From “tap-in” locations to hotspots; a tale of two cities

Cristian Jara-Figueroa[*]
MIT Media Lab

Manuel Aristaran[†]
MIT Media Lab.

(Dated: December 13, 2014)

In this work we estimate the origin destination matrix for the users of the public transportation system of the city of Bahía Blanca, Argentina. We use the methodology proposed in [1]. The method takes the boarding time and location of each user as input, together with the time and location of each of the buses of the fare collection system. Using the estimated origin destination matrix we propose a way to estimate the spatial distribution of the hotspots for Bahía Blanca and for Santiago. The measure comes from the assumption that the trips follow a gravity model. This allows us to locate the time dependent location of the cities' hotspots. We find that both cities have a highly centralized structure. The hotspots of Bahía Blanca remain somewhat constant, and the hotspots of Santiago evolve from one part of the city to the other.

I. INTRODUCTION

The origin destination matrix displays the number of trips going from each origin to each destination. It has many uses in planning the transportation system within a city.

The estimation of such matrix has been a growing field of study. One of the earliest models proposed to model the trip distribution was the so called gravity model. It assumes that the number of trips between two areas depends on a property of the area the trips are originating from, a property of the area the trips are going to, and some function of the cost to go from one location to another. Despite its simplicity, this model has been the starting point for many studies.

Because of the digital systems of fare collection included in many modern public transportation systems, it is easier to gather data on how people move around the city. This data allows us to build an origin destination matrix for the sector of the population that uses the system. The question now is not how to model this matrix, but what can we say about the structure of the city once we have the matrix.

Many fare collection systems only gather data on the boarding spatio-temporal location of each user. Such is the case of the “Transantiago” public transportation system and “Bahía Transporte SAPEM”. In this work we take the data on the spatio-temporal location of each boarding point for each user of the public transportation system of Bahía Blanca. The data was collected during a time window of one week. The transportation system of Bahía Blanca consists of 19 different bus lines, and no subway system.

II. OD ESTIMATION FROM “TAP-IN” DATA

A. OD estimation

We follow the procedure described in [1] to estimate the origin destination matrix from our dataset. We first estimate the alighting point of each commuter, and then decide wether each alighting corresponds to a change of bus or to the end of the trip.

The alighting estimation is performed by minimizing a cost function that depends on the position in the bus route, and the following boarding point. The authors in [1] propose a generalized time defined as

$$Tg_i = t_i + f_W \frac{d_{i-post}}{s_W}. \tag{1}$$

Where t_i is the time at the i-th position of the bus, d_{i-post} is the distance between the i-th position of the bus and the next boarding point, s_W is the average walking speed, and f_W is a coefficient that quantifies the preference of the subject to either walk or ride the bus. The problem now consists of minimizing Tg_i with the constraint that

$$d_{i-post} < d, \tag{2}$$

where d is the maximum allowed walking distance; taken as $700m$ for Bahía Blanca.

Once the alighting point has been estimated, we consider that the subject has reached its final destination when the time between the alighting and the next boarding, minus the time the user takes to get from its alighting point to its boarding point, is greater than 30 minutes. If the next trip occurs before that time period and the passenger boards a different line, the we merge the previous estimated alighting with this *leg* of the trip.

Because Bahía Blanca is a small city, we require that no trip takes longer than 2 hours. We also filter trips that are shorter than 5 minutes or longer than 1 hour, due to the features of the studied territory.

[*] crisjf@mit.edu
[†] arista@mit.edu

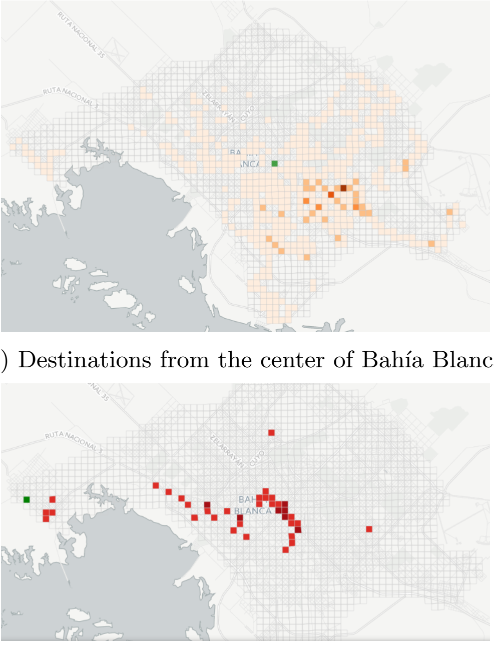
We generate a grid over the city, and aggregate trips according to their origin zone and their destination zone. Each zone has a square shape with 300 meters of side.

B. Results

From a total of $419,045$ tap-in records, and $6,319,982$ bus position records for the week from August 25th to August 29th 2014, this procedure was able to estimate a total of $300,312$ trips . The results are shown as an interactive visualization in *http://dump.jazzido.com/odviz/*.

We note that there is location in the center of the city where most of the trips are originated, see Figure (1a).

We also see that the area around downtown is the destination of most of the trips originated from the periphery. An example of this is shown in Figure (1b).



(a) Destinations from the center of Bahía Blanca.

(b) Example of the trips destinations from a location in the periphery of Bahía Blanca. Most of the commuters alight at the downtown area.

FIG. 1: The green area represents the source of the trips, the orange are the destinations. The darker the area, more trips have a destination in it.

III. USING HUMAN MOBILITY PATTERNS TO ESTIMATE THE IMPORTANCE OF A LOCATION.

Estimating the importance of a location is of particular interest for transportation planners and for starting business. People might want to know whether the precise location they are planning on opening a new business is of importance for the city or not. In what follows we propose a measure of the "charge" of each location within a city. This charge can be directly estimated from the origin destination matrix.

To test our method, we use the generated OD matrix for Bahía Blanca, and the available OD matrix for Santiago. Both matrices are aggregated by half hour, and therefore we can calculate the charge of each location by half hour.

A. Locating hotspots

1. *Travel charge*

Let us assume that the study under study is divided in N zones of about the same size, each characterized by its geometric center. For practical purposes these zonas can be transportation zones, as in the case of Santiago, or a square grid, as in the case of Bahía Blanca.

A gravity model [2] is typically used to generate a synthetic trip matrix matrix T, where T_{ij} is the number of trips originated at location i that have j as a destination. According to this type of models T_{ij} can be generated from the number of trips O_i originated at zone i as

$$T_{ij} = \frac{O_i q_j}{f(d_{ij})} \tag{3}$$

where d_{ij} is the distance between locations i and j, $f(d)$ is some function of d that is strictly increasing for large values of d, and q_j is a value that quantifies how attractive location j is for commuters. We will call this value the *travel charge* of location j, in analogy with Coulomb's Law.

Without loss of generality we write $f(d) = e^{-P(d)}$, and our model becomes

$$T_{ij} = O_i q_j e^{P(d)}. \tag{4}$$

Note that T_{ij} and O_i are directly obtained from the OD matrix. Our goal will be to estimate q_j, under the assumption that the gravity model holds. We must, however introduce an additional assumption on the form of f.

We can write Eq. (4) as

$$\log\left(\frac{T_{ij}}{O_i}\right) = P(d_{ij}) + \log q_j. \tag{5}$$

The two simplest options will be to assume $P(d)$ a polynomial of first order in d, and f becomes an exponencial, and $P(d)$ a polynomial of first order in $\log d$, and f becomes a power law. The choice will depend on the structure of the city we are studying; for Santiago we chose f to decay as a powerlaw, and for Bahia we choose f to be an exponential.

The coefficients of the polynomial $P(d)$ are the regression coefficients of $\log\left(\frac{T_{ij}}{O_i}\right)$ on d_{ij} and a constant, or on

$\log d_{ij}$ and a constant. Since the scale of q_i is not a fixed one, we can assume that $E\{\log q_j | d_{ij}\} = E\{\log q_j\} 0$. Note that this assumption, of the expected value and not the in-sample average, is just an assumption of the scale. A more general framework would lift this assumption and the final result would be modulated by $e^{-\log q_j}$.

Once f has been estimated from the data we can calculate the charge as

$$q_j = \frac{1}{T} \sum_{i=1}^{N} T_{ij} f(d_{ij}), \tag{6}$$

where $T = \sum_{ij} T_{ij}$ is the total number of trips. What this measure is doing is weighting the numbers of commuters that are heading to location j according to the distance they are traveling to get to j. A location can have a high charge when there are a lot of commuters arriving at the location, or when the commuters come from very distant locations within the city. Therefore this measure not also takes into account information about the number of incoming trips, but also about how far away are they coming from.

2. Spatial distribution of hotspots

One important thing to note is that large values of q_i give account for hotspots within the city; they are places that have high interest for commuters. The Venables index

$$V = \sum_{ij} s_i s_j d_{ij}, \tag{7}$$

introduced in [4] was used by in [3] to quantify the spatial distribution of the hotspots of a city. In Eq. (7) s_i is the share of individuals present at location i.

Motivated by this idea we propose a measure of the spatial distribution of the hotspots, calculated according to the travel charge of each location in the city.

$$V_\eta = \frac{\pi}{2} \sum_{ij} \frac{\eta_i}{\eta_T} \frac{\eta_j}{\eta_T} d_{ij}, \tag{8}$$

where $\eta_T = \sum_i \eta_i$ is the total charge of the city. The coefficients in front of the sum in (8) are taken such that if all the locations with $\eta_i \neq 0$ are located homogeneously distributed along the periphery of a circle of radius R, and they all have the same charge, then $V_\eta = 2R$ is the diameter of the circle. This scaling makes the value of V_η meaningful in terms of a distance and will allow us to compare different cities.

3. Results

Using the origin destination matrix for Santiago de Chile we fit a powerlaw on f and calculate the charge of each of the 795 areas. We do the same of the estimated origin destination matrix of Bahía Blanca, using an exponential for f. The results can be seen at *http://dump.jazzido.com/odviz/charge.html.*

We note that Santiago is a very centralized city during the entire day and that other locations of interest appear in sector Lo Barnechea, northeast of downtown, during the morning and in sector Puente Alto, far southeast from downtown, in the afternoon. It is important to note that despite been conditioned by the number of incoming trips, the charge of a zone is not directly dependent of it. For example, the sector of Lo Barnechea is an area not regarded as important when looking only at the number of incoming trips, but it has a high charge. This is because many of the people commuting to Lo Barnechea in the morning travel a long distance to get there. When taking a closer look at the trips that have as destination one of the zones in Lo Barnechea, we see that there are some commuters that board the public transportation system in the sector Puente Alto, very far away from Lo Barnechea. Future work could involve relating the average income of each one the zones with their hotspot status.

As an example we take a look at the evolution of the charge, and the incoming number of trips, for two hotspots of Santiago. In Figure (4) we see a morning hotspot, and an afternoon hotspot.

The macro indicators of total charge, number of trips, and Venables index are shown in Figure (5) for both Santiago and Bahía Blanca. The total charge of both cities peaks in the early morning, when a lot of people are heading from the outside of the city to the center, and then decays up to a steady state as the day passes.

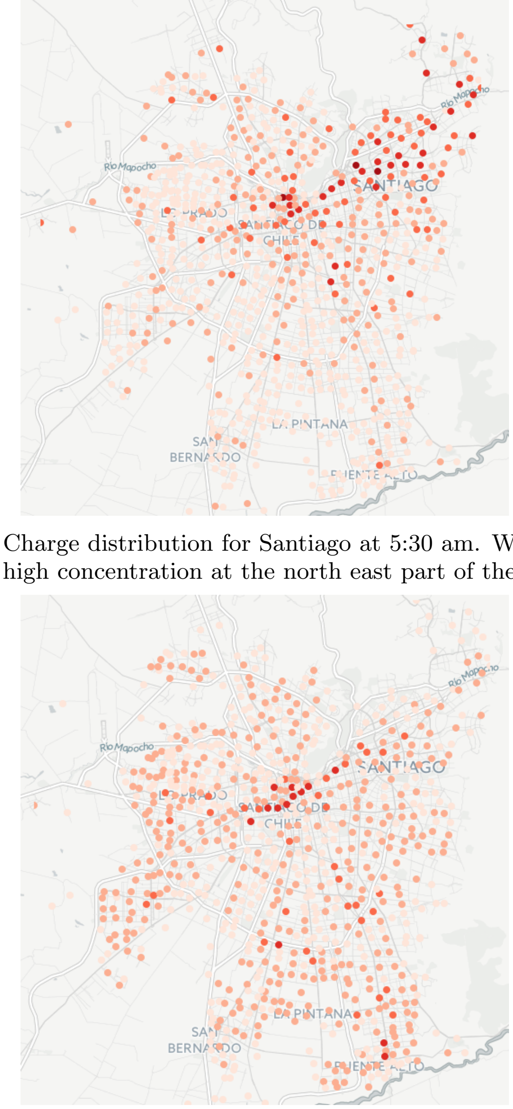
The total number of trips for Santiago, on the other hand, has a peak in the morning and other in the afternoon. We see an interesting behavior in Bahía Blanca, as the total number of trips has a third peak in the middle of the day; this is due to the many people that cut the day in half and go home for lunch, a very characteristic behavior of small cities in Argentina.

The Venables index increases as the day passes, meaning that the hotspots of the city are more and more spread. The habit of going home for lunch in Bahía Blanca is also captured by this index.

The analysis is not very powerful in this case, since both cities have a very centralize layout. It would be very interested analyzing a city that has many centers.

IV. CONCLUSION

In this work we have successfully built the origin destination matrix for the city of Bahía Blanca, starting from the data on the boarding time and location of each user. We would like to remark that this process is very simple and generalizable, however the parameters must be fitted according to the specific city the method is been applied to. Further restrictions on the model, like assum-



(a) Charge distribution for Santiago at 5:30 am. We see the high concentration at the north east part of the city.

(b) Charge distribution for Santiago at 6:00 pm. We see some hotspots appearing in the southern part of the city.

FIG. 2: Charge distribution for Santiago. The downtown area is a very large hotspot.

ing a minimal time duration for the trips, must also be location specific.

It must be considered that there is an inherent bias in measuring and characterizing city features from an origin destination matrix that only contains trips performed in the public transportation system. Further studies could take into account other sources of data, like inner city "tag systems", when they are available.

As we have shown, with the growing amount of available data, origin destination matrices are becoming much easier to estimate. This gives a big opportunity to analyze the inner structure of a city base on this data. In this article we were able to estimate the hotspots of the city, and their spatial distribution.

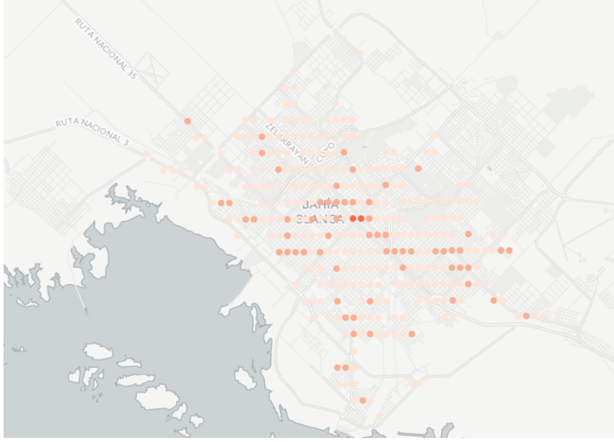


FIG. 3: Charge distribution for Bahía Blanca at 1.30 pm. The downtown area is clearly a big hotspot; this behavior remains constant throughout the day.

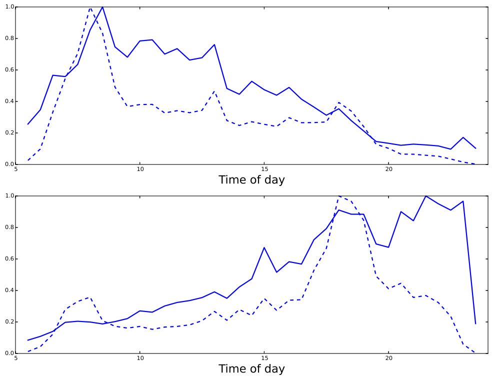


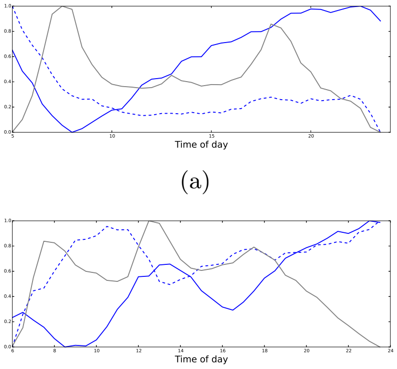
FIG. 4: Charge and incoming number of trips evolution. The solid line corresponds to the charge and the dashed line to the incoming number of trips. The upper panel is a hotspot location for the morning period, and the bottom panel is a hotspot location for the afternoon period. We see that despite the fact that the number of trips drive big changes in the charge, their behavior is different. The values of the y-axis are just referencial.

V. ACKNOWLEDGMENTS

We thank *Bahía Transporte SAPEM* and *Agencia de Innovaciǿn y Gobierno Abierto* of Bahía Blanca for providing the datasets.

[1] Marcela A. Munizaga and Carolina Palma, Transportation Research Part C 24 (2012) 918.

[2] Juan de Dios Ortúzar and Luis G. Willumsen, Modelling Transport, 4th edition, 2011 John Wiley & Sons, Ltd.



(a)

(b)

FIG. 5: Macro indicators of mobility. Gray: Total number of trips. Dashed blue: Total charge. Solid blue: Venables index. The values are only referencial since all the curves were scaled to fit the range between 0 and 1.

[3] Thomas Louail, Maxime Lenormand, Oliva García Cantú, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, Marc Barthelemy, arXiv:1401.4540.

[4] Pereira RHM, Nadalin V, Monasterio L, Albuquerque PHM (2013) Urban centrality: A simple index. Geographical Analysis 45: 7789.