# URBAN PASSENGER TRAFFIC ESTIMATION FROM AUTOMATED DATA COLLECTION SYSTEMS

A. *REMY - SNCF Innovation & Research Department, France-anais.remy@sncf.fr*

## ABSTRACT

Like many larges urban centers, Paris and its suburbs are faced to a strong population densification in recent years. The current situation is such that it's become a real challenge for public transport to assure a good quality service. In this context, SNCF Transilien, one of the (railway) public transport operators in region Île-de-France, asked Innovation and Research Department for a new tool in order to improve its analysis of passenger traffic.

For several years now, many technologies (smartcard, automatic passenger counter, etc.) provide huge volume of data particularly interesting for passenger traffic estimation.
In this work, we propose a method based on statistical extrapolation from automated data collection system. Indeed, data are often sparse due to non exhaustive equipment (e.g., all stations are not equipped with checkin/checkout). Thus, the historical data collected can be considered as a sample from which it is possible to infer the traffic. We also take benefice of merging different data sources to provide consolidated results.

The method has been implemented on large set of historical data (over two years) and results provided at different levels of aggregation (by station, by tariff, by day). This work shows that it is possible to mining simple and useful indicators from these complex and huge data.

Business model in public transport tend to be more and more based on actual passengers traffic. Automated data collection systems offer new insights to address this growing challenge in the transport field. Our research focuses on providing marketing value from these data by crossing intensive data management and classical statistical methods.

*Keywords: passenger traffic estimation, smartcard data, extrapolation, sampling theory*

# FROM DATA COLLECTED TO MARKETING VALUE

In 2010, SNCF Transilien daily transports about 2.8 millions passengers thanks to 14 commercial lines in Paris area. On this, various systems (ticket sales, smartcard checking, etc.) provide automatically many disaggregated data, i.e. each transaction is collected. The potential of these data is noteworthy from operator's viewpoint, especially because it deals with real (and non theoretical or simulated) mobility observed.

But in practice, what seems to be a great opportunity may quickly become a real headache… Indeed, these data are quite complex due to their heterogeneity (extract from several systems), huge volume and sparsity (sensors equipments are rarely exhaustive). So on, it is a challenge in transportation field to extract marketing value from this type of data.

This research contributes to develop a new tool to analyze passenger traffic from automated data systems. The objective is to provide coherent indicators at different levels: from totally disaggregated to more macro levels (by station, by tariff, by day, etc.).
In this paper, we present our work in progress:
- First, we remind briefly the concepts underlying passengers traffic and specify the terminology used;
- Traffic data are described and we introduce more precisely smartcard data;
- We present the extrapolation methodology in order to obtain aggregate traffic estimation. This approach is quite scalable with various sparse data and this is an innovative feature;
- Then, we expose a case study from large historical smartcard data.

# TERMINOLOGY

Let's return on the mobility phenomenon to analyze and define some concepts and relationships between them (brief zoom on object model [1]). At this step, we present "what we call mobility" later on. This is to distinguish from the issue "how can we measure the mobility?" addressed in the next section on data available.

We identify three keys objects on passenger traffic:
- Active cardholder: for example, a yearly pass equals to "one active cardholder" during twelve months. And a ticket for one single Origin-Destination (OD) equals also "one active cardholder" for this specific tariff.
- Trip: Origin-Destination carried by the same vehicle (in our case, train).
- Journey: Successive trips made by the same cardholder.

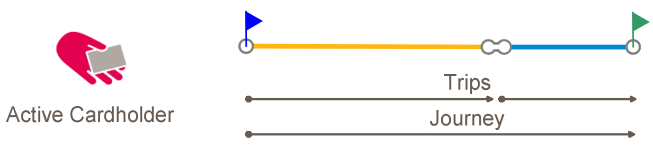


Figure 1 – Mains objects on passenger traffic

Several indicators do help to synthesize information: volume of actives cardholders, journeys, trips (journeys x connections), trips kilometers (journeys x kilometers travelled), trips minutes (journeys x travel times), etc.
These traffic indicators can be calculated on different levels of aggregation (according to spatial, temporal and tariff dimensions) adapted to marketing and financial reporting.

# TRAFFIC DATA

Passenger traffic measurement is a tricky issue. Can you imagine asking to all passengers (several millions per day) how many trips they have been carried (about tens of millions trips per month)? It was until now derived from some surveys or manual counts of passengers. But, this process is very expensive and provides relatively limited information (only the days of the surveys).

Automated data collection systems seem an attractive alternative because they produce continuous data. Moreover, this data are often precisely located in space and time (geolocated and timestamped data). The actual spatial spread of sensor systems (smartcard, automatic passenger counter, etc.) tends to reduce the cost of data collection to the benefit of more details data analysis.

Whatever the data source (manual or automatic), we collect two main types of information:
- Volumes of passengers -optionally oriented- on a specific network node: for instance, the number of passengers entered on a station or in a train.
- Volume of passengers -oriented- between two network nodes: for instance, the number of passengers carried on an Origin-Destination.
This second type of data is quite precise and more difficult to collect (famous OD matrix).

Smartcard data are a relevant illustration and have advantage of providing both types of traffic data (without cheating) discussed above:
- Volume of passengers on station = Check in/check out (see below figure A)
- Volume of passengers on OD = Linking of successive check in/check out by the same cardholder (see below figure B)

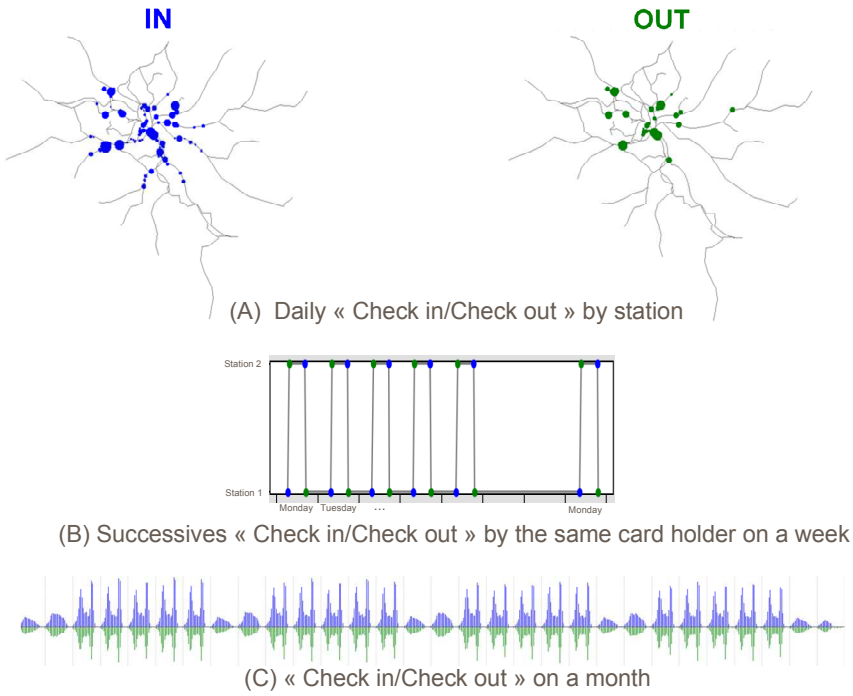


Figure 2 – Smartcard data illustration

These data are available only on stations and ODs with checking system.

# EXTRAPOLATION METHODOLOGY

A drawback of automated data collection systems is often to produce “incomplete” information. Indeed, data are recorded only where/when there are active sensors. For example in SNCF Transilien case, all stations are not equipped with checkin/checkout. Then, it is not trivial to aggregate these data at macro levels (e.g. on global network, by month, etc.). In this section, we discuss the extrapolation methodology developed in order to obtain global indicator at different levels of aggregation.

The methodology is based on sampling theory: statistical inference about the population using sample [2]. It is assumed (under coverage conditions) that data collected can be considered as a sample from which it is possible to infer the traffic. In this work, we have adapted this classical statistical approach to intensive data management [3].

The procedure is divided into two main steps:
1. Sampling design: We define the more realistic sampling design that allows saying “everything happens as if data were collected from a survey”. In our case, this sampling design is quite complex (stratified at several degrees) and takes into account spatio temporal structure of traffic data.
2. Calibration adjustment: The objective is to incorporate auxiliary information into the procedure. The principle is to calibrate extrapolation weights (thanks to optimization techniques) to satisfy some knows totals (e.g. volume of passengers that entered on a station, number of kilometers/connections on global network, etc.). Further more, this procedure allows improving the result by combining other data available.

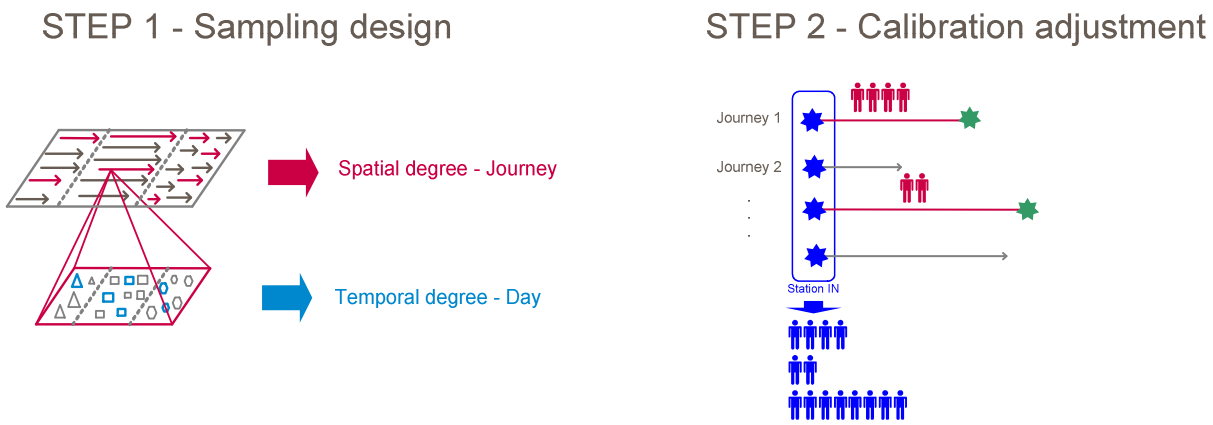


Figure 3 – Extrapolation methodology

Then, the formulas estimators are defined and the confidence intervals are expressed exactly or approximated according to the design complexity.

# CASE STUDY

This new methodology has been implemented for the SNCF Transilien railway operator over two years of historical data (from check in/check out systems, automatic passenger counter, etc.). This starts to become a pretty big dataset: for instance, more than 1 billion of smartcard transactions are recorded during this period. This work shows that it is possible to mining simple and useful indicators from these complex and huge data.

In this section, we focus on traffic estimation (without cheating) at several levels of aggregations (e.g. on global network, by month, etc.) and present some results and first examples of basic visualizations.

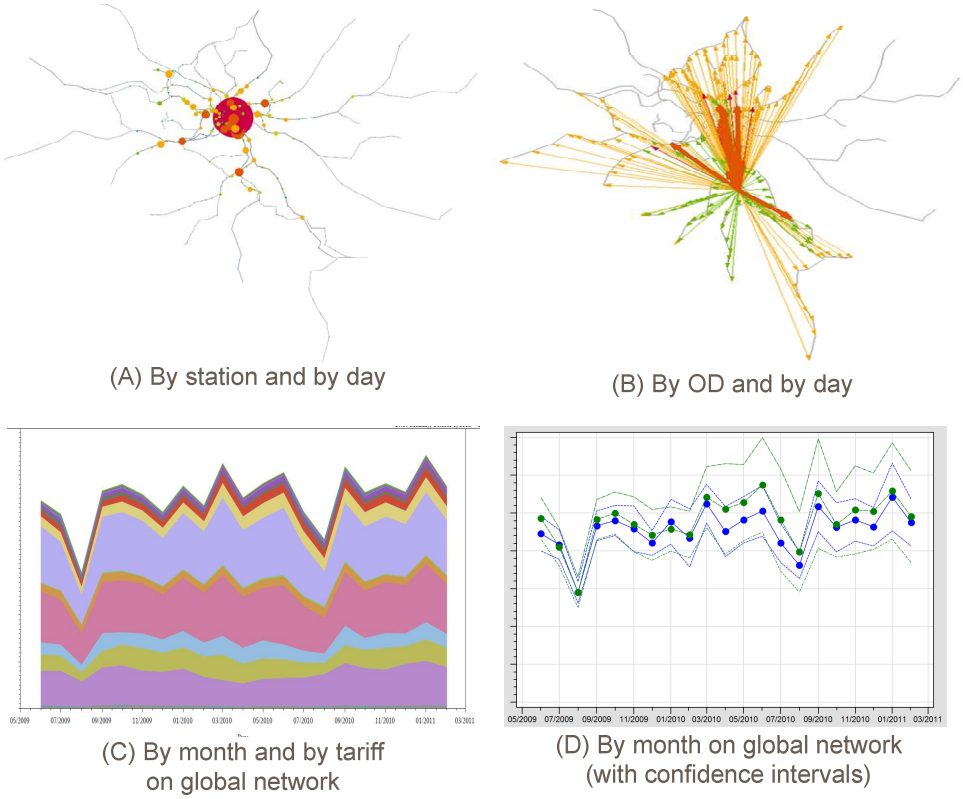


Figure 4 – Passengers traffic estimations at several aggregation levels

Traffic indicators are computed coherently from totally disaggregated to macro levels. It allows to gradually zoom in/out the data in order to provide details analysis and synthetic views.

The figures above explain three mains dimensions to explore:
- Temporal patterns to recognize from monthly or daily evolution (see above figures C and D)
- Spatial features on traffic estimations by station, OD (see above figures A and B).
- Tariff structure (see above figure C)

These new insights underline the need to have a tool for easily exploring traffic indicators at all the aggregate levels (including both observed and estimated data). Our methodological and technical support is still in progress to develop an innovative interface (thank to data visualizations [4]) adapted to different marketing analysis.

# CONCLUSION AND PERSECTIVES

This work succeeds to prove that it is possible to provide continuous and reliable (with confidence interval) information about mobility from various automated data collection systems. Taking benefits of these results, SNCF Transilien wants to use this knowledge of urban passenger traffic to improve its quality service. For example, indicators of punctuality weighted by passengers traffic (the number of passengers delays) could be automatically up dated.

The extrapolation methodology is quite scalable with various sensors data. For example, the methodology has been adapted to provide passenger traffic estimation (including cheating) from automatic passenger counter. In the same way, this approach could be extended to many other available traffic data (video, mobile, …).

This research aims to define a new methodology and implement it on historical data. A key issue is now to help marketing team to produce automatically and daily indicators (from proof of concept to operational tool).

# BIBLIOGRAPHY

[1] Pelletier, Trépanier, Morency, (2009). Smart card data use in public transit: A literature review, Transportation Research Part C
[2] Ardilly (2006). Les techniques de Sondages, Ed Technip
[3] Rémy, Chandesris (2012). Estimation des trafics voyageurs SNCF Transilien : un plan de sondage complexe et prise en compte d'échantillons multiples, 7ème colloque francophone sur les sondages
[4] Tufte (2001). The Visual display of quantitative information, Cheshire, CT, Graphics Press, 2e ed