Probability Aggregation Functions for Single-route Trip Estimation in Public Transportation

Rodrigo René Cura*[1], Romina Stickar*[2], Fernando Tohmé†[3] and Claudio Delrieux‡[4]

* *LINVI - Dpto. de Informática*
Univ. Nac. de la Patagonia San Juan Bosco
Brown 3051, Puerto Madryn, Argentina
[1]rodrigo.renecura@gmail.com
[2]romistickar@gmail.com
† *Departamento de Economía, Univ. Nac. del Sur, and CONICET*
‡ *Departamento de Ing. Eléctrica y Computadoras, Univ. Nac. del Sur, and CONICET*
Alem 1253, Bahía Blanca, Argentina
[3]ftohme@gmail.com
[4]cad@uns.edu.ar

***Abstract*—Different methods exist for estimating trips in public transportation systems. Some of the widely adopted estimation strategies are based on the traceability of passenger transfers. While these techniques have been effective in the case of some large cities, they fall short for trip estimation in smaller towns. This is because the underlying methodology leave aside single-route trips information, which are the most frequent in the latter case. In this paper we present a methodology based on combining probability distributions defined upon alternative partitions of the universe of possible trips. Our approach starts from intuitions similar to the Dempster-Shafer rule, but since in our application domain some of its basic assumptions fail to be satisfied, we estimate probability intervals by means of an aggregation procedure on the information of the different sources. Samples of those intervals yield the distributions to be used in simulation models of the behavior of the transportation system.**

***Resumen*—Existen diferentes métodos para estimar viajes en sistemas de transporte público. Algunas de las estrategias ampliamente adoptadas se basan en la trazabilidad de los transbordos de pasajeros. Si bien este grupo de técnicas ha resultado efectivo en el caso de ciudades de gran tamaño, no lo ha sido igualmente para poblaciones menores. La razón es que se deja de lado información sobre viajes de ruta simple, usuales en ciudades más pequeñas. En este trabajo presentamos una metodología basada en combinar distribuciones de probabilidad definidas sobre particiones alternativas del universo de viajes posibles. Aunque nuestro enfoque parte de intuiciones similares, se diferencia de la regla de Dempster-Shafer debido a que algunos supuestos de esta no pueden ser satisfechos. En vez de ello obtenemos intervalos de probabilidad agregando la información de las distintas fuentes. Muestras de estos intervalos generan las distribuciones a usarse en modelos de simulación del comportamiento del sistema de transporte.**

I. Introduction

Human mobility is perhaps the most important factor in urban planning, and therefore an adequate assessment of public transportation systems is essential for a comprehensive city organization. However, this assessment is sometimes difficult or confusing, in particular to estimate the present and future mobility requirements. For this reason, urban growth is usually not accompanied by a sensible reconfiguration of the public transportation facilities, which in turn leads to a significant decrease in the actual and perceived quality of the transportation services and the overall quality of life [1].

As part of the basic urban infrastructure, large cities have public transportation networks, in which several kinds of transportation modes may intervene, for instance subway, buses, or trains. The widespread use of smart card automated fare collection systems makes available relevant information that can be mined to estimate quite accurately urban mobility patterns. For example, if a user consistently uses the same card along a period of time, useful transportation parameters like daily uses, waiting times, and transfer times and places can be easily gathered. This information, collected in massive amounts, and together with reasonable assumptions, provides sufficiently detailed data about model parameters that can be used to assess the efficiency of the transportation network as a whole, as well as to spot specific bottlenecks, and to suggest policies to handle contingencies [2]–[4].

A key element in this analysis is the information about round-trips, since the transportation infrastructure in working days is clearly used mostly by workers, students, and in general by people subject to fixed schedules. Therefore, one-way only transportation information is of little or no use at all. Unfortunately, this is the only information available in small towns where users do not transfer between different lines, and round-trip assumptions are harder to sustain.

In those cases gathering precise and accurate transportation data (for instance, by using cameras) may be difficult and expensive, becoming a major limiting factor to the assessment of the effectiveness of a transportation network. For this reason, it is not uncommon to use indirect measures provided by data sources already available, for instance mobile phones [5] or georeferenced social network feeds [6], even though these sources were not initially intended to gather transportation information . Thus, adequate data fusion methodologies are required to combine these heterogeneous information sources into a unified model that may serve as a trustable aid in urban planification.

In this paper we develop a methodology for data fusion, which is based on a particular combination of probability distributions defined upon alternative partitions of the

universe of possible trips. The methodology is applied to a specific context in the Puerto Madryn city, Argentina. Our results show that, lacking the accurate information that can be obtained in large cities, our methodology may provide an adequate assessment of mobility patterns and the effectiveness of a public transportation infrastructure in small towns.

II. Preliminary definitions

Let $R = \{s_i\}_{i=1}^{|R|}$ be the class of possible stops. The set of possible trips is $T = \{(s_i, s_j) : i < j\}$, where a pair (s_i, s_j) indicates that s_i and s_j are, respectively, the boarding and descending stops[1]. Then, the set of possible trips can be partitioned in the two following ways:

Definition 1. *For* $b, a = 1, \ldots, |R|$

$$T_B = \{T_b\} \text{ where } T_b = \{(s_b, \cdot) : (s_b, \cdot) \in T\}$$

$$T_A = \{T_a\} \text{ where } T_a = \{(\cdot, s_a) : (\cdot, s_a) \in T\}$$

In words, $T_b \in T_B$ is the set of all possible trips that start at stop s_b. In the same way, $T_a \in T_A$ is the set of all possible trips that end at stop s_a. It is easy to see that, T_B and T_A constitute different partitions of T.

The available information sources are indeed associated to these partitions of T. We define $F = \{f_B, f_A\}$, with $f_B : T_B \to \mathbb{R}^+$ and $f_A : T_A \to \mathbb{R}^+$, as the set of information functions associated to the partitions[2]. In our particular case, $f_B(T_b)$ provides the number people boarding at stop s_b. Similarly $f_A(T_a)$ provides the number of people arriving at stop s_a.

Under the conditions already stated, we can regard the information structure as a directed weighted graph of the form $G = \langle R, T, w \rangle$ where w is an unknown function such that $w : T \to \mathbb{R}^+$ and $w(t)$ is the proportion or weight of the trips t over the total trips in T. The goal of this research is to present a methodology to find an estimation $w' \simeq w$ given the information gathered by the functions in F.

Since we focus on the graph structure, where w can be modeled as a probability function, we add the constraint $w : T \to [0, 1]$, with $\sum_{t \in T} w(t) = 1$. In addition, we use the notation $[w]$ to refer the adjacency matrix of G formed by the values of w.

III. Evidence aggregation

A possible approach to the approximation of w is through the application of *Dempster's Rule of Combination*, which allows to define mass functions up from information functions. In turn, the mass functions make it possible to construct belief functions [7].

[1]The condition that each pair of stops constitutes a potential trip holds in railroad lines or similar two-way modes of transportation with identical fixed stops. In small towns, typical bus routes have very similar round-trip routes. This condition is a reasonable approximation to the real world case we intend to analyze.

[2]The existence of more than one information source associated to the partition functions is feasible (and even desirable) but again for simplicity we will consider here only one information source.

Definition 2. *Given* $T_k \in 2^T$. *Let* $m_B : 2^T \to [0, 1]$ *and* $m_A : 2^T \to [0, 1]$ *such that*

$$m_B(T_k) = \begin{cases} \frac{f_B(T_k)}{\sum_{T_b \in T_B} f_B(T_b)} & \text{if } T_k \in T_B \\ 0 & \text{otherwise} \end{cases}$$

$$m_A(T_k) = \begin{cases} \frac{f_A(T_k)}{\sum_{T_a \in T_A} f_A(T_a)} & \text{if } T_k \in T_A \\ 0 & \text{otherwise} \end{cases}$$

Theorem 1. *Given* m_B *and* m_A *as defined above, then* m_B *and* m_A *are mass functions over* T.

Proof: We need to prove:

- M1. $m_B(\emptyset) = 0$ and $m_A(\emptyset) = 0$
- M2. $\sum_{T_k \in 2^T} m_B(T_k) = 1$ and $\sum_{T_k \in 2^T} m_A(T_k) = 1$

It is easy to see that M1 holds by definition. To see that M2 holds, notice that

$$\begin{aligned} \sum_{T_k \in 2^T} m_B(T_k) &= \sum_{T_k \in T_B} m_B(T_k) \\ &= \sum_{T_k \in T_B} \frac{f_B(T_k)}{\sum_{T_b \in T_B} f_B(T_b)} \\ &= \frac{1}{\sum_{T_b \in T_B} f_B(T_b)} \sum_{T_k \in T_B} f_B(T_k) \\ &= 1. \end{aligned}$$

With the same reasoning line, we may show $\sum_{T_k \in 2^T} m_A(T_k) = 1$. Then m_B and m_A are mass functions over T. ∎

Definition 3. *Given* $T_k \in 2^T$. *Let* $Bel_B : 2^T \to [0, 1]$ *and* $Bel_A : 2^T \to [0, 1]$ *such that*

$$Bel_B(T_k) = \sum_{U \subseteq T_k} m_B(U)$$

$$Bel_A(T_k) = \sum_{U \subseteq T_k} m_A(U)$$

Definition 4 (Dempster's Rule of Combination [8])**.** *Given* Bel_B *and* Bel_A *and their respective mass functions* m_B *and* m_A. *We define* $(m_B \oplus m_A)(\emptyset) = 0$ *and, for every* $T_k \in 2^T$ *such that* $T_k \neq \emptyset$

$$(m_B \oplus m_A)(T_k) = \sum_{\{U, V | U \cap V = T_k\}} \frac{m_B(U) m_A(V)}{c}$$

where $c = \sum_{\{U, V | U \cap V \neq \emptyset\}} m_B(U) m_A(V)$

Finally, we have to prove that $c > 0$ for all $T_k \in 2^T$, to make sure that Bel_B and Bel_A are combinable. Afterward, as $m_B(T_k) > 0$ iff $T_k \in T_B$, $m_A(T_k) > 0$ iff $T_k \in T_A$ and $T_B \cap T_A = \emptyset$, there does not exist T_k such that $m_B(T_k) > 0$ and $m_A(T_k) > 0$; then, $c = 0$ and Bel_B and Bel_A cannot be combined.

In conclusion, the Dempster-Shafer approach is unsuitable for our problem.

IV. Probabilities

In this Section we will put forward some definitions that settle down a relation between the weight function w and the information functions in F.

Definition 5. *Let $p_B : T \to [0,1]$ and $p_A : T \to [0,1]$ such that:*

$$p_B(t) = \frac{f_B(T_n)}{\sum\limits_{T_b \in T_B} f_B(T_b)} \frac{1}{|T_n|}, \text{ where } t \in T_n \in T_B \qquad (1)$$

$$p_A(t) = \frac{f_A(T_m)}{\sum\limits_{T_a \in T_A} f_A(T_a)} \frac{1}{|T_m|}, \text{ where } t \in T_m \in T_A \qquad (2)$$

The functions defined here are built taking the relative frequencies of the sets T_n, T_m and splitting them in equal parts over the elements of them. Applying the principle of indifference, we hold the assumption that the probabilities p_B and p_A are uniformly distributed over T_B and T_A respectively.

Theorem 2. *Given p_B and p_A as defined above, p_B and p_A are probability functions over T.*

Proof: We consider first p_B.

$$\begin{aligned}
\sum_{t \in T} p_B(i,j) &= \sum_{t \in T_k \in T_B} \frac{f_B(T_k)}{\sum\limits_{T_b \in T_B} f_B(T_b)} \frac{1}{|T_k|} \\
&= \sum_{T_k \in T_B} \sum_{t \in T_k} \frac{f_B(T_k)}{\sum\limits_{T_b \in T_B} f_B(T_b)} \frac{1}{|T_k|} \\
&= \sum_{T_k \in T_B} |T_k| \frac{f_B(T_k)}{\sum\limits_{T_b \in T_B} f_B(T_b)} \frac{1}{|T_k|} \\
&= \sum_{T_k \in T_B} \frac{f_B(T_k)}{\sum\limits_{T_b \in T_B} f_B(T_b)} \\
&= 1.
\end{aligned}$$

With the same reasoning line, we can show that $\sum\limits_{t \in T} p_A(i,j) = 1$. Then, p_B and p_A are probabilities over T. ∎

According to the literature on imprecise probabilities ([9], [10]), it is not feasible to define rational-valued probabilities based on incomplete or inaccurate knowledge. Hence, we may resort to other representations, for instance interval-valued probabilities. Following this strategy, we use the probabilities defined in Definition 5 to construct a interval $[p_B(t), p_A(t)]$ and the take $w'(t)$ such that $w'(t) \in [p_B(t), p_A(t)]$.

In consequence, $w'(t)$ is given by a cloud of possible distributions and to fully specify it an additional definition is required. Therefore, as presented in [11] we apply the linear opinions pool technique.

Definition 6. *Let $t \in T$, and let θ_B, $\theta_A \in \mathbb{R}^+$ such that $\theta_B + \theta_A = 1$. We define*

$$w'(t) = \theta_B p_B(t) + \theta_A p_A(t)$$

Additionally, in some cases we simply reference θ such that $\theta_B = \theta$ and $\theta_A = 1 - \theta$.

Example 1. *Suppose the following matrix, where each cell contains the number of trips performed from stop i (row) to stop j (column), and its respective $[w]$ as defined before.*

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} & 10 & 5 & 20 \\ & & 15 & 30 \\ & & & 10 \\ & & & \end{pmatrix} \end{array} \implies [w] = \begin{bmatrix} & \frac{10}{90} & \frac{5}{90} & \frac{20}{90} \\ & & \frac{15}{90} & \frac{30}{90} \\ & & & \frac{10}{90} \\ & & & \end{bmatrix}$$

The partitions T_B and T_A as defined before are:

$T_B = \{B_1, B_2, B_3, B_4\}$ with $B_1 = \{(1,2),(1,3),(1,4)\}$, $B_2 = \{(2,3),(2,4)\}$, $B_3 = \{(3,4)\}$, $B_4 = \emptyset$, and $T_A = \{A_1, A_2, A_3, A_4\}$, $A_1 = \emptyset$, $A_2 = \{(1,2)\}$, $A_3 = \{(1,3),(2,3)\}$, $A_4 = \{(1,4),(2,4),(3,4)\}$

Finally, applying Definition 5 with $\theta_B = \theta_A = \frac{1}{2}$, we obtain:

$$[p_B] = \begin{bmatrix} & \frac{35}{270} & \frac{35}{270} & \frac{35}{270} \\ & & \frac{45}{180} & \frac{45}{180} \\ & & & \frac{10}{90} \\ & & & \end{bmatrix}$$

$$[p_A] = \begin{bmatrix} & \frac{10}{90} & \frac{20}{180} & \frac{60}{270} \\ & & \frac{20}{180} & \frac{60}{270} \\ & & & \frac{60}{270} \\ & & & \end{bmatrix}$$

$$[w'] = \theta_B [p_B] + \theta_A [p_A] = \begin{bmatrix} & \frac{65}{540} & \frac{65}{540} & \frac{95}{540} \\ & & \frac{97.5}{540} & \frac{127.5}{540} \\ & & & \frac{90}{540} \\ & & & \end{bmatrix}$$

□

V. Results

To test the soundness of the model, we use the available data, like manual counting of boardings and alightings on the line of interest, origin-destination surveys and records of central stations. By feeding a simulation model with these datasets, we run 1000 simulations and build a hypothetical average complete trip matrix as shown in Fig. 1a. Precisely, the simulation is set on a bus-line with 68 stops (that is $|R| = 68$) and the number of possible trips are calculated by $\frac{|R|(|R|-1)}{2}$ giving a total of 2278. For each possible trip, the simulation generates values representing how many passengers made this particular one.

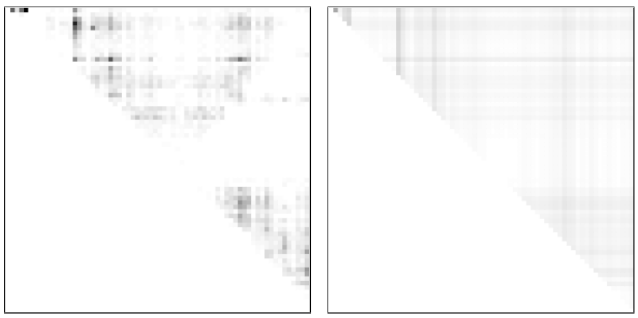
Since this information tends to be unavailable in practice, we construct the functions presented in definition 5 and compute w'. Figure 1b shows the result of the estimation. In addition, we apply a χ^2 goodness-of-fit test to check the model as follows.

A. χ^2 test

Over the results we perform a χ^2 goodness-of-fit test. Next, the computation of the χ^2 indicator is given by

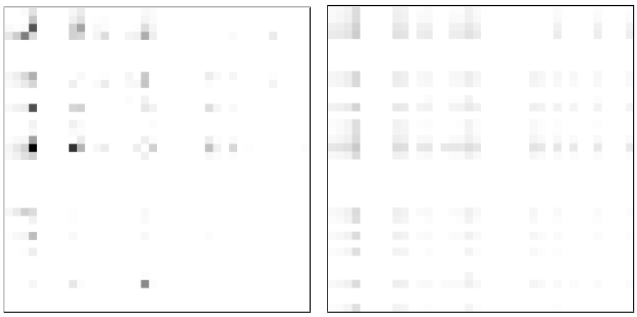
$$\chi^2 = \sum_{t \in T} \frac{(w(t) - w'(t))^2}{w'(t)}$$

Since $w'(t) > 0$ for all $t \in T$, we take it as denominator. The test yields $\chi^2 = 1.428425$, with degrees of freedom



(a) $[w]$ (b) $[w']$

Fig. 1: Trips matrix by stops. $\theta = 0.5$



(a) $[w]$ (b) $[w']$

Fig. 2: Trips matrix by neighborhoods. $\theta = 0.5$

$(2-1)(2278-1) = 2277$ and p-value $= 1$. Indeed, there is not evidence of difference between the distributions w' and w; then, we can say that the estimation passes the test.

In addition, in Fig. 2 we present the data grouped by neighborhoods. The results of a χ^2 test, are the same as the former case; still, in the figure we can discern a different matrix structure. It is possibly more appropriate to study the appropriateness of the simulation by examining the spectrum of this matrix, but such approach exceeds the scope of this paper.

Finally, it is worth noting that even if the results are positive according to the χ^2 test, the figures show particular differences between the distributions. For instance, white areas in the real scenarios (that is, without trips) correspond to areas with a slight noise in the estimation, indicating trips that the real-world sample fails to capture.

VI. Conclusion and future work

Our results suggest that the methodology proposed in this paper leads to a model that performs sufficiently good in the estimation of trips. In particular, it is statistically close to the information functions provided by real data sources.

In consequence, we highlight the following aspects of our methodology. First, unlike common practices that mainly rely on transfers between lines or modes of transport, our methodology is appropriate for use in small and medium scale cities in which the ratio of single-route trips is high. Second, it is affordable to gather the required data, which in many cases is already available, or otherwise the gathering process is easy to automate with low-cost devices. Lastly, the calculation and interpretation of the results is simple, which, together with the possibility of continuous feedback of the demand and changes in the transport system, makes our methodology actionable and easy to use for devising optimization or policy-making.

At the same time, we can envision possible improvements to the estimation. First, by adding information at the edge level, we can refine the assumption of homogeneous distributions made in definition 5; yet, this type of information can be costly. Second, a variable distribution of θ instead of a fixed one for all trips can be considered. Even though this could be difficult to check, this refinement is more feasible than the former one because the process is only needed during the set-up of the model. In other words, variable θ are static parameters that can be estimated through a deeper initial inquiry. Finally, higher abstractions may be foreseen, both in term of definitions and in the combination of the probabilities.

Acknowledgment

This work was partially funded by the grant *1st call for basic and applied research projects "University and Transport"* of the *National Ministry of Education*.

References

[1] J. de Dios Ortúzar and L. G. Willumsen, *Modelos de transporte*. Ed. Universidad de Cantabria, 2008, vol. 1.

[2] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[3] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014.

[4] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[5] S. Anapolsky, C. Lang, N. Ponieman, and C. Sarraute, "Exploración y análisis de datos de telefonıa celular para estudiar comportamientos de movilidad en la ciudad de buenos aires," in *XVIII CLATPU, Congreso Nacional de Transporte Público y Urbano*, 2014.

[6] A. J. Perez, L. D. Dominguez, A. J. Rubiales, and P. A. Lotito, "Optimización de matrices origen-destino estimadas a partir de datos georeferenciados en redes sociales," in *13º Simposio Argentino de Investigación Operativa*, 2015.

[7] G. Shafer *et al.*, *A mathematical theory of evidence*. Princeton university press Princeton, 1976, vol. 1.

[8] G. Shafer. (2015, Dec.) Dempster's rule of combination. [Online]. Available: doi:10.1016/j.ijar.2015.12.009 [Date accessed: May 20, 2016]

[9] H. E. Kyburg Jr, "Interval-valued probabilities," *Imprecise Probabilities Project*, 1998.

[10] P. Walley, *Statistical reasoning with imprecise probabilities*. Peter Walley, 1991, vol. 42.

[11] R. A. Jacobs, "Methods for combining experts' probability assessments," *Neural computation*, vol. 7, no. 5, pp. 867–888, 1995.