

# Analysis and clustering of massive public transport data

Pablo Gutiérrez-Asorey<sup>1</sup>  and Manuel Oviedo-de la Fuente<sup>1</sup> 

Universidade da Coruña, CITIC  
pablo.gutierrez, manuel.oviedo@udc.es

**Resumen** This work proposes a methodology for analyzing citizens' movement patterns in big cities by exploiting validation records from traveler cards in public transport networks. One important question that naturally arises about traveler movements is how many people travel across a line in a given time frame, that is, line load. However, answering such a simple question by exploiting the traveler card records in a traditional database proves costly both space and time-wise. In this work, we present a novel representation of boarding data that leverages Functional Data Analysis to efficiently answer queries about average traveler load for a given stop, day type, and time frame within the day.

**Keywords:** Databases · Big Data · transport networks · clustering.

## 1. Introduction

As a direct consequence of the widespread use of traveler cards to validate access to public transport networks, it is now possible to collect large volumes of data on travelers' movements. Transport administrators can exploit this data to improve the transport network, such as addressing connectivity problems between stops or overloading issues on some lines, stops, or hours. However, the vast amount of data that must be stored and analyzed makes direct analysis of the raw data overly complex and computationally intensive.

In this work, we focus on the problem of analyzing fluctuations in boarding public vehicles at any stop within a transport network over a given time period on a specific type of day (workdays, holidays, weekends, etc.). From this information, it is possible to derive the load of any specific line on a given day and time frame. Note that querying the traveler cards records in a traditional database to answer questions about the use of a single stop would require fairly complex queries; therefore, we propose modeling the temporal evolution of boardings at each stop using Functional Data Analysis (FDA) by boarding profiles as functional objects. This allows us to identify and represent typical behavioral patterns of stops, providing a compact description of the network by storing boarding profiles rather than the individual transaction data. This means, in essence, that instead of storing that for a stop  $A$  that has an average of 200 boardings at 8 : 30 AM, we store the fact that stop  $A$  exhibits a behavior of type  $\alpha$ , and said behavior has an average of 200 boardings at 8 : 30 AM.

In this paper, we describe our first approach to designing such a solution. In the following section 2, we describe our representation conceptually, whereas in Section 3 we describe our first approach to build such a representation. Section 4 then briefly discusses the results of our preliminary experiments.

## 2. Proposal and hypothesis

We chose to analyze the use of public transport, distinguishing between *work days*, *Fridays*, and *weekends*, as citizens' movement patterns vary notably across these categories. For each stop and day, we compute the number of boardings in half-hour intervals over 24 hours (i.e., 48 intervals) by preprocessing the traveler card records. We compute the average of these values across all the days of the same type. This results in a single distribution curve per stop and type of day, representing, for every half-hour interval, the average number of passengers boarding that stop on that type of day.

In other words, the distribution curve of a given stop and type of day is represented in the  $x$  axis by 48 periods of half an hour in a day, with the  $y$  axis representing the average number of people that boarded that stop in each half-hour period, for that type of day. To compare distribution curves across different stops, we convert the absolute counts (number of boardings) to percentages relative to the average total number of boardings over 24 hours for each stop and day type. This normalization removes differences in overall demand, allowing stops to be compared by the shape of their boarding profiles rather than their absolute passenger volume. This also makes it trivial to estimate the number of travelers who boarded at that stop on any given day and for any period.

Next, we consider that stops with similar circumstances ought to showcase similar patterns on their distribution curves. For example, it is reasonable to make the assumption that stops located in working-class neighborhoods would exhibit a peak of boardings early in the morning, when people take public transport to go to work. We want to group together stops with similar distribution curves, and for this purpose, we use FDA-based cluster analysis.

Functional data is defined as data consisting of observed functions or curves evaluated at some interval of time. An overview of statistical methods for FDA can be found at [6] and [1]. One notable field that has been considered from a functional perspective, albeit still a rather undeveloped research area, is that of cluster analysis and classification methods, ranging from approaches based on hierarchical clustering [4] to K-means clustering [2].

Using these techniques, we investigate two hypotheses. **The first hypothesis** is that there exist patterns in the fluctuations of boardings that are common to multiple stops and serve to characterize them. **The second hypothesis** is that it is possible to save storage space if, instead of storing the distribution curves of boardings for each stop and type of day, we store the centroid of its representative cluster and the average total number of boardings for each type of day. Such a representation should be able to answer queries about the expected stop load at any given stop and time frame using elementary operations.

### 3. Creating our representation

To test the previous hypothesis, we experimented with data on traveler card transactions for the month of February, 2019, for the public transport network of the city of Madrid (Spain). This data was graciously provided to us by the Madrid Regional Transport Consortium <sup>1</sup> as part of a joint R&D project with the Databases Laboratory of the University of A Coruña. We had a total of 106,820,280 records describing boardings, with 56,824,951 on *subway*, 13,274,164 on *train*, 949,924 on *trolley car* and 35,771,241 on *urban bus*.

First, we grouped all boarding records by *means of transport*, then by *dpoint*, and sorted the result by *datetime*. In February of 2019 there were boardings at 237 *subway* stops, 92 *train* stops, 52 *trolley car* stops and 10,043 *urban bus* stops. We calculated the average distribution curves for all of these stops on half-hour intervals, considering all *weekdays*, *Fridays*, and *weekends*, resulting in 3 curves per stop (a total of 31,272 curves, one per type of day). These curves were the input data for our cluster analysis.

Clustering smoothed curves rather than observed data typically produces better results, as demonstrated in [3]. However, we must consider that our data consist on what, in terms of statistics theory, are called *densities* (non-negative functional data curves) [5], hence we can not apply the more commonly used FDA methods for smoothing our data based on Hilbert spaces, as these can introduce negative values on the curves. Instead, we use the  $L_1$  metric, which guarantees that all resulting values are  $\geq 0$ . Next, we use curve depth (a statistical measure that quantifies how central or deep a data point is within a distribution or curve) to exclude the bottom 5% of curves before clustering, thereby enhancing the robustness and accuracy of our analysis. The excluded curves should still be analyzed separately.

For clustering the smoothed curves, we use the K-means method. We used the Silhouette method [7] to determine the  $k$  optimal number of clusters for the distributions of each means of transport and type of day. This technique explores different values of  $k$ , measuring how well each curve fits its assigned cluster relative to the others to find the best possible configuration of clusters.

### 4. Discussion

For this first analysis, we decided to separate the experiments by means of transport. This way, we also hoped to assess whether the means of transport carry considerable weight in determining whether a stop occurs. The results of our experiments suggest that the answer is negative. Nevertheless, the primary goal of these experiments was to validate whether our hypothesis (that there exist patterns in the fluctuations of boarding that are common across stops and serve to characterize them) holds.

The results we obtained support our first hypothesis, as clear behavioral patterns emerged in several transport modes. In particular, two very distinct

<sup>1</sup> <https://www.crtm.es/>

types of behaviors appeared very clearly on subway, train and trolley car, on both *work days* and *Fridays*: that of stops with most of their boardings early in the morning, and that of stops with more equally distributed boardings across the day, with smaller peaks on the morning, lunch time and late afternoon. *Weekends*, which showcase distinct behavioral patterns separate from weekdays, also fall within our expectations. Only in the bus network did our method fail to find patterns within a reasonable variance. The root cause is that bus networks are much more complex than other modes of transport. We concluded that a more in-depth analysis was necessary for characterizing the bus stops.

Our second hypothesis is supported by the fact that in our experiments, we obtained a total of 20 clusters versus 1,143 curves on three means of transport (subway, suburban train, and trolley car). This indicates that the proposed representation can substantially reduce the amount of information to be stored while preserving the primary behavioral characteristics of stops.

**Agradecimientos** This study was partially funded by: GRC: ED431C 2025/34, funded by GAIN/Xunta de Galicia, and by PID2022-141027NB-C21 (EarthDL): partially funded by MCIN/AEI/10.13039/501100011033 and EU/ERDF A way of making Europe. It was also partially funded by CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

## Referencias

1. Ferraty, F., Vieu, P.: Nonparametric functional data analysis: Theory and practice **51** (01 2006). <https://doi.org/10.1007/0-387-36620-2>
2. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. *JSTOR: Applied Statistics* **28**(1), 100–108 (1979)
3. Hitchcock, D., Booth, J., Casella, G.: The effect of pre-smoothing functional data on cluster analysis. *Journal of Statistical Computation and Simulation* **77**, 1043–1055 (12 2007). <https://doi.org/10.1080/10629360600880684>
4. Johnson, S.C.: Hierarchical clustering schemes. *Psychometrika* **32**(3), 241–254 (1967)
5. Petersen, A., Müller, H.G.: Functional data analysis for density functions by transformation to a hilbert space. *The Annals of Statistics* **44**(1) (feb 2016). <https://doi.org/10.1214/15-aos1363>, <https://doi.org/10.1214/2F15-aos1363>
6. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*. Springer (2005), <http://www.worldcat.org/isbn/9780387400808>
7. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987). [https://doi.org/https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7), <https://www.sciencedirect.com/science/article/pii/0377042787901257>