Towards an Affordable GIS for Analysing Public Transport Mobility Data: A Preliminary File Naming Convention for Avoiding Duplication of Efforts

Benito Zaragozí[1][a], Aaron Gutiérrez[1][b] and Sergio Trilles[2][c]

[1]*Departament de Geografia, Universitat Rovira i Virgili, C/Joanot Martorell, Vilaseca, Spain*

[2]*Institute of New Imaging Technologies, Universitat Jaume I, Av. Vicente Sos Baynat s/n, Castellón de la Plana, Spain*

Keywords: Smart Card Data, Public Transportation, Domain-specific Language, File Naming Convention, Medium-sized Data.

Abstract: Automated fare collection systems for public transport generate a large volume of information on the mobility of people in urban environments. New technologies associated with Big Data can facilitate the analysis of these data. However, the application of these technologies can be expensive and resource-demanding, especially in medium and small cities. This paper presents the case of the metropolitan transport authority of Tarragona, for which an affordable and extensible analysis system has been developed, based on relational databases and custom scripts. Among the technical problems that have had to be overcome, one of the first has been the unambiguous definition of the numerous queries required by mobility experts. For different reasons, mobility researchers request aggregate data queries from smart transport cards logs (e.g. providing a descriptive statement) and expect manageable tables to be analysed in a spreadsheet. To standardise the definition of queries, a domain-specific language as a file naming convention has been proposed with which database managers and mobility experts can communicate efficiently, avoiding confusion, duplication of efforts and other problems detected. The file naming convention has been applied as an early version within the defined use case to verify the viability of this idea.

1 INTRODUCTION

Data generated from Automated Fare Collection Systems (AFCS) using smart transport cards is characterised by its dynamism, as each time a user gets on a public transport vehicle, the system collects many data from its validation. As a result, millions of observations have been collected that provide valuable information for understanding passenger behaviours and can also help improve service quality (Kurauchi and Schmöcker, 2017). Since all transactions are gathered, there is great flexibility for studying any temporal and geographical extent (Morency et al., 2007). The analysis of such data is very common in the scientific literature (Pelletier et al., 2011; Bagchi and White, 2005). For example, smart travel card data has been used to identify different profiles of public transport users (Ma et al., 2013), to reconstruct source-destination matrices (Alsger et al., 2015) or to analyse tourist mobility patterns (Lu et al., 2019).

Smart cards can gather all transactions, so the size of data might become so large that after a short period of time, it is difficult to handle. From this point of view, smart card data can be seen as one sort of Big Data, and it is considered that, only by using specific technology, we can consider nearly the whole population data for analysing passengers behaviour. While in traditional data analysis, data sampling is required to select a small portion of data, data mining methods – such as clustering analysis – are often adopted to understand the global dataset characteristics. However, these procedures are far more complicated when data size increases. Generally, to apply big data analysis, some key points need to be considered, but the most crucial one is to decide how to store that data. Various types of databases could be employed to save that kind of data, including Hadoop Distributed File System (HDFS), Hbase, Apache Cassandra and Redis, just to name a few.

During the last years, many research works have

[a] https://orcid.org/0000-0003-2501-484X
[b] https://orcid.org/0000-0003-0557-6319
[c] https://orcid.org/0000-0002-9304-0719

Zaragozí, B., Gutiérrez, A. and Trilles, S.
Towards an Affordable GIS for Analysing Public Transport Mobility Data: A Preliminary File Naming Convention for Avoiding Duplication of Efforts.
DOI: 10.5220/0009766303020309
In *Proceedings of the 6th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2020)*, pages 302-309
ISBN: 978-989-758-425-1

applied the above mentioned technologies to analyse different amounts of data generated by smart cards systems. As an example, Apache Spark and HDFS were used to analyse 160 million of records from 20 months of transactions, between June 2014 and January 2016, generated by the Jakarta's Bus Rapid Transit in Indonesia (Prakasa et al., 2017). The same big data technologies were used in combination with Apache Kafka to support real-time analysis (Barth and Galante, 2016), building a so called Data Mining Framework for Bus Service Management (DMBSM). This system was tested using a reduced dataset of 3,186 records of smart cards in Shenzhen city (China), between October 22 and October 23, 2013. Another approach used SQL to extract passenger origin information and SPSS and Microsoft Excel to highlight connections among different transportation data sets, including the data generated by the smart cards (Istanbulkart) of the Bus Rapid Transit system (BRT) in Istanbul (Turkey) (Gokasar and Simsek, 2014). The same authors utilised a MySQL open-source database, MS Excel CSV file format and Rstudio software to study passenger origin information from the Istanbulkart data (Gokasar et al., 2015). This time the authors used a one-day dataset with 800,000 journey records. Another study analysed six months of smart card data from Brisbane — the third largest Australian city (Tao et al., 2014). The authors did not indicate the use of any specific technology to perform the analysis. In 2015, another study of the city of Montevideo (Uruguay) analysed smart card data logs comprising nearly 200 GB of data (Fabbiani et al., 2016). In this case, the authors used the *dispy* framework for creating and distributing parallel tasks among several computer nodes and *QGIS* software to manage geographic information and perform spatial analysis.

As can be seen, there are experiences in applying different robust systems in large and medium-sized cities, but there are also experiences in small cities or metropolitan areas. Obviously, larger cities have more resources to manage and analyse this data, so it is more viable to take advantage of the logging system to its full potential. However, in smaller cities, there are problems in adopting these new technologies. For these cases, collaborations between public transport authorities and research institutions have become a common strategy for analysing this specific type of data (Wu et al., 2015). This research work is an example of such type of collaboration, as it is the result of a research project carried out jointly by a research group of the Rovira i Virgili University, the Territorial Mobility Authority of Camp de Tarragona (ATMCdT), and the ATM in the Lleida area. In this research project, experts from different areas collaborate, including experts in mobility and transport studies, database managers, GIS analysts and technicians from the ATMCdT. As we explain later, good communication is essential in such a multidisciplinary research environment. In this context, the key objectives of this research are summarised are:

1. To study the main characteristics and the potential of the log data gathered by the ATMCdT.
2. To analyse the requirements for managing these data and the workflows between the different actors that need to cooperate to exploit these data.
3. To propose a Domain-Specific Language (DSL) for communicating ATMCdT technicians, database managers and mobility researchers that enhances code re-utilisation, saving time and effort for understanding the data. This DSL will be used as a file naming convention and can describe database queries in the context of public transport mobility.
4. To test and evaluate the proposed DSL in the context of a research project.

The remainder of this paper is organised as follows. The next section (Section 2 describes our case study, the main characteristics of the ATMCdT log data and the issues affecting its analysis. Section 3 proposes a definition of a DSL. This approach enables better communication between different actors, minimising the need for investing in new tools and avoiding the duplication of efforts. Section 4 draws conclusions and outlines further research directions for this affordable system.

2 CASE STUDY

The smart card data collected by the ATMCdT is of great value, as previously demonstrated in studies that analyzed the effectiveness and spatial coverage of the public transport system (Domènech and Gutiérrez, 2017) or other studies that focused on the use of public transport by tourists (Gutiérrez and Miravet, 2016a; Gutiérrez and Miravet, 2016b; Domènech et al., 2020). However, although these previous studies faced data management problems similar to those we will explain in this section, these issues tend to go unnoticed despite slowing down or making data analysis difficult.

2.1 Data Preprocessing

In this work, the Territorial Mobility Authority of Camp de Tarragona (ATMCdT), which is shown in

Figure 1, and ATM Lleida facilitate the log databases generated by their fare collection systems. This joint system is known as Fare Integration Management System (SGIT, according to its acronym in Catalan). This information system collects data for accounting purposes, so it should be understood that mobility studies were not contemplated in its initial design. The data collected includes –but it is not limited to– the exact day and time of travel, the stop where the passenger boarded, the company and the carrier that operates the transport, the municipality and the type of fare used in each transaction. Sometimes the destination stop can also be registered. The main advantage of the resulting databases is that the information is dynamic when it acquires a temporal dimension, and it allows to consult any registered time period. By contrast, these databases do not store data on the socio-economic profile of travellers, or these data can not be accessed due to legal restrictions.

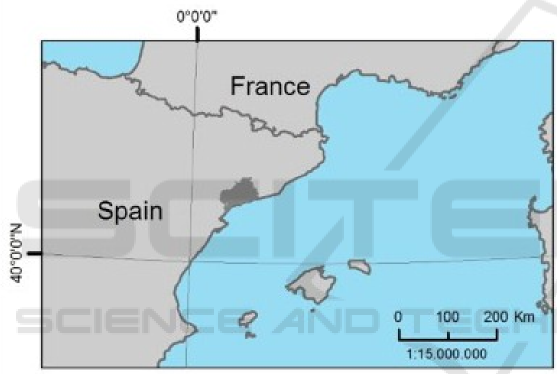

Figure 1: Location of the the Territorial Mobility Authority of Camp de Tarragona (ATMCdT) service area.

In the analysis of automated fare collection systems, one of the first steps after data collection is the data cleaning phase. There are different reasons for cleaning the dataset and improving the quality of data collection (Chandesris and Nazem, 2018). In this case, the log files analysed in this study were provided on a portable hard drive by ATMCdT. These records include all activity in 2018 for 135,365 different smart cards – 133,079 have been used more than once – and all single-trip tickets sold in the study area. This dataset weighs 6.1 GB in plain ASCII text format. During this period, the system collected 7,393,654 smart card travel transactions (rows) with 60 different attributes (columns) and 14,006,212 single-trip transactions with 22 attributes. This dataset did not include other operations that are also systematically recorded, such as card sales, recharges or cancellations, among others. The files generated by the ATMCdT platform are not standardised in terms of the type of information. We have reduced and cleaned all no relevant data for this study, performing a pre-processing step to create a relational database (SQLlite) and facilitate the first steps of analysis. This workflow, based on SQL databases, is very common in previous studies on transport smart cards, and it is very convenient to analyse datasets of a medium to big size (Li et al., 2018).

ATMCdT records have a very basic structure; they are sorted by date and time, the code of the smart card that performs the transaction, the type of transaction performed and other similar parameters. Given the research objectives, we excluded urban bus transactions from Tarragona and Reus and only filtered interurban transactions (5,414,028 transactions), selecting 9 out of 60 available attributes. Thanks to this selection and standardization process, we have significantly reduced the size of the database – from 9.1 GB to 1.5 GB, including indexes and support tables. More specifically, we discarded single-value columns, duplicated columns – due to legacy system changes – and derived columns (that is, aggregates that can be calculated from other attributes). The resulting database divides the records into different tables following a simple relational structure (see Figure 2), with the main transactions table linked to other tables built for normalisation purposes (agencies, fares, stops, municipalities, and routes). For clarity, the tables and columns were named following the General Traffic Feed Specification (GTFS), which is a well-known standard format for publishing public transit schedules. In addition to the ATMCdT data, the database was enriched with some layers of geographical information: (1) the municipalities, roads and shoreline were downloaded from the Cartographic and Geological Institute of Catalonia (www.icgc.cat) and adapted to the needs of the project, (2) the stops and routes were digitised manually, (3) official population data (https://www.idescat.cat/) and (4) ATMCdT zones.

2.2 Problem Statement

The design of the database is quite generic and is given by the characteristics of the raw data (see Figure 2). However, after the preprocessing step, there is a context that needs from a multidisciplinary approach. The research group is composed by two different types of profiles described as:

1. Mobility experts and researchers. The main objective of this profile is to interrogate the data management system (SQLite) by performing different types of queries, which include temporal and spatial dimensions. This profile has an inter-

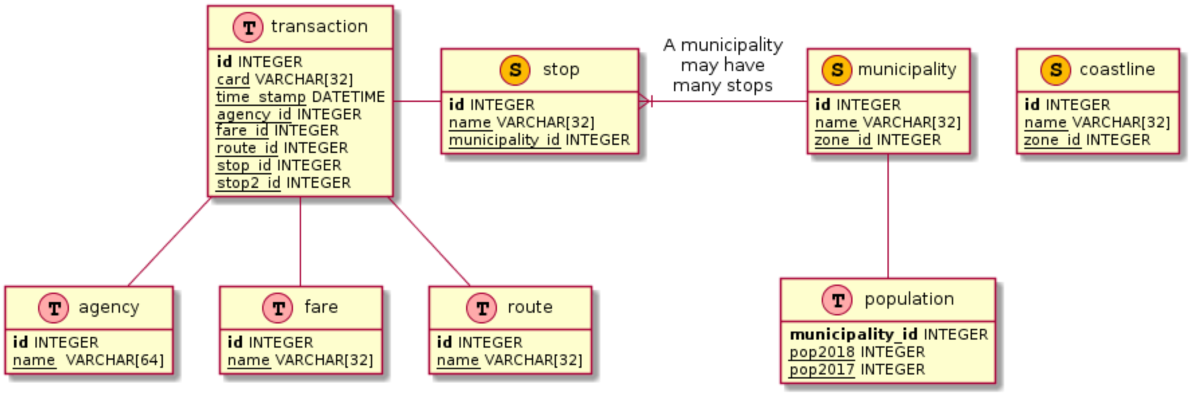

Figure 2: Entity relationship database model for storing ATMCdT Smart Card log data. We distinguish between regular attribute tables (*T*) and spatial tables (*S*).

est in the query results in order to perform further mobility analysis and participate in decision-making processes.

2. Spatial database manager. This profile is in charge of interacting with the data management system, coding SQL queries which satisfy the requests made by a mobility expert and execute them to get the results in a convenient data format (e.g. Spreadsheets).

This team structure presents a communication problem, between the mobility expert and the database administrator, which we discovered during the first phases of work. Both roles have expertise in its own domain but lack cross-domain knowledge to intuitively understand the objectives or the difficulties of a particular query. On the one hand, 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 difficult to express it in the most direct way to define an accurate SQL query. On the other hand, the database manager may have difficulties in understanding the specific purpose of a query. In order to obtain the correct result, a certain iterative process between both actors is usually necessary. This process is shown in Figure 3. On the left hand side, Figure 3 A) is a sequence diagram that formally describes the process. The mobility researcher must describe the query to the database manager actor as many times as necessary until obtaining the correct result. This process can be very tedious, repetitive or redundant, especially when the number of different queries increases.

In addition to the time spent in understanding the project needs, both actors (database manager and mobility researcher) can lose time in executing queries that are not interesting for the project. As in any analysis, experts in the field need to generate a lot of queries on databases, and not always is expected to get a beneficial outcome for decision making. Sometimes these queries do not vary substantially, so from the database manager's point of view, previous queries can be easily re-used to create new ones if they are documented correctly.

3 PROPOSED SOLUTION

According to the issues described in the previous section, the main bottleneck in the process appears when the quality of definition of the queries is poor or the target result is not correctly visualised. For this kind of problems, it would be possible to apply a technologically based solution. However, due to the limitations described above, developing or introducing new tools is not an option, but an undesired dependence.

As a solution to this problem, we propose the definition of a Domain-Specific Language (DSL) adapted to the needs of this type of applications. From here we will refer to this specification as the Mobility File Naming Convention (*MobilityFNC*). This approach which will be used as a convention to describe and identify the queries properly. Each query will be named using this convention and both files — the SQL query defined by the database experts and the results file — will be inseparable. This solution allows us to build a catalogue of already executed queries. Thus, before developing a new query, it will be possible to verify if this query has been previously coded or if there is a similar one to use as a model.

Figure 3.B presents the new workflow enhanced by *MobilityFNC*. As previously mentioned, the main advantage of *MobilityFNC* is to allow the possibility of building a query repository, with SQL scripts and their derived results (e.g. spreadsheets). Having the

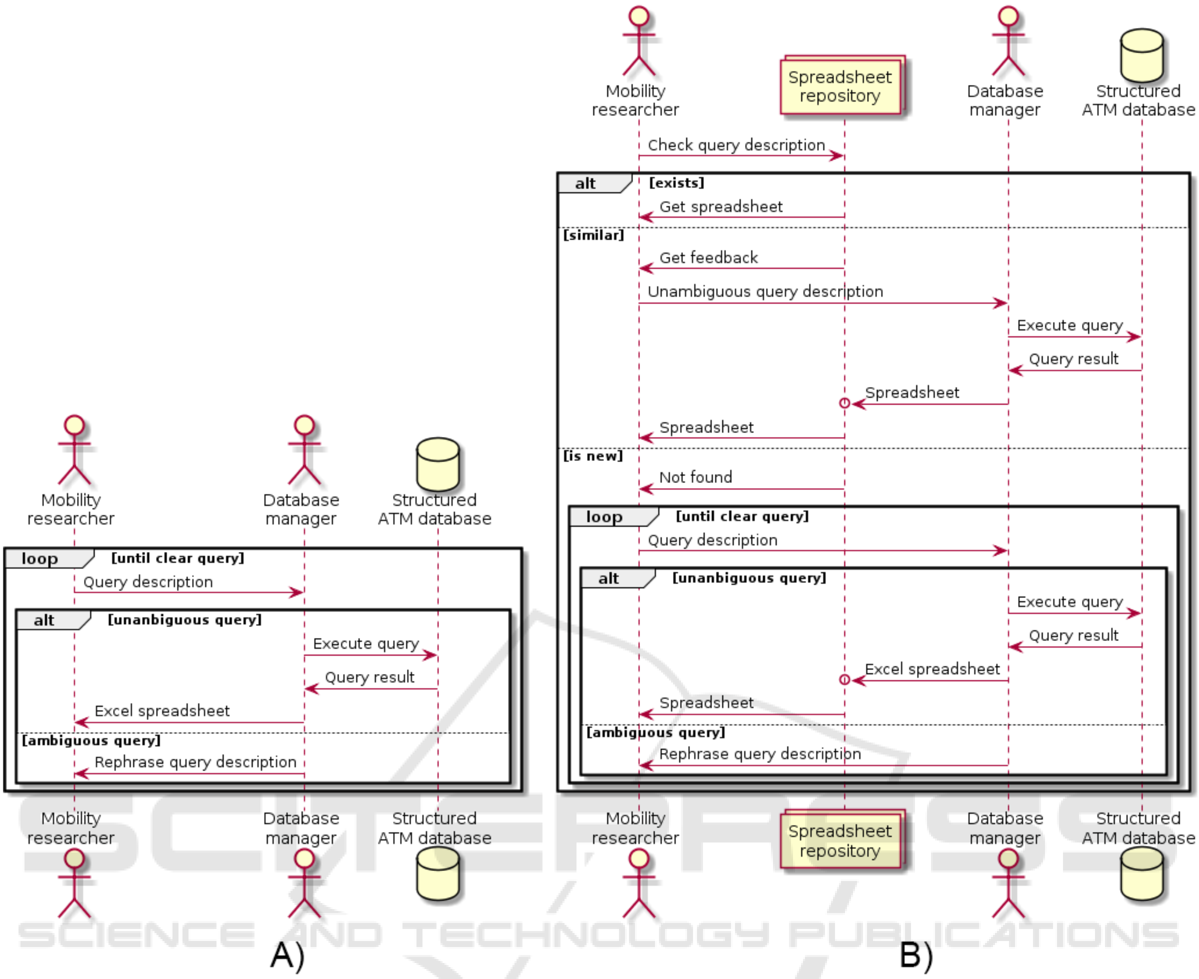

Figure 3: A) Sequence diagram showing the workflow between a mobility researcher and database manager in the context of a public mobility project; B) Sequence diagram showing the workflow between a mobility researcher and database manager using the *MobilityFNC* aproach.

access to that repository granted, there are three possible situations:

1. Self-sufficiency. If the query description already exists in the catalogue, mobility researchers could directly access it in the form of a spreadsheet (e.g. another researcher asked for the same query).
2. Feedback. If the query was not previously executed but there was any other similar query, in that case it would be easier to the database expert to adapt it (e.g. the new query refers to a different date, but apart from that it is identical).
3. Finally, If the query does not exist, the database manager will use the query description encoded with *MobilityFNC* and generate the SQL query, execute and incorporate it into the catalogue.

This interactive process between mobility researcher and database manager should not be tedious, repetitive or redundant, since the database manager would write the SQL expression following the proposed DSL, which should be always more concrete than a semi-structured text. *MobilityFNC* should be easily understood by researchers with no advanced experience in SQL. It is motivated to improve understanding between a mobility expert and a database manager. This specific language defines an structured and understandable way for mobility experts with which to consult information from a database without the need of using an advanced query language.

In the literature, there are many DSLs for different application domains. Among the best known we can highlight HTML, Unix shell scripts or GraphViz, among others. The main advantage of DSL over General Purpose Languages (GPL) is that DSLs are more attractive due to the proximity to the context by adopting more sensible programming features and systematic reuse (Van Deursen et al., 2000). However, designing and developing a DSL is not always meaningful, for example, SQL can be considered as a DSL that focuses on the domain of the database, but it is still too general and more complex than some scripting languages (Hudak, 1997).

MobilityFNC is used to name the files of the SQL queries with which it is translated. In this way, we can quickly identify if the SQL query has been previously encoded or may be helpful to encode a new one (e.g. if there are more queries involving the same information). Although it could be extended to different types of public mobility, such as car, train or plane, due to the selected use case, we base *MobilityFNC* on public bus mobility. Arguably similar to SQL, we will use *MobilityFNC* as a way for describing the *shape* and main elements of a table resulting from a query. Thus, we intend to maximise its compatibility and improve this proposal so it could be translated into valid SQL, at least for a previously known database model.

3.1 Lexicon

As already explained, a *MobilityFNC* expression will be used to name a file that stores a SQL query. Of course, depending on the operating system, certain characters are not allowed to be used for file naming. Following these requirements, *MobilityFNC* allows any byte except NULL, \, /, :, %, ?, ∗, ”, <, >, |. Another limitation is not to exceed 255 characters per query description, so it is important to avoid unnecessary redundancies. Based on this character restriction, a list of operators has been defined for *MobilityFNC*:

- Separate main blocks (source, filter, dimension and Operations) with “+”
- Add a new element at a same level with “−”
- Start a new level and add an element with “_”
- Separate rows and columns with “∼”
- Define a range in canonical form with “[␣]”
- Set a function or method with “{␣}”
- Determine an array of variables with “[␣,␣]”

In addition to the operators, there is also a selection of restricted words to encode queries. We distinguish 6 categories with selection of restricted words to be used as vocabulary to encode queries.:

- Aggregation. E.g. count, totals, subtotals, top.N, htotal, among others.
- Attributes. SQL table attributes shown in the ER model (Figure 2).
- Boolean. E.g. pop.between, pop.over, pop.less, pop.equal or applied to other associated attributes.
- Ranges. We use the -ly termination to refer to a known ranging (eg. monthly, yearly, etc), summerly or nonsummerly.
- Sources. ATM smart cards and TP (single tickets).
- Spatial. Coastal and ATMCdT zoning, but it could be exended with more spatial layers.

3.2 Syntax and Semantics

After defining all available pieces in the language, we establish the *MobilityFNC* grammar (Grammar 1) following the extended Backus-Naur (Wirth, 1996).

Grammar 1: Extended Backus-Naur grammar of *MobilityFNC*.

```
Filenane ::= Source_list,” + ”,[(Filters,” + ”)],Dimension,[(” + ”,Operations_list)],”.sql”

Source_list ::= ”log[”,Sources,”]”
Sources ::= source|(source,”,”,source)

Filters ::= Filter|(Filter,” − ”,Filter)
Filter ::= (”date[”,TempCardTypes,”]”)|(”cards[”,TempCardTypes,”]”)|(”municipality[”,DemSpatial,”]”)
TempCardTypes ::= ranges|(ranges,”,”,ranges)
DemSpatial ::= DemSpatialType|(DemSpatialType,”,”,DemSpatialType)
DemSpatialType ::= (spatial)|(boolean)

Dimension ::= Rows,” ”,Colums
Rows ::= (Row,” − ”,Row)|Row
Row ::= (ElementType,”[_”,atributeFeature,”]”)|ElementType|Filter
Columns ::= (Column,” − ”,Column)|Column
Column ::= (ElementType,”[_”,atributeFeature,”]”)|ElementType|Filter
ElementType ::= attribute|aggregation

Operations_list ::= attribute,”{”,Operations,”}”
Operations ::= aggregation|(aggregation,”,”,aggregation)
```

An encoded filename using *MobilityFNC* has the structure shown in two examples defined in Figure 4. The query description following this convention is divided into four different parts separated by sum signs. The first one defines the source(s) of interest to query. In the current project, only two different sources are available (ATM cards and TP single-ride tickets). The second block comprises the main filters to apply. It supports temporal filters, select types of smart cards and some spatial filters (depending on the spatial layers available). In the third block, the dimensions of the resulting table will be indicated, dividing into rows and columns. It will contain attributes of the table or aggregates. Considering the limited number of characters of a filename, this third block will support to apply filters on the attributes to avoid any redundany in the filters block. Finally, the last block

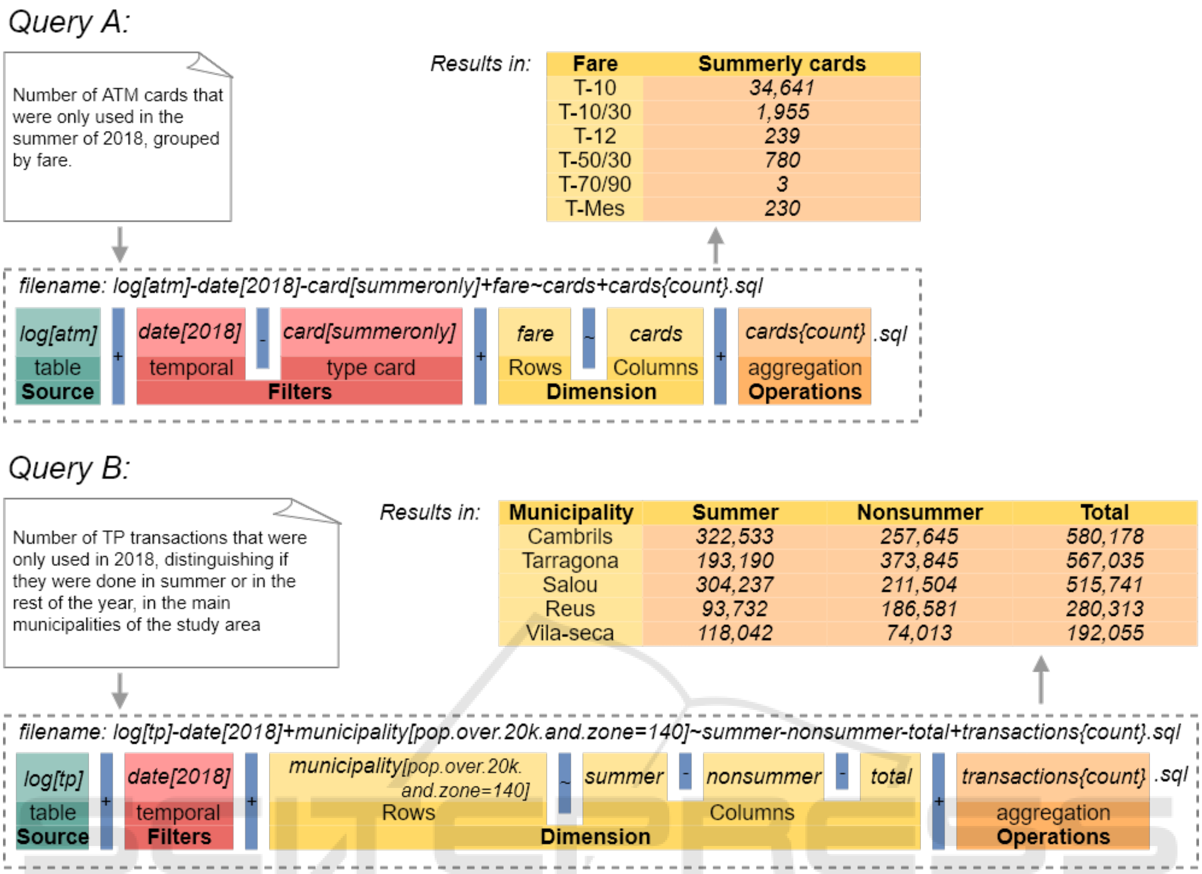

Figure 4: Examples of two filenames of queries encoded with *MobilityFNC*. .

will keep the aggregation operations to be applied to the body of the resulting table.

3.3 *MobilityFNC* Examples

In our project we were able to define 56 different query names using *MobilityFNC*, which helped us to better define our proposal. In this subsection, we illustrate two different examples of queries definitions, their structures and the intended results (Figure 4):

- Query A. The first example of a description query (Figure 4, Query A) shows a more straightforward query example where lists the number of ATM-CdT cards that were only used in the summer of 2018, grouped by fare. This query is useful to study those mobilities that only take place during the summer season. Inn some cases it could be a proxy to identify tourist mobilities.
- Query B. Another more complex example is shown in second place. It extracts the number of TP transactions (single-ride tickets) that were used in 2018, distinguishing if they were done in summer or the rest of the year, in the main municipalities of the study area. The use of Boolean filters in the dimension block is to avoid redundancy as municipality should be duplicated in the filters block.

4 CONCLUSIONS AND FUTURE WORK

We have proposed a solution to a communication issue that we detected in our multidisciplinary research group. The solution consists in a file naming convention that attempts to optimise the communication workflow between researchers in a project analysing public transportation smart card data. Our proposal, *MobilityFNC*, is currently in an early stage, it already supports mobility concepts and can apply temporal and geospatial filters. *MobitityFNC* is designed for mobility experts to be useful in the process of generating SQL queries by database administrators or developers. The SQL query is stored as content in the same file. This feature allows database managers to create SQL queries unambiguously, as well as to assist in query cataloguing and reuse.

As future work, we are aware of the limitations of the defined convention in its current state, so that our first intention is to extend the DSL definition to generalise it within the domain of mobility, and to be applied in more use cases. Finally, the last next step is the functionality to automatically generate SQL queries from the *MobilityFNC* format.

ACKNOWLEDGEMENTS

Research funded by the Spanish Ministerio de Ciencia e Innovación [grant number CSO2017-82156-R], the AEI/FEDER,UE, the Departament d'Innovació, Universitats i Empresa, Generalitat de Catalunya [grant number 2017SGR22] and the Escola d'Administració Publica de Catalunya, Generalitat de Catalunya [grant number 2018 EAPC 00002]. Sergio Trilles has been funded by the postdoctoral programme PINV2018-Universitat Jaume I (POSDOC-B/2018/12) and research stays programme PINV2019-Universitat Jaume I (E-2019-31).

REFERENCES

Alsger, A. A., Mesbah, M., Ferreira, L., and Safi, H. (2015). Use of Smart Card Fare Data to Estimate Public Transport Origin – Destination Matrix. *Transportation Research Record: Journal of the Transportation Research Board*, 2535(1):88–96.

Bagchi, M. and White, P. R. (2005). The potential of public transport smart card data. *Transport Policy*, 12:464–474.

Barth, R. S. and Galante, R. (2016). Passenger density and flow analysis and city zones and bus stops classification for public bus service management. In *SBBD*, pages 217–222.

Chandesris, M. and Nazem, M. (2018). Workshop synthesis: Smart card data, new methods and applications for public transport. *Transportation Research Procedia*, 32:16–23.

Domènech, A. and Gutiérrez, A. (2017). A GIS-Based Evaluation of the Effectiveness and Spatial Coverage of Public Transport Networks in Tourist Destinations. *ISPRS International Journal of Geo-Information*, 6(3):83.

Domènech, A., Miravet, D., and Gutiérrez, A. (2020). Mining bus travel card data for analysing mobilities in tourist regions. *Journal of Maps*, 16(1):40–49.

Fabbiani, E., Vidal, P., Massobrio, R., and Nesmachnow, S. (2016). Distributed big data analysis for mobility estimation in intelligent transportation systems. In *Latin American High Performance Computing Conference*, pages 146–160. Springer.

Gokasar, I. and Simsek, K. (2014). Using "big data" for analysis and improvement of public transportation systems in istanbul. In *Ase Bigdata/Socialcom/cybersecurity Conference, Stanford University, May 27-31, 2014*. Academy of Science and Engineering (ASE), USA,© ASE 2014.

Gokasar, I., Simsek, K., and Ozbay, K. (2015). Using big data of automated fare collection system for analysis and improvement of brt-bus rapid transit line in istanbul. In *94th Annual Meeting of the Transportation Research Board, Washington, DC*.

Gutiérrez, A. and Miravet, D. (2016a). 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*, 89(65):65–89.

Gutiérrez, A. and Miravet, D. (2016b). The determinants of tourist use of public transport at the destination. *Sustainability (Switzerland)*, 8(9):1–16.

Hudak, P. (1997). Domain-specific languages. *Handbook of programming languages*, 3(39-60):21.

Kurauchi, F. and Schmöcker, J.-D., editors (2017). *Public Transport Planning with Smart Card Data*. CRC Press, 1 edition.

Li, T., Sun, D., Jing, P., and Yang, K. (2018). Smart card data mining of public transport destination: A literature review. *Information*, 9(1):18.

Lu, Y., Mateo-Babiano, I., and Sorupia, E. (2019). Who uses smart card? Understanding public transport payment preference in developing contexts, a case study of Manila's LRT-1. *IATSS Research*, 43(1):60–68.

Ma, X., Wu, Y.-j., Wang, Y., Chen, F., and Liu, J. (2013). Mining smart card data for transit riders ' travel patterns. *Transportation Research Part C*, 36:1–12.

Morency, C., Trépanier, M., and Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203.

Pelletier, M. P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Prakasa, B., Putra, D. W., Kusumawardani, S. S., Widhiyanto, B. T. Y., Habibie, F., et al. (2017). Big data analytic for estimation of origin-destination matrix in bus rapid transit system. In *2017 3rd International Conference on Science and Technology-Computer (ICST)*, pages 165–170. IEEE.

Tao, S., Corcoran, J., Mateo-Babiano, I., and Rohde, D. (2014). Exploring brt passenger travel behaviour using big data. *Applied geography*, 53:90–104.

Van Deursen, A., Klint, P., and Visser, J. (2000). Domain-specific languages: An annotated bibliography. *ACM Sigplan Notices*, 35(6):26–36.

Wirth, N. (1996). Extended backus-naur form (ebnf). *Iso/Iec*, 14977(2996):2–1.

Wu, H., Tan, J.-A., Ng, W. S., Xue, M., and Chen, W. (2015). 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*, pages 1093–1098. ACM.