journey, $J_{i,t,v}$, begins, the initial risk potential value is assigned to all the seats in the vehicle. This initial value is the minimum, as it is assumed that there are no passengers in the vehicle.
At each stop the vehicle makes during the vehicle journey:

for each user that alights from the vehicle **do**
  Their seat ap is vacated and the minimum risk potential value is assigned.
  for each seat in its affected seats list **do**
    if the seat is occupied **then**
      The risk potential of the newly vacated seat ap increases.
    end if
  end for
  for each user boarding the vehicle **do**
    They are randomly assigned one of the seats with the lowest risk potential on the vehicle.
    for each seat af in its affected seats list **do**
      if the seat af is free **then**
        The risk potential of seat af increases.
      end if
    end for
  end for
end for

which takes into account certain parameters and is based on the estimation of these events made by the interaction generation process. In schematic terms, it performs three tasks: generation of the different sets of input data for the modelling, modelling of each of these sets and, finally, creation of reports with the results.

1) GENERATION OF THE DATASET TO BE MODELED

This task is conditioned by different parameters. Noteworthy among the parameters that determine the spatial and temporal limits of the scope of the study is that related to the discretisation of the duration of the interactions: the data records of the estimated interactions have a temporal granularity of 1 minute, but the analysis can be carried out with a greater granularity — 5 minutes, 10 minutes, and so on — depending on the type of routes or the ultimate objective of the study.

Therefore, the interaction events on vehicle journey $J_{i,t}$, that is, each field of record $E_{i,t}$, are accumulated in intervals of k minutes, giving rise to an array of n integer values, $E_{i,t}[n]$. A second relevant parameter is that which determines the number of generations of estimated interactions to be considered at this stage. If there are more than one, the final array $\hat{E}_{i,t}[n]$ will be calculated as the arithmetic mean of the records created for each vehicle, that is, if G is the number of generations to be processed and $\hat{E}_{g,t}[n]$ corresponds to the estimated events in generation g, then the final interaction record will be:

$$\hat{E}_{i,t}[n] = \frac{\sum_{g=1}^{G} E_{g,i,t}[n]}{G} \tag{1}$$

Finally, if in period T there have been N vehicle journeys, at moments of time $t_1, t_2, \ldots, t_N$ of vehicle journeys on route i, then the overall representation of the interaction events of that route in that period, $E_{i,T}$, is obtained from the expression:

$$E_{i,T}[n] = \frac{\sum_{n=1}^{N} \hat{E}_{i,t_n}[n]}{N} \tag{2}$$

That is, it is obtained by dividing the estimated number of interactions in all vehicle journeys by the number of completed journeys.

2) MODELING

The objective of this stage of the methodology is to obtain information to assess the risk of infection on the different routes of the transport network, based on the interaction event records $E_{i,T}$ described above. From the definition of the data record $E_{i,T}[n]$ expressed in (2), epidemiological information of interest can be extracted for each route for period T. Specifically, $\mathrm{ME}_{i,T}$ which is the estimated number of interaction events on route i will be determined by (3), where n is the number of elements in the record. The maximum value of $E_{i,T}[n]$ reflects the most likely interaction event duration. For close interaction events, which as mentioned above depend on epidemiological parameter v_2, if k is the duration of the interval used to define the $E_{i,T}$ records, then index w of the $E_{i,T}$ record to which close interaction events correspond is obtained by (4). For example, for COVID-19, this index would have a value of 4, since the value of v_2 for this disease is 15 minutes and k is 5 minutes. $\mathrm{CE}_{i,T}$, that is, the average number for vehicle journeys on route i, will be determined by (5).

$$\mathrm{ME}_{i,T} = \sum_{i=1}^{n} E_{i,T}[i] \tag{3}$$

$$w = \mathrm{INT}(v_2/k) + 1 \tag{4}$$

$$\mathrm{CE}_{i,T} = \sum_{i=w}^{n} E_{i,T}[i] \tag{5}$$

The technique chosen to obtain patterns for the interaction events was the clustering of similar objects, for which the specific algorithm and the metric used to validate the results can also be parameterised. Since there may be a considerable disparity of routes in a transport network, conditioned both by type and duration, it is also possible to parameterise a prior classification of the routes into R sets, for example by the planned duration of the vehicle journeys, in order to minimise the distortion of the distances in this clustering procedure. This classification yields different $E_{R,T}$ sets that will be the input data for each clustering process.

Another important parameter of this modelling phase is the total number of clusters to be generated in each $E_{R,T}$ set. The main consequence of this parameterisation process is that the number of clusters will probably not be optimal for all cases, but it is necessary since one of the main objectives of this study is to compare the interactions resulting from the application of different seat allocation strategies.

Once each $E_{R,T}$ dataset and the number of clusters to be generated have been determined, the chosen algorithm will be run, resulting in different clusters where the elements $E_{i,T}$ that are part of the cluster are similar, and where the centroid of each cluster represents the elements that are part of the cluster. In the methodology, the interaction event record for the centroid of each cluster obtained is represented by $C_{R,T,l}$, where subscript l is the cluster identifier.

3) ANALYSIS OF RESULTS: GENERATION OF KNOWLEDGE

Once the clustering and evaluation procedures have been carried out for each set of interaction data, determined by the prior classification of the routes and by the seat allocation policy, we proceed to the analysis of the results, which may vary in nature. First, there are the centroids of each of the clusters, for which the record $C_{R,T,l}$ is formed by the average number of interactions for each of the defined intervals of duration, determined by parameter k. Each centroid, together with the routes similar to it, provide relevant information on the average interactions of different durations. More specifically, by applying (3)–(5) with the data records of each centroid, information becomes available for all the routes belonging to the same cluster.

IV. RESULTS

The proposed methodology was applied to the intercity PRTS on the island of Gran Canaria (Canary Islands, Spain). This transport system is operated by the company *Global Salcai-Utinsa*, which annually transports around 20 million passengers and covers 25 million kilometres. The time period studied was the month of December 2019, two months before the COVID-19 pandemic was declared. The decision to select this month was made because in this period demand was not affected by the travel restrictions imposed by the health authorities as a result of the state of emergency.

TABLE 4. Some entities and instances of each uploaded to the TSG.

Entity	Meaning	Number of instances
N	Set of transport network nodes	2586
A	Set of arcs on the transport network directly linking two nodes	6155
J_T	Set of vehicle journeys that have been completed in time period T	70 734
v	Public transport vehicle	443

TABLE 5. Duration and total number of routes in each category.

Route group	Duration d (min)	Number of routes
R_1	$d < 25$	109
R_2	$25 \leq d < 34$	107
R_3	$34 \leq d < 47$	106
R_4	$d \geq 47$	118

A relational database was used to implement the methodology, with the relevant data required for this study from the operator's transport database, Neo4j, to implement the graph database used by the methodology, and the RStudio development environment [64] for programming the procedures used in the data preparation and modelling stages.

In the study period, 440 different routes were identified on the transport network, with a total of 70 734 vehicle journeys made. The number of passenger trips made in this period was 2 260 744. Of these trips, 1 101 338 recorded the origin stop and the destination stop, and 1 159 406 did not, so the process of estimating the destination stop described in Section III-B1 was applied to this set of trips. As a result of this process, an estimation of the destination stop could be completed on 860 909 trips; this was not possible on 298 497 trips. Finally, the process of selection and filtering of records resulted in a total of 1 797 107 trips being loaded into the TSG, and these were used to estimate passenger interactions according to the two seat assignment policies described in Section III-B2. Table 4 illustrates these data by associating them with the entities defined in the formalisation described in Section III-A.

Once a complete set of transport activity data was obtained and represented in the TSG, the remaining processes of the methodology were implemented by adopting a series of decisions based on aspects related to the transport network, epidemiological aspects and the modelling technique used.

In relation to the transport network, firstly, the routes were classified depending on the time taken to complete them, generating four subsets, four categories of routes R_1, R_2, R_3 and R_4 with the following characteristics: subset R_1 contains routes which take less than 25 minutes to complete, R_2 routes which take more than or equal to 25 minutes and less than 35 minutes, R_3 routes which take between 35 and 47 minutes, and R_4 routes which take more than or equal to

TABLE 6. Total number of routes in each category.

Area	Number of routes
N	148
S	201
C	43
M	48

47 minutes to complete. The maximum duration of a route in the transport network is 137 minutes.

The duration and number of routes in each group is shown in Table 5. Thus, the number of interaction event duration intervals is the same for all routes belonging to the same subset. Secondly, and also related to the routes, these have been subdivided into four groups, according to the geographical area through which they pass: N for the routes that run through the north, S for those that run between the capital, the east and the south, C for those that run between the capital and the central area, and M for the routes that, without passing through the capital, run between the south, the north and the centre.

The reason for this decision is to analyse the patterns of interaction events according to the geographical areas through which the route services pass. The total number of routes in each area is shown in Table 6.

As for other parameters, epidemiological parameter v_1 was set to 2 metres, v_2 to 15 minutes, the safety distance of the EP policy to 0.5 metres, and that of the MRP policy to 2 metres, and the duration of interactions was discretised into k intervals of 5 minutes. For each policy, 3 generations of data were estimated.

Lastly, a clustering modelling technique was chosen to identify the possible interaction profiles. The specific technique used was the k-means algorithm, a widely used unsupervised algorithm that appears to give partitions which are reasonably efficient in the sense of within-class variance, is easily programmed and is computationally economical [65]. The process subdivides the n input data records into k partitions where each is associated with the partition nearest to its mean, where the mean of each partition is its significant element and its centroid, the profile that characterises it. To evaluate the quality of the clusters that were obtained the silhouette was used [66]. This value measures the degree of cohesion of the elements that make up the cluster, so that the greater the cohesion of a cluster, the closer its centroid will be to each element of the set, and therefore the more representative it will be. It takes values in the interval $[-1, 1]$, so that the value -1 indicates a cluster with the lowest degree of cohesion and the value 1 indicates a cluster with the highest degree of cohesion. For the sake of clarity in presentation, the number of clusters for all datasets is set to 3. Having described how the methodology was implemented, the results are presented below. Figs. 5–8 show the clusters C_1, C_2, C_3 obtained with the data generated by applying the EP (Figs. 5(a)–8(a)) and MRP (Figs. 5(b)–8(b)) seat allocation policy in each of the four defined route categories.

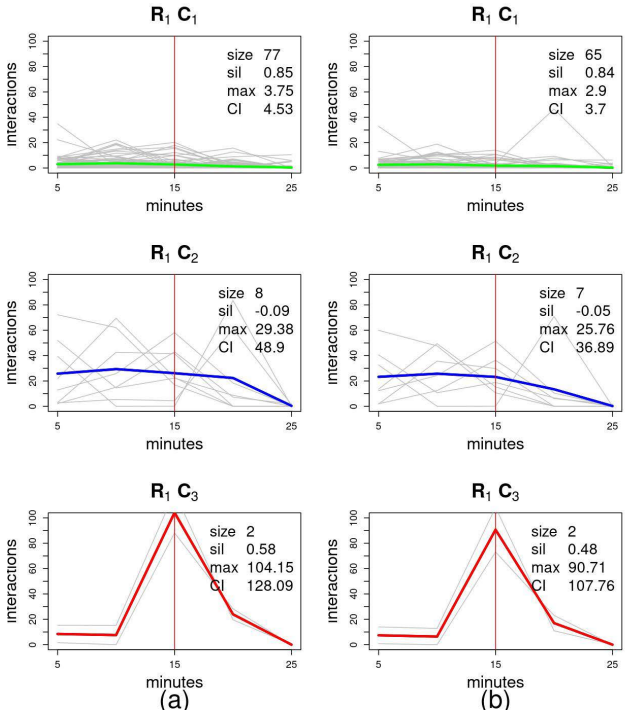

FIGURE 5. Plots of clusters and centroids obtained by applying (a) policy EP and (b) policy MRP to group of routes R_1. In these plots, the centroids of clusters C_1, C_2 and C_3 are represented by green, blue and red curves respectively.

The same criterion of presentation has been used in all of them: each column represents the three clusters obtained for each policy, ordered by the number of routes they contain. In each cluster, the curve representing the centroid obtained by applying the k-means algorithm is drawn. In the k-means algorithm, the centroid of a cluster represents its most significant value and corresponds to the mean value of the elements that form the cluster. The cluster with the green centroid is the most numerous, the cluster with the blue centroid is the second most numerous, and the cluster with the red centroid is the least numerous. In all the graphs, the horizontal axis represents the discretised duration of the average number of interactions per vehicle journey. The red vertical line identifies the boundary of the mean number of events lasting 15 minutes or more, above which close interactions are considered. In addition, the legend of each of the graphs includes four values that are considered significant for analysis purposes: the total number of routes belonging to the cluster (size), the value of its silhouette (sil), which quantifies the coherence of the cluster, the maximum value of average interactions of the profile obtained (max), and finally, the sum of its average interactions with a duration greater than or equal to 15 minutes, which may be considered a metric for quantifying the total number of close interactions (CI) that may occur in each cluster.

The plots in Fig. 5(a) show the results of the data clustering procedure for the R_1 set of routes (routes which take less than 25 minutes to complete), when interactions were estimated

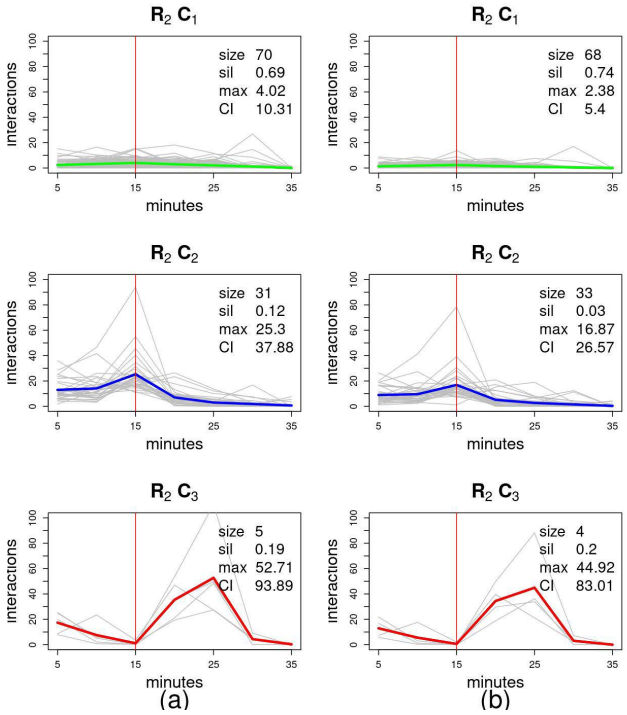

FIGURE 6. Plots of clusters and centroids obtained by applying (a) policy EP and (b) policy MRP to group of routes R_2. In these plots, the centroids of clusters C_1, C_2 and C_3 are represented by green, blue and red curves respectively.

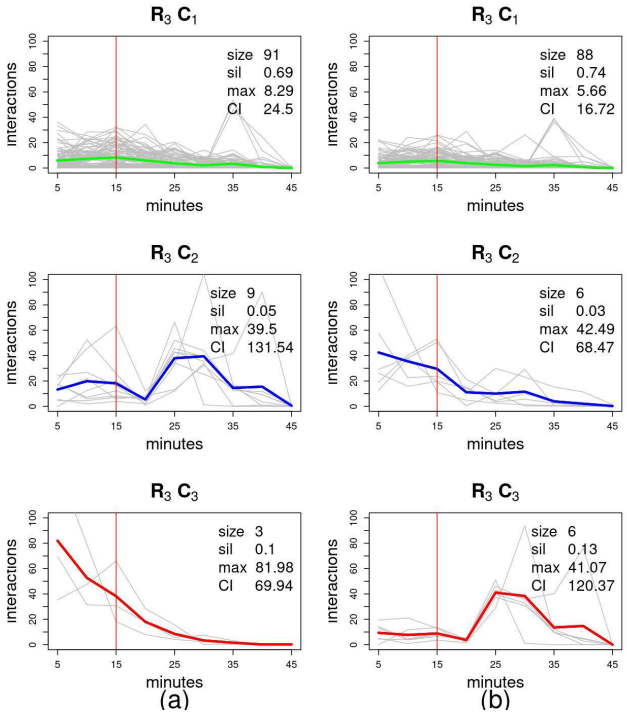

FIGURE 7. Plots of clusters and centroids obtained by applying (a) policy EP and (b) policy MRP to group of routes R_3. In these plots, the centroids of clusters C_1, C_2 and C_3 are represented by green, blue and red curves respectively.

by applying the EP seat allocation policy. Of the total in this category, the estimation process resulted in some interaction on 87 routes, representing 80% of the routes. In the remaining 20%, no interaction record was generated in the three simulations performed, as these were routes with a low number of vehicle journeys and passengers. As mentioned above, the three clusters generated are presented in order of size from largest to smallest. In this case, cluster C_1 contains approximately 88% of the routes and is quite cohesive, with the highest silhouette value of the three. As for the curve representing its centroid, with the values (3.01 3.75 2.79 1.31 0.43), it can be observed that it is nearly a horizontal line, reaches its maximum value of 3.75 when the relative interactions per vehicle journey have a duration of 10 minutes and, when only narrow interactions are considered, it is characterised by the value 4.53, corresponding to the total number of interactions with a duration greater than or equal to 15 minutes. Cluster C_2 contains routes on which interactions exhibit disparate behaviour — its silhouette value is very low — unlike cluster C_3 which, with only two routes, contains those with the highest number of close interactions in the set, a total of 128 per vehicle journey.

The plots in Fig. 5(b) represent the results when the MRP seat assignment policy is applied to the same set of routes, and significant differences are observed with respect to the EP policy. The first is that the number of routes with estimated interactions decreases from 80% to 67%, that is, out of 109 routes in the set, records are generated in 74 routes. The second is that the maximum centroid values decrease by 22% in C_1, the largest cluster, and by about 12% in C_2 and C_3 respectively. And the third, closely related to the preceding observation, is that the values characterising the centroids also decrease in C_1, C_2 and C_3, by 20%, 24% and slightly more than 16% respectively. Again, cluster C_1 has the highest coherence and C_2 contains the most disparate route profiles.

Plots (a) and (b) in Fig. 6 show the results of clustering the R_2 category data (routes which take more than or equal to 25 minutes and less than 35 minutes) using the two defined policies. In this case, there is hardly any reduction in the total number of routes affected by interactions, but there is a significant reduction in the estimated close interactions per vehicle journey in the results in (b) compared to those in (a), which is around 47% in the largest cluster C_1, 33% in cluster C_2 and 11% in cluster C_3.

The results for set R_3 (routes which take between 35 and 47 minutes), with 106 routes, are shown in Fig. 7. In this case, between 3 and 5 routes have no estimated interactions, and in cluster C_1 a 31% reduction in interactions is observed when the MRP seat assignment policy is applied. In clusters C_2 and C_3 there is a regrouping of routes, all of them with a rather low coherence.

Finally, the results for the 118 routes in the last set R_4 (routes which take more than or equal to 47 minutes to complete), which contains the routes with the longest journey times of more than 47 minutes, are presented in Fig. 8.

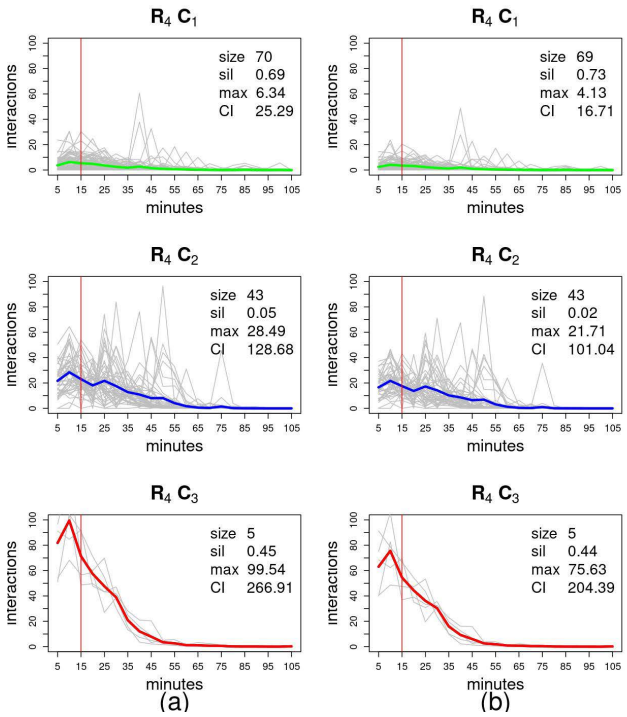

FIGURE 8. Plots of clusters and centroids obtained by applying (a) policy EP and (b) policy MRP to group of routes R_4. In these plots, the centroids of clusters C_1, C_2 and C_3 are represented by green, blue and red curves respectively.

TABLE 7. Total number of routes in each area and group without (w/o) and with estimated interactions in each policy.

		Area							
		N		S		C		M	
	Route	w/o	with	w/o	with	w/o	with	w/o	with
EP	R_1	2	40	10	37	9	4	1	6
	R_2		45	1	43		9		9
	R_3		36	3	38		11		18
	R_4		25		69		10		14
MRP	R_1	9	33	14	33	9	4	3	4
	R_2	1	44	1	43		9		9
	R_3	2	34	4	37		11		18
	R_4		25		69		10	1	13

Except for one, all of them give rise to some interaction in the different simulations, and when comparing the results obtained by the two seat allocation policies, again in the most conservative — the MRP policy — there is a significant reduction in the number of interactions compared to the EP, which is around 34% in the first cluster, containing about 60% of the routes in this set, 21% in the second cluster and 23% in the third cluster.

Tables 7 and 8 show these results for the different geographical areas into which the transport network was subdivided, also distinguishing between the two policies applied. As a first approximation, Table 7 shows the total number of routes in each area and each group for which interactions were not estimated and for which interactions were estimated, depending on the policy applied. It can be seen that there is no significant decrease in the number of routes on which

TABLE 8. Distribution of the routes of each group in each area and each cluster.

		Area											
		N			S			C			M		
	Route	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
EP	R_1	40			28	7	2	3	1		6		
	R_2	38	7		20	18	5	6	3		6	3	
	R_3	35	1		27	8	3	11			18		
	R_4	20	5		26	38	5	10			14		
MRP	R_1	33			25	6	2	3	1		4		
	R_2	36	8		20	19	4	6	3		6	3	
	R_3	33	1		26	5	6	11			18		
	R_4	20	5		26	38	5	10			13		

interactions are not estimated when the more conservative seat allocation policy is applied, with the exception of the R_1 route category in the northern part of the transport network, where the number of routes with interactions decreases by just over 17%, from 40 to 33.

Table 8, by contrast, shows the distribution of the routes in each of the geographical areas and each category in the clusters obtained. Although no substantial decreases are observed when applying the different policies, it does reflect data concerning the type of route in each area of the transport network, such as, for example, the fact that almost half of the routes in the south zone have a profile with a high number of close interactions.

V. DISCUSSION

The estimated interactions, as presented in this paper, provide new knowledge in two ways: on the one hand, about the interactions that may be occurring in the transport network, and on the other hand, the extent to which these are affected by applying different seating policies. This provides a way of measuring the effect of implementing rules or procedures to determine passenger locations in order to reduce contact between people. It should be noted that the results refer to estimated interactions over the entire study period, without distinguishing between different types of day (e.g. working or non-working) or between different time bands, which is a higher level of detail and is covered by the proposed methodology.

The EP policy, where a passenger prefers to sit in a seat where the surrounding seats are unoccupied, determines the minimum threshold of interactions in systems where no seat allocation is applied, as it does not take into account people travelling together or the preferences of certain age groups. For this reason, the results obtained by applying this policy can be considered a measure of the interactions that, at the very least, are occurring in the vehicle journeys, both at network level and at the level of individual routes. From the results obtained with the records of the three simulations carried out with this policy, it can be seen that in Table 7, of the 440 routes of the transport network, in 26 no interaction is estimated, which represents 6%, and it is area C which has the highest proportion of routes with no interactions, more than 25%. In general, these are routes with a low number of

TABLE 9. Averages obtained by applying the two policies in cluster C_1.

	Size			Close Interactions (CI)		
	EP	MRP	reduction	EP	MRP	reduction
R_1	77	65	15%	4.53	3.7	18.3%
R_2	70	68	2.8%	10.31	5.4	47.6%
R_3	91	88	3.2%	24.5	16.72	31.75%
R_4	70	69	1.4%	25.29	16.71	33.92%

passengers and vehicle journeys, and almost all of them have short routes, with journeys of less than 25 minutes.

In Table 8, of the 414 routes with estimated interactions, firstly, area N stands out, with a generally low interaction profile, since more than 90% of its routes are grouped in C_1. Secondly, area S, where the routes with the longest duration, those included in R_3 and R_4, have a greater weight and where the profiles with the highest number of close interactions are also found; more than 80% of the total number of routes grouped in C_2 and C_3 are in this area. Finally, areas C and M, with a smaller number of routes, of medium-long duration and which, for the most part, are grouped in the clusters with the lowest interaction. As for the routes in the clusters with the longest interactions, for example, those found in area C_3 of all the groups of routes, different types of routes can be observed, some with less than 20 vehicle journeys in the month of the study and others with more than 1000 vehicle journeys. To be able to draw conclusions in these cases, it would be necessary to apply greater temporal granularity to the records for the period of study (at the level of days of the week and/or time bands) in order to identify the possible causes.

As for the effects of applying a more conservative seat allocation policy, in order to minimise interactions between passengers, this methodology proposes a way of quantifying it, based on two metrics associated with the clusters that are generated: the total number of routes grouped in each cluster (size) and the average number of estimated close interactions (CI). As an example, and by way of summary, Table 9 shows those obtained in cluster C_1, the most numerous cluster as it contains 75% of the affected routes, where it can be seen that, while the number of affected routes decreases significantly only in the shorter routes, the reduction in the number of close interactions is significant in all types of routes, especially among those with a duration of between 25 and 34 minutes.

VI. LIMITATIONS OF THE STUDY

This section describes the limitations of this study. The first is that it assumes that there is a risk of infection between two people when they are in close contact, and does not consider the risk of transmission by aerosol or fomite. Therefore, the methodology used could only be applied in the case of diseases where the main mode of transmission is close contact, as is the case with COVID-19 [67]. A second limitation is that it is applied in intercity road transport systems and assumes that all passengers are seated. For this type of transport system, this assumption is not a serious limitation, since standing is usually not permitted for safety reasons. The methodology followed would not, however, be applicable to the case of urban public road transport, where standing is permitted and is common. In the context of a pandemic, it is common to limit vehicle occupancy in this type of transport using criteria that are not based on objective parameters. The proposed methodology could therefore be applied to obtain information that would facilitate the planning of transport services with the aim of reducing the risk of infection based on a calculation of capacity using objective parameters, as opposed to simply reducing capacity by an arbitrary amount. Another limitation is that it is assumed that there is a risk of infection in vehicles when two passengers are on the same vehicle at the same time. Therefore, the presence of two passengers at the same stop on the transport network has not been considered. In the case of intercity public road transport, this limitation is of relative importance for two reasons. The first reason is that this type of transport is planned around timetables, which means that passengers arrive at a stop a few minutes before catching the vehicle in which they will be travelling, and it is not common for them to spend long periods of time at the stops. The second is that most of the stops on this type of transport system are located outdoors, thus reducing the risk of infection. The final limitation is that since the passenger's seat in the vehicle is not known, the location of the passenger was simulated based on a seating allocation policy. The importance of this limitation is also relative, since the objective of the study was to learn on which routes and at what times the risk of infection is greatest. In this study, the policy applied was an EP policy, the aim of which is to approximate the passenger's seating behaviour. In reality, close interactions are likely to be greater, as the possibility that passengers may be travelling together is not taken into account. However, for the purposes intended, this limitation does not invalidate the information obtained. Moreover, by simulating the location of passengers in vehicles, it is possible to assess the impact of different seating strategies designed to minimise the risk of infection and maximise the available vehicle capacity.

VII. CONCLUSION

This article presents the results of a research project designed to gather information about the risk of infection on the routes of an intercity road transport system. This information can be used to identify the routes with the highest risk and to assess the impact of different measures to minimise this risk. To achieve this objective, a data mining methodology was used. The results were obtained by analysing a real case of a transport system where the data from an intercity transport operator on the island of Gran Canaria was analysed for the month of December 2019.

The results provide new insights into the interactions that occur between passengers in a public transport network, useful both for epidemiological control by health authorities and for the transport operator when implementing effective measures to reduce the risk of infection. Specifically, the effects of two seat allocation policies were analysed. The first of these policies is an approximation of the usual behaviour

of passengers when choosing their seat in the vehicle, and the second is a strategy that aims to minimise the risk of infection. The methodology used to obtain these results was parameterised in accordance with epidemiological aspects and entities related to transport activity. To be precise, the definition of close contact for COVID-19 was used, together with the duration of the routes analysed and the geographical area in which they operate. Given the fact that the parameters of the methodology can be adapted, it could be applied to other diseases and use other transport-related aspects, such as the type of route, time bands, periods of time, etc. This is made by possible by the fact that the initial transport activity data can be used to generate a coherent and robust data set structured in the form of a graph. In order to obtain information about the interactions that occur on the transport system, the k-means classification technique was used to extract information from the resulting clusters and their centroids.

ACKNOWLEDGMENT

The authors wish to express their gratitude to Salcai Utinsa S. A. (GLOBAL) (one of the main road transport company that operates in Gran Canaria) for their collaboration in providing all the data used to develop this research work.

REFERENCES

[1] K.-Y. Wang, "How change of public transportation usage reveals fear of the SARS virus in a city," *PloS One*, vol. 9, no. 3, pp. 1–10, 2014.
[2] J. DeWeese, L. Hawa, H. Demyk, Z. Davey, A. Belikow, and A. El-Geneidy, "A tale of 40 cities: A preliminary analysis of equity impacts of COVID-19 service adjustments across North America," *Findings*, Jun. 2020, doi: 10.32866/001c.13395.
[3] E. Jenelius and M. Cebecauer, "Impacts of COVID-19 on public transport ridership in Sweden: Analysis of ticket validations, sales and passenger counts," *Transp. Res. Interdiscipl. Perspect.*, vol. 8, Nov. 2020, Art. no. 100242, doi: 10.1016/j.trip.2020.100242.
[4] A. Tirachini and O. Cats, "COVID-19 and public transportation: Current assessment, prospects and research needs," *J. Public Transp.*, vol. 22, no. 1, pp. 1–21, 2020.
[5] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. L. Spada, M. Mirmozafari, M. Dehghani, and A. Sabet, "Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment," *IEEE Access*, vol. 8, pp. 109581–109595, 2020.
[6] A. Corsi, F. Souza, and R. Pagani, "Big data analytics as a tool for fighting pandemics: A systematic review of literature," *J. Ambient Intell. Humanized Comput.*, vol. 12, pp. 9163–9180, Oct. 2021.
[7] W. F. Wells, "On air-borne infection: Study II. Droplets and droplet nuclei," *Amer. J. Epidemiol.*, vol. 20, no. 3, pp. 611–618, Nov. 1934.
[8] M. E. Halloran, I. M. Longini, A. Nizam, and Y. Yang, "Containing bioterrorist smallpox," *Science*, vol. 298, no. 5597, pp. 1428–1432, 2002.
[9] N. M. Ferguson, "Planning for smallpox outbreaks," *Nature*, vol. 425, pp. 681–685, Oct. 2003.
[10] I. M. Longini, M. E. Halloran, A. Nizam, and Y. Yang, "Containing pandemic influenza with antiviral agents," *Amer. J. Epidemiol.*, vol. 159, no. 7, pp. 623–633, Apr. 2004.
[11] J. Wallinga, P. Teunis, and M. Kretzschmar, "Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents," *Amer. J. Epidemiol.*, vol. 164, no. 10, pp. 936–944, Sep. 2006.
[12] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations," *J. Math. Biol.*, vol. 28, no. 4, pp. 365–382, 1990.
[13] L. Danon, "Networks and the epidemiology of infectious disease," *Interdiscipl. Perspect. Infectious Diseases*, vol. 2011, Mar. 2011, Art. no. 284909, doi: 10.1155/2011/284909.
[14] N. M. Ferguson, "Strategies for containing an emerging influenza pandemic in southeast Asia," *Nature*, vol. 437, no. 7056, pp. 209–214, Sep. 2005.
[15] I. M. Longini, "Containing pandemic influenza at the source," *Science*, vol. 309, no. 5737, pp. 1083–1087, 2005.
[16] J. Mossong, "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Med.*, vol. 5, no. 3, p. e74, Mar. 2008, doi: 10.1371/journal.pmed.0050074.
[17] P. Klepac, S. Kissler, and J. Gog, "Contagion! the BBC four pandemic—The model behind the documentary," *Epidemics*, vol. 24, pp. 49–59, 2018.
[18] N. Lapidus, "Factors associated with post-seasonal serological titer and risk factors for infection with the pandemic A/H1N1 virus in the French general population," *PloS One*, vol. 8, no. 4, pp. 1–8, Apr. 2013.
[19] M. Ajelli and M. Litvinova, "Estimating contact patterns relevant to the spread of infectious diseases in Russia," *J. Theoretical Biol.*, vol. 419, pp. 1–7, Apr. 2017.
[20] K. Leung, M. Jit, E. H. Y. Lau, and J. T. Wu, "Social contact patterns relevant to the spread of respiratory infectious diseases in Hong Kong," *Sci. Rep.*, vol. 7, no. 1, p. 7974, Aug. 2017, doi: 10.1038/s41598-017-08241-1.
[21] Y. Ibuka, Y. Ohkusa, T. Sugawara, G. B. Chapman, D. Yamin, K. E. Atkins, K. Taniguchi, N. Okabe, and A. P. Galvani, "Social contacts, vaccination decisions and influenza in Japan," *J. Epidemiol. Community Health*, vol. 70, no. 2, pp. 162–167, 2016.
[22] Y.-C. Fu, D.-W. Wang, and J.-H. Chuang, "Representative contact diaries for modeling the spread of infectious diseases in Taiwan," , vol. 7, no. 10, pp. 1–7, Oct. 2012.
[23] P. Horby, "Social contact patterns in Vietnam and implications for the control of infectious diseases," *PloS One*, vol. 6, no. 2, pp. 1–7, Feb. 2011.
[24] M. C. Kiti, T. M. Kinyanjui, D. C. Koech, P. K. Munywoki, G. F. Medley, and D. J. Nokes, "Quantifying age-related rates of social contact using diaries in a rural coastal population of Kenya," *PloS One*, vol. 9, no. 8, pp. 1–9, Aug. 2014.
[25] S. P. Johnstone-Robertson, "Social mixing patterns within a south African township community: Implications for respiratory disease transmission and control," *Amer. J. Epidemiol.*, vol. 174, no. 11, pp. 1246–1255, Nov. 2011.
[26] C. G. Grijalva, N. Goeyvaerts, H. Verastegui, K. M. Edwards, A. I. Gil, C. F. Lanata, and N. Hens, "A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of Peru," *PloS One*, vol. 10, no. 3, pp. 1–14, Mar. 2015.
[27] G. E. Potter, "Networks of face-to-face social contacts in Niakhar, Senegal," *PloS One*, vol. 14, no. 8, pp. 1–22, Aug. 2019.
[28] A. Latsuzbaia, M. Herold, J.-P. Bertemes, and J. Mossong, "Evolving social contact patterns during the COVID-19 crisis in Luxembourg," *PloS One*, vol. 15, no. 8, pp. 1–13, Aug. 2020.
[29] J. Zhang, M. Litvinova, Y. Liang, Y. Wang, W. Wang, S. Zhao, Q. Wu, S. Merler, C. Viboud, A. Vespignani, and M. Ajelli, "Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China," *Science*, vol. 368, no. 6498, pp. 1481–1486, Jun. 2020.
[30] C. I. Jarvis, K. Van Zandvoort, A. Gimma, K. Prem, P. Klepac, G. J. Rubin, and W. J. Edmunds, "Quantifying the impact of physical distance measures on the transmission of COVID-19 in the U.K.," *BMC Med.*, vol. 18, no. 1, p. 124, May 2020, doi: 10.1186/s12916-020-01597-8.
[31] J. A. Backer, L. Mollema, E. R. A. Vos, D. Klinkenberg, F. R. M. van der Klis, H. E. de Melker, S. van den Hof, and J. Wallinga, "Impact of physical distancing measures against COVID-19 on contacts and mixing patterns: Repeated cross-sectional surveys, the Netherlands, 2016–2017, April 2020 and June 2020," *Euro Surveill*, vol. 26, no. 8, 2021, Art. no. 2000994, doi: 10.2807/1560-7917.ES.2021.26.8.2000994.
[32] D. M. Feehan and A. S. Mahmud, "Quantifying population contact patterns in the United States during the COVID-19 pandemic," *Nature Commun.*, vol. 12, no. 1, p. 893, Feb. 2021, doi: 10.1038/s41467-021-20990-2.
[33] P. Coletti, J. Wambua, A. Gimma, L. Willem, S. Vercruysse, B. Vanhoutte, C. I. Jarvis, K. Van Zandvoort, J. Edmunds, P. Beutels, and N. Hens, "Comix: Comparing mixing patterns in the Belgian population during and after lockdown," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Dec. 2020.
[34] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke, "Strategies for mitigating an influenza pandemic," *Nature*, vol. 442, no. 7101, pp. 448–452, Jul. 2006.
[35] S. Eubank, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, pp. 180–184, May 2004.
[36] L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, and R. C. Brunham, "Network theory and SARS: Predicting outbreak diversity," *J. Theor. Biol.*, vol. 232, no. 1, pp. 71–81, Jan. 2005.

[37] T. Harko, F. S. N. Lobo, and M. K. Mak, "Exact analytical solutions of the susceptible-infected-recovered (SIR) epidemic model and of the SIR model with equal death and birth rates," *Appl. Math. Comput.*, vol. 236, pp. 184–194, Mar. 2014.
[38] E. Volz and L. A. Meyers, "Susceptible–infected–recovered epidemics in dynamic contact networks," *Proc. Royal Soc. B, Biol. Sci.*, vol. 274, pp. 2925–2934, Dec. 2007.
[39] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks," *J. Theor. Biol.*, vol. 271, no. 1, pp. 166–180, Feb. 2011.
[40] C. Cattuto, W. Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, "Dynamics of person-to-person interactions from distributed RFID sensor networks," *PloS One*, vol. 5, no. 7, pp. 1–9, 2010.
[41] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 51, pp. 22020–22025, Dec. 2010.
[42] L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Ravà, C. Rizzo, and A. E. Tozzi, "Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors," *PLoS ONE*, vol. 6, no. 2, Feb. 2011, Art. no. e17144, doi: 10.1371/journal.pone.0017144.
[43] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, "High-resolution measurements of Face-to-Face contact patterns in a primary school," *PLoS ONE*, vol. 6, no. 8, Aug. 2011, Art. no. e23176, doi: 10.1371/journal.pone.0023176.
[44] A. Stopczynski, A. S. Pentland, and S. Lehmann, "Physical proximity and spreading in dynamic social networks," 2015, *arXiv:1509.06530*.
[45] M. Génois and A. Barrat, "Can co-location be used as a proxy for face-to-face contacts?" *EPJ Data Sci.*, vol. 7, no. 1, p. 11, May 2018, doi: 10.1140/epjds/s13688-018-0140-1.
[46] A. Barnawi, P. Chhikara, R. Tekchandani, N. Kumar, and B. Alzahrani, "Artificial intelligence-enabled Internet of Things-based system for COVID-19 screening using aerial thermal imaging," *Future Gener. Comput. Syst.*, vol. 124, pp. 119–132, Nov. 2021.
[47] A. Kumar, K. Sharma, H. Singh, S. G. Naugriya, S. S. Gill, and R. Buyya, "A drone-based networked system and methods for combating coronavirus disease (COVID-19) pandemic," *Future Gener. Comput. Syst.*, vol. 115, pp. 1–19, Feb. 2020.
[48] S. Chang, "Mobility network models of COVID-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, Jan. 2021.
[49] W. Kermack and A. McKendrick, "A contribution to the mathematical theory of epidemics," Roy. Soc., London, U.K., Tech. Rep., 1927, pp. 700–721, vol. 115, no. 772.
[50] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact," *IEEE Access*, vol. 8, PP. 90225–90265, 2020.
[51] S. Merler and M. Ajelli, "The role of population heterogeneity and human mobility in the spread of pandemic influenza," *Proc. Roy. Soc. B, Biol. Sci.*, vol. 277, no. 1681, pp. 557–565, Feb. 2010.
[52] P. Cooley, S. Brown, J. Cajka, B. Chasteen, L. Ganapathi, J. Grefenstette, C. R. Hollingsworth, B. Y. Lee, B. Levine, W. D. Wheaton, and D. K. Wagener, "The role of subway travel in an influenza epidemic: A new York city simulation," *J. Urban Health*, vol. 88, no. 5, pp. 982–995, Oct. 2011.
[53] L. Goscé and A. Johansson, "Analysing the link between public transport use and airborne transmission: Mobility and contagion in the London underground," *Environ. Health*, vol. 17, no. 1, pp. 1–11, Dec. 2018.
[54] J. Troko, P. Myles, J. Gibson, A. Hashim, J. Enstone, S. Kingdon, C. Packham, S. Amin, A. Hayward, and J. N. Van-Tam, "Is public transport a risk factor for acute respiratory infection?" *BMC Infectious Diseases*, vol. 11, no. 1, pp. 1–6, Dec. 2011.
[55] K. Luo, "Transmission of SARS-CoV-2 in public transportation vehicles: A case study in Hunan province, China," *Open Forum Infectious Diseases*, vol. 7, no. 10, pp. 1–5, 2020.
[56] Y. Shen, C. Li, and H. Dong, "Community outbreak investigation of SARS-CoV-2 transmission among bus riders in eastern China," *JAMA Internal Med.*, vol. 180, no. 12, pp. 1665–1671, 2020.
[57] M. Hu, "Risk of coronavirus disease 2019 transmission in train passengers: An epidemiological and modeling study," *Clin. Infectious Diseases*, vol. 72, no. 4, pp. 604–610, 2021.
[58] M. Severo, A. I. Ribeiro, R. Lucas, T. Leão, and H. Barros, "Urban rail transportation and SARS-Cov-2 infections: An ecological study in the Lisbon metropolitan area," *Frontiers Public Health*, vol. 9, pp. 1–8, Feb. 2021.
[59] *Centers for Disease Control and Prevention. Appendices: Appendix A—Glosary of Key Terms*. Accessed: Mar. 5, 2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/appendix.html#Key-Terms
[60] D.-T. Dinh, V.-N. Huynh, and S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Inf. Sci.*, vol. 571, pp. 418–442, Sep. 2021.
[61] R. Deb and A. W.-C. Liew, "Missing value imputation for the analysis of incomplete traffic accident data," *Inf. Sci.*, vol. 339, pp. 274–289, Apr. 2016.
[62] D. Li, Y. Lin, X. Zhao, H. Song, and N. Zou, "Estimating a transit passenger trip origin-destination matrix using automatic fare collection system," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, vol. 6637, 2011, pp. 502–513.
[63] L. He and M. Trépanier, "Estimating the destination of unlinked trips in transit smart card fare data," *Transp. Res. Rec.*, vol. 2535, no. 1, pp. 97–104, 2019.
[64] *RStudio: Integrated Development Environment for R*. Accessed: Feb. 2, 2022. [Online]. Available: http://www.rstudio.com/
[65] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probabilities*, 1967, pp. 281–297.
[66] 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.

TERESA CRISTÓBAL received the B.S. degree in computer science, the M.S. degree in intelligent systems and numeric applications in engineering, and the Ph.D. degree in computer science from the University of Las Palmas de Gran Canaria, Canary Island, Spain, in 1990, 2014, and 2019, respectively.

Since 2012, she has been a Research Assistant with the Institute for Cybernetic, University of Las Palmas de Gran Canaria. Her research interests include development of intelligent transport systems for public transport and using data mining based models for public information services.

ALEXIS QUESADA-ARENCIBIA received the Graduate Engineering and Ph.D. degrees in computer science from the University of Las Palmas de Gran Canaria (ULPGC), in 1997 and 2001, respectively.

He is currently a Doctor-Employed Teacher with the Computer Science and Systems Department, ULPGC. He is also a Research Member with the ULPGC Research Institute "University Institute for Cybernetics (IUCTC)". His research interests include cybernetics, robotics, artificial vision, and intelligent transport systems. He is one of the Chair Members of the International Conference EUROCAST.

GABRIELE S. DE BLASIO was born in Gran Canaria, Spain, in 1961. He received the B.Sc. degree in physics from the Complutense University of Madrid, Spain, in 1984, and the Ph.D. degree in computer science from the University of Las Palmas de Gran Canaria (ULPGC), in 2009.

He has been a Researcher with the Institute of Cybernetics Sciences and Technologies (IUCTC), ULPGC, since 1991. He is currently an Associate Professor with the Computer Science Department, ULPGC. His research interests include ubiquitous computing, indoor positioning systems, intelligent transportation systems, biocybernetics, and systems theory. He has served on the Organizing Committee for the International Conference EUROCAST.

GABINO PADRÓN received the B.S. and Ph.D. degrees in computer science from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1990 and 2015, respectively.

Since 1989, he has been a Professor with the Informatics and Systems Department, University of Las Palmas de Gran Canaria. His research interests include ubiquitous computing, intelligent transport systems, data mining, and technologies for education.

FRANCISCO ALAYÓN was born in Las Palmas de Gran Canaria, Spain, in 1964. He received the B.S., M.S., and Ph.D. degrees in computer engineering from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1989 and 2007, respectively.

Since 1989, he has been a Professor with the Informatics and Systems Department, University of Las Palmas de Gran Canaria. He is the author of more than 50 articles and 20 inventions. He holds one patent. His research interests include passenger transport systems focuses in transport network planning, communications systems, and integration of the transport vehicle devices in the company's data networks.

CARMELO R. GARCÍA received the B.S. and Ph.D. degrees in computer science from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1989 and 1995, respectively.

Since 1987, he has been a Professor with the Informatics and Systems Department, University of Las Palmas de Gran Canaria, where he is currently the Director. His research interests include ubiquitous computing, intelligent transport systems, data mining, and technologies for education.

Dr. García was a finalist of the Spain University Foundation National Award in the Technological Transfer Modality, in 2005. He was a recipient of the 11th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI) Paper Award, in 2017.