ESTIMATION OF A ZONAL ORIGIN-DESTINATION MATRIX BASED ON SMARTCARD TRIP DATA: A PLANNING SOFTWARE APPLICATION

Sebastián Tamblay, Pontificia Universidad Católica de Chile, stm@ing.puc.cl
Juan Carlos Muñoz, Pontificia Universidad Católica de Chile, jcm@ing.puc.cl
Juan de Dios Ortúzar, Pontificia Universidad Católica de Chile, jos@ing,puc.cl

ABSTRACT

We present an extension of the methodology presented by Tamblay et al. (2016), which allows to estimate a zonal origin-destination (O-D) matrix for a city, given a stop-to-stop trip matrix. The extension focuses on three improvements: addressing data limitations, enhancing the model formulation, and updating Santiago's zonal system, including a new methodology for access links construction. The inclusion of the resulting matrix in the new transit planning tool that is being developed (STEPTRANS) will allow the evaluation of major changes to the transport system, as a zonal O-D matrix allows modelled users to choose their initial and final stops.
Keywords: public transport; zonal origin-destination matrix; smartcard data.

RESUMEN

Presentamos una extensión de la metodología presentada en Tamblay et al. (2016), que permite estimar una matriz origen-destino (O-D) zonal para una ciudad, dada una matriz de viajes paradero-a-paradero. La extensión comprende tres mejoras: aborda limitaciones de datos, mejora la formulación del modelo, y actualiza la zonificación de Santiago, incluyendo una nueva metodología para construir arcos de acceso. Incluir esta matriz en la nueva herramienta de planificación de transporte público que se está desarrollando (STEPTRANS) permitirá evaluar cambios mayores en el sistema, porque la matriz zonal permite a los usuarios modelados escoger sus paraderos de inicio y fin del viaje.
Palabras claves: transporte público, matriz origen-destino zonal; tarjetas inteligentes.

1 INTRODUCTION

Recent years have seen an increasing trend for implementing information technologies, such as Automated Fare Collection (AFC) and Automatic Vehicle Location (AVL), in public transport systems. These technologies provide valuable information about trip patterns and, with adequate computational tools, this new massive data can be processed and serve as input for planning decisions.

A key planning consideration when designing transit services is the projected demand on each route, as this is a necessary input to determine frequencies and vehicle capacities. Predicting demand changes in complex networks requires a thorough modelling of the public transport system and their users' decisions.

Aiming to aid in this process, Project D10I1049-D10E1049 of the *Fondo de Fomento al Desarrollo Científico y Tecnológico* (FONDEF) is developing a public transport planning computational tool: STEPTRANS (Spanish initials for tactical-strategical planning software). This tool seeks to predict the reactions of public transport users to network changes, allowing better design and operation of the transport system (Raveau et al., 2017).

As one of its main inputs, this tool needs a reliable origin-destination (O-D) matrix for the city of interest, which is assigned to the transit network using state of practice route choice models. STEPTRANS's basic O-D matrix is the "stop-to-stop" trip matrix developed by Munizaga and Palma (2012) (and enhanced in Devillaine et al., 2012; and Munizaga et al., 2014), which identifies boarding stops and estimates alighting points using AFC and AVL.

However, this stop-to-stop matrix is not really adequate to estimate the impact of projects that could change the stop where users start or end their trips; *e.g.*, extending routes, relocating bus stops, designing new routes or changing the fare structure. To study this kind of changes, a zonal O-D matrix that allows modelled users to choose their initial and final stops is needed.

In this line, Tamblay et al. (2016) presented a methodology to estimate a zonal origin-destination matrix for a city, given a stop-to-stop trip matrix. The approach infers a probability for the zone of origin and destination for each trip between two stations, using a discrete choice model. Model inputs are land use and public transport information, for any period of interest in any city. The methodology was applied to estimate a model for Santiago de Chile's morning peak period.

This study presents a series of extensions to the previous methodology, focusing on three major improvements: (i) addressing data limitations, (ii) enhancing the model formulation, and (iii) updating Santiago's zonal system, including a new methodology for access links construction.

The remainder of this paper is structured as follows. Section 2 presents a brief summary of Tamblay et al. (2016) methodology, along with the data improvements and the proposed enhanced model formulation. Section 3 introduces and describes model inputs for the case of Santiago de Chile's; additionally, it describes the proposed methodology for access links construction. Then, Section 4 presents the basic zonal inference model, along with a validation analysis and the results of the

proposed model extensions. Finally, in Section 5 we draw our conclusions, present future research guidelines, and provide policy applications of this study.

2 MODEL FORMULATION

This section presents the zonal inference model developed by Tamblay et al. (2016), which assigns a probability for the zone of origin and destination for each entry in a stop-to-stop trip matrix. In other words, the model infers the access (and egress) walks for each trip between two bus stops (or metro stations), considering land use and network information of zones within walking distance. Furthermore, this section also presents the proposed enhanced model formulation.

2.1 Zonal inference model

The proposed zonal inference model allows us to obtain the probability that an observed trip using access stations k and l (as their boarding and alighting points, respectively), was originated in zone i and had zone j as its destination. This probability is defined as $\text{Prob}\left(^{ij}/_{kl}\right)$.

In order to obtain these probabilities for each bus stop, metro station, and zone, we follow Daly (1982) and formulate this particular gravitational model:

$$T_{ij}^{kl} = A_i B_j f_{ij}^{kl} \quad \forall (i,j)_\wedge (k,l) \tag{1}$$

Subject to the constraints:

$$\sum_m \sum_n T_{mn}^{kl} = T_{kl} \quad \forall (k,l) \tag{2}$$

where A_i and B_j are measures of the generating and attracting powers of zone i and j, linked to the population and land use information for these zones; T_{ij}^{kl} is the (unknown) number of trips made from zone i to zone j that use access stations k and l as their boarding and alighting point, respectively; f_{ij}^{kl} is an inverse measure of the "cost" of choosing those detentions to make the trip; T_{kl} is defined as the observed trips between public transport detentions k and l, which is supposed known for every k and l; lastly, sets m and n are the zones of influence for detentions k and l, respectively (i.e., zones within walking distance from each bus stop or metro station).

Then, by defining:

$$V_{mn/kl} = ln(A_m B_n f_{mn}^{kl}) \tag{3}$$

The latter can be represented by a disaggregate Logit model (Daly, 1979; McFadden, 1979), as follows:

$$Prob\left(^{ij}/_{kl}\right) = \frac{e^{V_{ij/kl}}}{\sum_m \sum_n e^{V_{mn/kl}}} \tag{4}$$

Different model specifications can be estimated using survey data and the Maximum Likelihood method (Ortúzar and Willumsen, 2011).

It is worth noting that this Logit model including size variables is obtained without any loss of generality and imposing no restrictions to the specifications of the cost function implied (Daly, 1982).

Finally, defining T_{ij} as the trips from zone i to j, we can reconstruct each element of the desired zonal O-D matrix from the observed travel behaviour between access stations as shown by Equation (7):

$$T_{ij} = \sum_k \sum_l Prob\left({}^{ij}/_{kl}\right) * T_{kl} \quad \forall(i,j) \qquad (5)$$

2.2 Enhanced model formulation

First, the new methodology uses, as its estimation dataset, the latest large scale O-D survey available for Santiago de Chile (SECTRA, 2015). This is a significant improvement, as the previous work used a smaller survey with lesser coverage and not validated by the Chilean authorities. Using the new O-D survey has the added advantage that it has data for the whole day (and not only for the morning, as the previous dataset), enabling the estimation of the model for new periods: afternoon peak and morning off-peak. Additionally, O-D surveys are periodically conducted and available for both Santiago and other large cities of the country. This way, developing a process that enables to estimate the model directly from this kind of survey will simplify Santiago's model updates and potential STEP implementations in other cities.

Additionally, using the O-D survey as the new dataset enables the estimation of a weighted Logit model, allowing the inclusion of basic socio-demographic and trip patterns corrections, thus decreasing selection bias.

On the other hand, as advised in specialized literature, we analysed distinguishing travellers attracted to different destination alternatives. In this line, considering that students in Santiago are given special smart-cards (which allow them to pay a third of the regular fare), we studied models in which student cards received a different treatment. This way, we included an interaction between the total educational squared metres and smart-card type, using a dummy variable with unitary value for student cards. This student differentiation approach was presented first by Tamblay et al. (2015).

Finally, the base model's inverse cost function (f_{ij}^{kl}) was related to the access and egress links distances, from the observed stops *k* and *l* to zones *i* and *j*, respectively (Tamblay et al., 2016). However, the users' perceived cost of choosing particular stops to make a trip between two zones encompass a process significantly more complex than a simple distance measure. This way, we included new cost measures, using STEPTRANS's route choice model utilities as an indicator of the users' perceived cost of choosing these particular stops to make a trip between those two zones.

This way, a route and access station hierarchical Logit choice model, currently under development (with preliminary results presented in Abud, 2015), was included in the calibration procedure. In this choice model, users decide which route and bus stops (or metro stations) they will use in order to travel between two particular zones.

3 DATA AND MODEL INPUTS FOR SANTIAGO, CHILE

This section presents the different sources of information that were gathered and unified in order to estimate and apply the zonal inference model for Santiago's morning peak. A brief description of data origins and processing, along with new methodology for access links construction, is explained next.

3.1 Smart-card origin-destination data

The zonal inference model works by assigning observed trips between two public transport stops to nearby zones, through a Logit model. In this study, we considered observed trips as those belonging to the stop-to-stop trip matrix estimated through the methodology presented in Munizaga and Palma (2012) (and enhanced in Devillaine et al., 2012; and Munizaga et al., 2014), for April 2013 Santiago's network. These trips were used in T_{kl} variables defined in the previous section.

It is important to note that the way these matrices were built translates into some limitations for this study. Particularly, the assigned stop-to-stop matrix does not explicitly includes non-integrated modes (such as share taxis and private vehicles) nor fare evasion, which distorts some of its trips; this limitation is extended to the zonal O-D matrix when assigned by the zonal inference model.

3.2 Land use information

Generating and attracting powers of each zone (A_i and B_j, respectively) are linked to the land use information of each one of them. The latter allows the model to capture that residential zones generate a larger amount of trips or business centres attract more trips, for instance.

Land use information was included for Santiago de Chile for 2014, which contains the number of units and total squared meters dedicated to each land use classification in each zone. Some of the most relevant classifications available are: residential, commercial, educational, industrial, health and offices buildings. This information is gathered and updated periodically by the Chilean Internal Revenue Service (SII, 2014) with taxation and regulatory purposes, meaning that it is available to transportation planners at practically no cost.

3.3 Latest large scale O-D survey

In order to estimate the zonal inference model through the maximum likelihood procedure, survey entries were included. As previously mentioned, this study uses the latest large scale O-D survey available for Santiago de Chile, as its estimation dataset for the three periods modelled (SECTRA, 2015).

After digitation and processing, a total of 2033 valid entries (i.e., which had clearly identified routes, origins, and destinations) remained for the morning peak period. Similarly, 1051 valid entries remained for the afternoon peak and 1039 for the morning off-peak period.

3.4 Zonal system and access links

The proposed model allocates observed trips into nearby zones, following the probabilities resulting from the Logit calculations. This way, one of the key elements to determine is the city's zonal system that will be used. The chosen zoning system is based in the one proposed in SECTRA (2015) O-D survey, but some zones were considered too large and were divided, resulting in an original zonal system with 1,171 zones.

In order to allow the representation of the alternatives considered by users and further application of the zonal inference model, Tamblay et al. (2016) presented a methodology to add access links from zonal centroids to bus stops and metro stations. This was achieved with two sequential tasks: (i) decide which stops were connected by access links to each zone; and (ii) assign each one of them a distance value.

The first was solved by connecting centroids to bus stops and metro stations that were inside the zonal boundaries, but also inside an imaginary extension of its borders defined as the influence radius. Influence radii were obtained from Abud (2015) survey data, resulting in a measure of 250 metres for bus stops and 750 metres for metro stations.

Once defined which zones were connected to which stops (both for the software and zonal inference model), we proceeded to assign a distance value to each access link, considering land use variables from the city blocks of each zone. The purpose was to obtain accurate values for access times as each zone is not necessarily homogenous and, for instance, population may be concentrated in particular areas of the zone.

This way, each city block M was assigned a point in its geometrical centre with two different weights linked to land use characteristics; one for its potential as a trip generator (w_M^g , linked to residential land use, i.e., housing squared metres), and the other as a trip attractor (w_M^a , linked to commercial, industrial, residential, health, educational, and offices total squared metres).

It is worth highlighting that the relative attraction importance of each land use type was obtained from preliminary zonal inference models, through an iterative procedure (rapidly convergent) in which the model attraction parameters were used to build new access arcs, which in turn lead to new models with updated parameters. This iterative procedure was first presented in Tamblay et al. (2015).

Then, distances are measured from each one of these points to each public transport detention inside the zone; finally, weighted averages of these distances are calculated using the generating and attracting weights of each city block. Thus, two distances are calculated for each detention: one corresponding to the generator link and other to the attractor link of the zone. Note that this methodology allows access time from zone i to station k to be different from the access time from station k to zone i, giving a better representation of heterogeneous zones (where population and trip attractors are located in different sectors).

4 FINAL MODEL AND ZONAL MATRIX ESTIMATION

This sections presents model results, obtained using BIOGEME (Bierlaire, 2003), for the three modelling periods: morning peak (trips ending between 8:00 and 9:00), afternoon peak (trips starting between 18:00 and 19:00), and morning off-peak (trips with mean time between 10:00 and 12:00).

First, the base model for each period is presented. Then, results are validated by comparing them with alternative simpler approaches. Third, model extensions are discussed and presented, including student cards differentiation and the new cost measure proposed in this study.

4.1 Basic zonal inference models

Different specifications for the zonal inference model were estimated and compared, following the maximum likelihood procedure and a Weighted Multinomial Logit formulation. Weights were obtained from SECTRA (2015) O-D survey and the proposed utility function follows Tamblay et al. (2016), with the general form:

$$V_{mn/kl} = l\,n(A_m) + ln(B_n) + \ln(f_{mn}^{kl}) \tag{6}$$

More specifically, the proposed functions were expressed in the following manner:

$$V_{mn/kl} = \theta_O * \ln\left(\theta_{u_{h_O}} * U_{h_{Om}}\right) + \theta_D * \ln\left(\sum_P \theta_{m2_{P_D}} * M2_{P_{Dn}}\right) + \ln\left(\frac{e^{(\beta_O * d_{m-k})}}{(d_{m-k})^{\gamma_O}}\right) + \ln\left(\frac{e^{(\beta_D * d_{l-n})}}{(d_{l-n})^{\gamma_D}}\right) \tag{7}$$

Where $M2_{H_{O_m}}$ are the total housing squared metres in the zone of origin m; $M2_{P_{Dn}}$ are the total squared metres of land use classification P in destination zone n; d_{m-k} and d_{l-n} are access distances in kilometres from origin and destination zones to the chosen initial and final detentions, respectively; finally, θ, β, and γ are parameters to be estimated. The first and second sum terms are related to the generating and attracting powers of zones m and n, respectively; while the last two correspond to the inverse cost measure f_{mn}^{kl}, where we used a gamma function and explicitly allowed different impacts of access and egress distances on utility.

It is important to note that, for identifiability reasons, we cannot estimate a parameter for each size variable in origin and destination, so two of them are fixed to a value of one. Furthermore, both θ_O and θ_D are fixed to unitary values, in order to ensure that the model is independent from the zonal system used for estimation (Daly, 1982).

Table 1, 2, and 3 shows the estimation results for this formulation, for the morning peak, afternoon peak, and morning off-peak, respectively.

Table 1: Zonal inference model for the morning peak period

Associated to	Parameter	Value	t test
Generation (origin) A_i	λ_O	1	Fixed
	$\theta_{m2_{C_O}}$ - Commercial m^2	0	Dropped
	$\theta_{m2_{E_O}}$ - Educational m^2	0	Dropped
	$\theta_{m2_{H_O}}$ - Habitational m^2	1	Fixed
	$\theta_{m2_{I_O}}$ - Industrial m^2	0	Dropped
	$\theta_{m2_{O_O}}$ - Offices m^2	0	Dropped
	$\theta_{m2_{S_O}}$ - Health m^2	0.903	1.81
Attraction (destination) B_j	λ_D	1	Fixed
	$\theta_{m2_{C_D}}$ - Commercial m^2	0.348	2.31
	$\theta_{m2_{E_D}}$ - Educational m^2	0.534	2.32
	$\theta_{m2_{H_D}}$ - Habitational m^2	0.428	2.86
	$\theta_{m2_{I_D}}$ - Industrial m^2	0.487	2.18
	$\theta_{m2_{O_D}}$ - Offices m^2	0.596	2.55
	$\theta_{m2_{S_D}}$ - Health m^2	1	Fixed
Cost f_{ij}^{kl}	β_O	0	Dropped
	γ_O	2.42	26.53
	β_D	-2.60	-3.75
	γ_D	1.38	3.94
Fit	**Null log-likelihood**	**-5519.6**	
	Final log-likelihood	**-4256.0**	
	Likelihood ratio test	**2527.2**	
	Valid entries	**2033**	

Table 2: Zonal inference model for the afternoon peak period

Associated to	Parameter	Value	t test
Generation (origin) A_i	λ_O	1	Fixed
	$\theta_{m2_{C_O}}$ - Commercial m^2	0.440	2.56
	$\theta_{m2_{E_O}}$ - Educational m^2	0	Dropped
	$\theta_{m2_{H_O}}$ - Habitational m^2	0.330	2.90
	$\theta_{m2_{I_O}}$ - Industrial m^2	0.326	1.86
	$\theta_{m2_{O_O}}$ - Offices m^2	0.623	2.47
	$\theta_{m2_{S_O}}$ - Health m^2	1	Fixed
Attraction (destination) B_j	λ_D	1	Fixed
	$\theta_{m2_{C_D}}$ - Commercial m^2	0	Dropped
	$\theta_{m2_{E_D}}$ - Educational m^2	0	Dropped
	$\theta_{m2_{H_D}}$ - Habitational m^2	1	Fixed
	$\theta_{m2_{I_D}}$ - Industrial m^2	0	Dropped
	$\theta_{m2_{O_D}}$ - Offices m^2	0	Dropped
	$\theta_{m2_{S_D}}$ - Health m^2	2.01	2.72
Cost f_{ij}^{kl}	β_O	0	Dropped
	γ_O	1.92	19.14
	β_D	-2.13	-2.48
	γ_D	1.31	2.84
Fit	**Null log-likelihood**	**-3070.2**	
	Final log-likelihood	**-2477.6**	
	Likelihood ratio test	**1185.4**	
	Valid entries	**1051**	

Table 3: Zonal inference model for the morning off-peak period

Associated to	Parameter	Value	t test
Generation (origin) A_i	λ_O	1	Fixed
	$\theta_{m2_{C_O}}$ - Commercial m^2	0.205	2.28
	$\theta_{m2_{E_O}}$ - Educational m^2	0	Dropped
	$\theta_{m2_{H_O}}$ - Habitational m^2	1	Fixed

	$\theta_{m2_{I_O}}$ - Industrial m^2	0	Dropped
	$\theta_{m2_{O_O}}$ - Offices m^2	0	Dropped
	$\theta_{m2_{S_O}}$ - Health m^2	1.70	2.46
Attraction (destination) B_j	λ_D	1	Fixed
	$\theta_{m2_{C_D}}$ - Commercial m^2	1	Fixed
	$\theta_{m2_{E_D}}$ - Educational m^2	0.684	2.66
	$\theta_{m2_{H_D}}$ - Habitational m^2	0.375	4.93
	$\theta_{m2_{I_D}}$ - Industrial m^2	0	Dropped
	$\theta_{m2_{O_D}}$ - Offices m^2	0.326	2.15
	$\theta_{m2_{S_D}}$ - Health m^2	0	Dropped
Cost f_{ij}^{kl}	β_O	-2.36	-2.30
	γ_O	1.42	2.69
	β_D	0	Dropped
	γ_D	2.36	21.39
Fit	**Null log-likelihood**	**-2933.2**	
	Final log-likelihood	**-2281.7**	
	Likelihood ratio test	**1302.8**	
	Valid entries	**1039**	

From results above we observe that all included parameters have the expected signs, and most of them are statistically significant. A more detailed look highlights that in the morning peak habitational and health buildings generate trips; the first one is obvious, while the latter is explained mainly by early appointments and work shift changes. In this period, all attracting variables resulted significant, with offices and health being the most important. On the other hand, in the afternoon peak we can observe an inverse trend, in which habitational and health buildings become the more important trip attractors. Educational land use also loses its significance, as most schools and universities classes end earlier than the afternoon peak. Finally, the morning off-peak presents a similar situation to its peak counterpart; however, industries and health loses some attracting power, while commercial building gain. It is surprising that in this period health land use didn't resulted significant, while in other periods resulted so important. A possible explanation is that the estimation dataset could have few health purpose trips, insufficient to capture their attraction effect precisely. Anyhow, this subject is being addressed and models are being checked for input inconsistencies.

Regarding the inverse cost function, parameter β_O and β_D dropped out (in different periods) due to its low significance and the fact that it is not fundamental (cost function maintains the expected trend without it).

4.2 Validation analysis

Results validation was done by comparing the model's prediction to two alternative zonal assignments: equiprobable, which means assigning the trips in each stop with the same probability to each zone connected with an access link to it; and stop's physical zone assignment, which means assigning all trips to the zone where the stop is. The model was expected to behave better than these alternatives, as the first one is a very simple solution and the second one should fail often, as zones are relatively small and most zonal boundaries are defined in important streets.

The three alternative assignments (inference model, equiprobable, and stop's physical zone) were applied to each survey origin and destination stops, obtaining three predicted O-D matrices, each following one of these approaches. These predicted matrices were then compared to the survey's actual O-D matrix, and percentage absolute differences were obtained for each alternative (Table 4). Results show that the proposed model allows for better results than applying alternative solutions.

Table 4: Percentage absolute differences, proposed model comparison with alternatives

	Morning peak		Afternoon peak		Morning off-peak	
	Origin	Destination	Origin	Destination	Origin	Destination
Stop's physical zone	75.8%	70.6%	77.5%	84.3%	81.8%	71.4%
Equiprobable	61.1%	69.4%	76.1%	74.5%	71.5%	76.0%
Inference model	45.1%	55.8%	62.4%	60.6%	59.7%	57.4%

4.3 Extensions results

The proposed student smartcard differentiation was studied, using the formulation first proposed by Tamblay et al. (2015) and detailed in Section 2.2. However, in contrast with the significant results obtained in the previous study, the new models could not capture significant effects of the interaction.

Better results were obtained with the new cost measure proposed as a model extension in Section 2.2. In this formulation, f_{ij}^{kl} variables will be related to the probability that a trip originated in zone i with its destination in zone j uses access stations k and l, as their boarding and alighting points (defined as $Prob\left({}^{kl}/_{ij}\right)$), since when a route alternative is perceived as costly the probability of choosing its related detentions must be low for a rational user (and vice versa for low cost routes). As previously stated, these probabilities are obtained from STEPTRANS's route and access station choice model (with preliminary results presented in Abud, 2015).

This way, access and egress distances are already included in $Prob\left({}^{kl}/_{ij}\right)$ and are therefore removed from f_{ij}^{kl}, resulting in a new formulation as shown in Equation (8).

$$f_{ij}^{kl} = Prob\left({}^{kl}/_{ij}\right)^{\gamma_P} * e^{\beta_P * Prob\left({}^{kl}/_{ij}\right)} \quad (8)$$

The morning peak model was re-estimated using this new formulation, and results are shown in Table 5. Results show a significant likelihood gain, when compared with the basic model that used a simpler cost function. Indeed, the likelihood ratio test improved from 2527.2 to 4225.9. It is important to note that 2 observations dropped (going from 2033 to 2031 valid entries), as the route choice model assigned a null probability that a trip between those zones would use the chosen stops.

Table 5: Zonal inference model for the morning peak period

Associated to	Parameter	Value	t test
Generation (origin) A_i	λ_O	1	Fixed
	$\theta_{m2_{C_O}}$ - Commercial m^2	0	Dropped
	$\theta_{m2_{E_O}}$ - Educational m^2	0	Dropped
	$\theta_{m2_{H_O}}$ - Habitational m^2	1	Fixed
	$\theta_{m2_{I_O}}$ - Industrial m^2	0	Dropped
	$\theta_{m2_{O_O}}$ - Offices m^2	0	Dropped
	$\theta_{m2_{S_O}}$ - Health m^2	1.36	2.15
Attraction (destination) B_j	λ_D	1	Fixed
	$\theta_{m2_{C_D}}$ - Commercial m^2	0.646	2.32
	$\theta_{m2_{E_D}}$ - Educational m^2	0.204	1.67
	$\theta_{m2_{H_D}}$ - Habitational m^2	0.175	2.56
	$\theta_{m2_{I_D}}$ - Industrial m^2	0.0321	0.76*
	$\theta_{m2_{O_D}}$ - Offices m^2	1.37	2.41
	$\theta_{m2_{S_D}}$ - Health m^2	1	Fixed
Cost f_{ij}^{kl}	β_P	-17.4	-10.20
	γ_P	29.8	12.84
Fit	**Null log-likelihood**	**-5515.7**	
	Final log-likelihood	**-3402.7**	
	Likelihood ratio test	**4225.9**	

Valid entries	**2031**

5 CONCLUSIONS

In conclusion, we have presented an extension for the methodology presented by Tamblay et al. (2016), that allows us to estimate a zonal origin-destination matrix from an observed stop-to-stop O-D matrix. The extensions focused on three major improvements: (i) addressing data limitations, (ii) enhancing the model formulation, and (iii) updating Santiago's zonal system, including a new methodology for access links construction.

The extended methodology is applied to Santiago's transit system, including its morning peak, afternoon peak, and morning off-peak periods. However, the methodology and model are completely general and could be used in any public transport system with sufficient information.

By including a more massive and complete data source, the latest available O-D survey for Santiago (SECTRA, 2015), many of the limitations of the base model (Tamblay et al., 2016) were solved. However, the resulting zonal matrix is still an assignment of the stop-to-stop trip matrix and does not explicitly include non-integrated modes (such as share taxis and private vehicles) nor fare evasion. The latter is being addressed in another study, also related to the new computational tool being developed (STEPTRANS).

Finally, the paper highlights the policy implications of the research. The inclusion of the resulting matrix in STEPTRANS will allow the evaluation of major changes to the transport system. Previously, transport projects that modified the stop where users started or ended their trips could not be properly evaluated, because the trip assignment stage was based on a stop-to-stop matrix. For example, it was not possible to know how many trips would start or end in a new bus stop. So, the proposed methodology and model becomes a key ingredient of the public transport planning software for Santiago, allowing the analysis of trips at a new zonal level, resulting in better and more flexible predictions.

ACKNOWLEDGEMENTS

This research was supported by FONDEF D10I1049 and FONDEF D10E1049 "Una herramienta táctico-estratégica de gestión y planificación de sistemas de transporte público urbano".

REFERENCES

Abud, I., 2015. *Modelos de Elección de Paradero, Modo y Ruta para Herramientas de Planificación de Transporte Público*. Tesis de Magíster en Ciencias de la Ingeniería, Pontificia Universidad Católica de Chile.

Bierlaire, M., 2003. *BIOGEME*: A free package for the estimation of discrete choice models. *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland.

Daly, A. J., 1979. Some developments in transport demand modelling. In *Behavioural Travel Modelling* (Edited by D.A. Hensher and P.R. Stopher). Croom Helm, London.

Daly, A. J., 1982. Estimating choice models containing attraction variables. *Transportation Research Part B: Methodological*, *16*(1), 5-15.

Devillaine, F., Munizaga, M., and Trepanier, M., 2012. Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board, 2276* (1), 48-55.

McFadden, D., 1979. Quantative methods for analysing travel behaviour of individuals. In *Behavioural Travel Modelling* (Edited by D.A. Hensher and P.R. Stopher). Croom Helm, London.

Munizaga, M., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, *24*, 9-18.

Munizaga, M., Devillaine, F., Navarrete, C., Silva, D., 2014. Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, *44*, 70-79.

Ortúzar, J. de D., Willumsen, L. G., 2011. *Modelling Transport* (Fourth Edi.). Chichester: John Wiley & Sons, Ltd.

Raveau, S., Munoz, J. C., Prato, C. G., Soto A., Tamblay, S. & Iglesias, P., 2017. A behavioural planning tool for modelling public transport systems. *Submitted to TransitData2017, Santiago, Chile.*

SECTRA, 2015. *Informe Final del estudio "Encuesta Origen Destino de Viajes 2012"*. Secretaría Interministerial de Planificación de Transporte, Santiago.

SII, 2014. *Roles de la Región Metropolitana de la Segunda Serie No Agrícola y Superficie Construida.* Datos obtenidos en virtud de la Ley Nro. 20.285. Servicio de Impuestos Internos, Santiago.

Tamblay, S., Galilea, P., & Muñoz, J.C., 2015. Estimation of a Zonal Origin-Destination Matrix from Observed Public Transport Trips for Santiago de Chile. *Actas del XVII Congreso Chileno de Ingeniería de Transporte. Concepción, Chile.*

Tamblay, S., Galilea, P., Iglesias, P., Raveau, S. & Muñoz, J.C., 2016. A zonal inference model based on observed smart-card transactions for Santiago de Chile. *Transportation Research Part A: Policy and Practice*, *84*, 44-54.