energies

MDPI

Article

Development of a Common Framework for Analysing Public Transport Smart Card Data

Benito Zaragozí [1], Sergio Trilles [2], Aaron Gutiérrez [1,*] and Daniel Miravet [3,4]

[1] Departament de Geografia, Universitat Rovira i Virgili, C/Joanot Martorell, 43480 Vilaseca, Spain; benito.zaragozi@urv.cat
[2] Institute of New Imaging Technologies (INIT), Universitat Jaume I, Av. Vicente Sos Baynat s/n, 12071 Castelló de la Plana, Spain; strilles@uji.es
[3] Consortium of Public Transport of Camp de Tarragona, C/d'Anselm Clavé 1, 43004 Tarragona, Spain; daniel.miravet@urv.cat
[4] Research Centre on Economics and Sustainability (ECO-SOS), Department of Economics, Universitat Rovira i Virgili, 43204 Reus, Spain
* Correspondence: aaron.gutierrez@urv.cat; Tel.: +34-977-558-147

Abstract: The data generated in public transport systems have proven to be of great importance in improving knowledge of public transport systems, being very valuable in promoting the sustainability of public transport through rational management. However, the analysis of this data involves numerous tasks, so that when the value of analysing the data is finally verified, the effort has already been very great. The management and analysis of the collected data face some difficulties. This is the case of the data collected by the current automated fare collection systems. These systems do not follow any open standards and are not usually designed with a multipurpose nature, so they do not facilitate the data analysis workflow (i.e., acquisition, storage, quality control, integration and quantitative analysis). Intending to reduce this workload, we propose a conceptual framework for analysing data from automated fare collection systems in mobility studies. The main components of this framework are (1) a simple data model, (2) scripts for creating and querying the database and (3) a system for reusing the most useful queries. This framework has been tested in a real public transport consortium in a Spanish region shaped by tourism. The outcomes of this research work could be reused and applied, with a lower initial effort, in other areas that have data recorded by an automated fare collection system but are not sure if it is worth investing in exploiting the data. After this experience, we consider that, even with the legal limitations applicable to the analysis of this type of data, the use of open standards by automated fare collection systems would facilitate the use of this type of data to its full potential. Meanwhile, the use of a common framework may be enough to start analysing the data.

Keywords: public transport; smart card data; geodatabase; open science

Citation: Zaragozí, B.; Trilles, S.; Gutiérrez, A.; Miravet, D. Development of a Common Framework for Analysing Public Transport Smart Card Data. *Energies* **2021**, *14*, 6083. https://doi.org/10.3390/en14196083

Academic Editor: Francesco Bellotti

Received: 6 August 2021
Accepted: 18 September 2021
Published: 24 September 2021

1. Introduction

During the last three decades, the public transport systems of numerous cities and metropolitan areas have generated increasing volumes of data that may be of great interest in the analysis of all kinds of mobility issues [1]. Automated Fare Collection (AFC) systems are usually based on the use of smart travel cards, above all to speed up access to public transport [2]. However, these systems also collect data that can be useful for dealing with different management needs, such as providing the data necessary to access public subsidies, adjusting supply and demand in specific scenarios and in many other measures that ensure the sustainability of the public transport system through rational management [3].

The data collected by AFC systems range from card reloading operations to the validations that each user performs when using public transport. There are differences among the information recorded of each AFC system. The most basic information recorded

for each transaction on board is card code, fare type, route, transport mode (e.g., bus, train and metro), stop and exact date and time of the transaction. This rich and diverse source of information is potentially valuable for understanding user needs and thus improving the quality of service. However, even though there are already many examples of this data being used, there is still much work to do [4].

The main purpose of an AFC system is to generate data for accounting purposes. However, once its primary function is accomplished, these data can be costly to handle if other uses are not foreseen. As AFC systems collect all the transactions carried out, there is considerable flexibility for studying any period, or working at different geographical scales [5]. There are numerous precedents where this type of data has been analysed using a wide variety of methodologies and responding to different needs [6,7]. For example, smart travel card data have been used to identify different profiles of public transport users [8], reconstruct origin-destination matrices [9], examine the relationships between crime and the location of bus stops [10], evaluate the benefits of a major transport network improvement [11], or analyse patterns of tourist mobility [12].

Despite the growing list of identified benefits that has demonstrated the value of analysing AFC data in detail, little progress has been made in importing this type of analysis to most public transportation agencies [13]. Before considering the analysis of AFC data, public transport organisations and mobility researchers must consider the data analysis costs (in-house or third-party), additional surveying costs and marketing costs when changes are introduced in response to the patterns found in data [7]. The data collected using smart travel cards can be seen as a type of Big Data, where finding the *data value* is essential before deploying complex data management systems. In this way, the main focus of this research is trying to bridge the gap that exists between research with this type of data and its subsequent application in a real context.

Considering the requirements to analyse this data source, the main objective of this research work is to propose a framework to facilitate the preparation and analysis of AFC data, allowing the sharing and reuse of analysis methods (even when the access to data is limited because of privacy regulations). To achieve this, the following specific objectives need to be addressed.

1. Design a database model for general mobility information purposes that could solve most of the expected queries in a straightforward manner.
2. Implement the model using Free and Open-Source Software (FOSS) and enable a collaboration framework for exploring the data in a multidisciplinary research team.
3. Deploy the database in a real case study for testing the flexibility of this framework.
4. Evaluate the suitability and reproducibility of this system.

While being aware that no solution could cover all use cases, the following proposal seeks to exemplify how studies that analyse AFC data can better detail their methodology if they share data models and code, making the research somewhat more reproducible, even when the data cannot be shared publicly.

In the following section, it is described how data collected by AFC systems, using smart travel cards, is being analysed in different types of studies. Section 3 describes the system architecture of the proposed solution. Section 4 shows how the system was implemented in a real case study and describes its capabilities to query and analyse public transport data. Finally, Section 5 summarises our main findings, evaluates the system capabilities, draws conclusions and outlines future research directions for further developing the proposed system.

2. Automated Fare Collection Systems Using Smart Travel Cards

To address the first objective of this research, it is necessary to carry out a review of previous works and experiences in which the records of an AFC have been analysed for addressing different problems. From these previous works, it will be possible to extract common elements to establish a framework and tools that can be adopted in future studies.

We distinguish between the details of the mobility studies themselves and technologies or platforms that facilitate such analyses.

2.1. Key Concepts and Principles

AFC systems through smart travel cards became popular during the 1990s, but their proliferation managing payments in public transport networks around the world took place, especially during the 2000s. Octopus card in Hong Kong, Navigo in Paris, Compass card in Vancouver, Oyster in London, Bip! in Santiago, Troika card in Moscow or OV-Chip in the Netherlands are well-known examples of the implementation of this systems. Smart travel card data offer significant advantages in comparison with traditional data sources (e.g., travel diaries or surveys). Various literature reviews coincide in underlining the opportunities that smart travel card data provide for transport and mobility studies [4,6]. First of all, the main strength is that this source of data comprises the whole universe of public transport users in a specific area, in contrast to the sample used for traditional surveys. Second, smart travel card data enable analysis at different territorial and temporal scales, as all the travels reported are timestamped and can be georeferenced. Third, it is possible to make longitudinal studies at the individual level as each transaction is linked with a card. This enables complex and detailed studies of travel behaviour and demand forecasts for multiples periods. Fourth, it is possible to make interannual studies that would enable investigating at different spatial and temporal (or individual) scales the evolution of mobility and public transport demand.

Each of the four key points previously mentioned could be revisited by incorporating the potentialities that this source of data offers for particular research areas, for example, in tourism studies [14]. First, traditional surveys tend to be focused on the daily mobility of the resident population, and the mobilities of visitors during their stay are usually under-represented or ignored. By contrast, the data provided by smart travel cards include all the public transport users (local population, but also visitors). Second, the ability to produce multi-temporal analyses enables studying the effects of tourism activity in specific areas and periods. For instance, the effects on public transport demand for seasonal tourism activity in coastal or mountain destinations could be studied in detail [15]. Third, using longitudinal analysis, it is possible to differentiate tourists and frequent users of public transport services. Thus, it is possible to explicitly study the size and characteristics of the use of public transport by tourists in certain territories. Fourth, it is also possible to study the interannual evolution of public transport demand among tourists in a specific area and time. This aspect is relevant for evaluating the effects of policies for increasing the sustainable mobility of tourists.

The use of smart travel card data for research also implies some difficulties. AFC through smart travel cards continuously collects daily traveller transactions. Thus, the data volume could become very large and challenging to handle [16]. As mentioned above, smart travel card data can, therefore, be regarded as one type of Big Data [17]. Moreover, these databases are created for fare collection and management, not for research or general information purposes. This implies that substantial data cleaning and data processing and enrichment is required [18]. This task differs according to the characteristics of the data source. Every AFC has a different structure. However, some basic information is recorded by all the systems: the identity of the card, card or fare type, date and time of the transaction, location of the transaction (i.e., bus or train station, or on-board at a bus or light train). Another kind of information is also registered in some cases: vehicle identification, the stop and route number or code, the travel direction, and in fewer cases, if the transaction is an originating trip leg or a transfer from an earlier trip leg [19].

Apart from the already described management issues, there are also some limitations due to the inherent characteristics of these datasets. The main challenge is the lack of sociodemographic data associated with each smart travel card [17]. For this reason, an emerging research line is how to enrich these datasets with indirect information about passenger profiles. Some studies estimate passenger profiles [20], while others combine the

data from smart travel cards with passenger surveys [21]. Another challenge is the lack of information about travel purpose and the final destination. Different researchers have developed methods to estimate the origin–destination matrices [19] and the use of the trip (mainly for commuters) [22].

The authors of [6] identified three main research purposes in studies using smart travel card data: (1) strategic-level studies: long-term network planning, passenger behaviour analysis and demand forecasting; (2) tactical-level studies: longitudinal studies oriented to identify patterns in travel behaviour to adjust transport services; and (3) operational-level studies: supply-and-demand indicators. Identify the travel behaviour of public transport passengers is a common objective in all studies. More recently, a more specific proposal in [23] classified studies in the sub-field of the study of travel behaviour through smart travel card data in three main domains: (1) studies focused on understanding the data; and supposing a manipulation of the data to extract indicators to show what happens on the analysed transport network; (2) studies aiming to explain travel behaviour, and implying the use of various external sources of data according to the study objective; and (3) studies oriented to support decision-making (mainly for demand forecast and transport planning).

The vast majority of the studies mentioned above are focused on the daily mobility of the resident population, but more research gaps are waiting to be addressed. For example, up until now, few studies have identified and characterised tourists or specific groups of travellers among the rest of public transport network users. To name a few, there are examples for London [24] or Singapore [12]. As can be seen, there is a wide diversity of study areas, differently implemented AFC systems, and an increasing number of research questions. In this context, a new platform for analysing data from smart travel cards should be designed for general information purposes, considering scalability and flexibility as its main characteristics.

2.2. Experiences Using Different Technological Solutions

As already explained, AFC systems collect data continuously over long periods, recording years and even decades of transactions generated by a large number of users. For this reason, this data source can be considered as one type of Big Data [17]. This implies difficulties for the public transport agencies when managing or analysing these volumes of data, which generally requires an adequate technological solution. During the last decade, many studies used big data technologies to manage AFC systems around the world. There are currently several types of database management systems that can be suitable for handling these large volumes of information (e.g., Hadoop Distributed File System (HDFS), Hbase, Apache Cassandra, Redis or distributed object-relational databases such as Postgres, to name a few) [25].

A recent study used Big Data technologies (Apache Spark and Hadoop) to analyse more than 160 million transactions generated by the Jakarta's Bus Rapid Transit (Indonesia) [26]. Another study used the dispy framework for parallel computing and QGIS (https://www.qgis.org/, accessed on 26 July 2021) to perform spatial analyses over a dataset containing more than 200 GB of smart travel card data transactions from the AFC system in Montevideo (Uruguay) [27]. In [28], a smart urban transportation management is presented. The proposed solution is featured as a scalable, flexible and dynamic platform. This platform is built using technologies such as Apache Spark, Apache Hadoop and Ophidia [29]. Also using Big Data technologies, a real-time analysis framework was proposed [30] for this type of data, and the authors performed a test on a reduced one-day dataset from Shenzhen (China). However, when a real-time analysis is not required or the amount of data to be analysed is smaller, other approaches are used. One common approach is to combine Structured Query Language (SQL) databases with more specific analytical software (e.g., SPSS, Microsoft Excel or RStudio) [31]. Considering the reviewed scientific literature, the use of SQL databases for these purposes is certainly recurrent. In a recent systematic literature review, seven out of nine of the documents that reported the tools chosen for the analysis of smart travel card data used an SQL database [32] by

itself or in combination with other tools. Despite this, none of the research works carried out in the systematic review offer details of the database system in terms of design and implementation.

There are numerous experiences in the application of different database management systems in large cities and metropolitan areas. There are also examples at other smaller scales, where these data also fulfil mainly accounting functions. Logically, larger cities have more resources to continue exploiting these data but, in smaller cities, it is more challenging to invest in adopting new technologies, especially if the benefit of utilising this information has yet to be demonstrated. In these cases, the use of free and open-source software and reproducible management models, within the collaboration between public transport authorities and mobility research groups, has become an interesting strategy to analyse this specific type of data [33]. More recently, some studies have pointed out that the use of standards in AFCs and the subsequent data analysis could provide significant advantages and favour the use of this type of data in all types of applications [13,34], such as those seen in the previous sections.

2.3. Aspects to Improve

As explained in this section, many studies have recently appeared that use smart card data to achieve different purposes that go beyond the merely accounting purpose for which these data were originally collected. These types of applications are on the rise, but they are isolated applications, dependent on access to data and the availability of resources for analysis. Furthermore, due to their particular nature, the studies reviewed in the previous subsection are not easily reproducible. The reasons are that the data structures of the AFC are very particular, and very few authors may work with data from different AFC systems. Although there are reasons why these data cannot be shared, and the tools for their analysis should be tailored in each case, it is reasonable to think that these studies would benefit if a common framework or platform were available, so that working methods and tools could be shared. In the following section, we propose a first implementation of such a framework.

Once that framework is proposed, we also develop a case study showing how the framework addresses all the phases of a knowledge discovery process (selection, preprocessing, transformation, data mining, interpretation and decision-making). Among the case studies presented above, the analysis of specific profiles of tourists is one of the least worked due to several difficulties mentioned above (e.g., the lack of socioeconomic information of travellers). Therefore, this seems like a suitable case study to test the proposed framework, as it involves the most common problems that this type of study generally faces.

3. Proposed Architecture

As mentioned above, there is currently a wide variety of FOSS projects that could be used to implement a solution like the one outlined in the objectives of this project. In short, a framework designed to ingest a medium or large volume of data from the logs created by an AFC system and facilitate its exploitation through SQL queries is necessary. It must be a database with the essential spatial component to contextualise this type of information. In addition, a collaborative work protocol must be established in which the most interesting queries made to the database can be identified, and a duplication of effort is avoided. Finally, it must be possible to export the results of the database queries to formats in which the end users—in this case the experts in mobility and public transport—can carry out a more specific analysis of the information. To achieve this, a software architecture that satisfies all the mentioned requirements has been proposed. The designed architecture consists of three layers: content, services and applications. This design is detailed in Figure 1, which shows how each layer has been deployed using FOSS and the roles assigned in each part of the architecture.

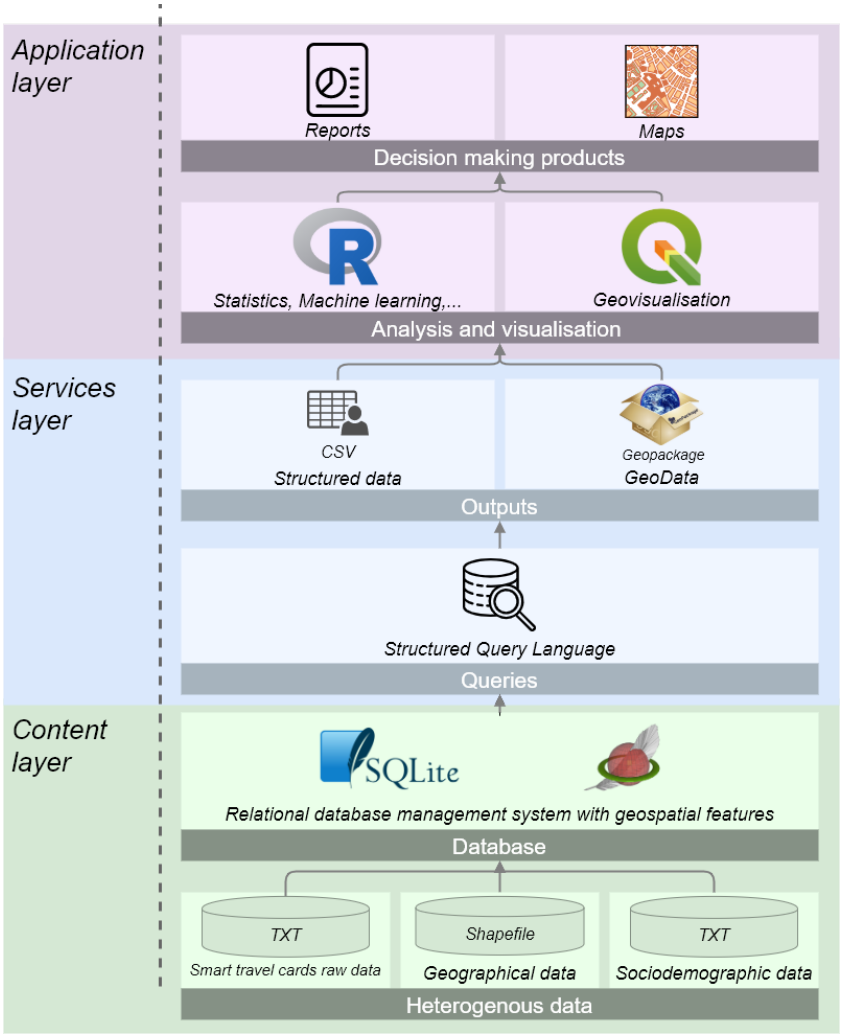

Figure 1. Proposed system architecture for the analysis of smart travel card data.

The presented architecture aims to cover the two leading roles involved in this kind of project: (1) mobility experts and researchers, who are in charge of defining the requirements to formulate each query. Later, they also analyse the query results using advanced techniques to obtain valuable knowledge that can be used in decision-making. This role is at the application layer. (2) Spatial database managers, who are involved in the content layer, maintaining databases and overseeing Extract Transform Load (ETL) processes. In addition, at the service layer, they will be in charge of coding and launching the SQL queries based on the requirements defined by the mobility experts. The following subsections detail how each of the presented layers has been implemented.

3.1. Content Layer: Database Design

As explained in Sections 1 and 2, different types of data need to be managed when analysing smart travel card data. First, AFC systems usually collect all the transactions generated by smart travel card users. In most cases, the data provider serves these as semi-structured data, and an ETL process is necessary to integrate the data accurately. Other data types are added to the same content layer, including Geographic Information System (GIS) layers and sociodemographic data of the study area. During the ETL process, all data sources must be loaded into a database. One solution is to choose a Relational Database Management System (RDBMS) with support for storing and analysing geospatial data.

Among the available RDBMS offering spatial support, only a few FOSS object-relational databases provide the necessary capabilities: highlighting SQLite with Spatialite support for embedded databases, and PostgreSQL with the PostGIS spatial extension for

server or cloud-based implementations. First, SQLite/Spatialite is a popular RDBMS and serves as the the basis for important formats such as the OGC Geopackage (https://www.geopackage.org/ (accessed on 26 July 2021)). PostgreSQL is a more versatile platform that supports most of the significant features of the current SQL standard [35] and a long list of additional supported features. No current version of any RDBMS claims full conformance with the SQL standard [36], but this level of compliance is enough to easily enable most of the SQL queries developed in other platforms to run on PostgreSQL. Several configurations and features of PostgreSQL can also be used in Big Data applications (including Greenplum, PostgresXL and JSONB) and it works fine with other open Big Data technologies such as Apache Spark. Both options, Spatialite and PostGIS, include various geospatial functionalities such as geoprocessing vectorial geometries, raster operations and the possibility of running native network analysis in an SQL environment. Considering the data types to be stored, SQLite/Spatialite is adequate to analyse data collected in a medium or small area. However, being an SQL-based platform, developments done using SQLite can be easily migrated and scaled to more robust on-server solutions using Postgres/PostGIS.

The proposed design for the database is shown in Figures 2 and 3. This database model is quite generic and is based on the typical data collected in raw log data, the necessary context information, and the most common queries as identified in previous studies (see Section 2). For the sake of clarity, when it was possible, entities were named following the General Transit Feed Specification (GTFS), which is a popular standard format for publishing public transport schedules (https://gtfs.org/reference/static (accessed on 26 July 2021)).

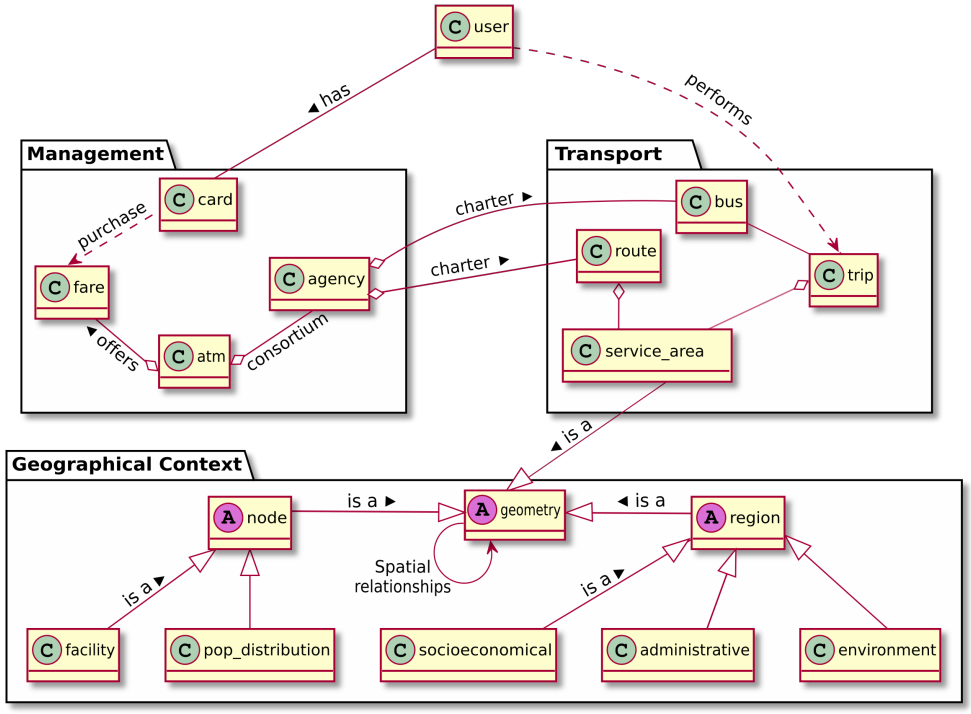

Figure 2. Conceptual data model describing the entities and relationships participating in an AFC. This model includes the possibility of providing a geographical context which is necessary for answering different management and research questions.

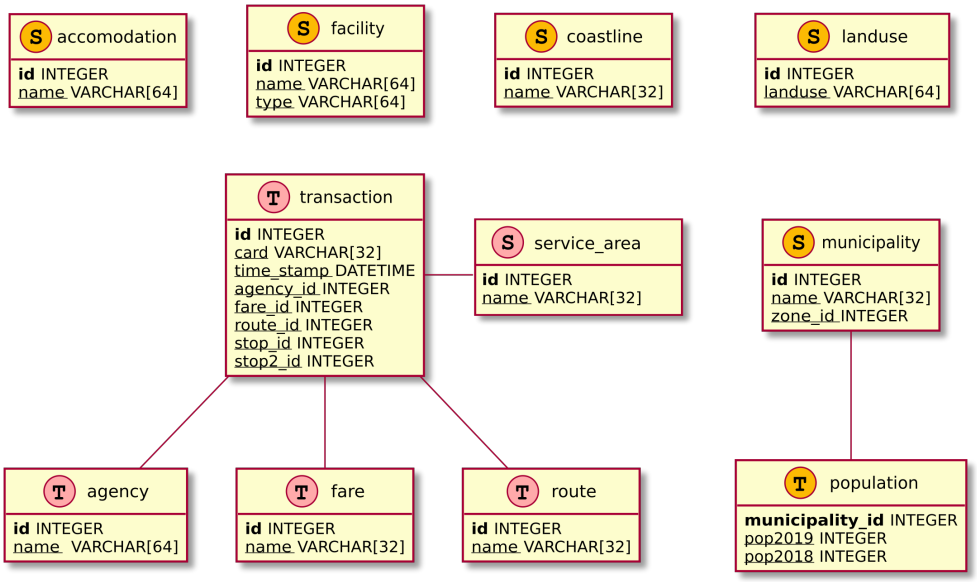

Figure 3. Logical database model for storing relevant smart travel card log data and context geographical and socio-demographic data. This model is extensible by adding new thematic and spatial layers. A reference symbol separates spatial tables (*S*) from regular attribute tables (*T*). Different colours separate data provided by an AFC (Red) from geographical data added for contextualising the system (orange). Adapted from the work in [37].

Figure 2 shows a conceptual model that is an effort to synthesise a specific case study but representative of the common elements of an AFC. The description of an AFC has been well developed in several previous studies [15,38–40] and will be completed in the case study we propose in Section 4. The conceptual model comprises four main parts that we separate between (1) the user of the service, (2) the management of transactions and the provision of the service, (3) the structure and layout of the transmission network and (4) the geographical context. Although standardised symbology has been used, there are some elements of this diagram that need to be explained. Yellow rectangles represent classes or entities of some kind, but classes represented with a letter "A" are abstract classes, in this case, referred to the spatial representation of different types of territorial and socio-economic information. These classes are closely related to the standards by which geometries are represented in GIS today, but this model is not limited to any particular standard. Dashed lines indicate that the relationship may not always occur. For example, a user may have a card and not use it for a significant period, or not even recharge it. Continuous lines without any arrowhead indicate wider relationships in both directions, while in other cases the relationship has an explicit directionality. The rhombus-shaped arrowheads indicate aggregation relationships. For example, a transport agency may have one or more buses in its fleet and manage one or more routes in the territorial system. On the contrary, a route can not be managed by more than one agency simultaneously. In the same way a transport consortium ("atm" in the diagram) regulates the offer of a certain number of fares, and the transport agencies have to adapt to the different fares proposed by the "atm". Finally, in the geographical context package, there are only two spatial representations at two levels: if they have a known extension (region, polygon) or if only their location is known (node, point). In GIS it is usual also to have representations in the form of linestrings, but in this case, it would not be correct since the exact information of the routes depends on information that is not collected by an AFC. service_areas usually represent bus stops but sometimes, because of data privacy regulations, they represent aggregates of more than one bus stops (e.g., census districts or municipalities). The concern for privacy is due to the possibility that in bus stops with few validations the

identification of the user may be exposed or easily discoverable. Then, routes are considered a sequence or aggregation of service_areas. The same goes for trips, which are also a sequential aggregation of service_areas even if traffic has been redirected for some reason. Geometrically, a service_area can be a node (e.g., bus stop) or a region (e.g., census district or neighbourhood), and it is the only information provided by an AFC that can be georeferenced without much difficulty. Nodes are related to each other or to regions by means of spatial relationships (e.g., intersection, overlap, proximity, among others).

Based on the conceptual model (see Figure 2), a more concrete logical database model is designed for general information purposes. In Figure 3, the diagram now uses yellow rectangles to represent tables with their columns. This model shows better how data would be organised into an object-relational database system. The logical model is built around the transaction table that includes all the basic information about transactions collected by an AFC system. Thus, this table records all the "data transactions" generated in the management and transport packages (management and transport transactions). Some elements from the previous diagram are missing due to a lack of available data. User cannot be completely identified because on many occasions, multiuser cards or single trip tickets can be used or the same user could hold different cards or fares for some reason. The bus or vehicle is also difficult to identify. As the original purpose of the AFC is accounting, vehicle registration is less important. Sometimes a code is registered to identify the validation machine that certifies the transaction, but these machines can be exchanged between buses or replaced, so the codes would only be valid for a certain period. If necessary, a new reference table could be added to refer to the vehicle. Normalisation is achieved by splitting several lookup tables (agency, fare or service_area), and this also decreases database size. The service_area table stores the most essential geographical information in an AFC. Other information, such as the postal address of smart travel card users could be georeferenced, but that information would be confidential and not available by all means. Using service_area as a spatial layer, there is a broad opportunity for creating all the necessary spatial joins to other useful geographical datasets. In Figure 3, only administrative boundaries (municipality) and population data were joined in the model. However, through the administrative codes, it would be possible to link almost any sociodemographic information provided by official sources (i.e., labour market statistics). The spatial component allows the integration of almost any territorial variable. In Figure 3 the coastline is imported as a spatial layer, so variables related to the distance to the coast could be spatially joined to each smart travel card transaction. Many other layers could be of interest for different purposes explored in the scientific literature (i.e., land uses, streets and social network data, to name a few).

3.2. Services Layer: Queries Nomenclature

This layer is used to query the relational database generated in the content layer. SQL-encoded queries will be tailored to the goals set at the application layer. The team structure defined above reveals a communication problem that may arise when the same person does not perform the mobility expert and the database administrator roles. Both roles have wide expertise in their domain but lack cross-domain knowledge to intuitively understand the objectives, or the difficulties, of a particular query. The mobility expert does not have the experience of working with a relational database, so when requesting a new query from the database manager, it may be not easy to express it in the most direct way to define an accurate SQL query. The database manager may have difficulties in understanding the specific purpose of a query. To obtain the correct result, a certain iterative and interactive process between both actors is usually necessary. The transport and mobility researchers must describe the query to the database manager actor often as necessary until obtaining an unambiguous query definition. This process can be very tedious, repetitive or redundant, especially when the number of team members or different queries increases. It should be remembered that here we are dealing with different roles and not with individuals, so it is reasonable to say that sometimes an individual can develop both roles, and therefore

this communication problem would not exist. In either case, a single individual will need to keep a record of queries already made to avoid duplicating efforts and to retain code samples to facilitate future queries.

As a solution to this problem, it was proposed the definition of a Domain-Specific Language (DSL) adapted to the needs of this type of applications [37]. This DSL is used as a convention to describe and identify the queries accurately. Figure 4 presents the new workflow achieved using the proposed DSL, namely, MobilityFNC [37]. Each query is named using this convention, and both files—the SQL query defined by the database experts and the results file—are inseparable. The key element in this solution is that the common nomenclature enables the creation of a repository of the query outputs, and so avoiding the duplication of efforts when creating new queries.

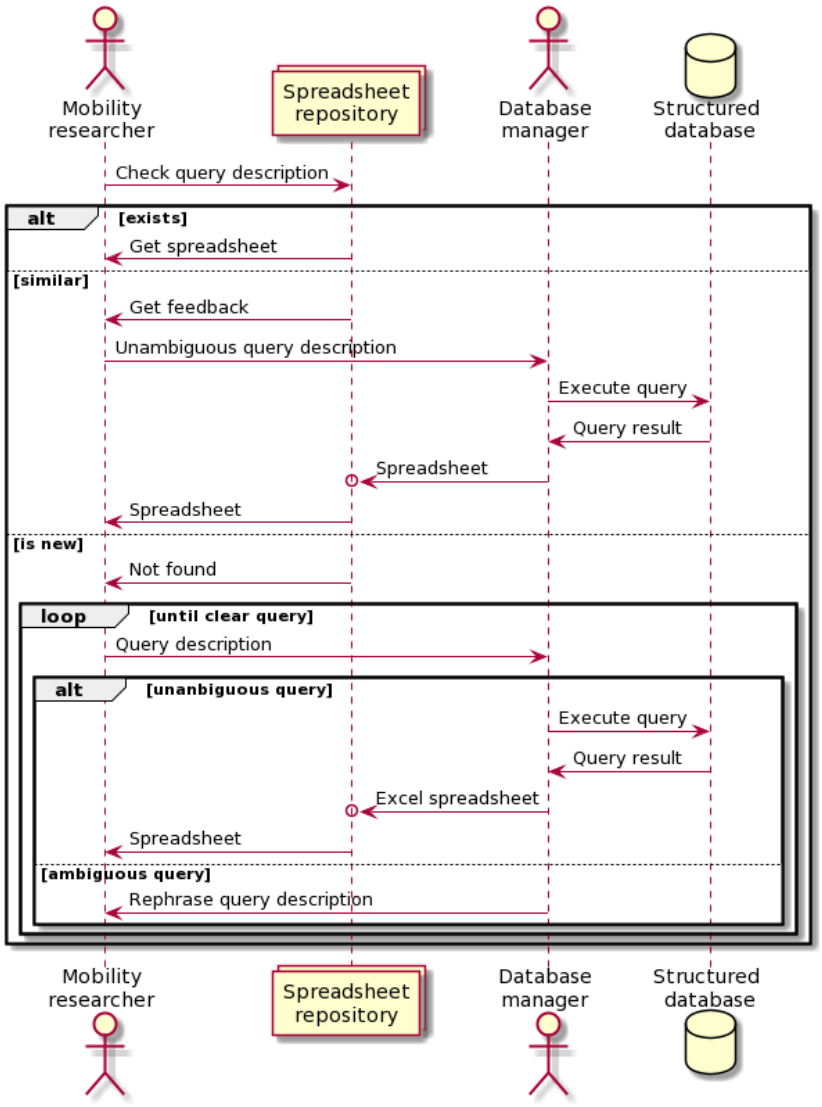

Figure 4. Sequence diagram showing the information workflow between a transport and mobility researcher and database manager using the *MobilityFNC* approach. Adapted from the work in [37].

Query results are exported using two well-known formats: Comma-Separated Values (CSV) and OGC Geopackage. The selection of the data format will be conditioned to the type of query result. If the result data contains a geospatial component, the OGC Geopackage format is used; otherwise, the output is encoded using CSV format. CSV is extensively used in any discipline and can be opened using common programs such as word processors and spreadsheets. Otherwise, the OGC Geopackage format is a universal

format for transferring vector and raster spatial data and is supported by Open Geospatial Consortium (OGC) [41]. OGC Geopackage is supported natively by SQLite/Spatialite.

3.3. Application Layer: Decision Making

In the application layer, the outputs from the service layer are used to perform analyses and visualisations according to the hypotheses or questions defined. Note that the visualisations must be adapted to the level of knowledge of the end user.

Currently, there are different analytical methods capable of assisting in public transport data analysis for different purposes. Examples of these include machine learning or data mining algorithms [42]. These methods can be useful to analyse this kind of data and obtain some more specific results, such as prevalent routes, activity patterns, relationships between territorial resources and use of public transport.

At the first level of complexity, basic statistical methods are located for analysing passenger trip patterns. These include frequency analysis, analysis of variance, and related spatial and temporal correlations between trips [43]. On a more sophisticated level, other descriptive or predictive techniques can be used (e.g., clustering, factorial analysis, association rules, classification or regression). Clustering techniques stand out above other techniques when it comes to analysing smart travel card data [44,45]. These are applied to extract useful patterns from datasets containing spatiotemporal data [46]. For example, the K-means clustering algorithm is well suited for recognising abnormal behaviour, or for identifying common trip patterns at the spatial and temporal level [47]. Many FOSS tools enable the utilisation of this type of analysis; one of the best known and powerful is the *R* platform [48]. The *R* platform offers many libraries containing a wide variety of machine learning algorithms and techniques designed to apply filters, dimensional reduction, clustering or other methods (see CRAN task views [49,50]).

The results of these analyses can be presented in various forms, but these are commonly used to create visualisations of information that facilitate understanding of complex patterns. This is especially true when analysing patterns present in the smart travel card data. There are different types of visualisations, and depending on the data or knowledge to represent, one type may be more suitable than another. Some of the most common types or categories of data visualisation are temporal (i.e., lines chart, stacked area chart or bar chart), multidimensional (i.e., histogram or scatter plot), hierarchical (i.e. tree diagram, sunburst diagram, dendrogram), networks (i.e., alluvial diagram, node-link diagram or words cloud) and geospatial (i.e., points map, heatmap or choropleths map) [51,52], although a tabular representation can also be considered as a visualisation type.

These types of visualisation are mixed to create reports adapted to end user needs [53]. A recently used form is the creation of visualisation panels or dashboards [54,55]. It is a composition of different types of visualisations necessary to achieve one or more objectives. The different charts are consolidated and organised on a single screen so that the information can be monitored at a glance. Many times these control panels work in real-time as they have a direct connection to data sources or analysis procedures. These panels are very helpful for mobility experts or rulers as they easily and directly help to know the current state to make the appropriate decisions or legislate based on the results.

One of the most widely used types of visualisations in public transport are maps as both smart travel card data and analysis results contain a geospatial component. R also has many options for generating maps but, one of the most used programs to generate this kind of representations, within the FOSS ecosystem, is QGIS. QGIS provides a wide range of capabilities, including viewing, managing, editing, analysing geospatial data and composing maps. These visualisations will be used, together with other charts and graphics, to generate advanced reports. The final objective is to provide necessary evidence in decision-making processes by public and private administrators of transport infrastructure.

4. Case Study: The Territorial Mobility Authority of Camp de Tarragona

The main purpose of this case study is to show a real application of the proposed framework (Figure 1), developing the three layers of the proposed system architecture, from data acquisition to the development of useful reports. However, as a secondary objective of this case study, a data mining analysis and some data reports were produced. This case study develops the methodology applied in a recent study [14], so that only a brief geographical context and brief problem description are presented below.

The data for this case study are referred to the area of the Camp de Tarragona (Catalonia, north-east Spain). The Territorial Mobility Authority of Camp de Tarragona (ATMCdT, according to its acronym in Catalan) has been running an AFC system for more than ten years, providing bus and train service for an area of 2998 km^2 and a population of 626,277 inhabitants https://www.idescat.cat/ (accessed on 26 July 2021). In Figure 5, there is a general context of the area. Among other geographical particularities, this region is shaped by coastal tourism. This drives an unbalanced public transport (especially during the summer season). In some coastal municipalities, the demand for public transport in the summer can increase by a factor of six or eight compared to the regular use that occurs in the winter months. This type of phenomenon makes it necessary to have the best possible information to regulate the supply–demand binomial, and thus not risk the quality of the service and the sustainability of the system for both, tourists and residents [15,38–40].

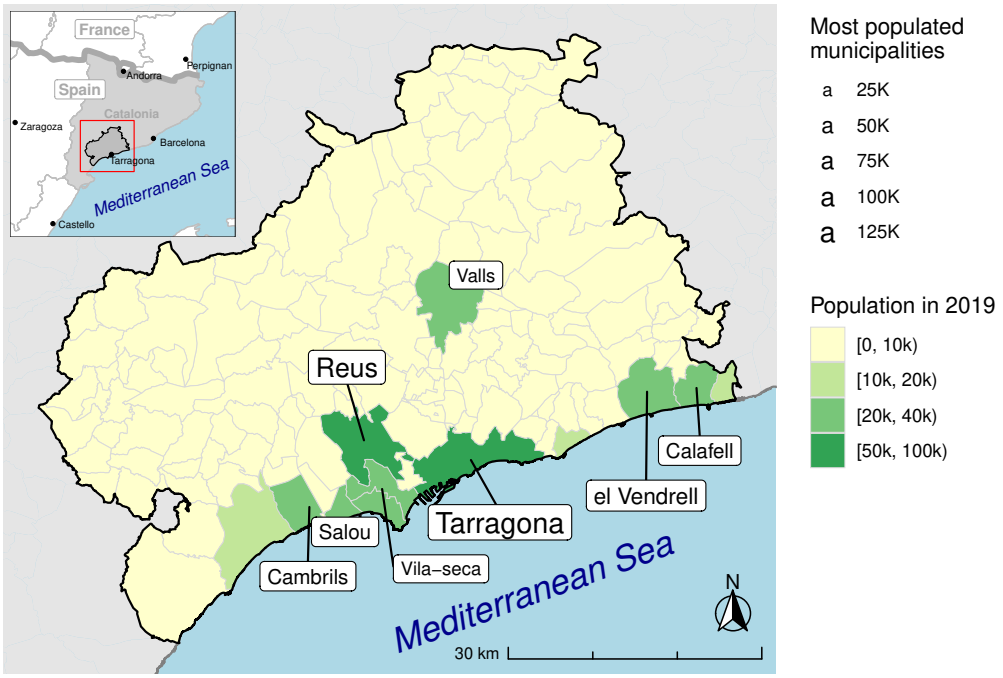

Figure 5. Context of the Territorial Mobility Authority of Camp de Tarragona (ATMCdT) service area.

In the following subsections, there is an explanation of the general process of how ATMCdT data were preprocessed, filtered and analysed, in order to find interesting activity patterns to explain the structure of those mentioned above unbalanced public transport demand. The case study is focused on passengers travelling only during summer 2019 (from 21 June to 22 September, both included). We assume that these seasonal passengers that concentrate all their transaction in summer should be mostly related to tourism activity in the region. For this reason, we filtered the smart travel cards that have made 100% of their transactions in summer and excluded the others. That group includes 37,054 cards and 685,294 transactions. T-10 cards represent 93.2% (34,552 cards) of that group and 87.4% of the transactions (599,305). T-10 and T-70/90 are the only multipersonal fare types in the ATMCdT system, so they allow travelling in groups (consecutive transactions when

boarding). However, T-70/90 is a fare for large families and is not as interesting as the T-10 in the study of tourism.

The standard T-10 card covers ten transactions for a maximum of a year. It could be recharged with additional ten transactions, up to a maximum of 30. The rest of ATM fares are unipersonal and imply benefits for cardholders if they frequently use public transport services during the year. Any analysis of group travelling should take into account these parameters. On the other hand, single-ticket passengers' patterns are not analysed. They are not using a smart travel card, and it is not possible to develop any longitudinal analysis of their mobility patterns. These decisions have been explained in depth in [14].

4.1. Planning

The planning phase for implementing the proposed framework in this case study was developed through several stages that ensured the achievement of the research objectives. As explained above, the ATMCdT, public administrations, local stakeholders, and transport and mobility researchers are interested in analysing this source of information in the most transparent, practical and inexpensive way. For this reason, the ATMCdT facilitated the inventory of existing data. More than ten years of records are stored, but there is no technical documentation about the original design or the changes experienced by the system. ATMCdT provided a limited extraction of the dataset for 2019 for this test, excluding all operations that do not represent a trip, or a part of the trip chain (e.g., management transactions). In this stage, the preliminary database design presented in Figure 3 was used.

4.2. Data Collection

The ATMCdT facilitated an anonymised extraction of the attributes necessary for implementing the proposed database model. For example, all sensitive attributes, such as the card code, have been obfuscated so that the data cannot be traced. The ATMCdT system is also known as Fare Integration Management System (SGIT, according to its acronym in Catalan). As in other cases commented in Section 2, the SGIT collects data for accounting purposes only. Their use in mobility or transport geography studies was not considered in its original design. The data collected include all the transactions made using smart travel cards with different fare types, but also adds all the transactions made using single-trip tickets. These logs contain the exact day and time of each transaction, the service area where the passenger boarded, the agency that operates the transport, the municipality and the type of fare used in each transaction. All data gathered by SGIT can be accessed using an ad hoc-designed software that allows the technicians who work at the ATMCdT to download data reports in spreadsheet format (CSV). ATMCdT staff usually analyse this data using highly resource-demanding nested spreadsheets and pivot tables.

In addition to the extraction of the ATMCdT data, the database was completed with some geographical information including (1) the administrative boundaries, roads and coastline that were downloaded from the Cartographic and Geological Institute of Catalonia (https://www.icgc.cat/ (accessed on 26 July 2021)) and adapted to the project needs; (2) official demographic data for 2019 was retrieved in CSV format (https://www.idescat.cat/ (accessed on 26 July 2021)); and (3) ATMCdT zones were made by combining administrative boundaries.

During the first stages of the implementation process, it was decided that the above mentioned open standard formats were suitable to vertebrate the information flow: (1) CSV for accessing SGIT data and exporting query results and (2) Geopackages to store base spatial layers and spatial query outputs. After all these considerations, a tailored data download containing all the necessary information was performed. Finally, near 1.5 GB of the data extraction facilitated by ATMCdT in CSV format and spatial layers was downloaded and prepared to be loaded into a spatial database.

4.3. Extract, Transform and Load

In the analysis of AFC systems, this is an important step that comes right after data collection. Considering the database model proposed in Figures 2 and 3, we chose a portable combination using SQLite/Spatialite and then we wrote all the necessary scripts and Data Definition Language (DDL). These scripts were intended for (1) filtering only those fare types that could be tracked for activity patterns (e.g., obviously excluding single-trip transactions);=, (2) selecting only those relevant attributes defined as useful by the model (9 out of the 11 attributes facilitated by ATM), (3) creating an empty dummy database following the database model, (4) loading all datasets into a relational database and (5) setting up the database for increasing the database performance (e.g., applying normalisation, creating indexes and useful views). After this ETL process, a slight reduction in data size has been achieved. Most of this process was performed using SQL expressions that could be adapted to automate the ETL process in different scenarios. All these scripts are published in a dedicated code repository (https://github.com/gratet/sgit-explorer (accessed on 26 July 2021)).

4.4. Mining Public Transport Users Activity Patterns

According to the objectives of this research, and following the parameters of this particular case study, the analysis focused on smart cards only used during the summer period. After a SQL filtering process, the data to be analysed included 37,054 different smart travel cards which made 685,294 transactions. These data can be considered a reliable representation of the set of tourists and visitors that mostly contribute to the aforementioned supply-demand decompensation (between the winter and summer seasons) [14]. Figure 6 shows the spatial distribution of these transactions.

Other more specific SQL queries were written for extracting fourteen simple but meaningful card descriptors, which referred to the level of activity and the spatial concentration of each user. These simple descriptors are the result of SQL aggregation queries that could be calculated for other smart travel card dataset containing the minimum information described in the database model (Figure 3). These indicators consisted in the total number of transactions, the number of municipalities or zones visited, the number of days and months the card was active, the number of days the card was enabled—as the difference between the last day the card was used and when it was used for the first time, the minimum, average and the maximum group size (only for T-10 and T-70/90 cards), the percentage of trips concentrated in two or three municipalities, the number of routes used, the percentage of transactions made from main cities or the most touristic ones and the percentage of trips concentrated in weekdays or weekends. Finally, considering the similarity between some of these variables, a correlation analysis was performed, showing that some variables were highly correlated (*over* $\pm$ 0.7). After removing those redundant variables, only seven indicators were kept.

SQL makes it easy to obtain descriptive statistics from a ready-to-use database. However, in order to extract further knowledge, such as group behaviour, or frequent activity patterns, more advanced data analysis techniques need to be used [46]. In this context, clustering techniques obtained good results in studying the activity patterns hidden in smart travel card datasets [56–61]. In this case study, a model-based clustering analysis [62] was performed using RStudio and, to get an appropriate result, solutions between two and ten clusters were calculated. As a feature of model-based clustering techniques, an analytic hierarchy process, based on the fit indices Akaike Information Criterion (AIC), Approximate Weight of Evidence Criterion (AWE), Bayesian Information Criterion (BIC), Classification LikelihoodCriterion (CLC) and Kullback Information Criterion (KIC), was performed, suggesting that the best solution is a highly constrained but also parsimonious model with seven clusters [62]. Other parameters such as the probability of a card to be classified in a cluster (min, max), the entropy of the model or the number of cards in each cluster could be considered for selecting the most appropriate solution [14], but it was not necessary for this demonstration. Finally, one advantage of model-based clustering over other techniques is

that the result includes measurements of the accuracy of each model, and the importance of each variable can be evaluated. The R script for performing this clustering analysis can also be found in the same code repository (https://github.com/gratet/sgit-explorer (accessed on 26 July 2021)).

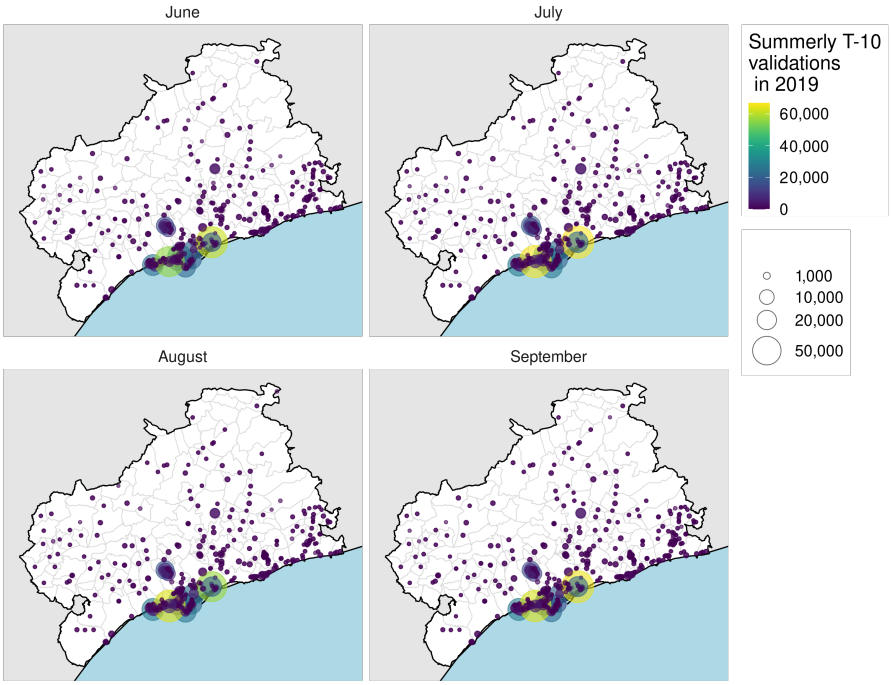

Figure 6. Summer-only T-10 transactions per month and stop, during summer of 2019.

Table 1 shows descriptive statistics of the seven clusters solution. All clusters included more than 1% of cards and cluster #4, showing average values for all variables, kept almost one-third of cards, representing the most common usage of the summer-only smart cards. The other clusters also show other interesting patterns in the use of the multi-person rate in summer. For example, cluster #1 characterised by a high number of trips, but the cards are barely active for a few days and are used mainly for group trips. This cluster #1, which only includes 295 cards, stands out as the only one in which the card is never used individually and is used by groups of up to 30 travellers (the maximum allowed by the T-10 fare).

Cluster #3 shows a more prolonged activity throughout the summer period, in which the cards are always used for more than two weeks and usually more than a month and a half. During this period, it is normal for the number of trips to be higher than that of other users and for the geographical distribution of these trips to be somewhat wide.

Table 1. Descriptive statistics of the seven cluster solution.

	1 (*n* = 295)	2 (*n* = 15,010)	3 (*n* = 3002)	4 (*n* = 13,086)	5 (*n* = 3115)	6 (*n* = 2546)	Total (*n* = 37,054)
transactions							
Mean (SD)	27.8 (21.2)	22.9 (15.0)	40.8 (39.8)	11.9 (8.53)	9.34 (5.67)	10.1 (9.74)	18.5 (18.3)
Median [Min, Max]	20.0 [8.00, 129]	20.0 [2.00, 187]	29.0 [2.00, 752]	10.0 [1.00, 116]	9.00 [1.00, 50.0]	8.00 [1.00, 106]	12.0 [1.00, 752]
visited_municipalities							
Mean (SD)	1.77 (0.797)	3.56 (0.725)	3.19 (1.17)	2.12 (0.623)	2.04 (0.682)	1.63 (0.664)	2.75 (1.04)
Median [Min, Max]	2.00 [0.00, 4.00]	4.00 [1.00, 7.00]	3.00 [1.00, 8.00]	2.00 [1.00, 4.00]	2.00 [0.00, 4.00]	2.00 [0.00, 4.00]	3.00 [0.00, 8.00]
active_days							
Mean (SD)	1.61 (1.16)	4.86 (2.57)	16.1 (14.4)	2.29 (1.33)	1.86 (0.961)	4.45 (4.00)	4.56 (5.88)
Median [Min, Max]	1.00 [1.00, 7.00]	4.00 [1.00, 24.0]	11.0 [2.00, 90.0]	2.00 [1.00, 20.0]	2.00 [1.00, 10.0]	3.00 [1.00, 24.0]	3.00 [1.00, 90.0]
active_period							
Mean (SD)	2.64 (4.64)	9.55 (6.88)	54.3 (15.0)	4.06 (3.94)	3.70 (5.10)	12.5 (11.3)	10.9 (15.1)
Median [Min, Max]	1.00 [1.00, 43.0]	7.00 [1.00, 43.0]	53.0 [24.0, 95.0]	3.00 [1.00, 39.0]	2.00 [1.00, 50.0]	9.00 [1.00, 45.0]	6.00 [1.00, 95.0]
avg_group_size							
Mean (SD)	11.0 (4.10)	2.12 (1.01)	1.09 (0.856)	2.82 (1.27)	2.84 (1.44)	1.17 (0.912)	2.35 (1.53)
Median [Min, Max]	10.0 [8.00, 30.0]	2.00 [0.00, 8.00]	1.00 [0.00, 7.00]	3.00 [0.00, 7.00]	3.00 [0.00, 9.00]	1.00 [0.00, 7.00]	2.00 [0.00, 30.0]
weekends							
Mean (SD)	15.0 (32.1)	27.2 (18.1)	22.1 (19.7)	5.57 (12.5)	86.9 (16.5)	8.74 (15.5)	22.8 (27.2)
Median [Min, Max]	0.00 [0.00, 100]	27.0 [0.00, 92.0]	20.0 [0.00, 100]	0.00 [0.00, 53.0]	100 [50.0, 100]	0.00 [0.00, 80.0]	16.0 [0.00, 100]
transactions_cgc							
Mean (SD)	73.1 (34.8)	66.8 (17.3)	52.4 (32.5)	78.2 (22.7)	69.0 (29.5)	6.28 (12.1)	65.7 (28.2)
Median [Min, Max]	100 [0.00, 100]	68.0 [0.00, 100]	53.0 [0.00, 100]	80.0 [0.00, 100]	66.0 [0.00, 100]	0.00 [0.00, 55.0]	68.0 [0.00, 100]

Next, cluster #2 stands out for being the one with the lowest spatial concentration—visiting more than three municipalities, but not exclusively in the Cambrils-Salou-Vilaseca area. Clusters #2 and #4 represent the most common usages of T-10 cards by tourists and visitors. Clusters #5 to #6 correspond to those users who barely spent ten trips, which is the minimum number of trips that any fare can charge. It is normal that the lower the number of trips, the greater the spatial concentration of travellers and, for example, clusters #5 and #6 concentrated practically all of their trips in three or two municipalities. In cluster #6, it can be seen that only 6.28% of the trips were made in those municipalities where tourist activity is greater. These clusters could be further refined, and the variables could be selected with different criteria. For example, the variables that describe the number of consecutive trips that a card makes are very similar. However, this description of the data makes it possible to highlight behaviours in the use of the cards that do not resemble what is expected for tourists and visitors. For example, it could be the case that a certain buses get full using less than three cards. It could also be useful to characterise better the spatial concentration of trips to size the offer. If there are service_areas or routes that concentrate most of the trips in a certain time slot, it would be interesting to reshape the service at those areas or routes. Finally, the specific needs related to the transport title that best suits public transport users' needs can be profiled in terms of spatial distribution and time evolution. As a result, the choice of the location of the sale points and forecasts of the specific needs of the demand can be more accurate.

To visualise the results of the clustering analysis, the R platform also provides numerous visualisation tools, including some GIS capabilities. The application layer of the proposed system architecture can rely on it. The activity patterns were made comparable by extracting the most representative cards of each cluster. This was accomplished by filtering one-hundred cards according to the probability with which they were classified in their cluster. Plotting these patterns in a calendar-based visualisation can help to explain the activity patterns of each cluster. This was done using the sugrrants package [63]. These plots allow comparing temporal patterns at an hourly and daily temporal resolution. Many different data views could be created, even a data dashboard containing different plots and tables.

As part of the data mining process, it could be seen that cluster #1 (roaming groups) comprises the most irregular patterns or that clusters #4 and #5 are highly variable in the dichotomy between weekdays and weekends. As an example of this type of visualisation, Figure 7 shows a frame-calendar of the activity patterns of the top-100 cards in cluster #4. As mentioned above, it is not an objective of this work to analyse these results in detail. However, some general patterns can be appreciated in Figure 7. There is a progressive increase in activity at the beginning of the summer season and a decrease in activity at the end of the season, respectively. Finally, different activity patterns during weekends and holidays (i.e., on 15 August or 11 September) can be seen.

Regarding the patterns revealed in the clustering analysis, it can be seen that several clusters show similar activity patterns (e.g., one main activity peak in the morning and a secondary one in the evening), only with different intensities and lengths of stays. As already said, this kind of patterns have been further studied and explained by transport and mobility researchers in a more detailed data analysis [14].

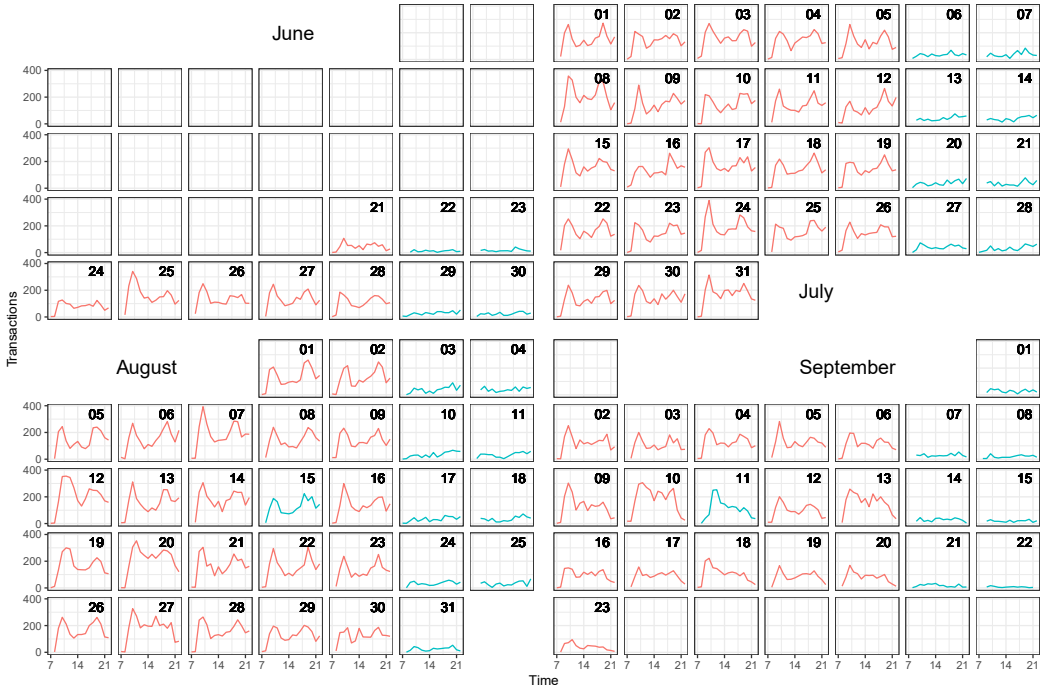

Figure 7. Cluster #4 activity calendar (n = 13,086). Red activity patterns mean weekday and blue activity patterns represent weekend or holiday.

5. Concluding Remarks

In Sections 1 and 2, it is explained how smart travel card data can play an important role in improving management and decision-making in public transport systems. However, we agree with other authors that many transport authorities and operators tend to make poor use of smart travel card data despite recognising the value of its potential applications [17]. The difficulties entailed by the need to cope with big datasets are often highlighted as a significant challenge that prevents the use of the data [64]. In some cases, this barrier has been successfully circumvented using win-win collaboration agreements between research centres and transport authorities and operators [33]. Indeed, this sort of collaboration has enabled the exploitation of the data gathered in the SGIT system of ATMCdT. A system architecture that addresses all the shortcomings described when analysing smart travel card datasets generated in small to medium-sized transport systems is described and implemented.

This research describes the design and development of a GIS database for managing public transport smart travel card data, along with the necessary tools and resources required to complete useful data mining analyses. A background about how these data are collected and analysed for different purposes is provided. The database model is then designed and implemented as a prototype that ensures the usability of this solution from the most used GIS and analytical applications (QGIS, ArcGIS, R or Microsoft Excel, among others). Finally, this model is applied to the ATMCdT case study for showing its suitability for solving a typical analysis involving smart travel card data. In addition to what other previous researches in this field provide, our proposal explicitly offers a schema that includes different analytical functionalities such as different spatio-temporal data aggregations of results, the possibility of piping the outputs to other more advanced analyses and a portable solution made in plain SQL.

Using a case study, it has been shown how this architecture enables a knowledge discovery process that can be useful to stakeholders such as ATMCdT. The purpose of the analysis presented in Section 4 was to demonstrate the implementation of the system using FOSS and the role of each system component. Furthermore, this case study is particularly interesting as it adds new evidence for a line of research that remains open. As other researchers have pointed out, there is a lack of adequate data to capture the evolution of

dynamics in the tourist destination characterised by seasonal patterns of tourist arrivals [65]. However, it should be remembered that these questions are beyond the objectives of this work. The clustering analysis presented in Section 4 should be reviewed and discussed in a specific study of mobility in tourist destinations. In this sense, a fully developed example of such a study can be found in [14].

In other previous studies, the methodological framework is not usually described, not addressing the reproducibility criteria that many authors have been defending in recent decades [66]. One of the main contributions of this work is to describe this framework in detail, using diagrams together with a textual description, so that other authors and users can start from a common base and reuse the efforts in the analysis of smart card data. The architecture and models proposed here are conceptual and agnostic in term of implementation, and thanks to the use of well-known standards, various technological solutions could be implemented. In any case, it can be added that the use of FOSS facilitates greater flexibility when proposing significant changes in a system [67], or for building more reproducible workflows [66]. As smart card data cannot usually be published (for example, due to privacy agreements), sharing the code for database queries and analyses will not make research fully reproducible. As an alternative, it would be possible to describe the data model in detail or to provide data samples with a demonstrative value. In addition to the description of the framework and the data model, another more concrete contribution of this work is the publication of the code repository that has been used for the proposed case study. If other researchers agree that the proposed framework meets their needs, our contribution could help them analyse an AFC log for the first time or, even better, the publicly available code repository could be used as a method sharing platform.

The geospatial data formats such as Spatialite or OGC Geopackages are becoming compelling standards for sharing and analysing geospatial data. These formats are highly integrated with many GIS tools and enable using virtually any tool necessary for processing transport smart travel card data logs (including spatial queries and operations, network and distance calculations, multimedia attachments and slope-based time calculations). This flexibility has only been tested in this research, and very simple spatio-temporal SQL aggregations have been performed in SQL. These aggregations guarantee to preprocess and filter a significant amount of data into smaller datasets that can be better explored in other analytical frameworks (R/Python). However, it is shown how the database model can be enriched by adding additional thematic and spatial datasets. In this way, further research will enable designing a complete database model to answer most of the questions that can be solved by analysing smart travel card data. Moreover, if these data formats are not robust enough to absorb all the necessary datasets, the use of SQL databases facilitates the portability of the model to other more suitable platforms (e.g., PostgreSQL/PostGIS). Along with the extended use of RDBMS, this portability is yet another reason why many researchers use SQL databases for analysing this type of data [57].

The proposed model is considered to be extensible but still closely linked to the first case study analysed. This solution is designed for research teams with the resource limitations presented in the introduction. Therefore, it is expected that more powerful (and resource-demanding) technological solutions will be adopted once the value of the data becomes clear. Finally, we consider that the proposed framework and the shared code are far from solving the particular difficulties that will arise when analysing the data of each AFC system. A solution that encourages the use of open standards and less restricted access to this type of data would be necessary, but in the meantime, looking for a common framework that facilitates the exploitation of these data may be a good option.

Author Contributions: Conceptualisation, B.Z.; Data curation, D.M. and B.Z.; Formal analysis, B.Z.; Funding acquisition, A.G.; Investigation, A.G.; Methodology, B.Z. and S.T.; Project administration, A.G.; Resources, S.T. and D.M.; Software, B.Z. and S.T.; Supervision, A.G.; Writing—original draft, B.Z. and S.T.; Writing—review and editing, A.G. and D.M. All authors have read and agreed to the published version of the manuscript.

Funding: Research funded by the "Fondo Supera Covid-19", created by the Santander Bank, CRUE Universidades Españolas and the Consejo Superior de Investigaciones Cienfíficas (CSIC) (project title: COVMOVTUR-COVID-19 and mobilities in tourist regions: change of patterns and their effect on physical and mental health of residents and visitors) and the Escola d'Administració Publica de Catalunya, Generalitat de Catalunya [grant number 2018 EAPC 00002]. Sergio Trilles has been funded by the postdoctoral Juan de la Cierva fellowship programme of the Spanish Ministry for Science and Innovation (IJC2018-035017-I).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anda, C.; Erath, A.; Fourie, P.J. Transport modelling in the age of big data. *Int. J. Urban Sci.* **2017**, *21*, 19–42. [CrossRef]
2. Green, J.; Chickola, L.; Emanuel, E.S.; Cruickshank, A. Automated Fare Collection System. US Patent 6,957,772, 25 October 2005.
3. Makarova, I.; Pashkevich, A.; Shubenkova, K. Ensuring Sustainability of Public Transport System through Rational Management. *Procedia Eng.* **2017**, *178*, 137–146. [CrossRef]
4. Kurauchi, F.; Schmöcker, J.D. (Eds.) *Public Transport Planning with Smart Card Data*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2017; p. 274. [CrossRef]
5. Morency, C.; Trépanier, M.; Agard, B. Measuring transit use variability with smart-card data. *Transp. Policy* **2007**, *14*, 193–203. [CrossRef]
6. Pelletier, M.P.; Trépanier, M.; Morency, C. Smart card data use in public transit: A literature review. *Transp. Res. Part C Emerg. Technol.* **2011**, *19*, 557–568. [CrossRef]
7. Bagchi, M.; White, P.R. The potential of public transport smart card data. *Transp. Policy* **2005**, *12*, 464–474. [CrossRef]
8. Ma, X.; Wu, Y.j.; Wang, Y.; Chen, F.; Liu, J. Mining smart card data for transit riders' travel patterns. *Transp. Res. Part C* **2013**, *36*, 1–12. [CrossRef]
9. Alsger, A.A.; Mesbah, M.; Ferreira, L.; Safi, H. Use of Smart Card Fare Data to Estimate Public Transport Origin—Destination Matrix. *Transp. Res. Rec. J. Transp. Res. Board* **2015**, *2535*, 88–96. [CrossRef]
10. Zahnow, R.; Corcoran, J. Crime and bus stops: An examination using transit smart card and crime data. *Environ. Plan. B Urban Anal. City Sci.* **2021**, *48*, 706–723. [CrossRef]
11. Arbex, R.; da Cunha, C.B.; Speicys, R. Before-and-after evaluation of a bus network improvement using performance indicators from historical smart card data. *Public Transport* **2019**, 1–19. [CrossRef]
12. Lu, Y.; Mateo-Babiano, I.; Sorupia, E. Who uses smart card? Understanding public transport payment preference in developing contexts, a case study of Manila's LRT-1. *IATSS Res.* **2019**, *43*, 60–68. [CrossRef]
13. Lawson, C.T.; Tomchik, P.; Muro, A.; Krans, E. Translation software: An alternative to transit data standards. *Transp. Res. Interdiscip. Perspect.* **2019**, *2*, 100028. [CrossRef]
14. Gutiérrez, A.; Domènech, A.; Zaragozí, B.; Miravet, D. Profiling tourists' use of public transport through smart travel card data. *J. Transp. Geogr.* **2020**, *88*, 13. [CrossRef]
15. Gutiérrez, A.; Miravet, D. The determinants of tourist use of public transport at the destination. *Sustainability* **2016**, *8*, 908. [CrossRef]
16. Faroqi, H.; Mesbah, M.; Kim, J. Applications of transit smart cards beyond a fare collection tool: A literature review. *Adv. Transp. Stud.* **2018**, *45*, 107–122.
17. Shmöcker, J.; Karauchi, F.; Shimamoto, H. An Overview on Opportunities and Challenges of Smart Card Data Analysis. In *Public Transport Planning with Smart Card Data*; Kurauchi, F., Schmöcker, J.D., Eds.; CRC Press: Boca Raton, FL, USA, 2017; pp. 1–14.
18. Chandesris, M.; Nazem, M. Workshop Synthesis: Smart card data, new methods and applications for public transport. *Transp. Res. Procedia* **2018**, *32*, 16–23. [CrossRef]
19. Hickman, M. Transit origin-destination estimation. In *Public Transport Planning with Smart Card Data*; Kurauchi, F., Schmöcker, J., Eds.; CRC Press: Boca Raton, FL, USA, 2017; pp. 15–35.
20. Kusakabe, T.; Asakura, Y. Combination of smart card data with person trip survey data. In *Public Transport Planning with Smart Card Data*; Kurauchi, F., Schmöcker, J., Eds.; CRC Press: Boca Raton, FL, USA, 2017; pp. 1–14.
21. Brakewood, C.; Watkins, K. A Method for Conducting Before-After Analyses of Transit Use by Linking Smart Card Data and Survey Responses. In *Public Transport Planning with Smart Card Data*; Kurauchi, F., Schmöcker, J.D., Eds.; CRC Press: Boca Raton, FL, USA, 2017; pp. 93–111.
22. Ali, A.; Lee, S. Destination and Activity Estimation. In *Public Transport Planning with Smart Card Data*; Kurauchi, F.; Schmöcker, J.D., Eds.; CRC Press: Boca Raton, FL, USA, 2017; pp. 37–53.
23. Ghaemi, M.S.; Agard, B.; Trépanier, M.; Partovi Nia, V. A visual segmentation method for temporal smart card data. *Transp. A Transp. Sci.* **2017**, *13*, 381–404. [CrossRef]

24. Ortega-Tong, M.A. Classification of London's Public Transport Users Using Smart Card Data. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2013.
25. Gudivada, V.N.; Rao, D.; Raghavan, V.V. NoSQL systems for big data management. In Proceedings of the 2014 IEEE World Congress on Services, Anchorage, AK, USA, 27 June–2 July 2014; pp. 190–197.
26. Prakasa, B.; Putra, D.W.; Kusumawardani, S.S.; Widhiyanto, B.T.Y.; Habibie, F. Big data analytic for estimation of origin-destination matrix in Bus Rapid Transit system. In Proceedings of the 2017 3rd International Conference on Science and Technology-Computer (ICST), Yogyakarta, Indonesia, 11–12 July 2017; pp. 165–170.
27. Fabbiani, E.; Vidal, P.; Massobrio, R.; Nesmachnow, S. Distributed big data analysis for mobility estimation in intelligent transportation systems. In *Proceedings of the Latin American High Performance Computing Conference, Mexico City, Mexico, 29 August–2 September 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 146–160.
28. Fiore, S.; Elia, D.; Pires, C.E.; Mestre, D.G.; Cappiello, C.; Vitali, M.; Andrade, N.; Braz, T.; Lezzi, D.; Moraes, R.; Aloisio, G. An Integrated Big and Fast Data Analytics Platform for Smart Urban Transportation Management. *IEEE Access* **2019**, *7*, 117652–117677. [CrossRef]
29. Fiore, S.; D'Anca, A.; Elia, D.; Palazzo, C.; Williams, D.; Foster, I.; Aloisio, G. Ophidia: A full software stack for scientific data analytics. In Proceedings of the 2014 International Conference on High Performance Computing & Simulation (HPCS), Bologna, Italy, 21–25 July 2014; pp. 343–350.
30. Barth, R.S.; Galante, R. *Passenger Density and Flow Analysis and City Zones and Bus Stops Classification for Public Bus Service Management*; Brazilian Symposium on Databases: Salvador, Brazil, 2016; pp. 217–222.
31. Gokasar, I.; Simsek, K. Using "Big Data" For Analysis and Improvement of Public Transportation Systems in Istanbul. In Proceedings of the Ase Bigdata/Socialcom/Cybersecurity Conference, ©ASE 2014, Stanford University, Stanford, CA, USA, 27–31 May 2014; Academy of Science and Engineering (ASE): Los Angeles, CA, USA, 2014.
32. Li, T.; Sun, D.; Jing, P.; Yang, K. Smart card data mining of public transport destination: A literature review. *Information* **2018**, *9*, 18. [CrossRef]
33. Wu, H.; Tan, J.A.; Ng, W.S.; Xue, M.; Chen, W. FTT: A system for finding and tracking tourists in public transport services. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Victoria, Australia, 31 May–4 June 2015; pp. 1093–1098.
34. Lovelace, R.; Parkin, J.; Cohen, T. Open access transport models: A leverage point in sustainable transport planning. *Transp. Policy* **2020**, *97*, 47–54. [CrossRef]
35. PostgreSQL 12 Documentation. Appendix D. SQL Conformance. 2020. Available online: https://www.postgresql.org/docs/12/features.html (accessed on 26 July 2021).
36. ISO Central Secretary. *Information Technology–Database Languages–SQL—Part 1: Framework*; Standard ISO/IEC TR 9075-1:2016; International Organization for Standardization: Geneva, Switzerland, 2016.
37. Zaragozí, B.; Gutiérrez, A.; Trilles, S. Towards an Affordable GIS for Analysing Public Transport Mobility Data: A Preliminary File Naming Convention for Avoiding Duplication of Efforts. In Proceedings of the 6th International Conference on Geographical Information Systems Theory, Applications and Management, Heraklion, Greece, 3–5 May 2019; SCITEPRESS—Science and Technology Publications: Setubal, Portugal; 2020; pp. 302–309. [CrossRef]
38. Gutiérrez, A.; Miravet, D. Estacionalidad turística y dinámicas metropolitanas: Un análisis a partir de la movilidad en transporte público en el Camp de Tarragona. *Revista de Geografía Norte Grande* **2016**, *89*, 65–89. [CrossRef]
39. Domènech, A.; Gutiérrez, A. A GIS-Based Evaluation of the Effectiveness and Spatial Coverage of Public Transport Networks in Tourist Destinations. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 83. [CrossRef]
40. Domènech, A.; Miravet, D.; Gutiérrez, A. Mining bus travel card data for analysing mobilities in tourist regions. *J. Maps* **2020**, *16*, 40–49. [CrossRef]
41. Open Geospatial Consortium. Available online: https://www.opengeospatial.org/standards (accessed on 26 July 2021).
42. Ghaemi, M.S.; Agard, B.; Nia, V.P.; Trépanier, M. Challenges in Spatial-Temporal Data Analysis Targeting Public Transport. *IFAC-PapersOnLine* **2015**, *48*, 442–447. [CrossRef]
43. Sezhian, M.V.; Muralidharan, C.; Nambirajan, T.; Deshmukh, S. Performance measurement in a public sector passenger bus transport company using fuzzy TOPSIS, fuzzy AHP and ANOVA—A case study. *Int. J. Eng. Sci. Technol. (IJEST)* **2011**, *3*, 1046–1059.
44. Briand, A.S.; Côme, E.; Trépanier, M.; Oukhellou, L. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transp. Res. Part C Emerg. Technol.* **2017**, *79*, 274–289. [CrossRef]
45. Morency, C.; Trépanier, M.; Agard, B. Analysing the variability of transit users behaviour with smart card data. In Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference, Toronto, ON, Canada, 17–20 September 2006; pp. 44–49.
46. El Mahrsi, M.K.; Come, E.; Oukhellou, L.; Verleysen, M. Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 712–728. [CrossRef]
47. Agard, B.; Partovi Nia, V.; Trépanier, M. Assessing public transport travel behaviour from smart card data with advanced data mining techniques. In Proceedings of the World Conference on Transport Research, Rio de Janeiro, Brazil, 15–18 July 2013.
48. Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **1996**, *5*, 299–314.
49. Hothorn, T. CRAN Task View: Machine Learning & Statistical Learning. 2020. Available online: https://cran.r-project.org/web/views/MachineLearning.html (accessed on 26 July 2021).

50. Friedrich Leisch, B.G. CRAN Task View: Cluster Analysis & Finite Mixture Models. 2020 Available online: https://cran.r-project.org/web/views/Cluster.html (accessed on 26 July 2021).
51. Khan, M.; Khan, S.S. Data and information visualization methods, and interactive mechanisms: A survey. *Int. J. Comput. Appl.* **2011**, *34*, 1–14.
52. Koua, E.L.; Kraak, M.J. Alternative visualization of large geospatial datasets. *Cartogr. J.* **2004**, *41*, 217–228. [CrossRef]
53. Lock, O.; Bednarz, T.; Pettit, C. The visual analytics of big, open public transport data—A framework and pipeline for monitoring system performance in Greater Sydney. *Big Earth Data* **2020**, *5*, 134–159. [CrossRef]
54. Sedrakyan, G.; Mannens, E.; Verbert, K. Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. *J. Comput. Lang.* **2019**, *50*, 19–38. [CrossRef]
55. Zheng, J.G. Data visualization in business intelligence. In *Global Business Intelligence*; Taylor & Francis: London, UK, 2017; pp. 67–82.
56. Trépanier, M.; Habib, K.M.; Morency, C. Are transit users loyal? revelations from a hazard model based on smart card data. *Can. J. Civ. Eng.* **2012**, *39*, 610–618. [CrossRef]
57. Liu, Y.; Cheng, T. Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transp. A Transp. Sci.* **2018**, *16*, 76–103. [CrossRef]
58. Liang, Q.; Weng, J.; Zhou, W.; Santamaria, S.B.; Ma, J.; Rong, J. Individual Travel Behavior Modeling of Public Transport Passenger Based on Graph Construction. *J. Adv. Transp.* **2018**, *2018*, 3859830. [CrossRef]
59. Manley, E.; Zhong, C.; Batty, M. Spatiotemporal variation in travel regularity through transit user profiling. *Transportation* **2018**, *45*, 703–732. [CrossRef]
60. Faroqi, H.; Mesbah, M.; Kim, J. Spatial-temporal similarity correlation between public transit passengers using smart card data. *J. Adv. Transp.* **2017**, *2017*, 1318945. [CrossRef]
61. Briand, A.S.; Côme, E.; El Mahrsi, M.K.; Oukhellou, L. A mixture model clustering approach for temporal passenger pattern characterization in public transport. In Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, 19–21 October 2015; pp. 37–50. [CrossRef]
62. Akogul, S.; Erisoglu, M. An approach for determining the number of clusters in a model-based cluster analysis. *Entropy* **2017**, *19*, 452. [CrossRef]
63. Wang, E.; Cook, D.; Hyndman, R.J. Calendar-Based Graphics for Visualizing People's Daily Schedules. *J. Comput. Graph. Stat.* **2020**, *29*, 490–502. [CrossRef]
64. Imai, R.; Iboshi, Y.; Nakamura, T.; Morio, J.; Makimura, K.; Hamada, S. Consideration on practical use of trail data acquired by smart card of transportation. In Proceedings of the JSCE Annual Meeting, Tokyo, Japan, 5–7 September 2012.
65. Wang, Z.; He, S.Y.; Leung, Y. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behav. Soc.* **2018**, *11*, 141–155. [CrossRef]
66. Zaragozí, B.M.; Trilles, S.; Navarro-Carrión, J.T. Leveraging Container Technologies in a GIScience Project: A Perspective from Open Reproducible Research. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 138. [CrossRef]
67. Zaragozí, B.; Giménez, P.; Navarro, J.T.; Dong, P.; Ramón, A. Development of free and opensource GIS software for cartographic generalisation and occupancy area calculations. *Ecol. Inform.* **2012**, *8*, 48–54. [CrossRef]