THE BOOTSTRAPPING APPROACH FOR INFERRING CONFIDENT FREIGHT TRANSPORT MATRICES

F.G. Benitez, L. Romero, N. Caceres, J.M. del Castillo

Transportation Engineering, Faculty of Engineering
University of Seville
Camino de los Descubrimientos, s/n, Seville 41092, Spain
E-mail corresponding author: benitez@esi.us.es

Keywords: Origin-destination matrix, updating, adjusting, freight, survey, transport, matrix estimation, bootstrap.

Brief professional biography of each author

Francisco G. Benitez was born in Seville, Spain, in 1956. He received the Ph.D. degree in Industrial Engineering from the Technical University of Madrid, in 1981. He was a Research Fellow at the Department of Engineering Science, University of Oxford in 1981-83, and a Fulbright Scholar and Visiting Associate at the California Institute of Technology 1984-1986. He is currently a Professor with the Department of Transportation Engineering, University of Seville, and Head of Transportation Engineering and Infrastructure Division. His research topics are transportation modelling, transmissions, GIS, numerical methods, ITS.

Email: benitez@esi.us.es

Luis M. Romero was born in Badajoz, Spain, in 1971. He received the Ph.D. degree in Industrial Engineering from the University of Seville, in 2007. He is currently a Senior Researcher with the Department of Transportation Engineering, University of Seville. His research topics are transport demand modelling, traffic flow, numerical methods, ITS and software design.

Email: l_m_romero@esi.us.es

Noelia Caceres was born in Don Benito, Spain, in 1980. She received the Ph.D. degree in Telecommunication Engineering from the University of Seville, in 2010. She is currently a Senior Researcher with the Department of Transportation Engineering, University of Seville. Her research interests include transport demand modelling, traffic flow, software design and ITS.

Email: noeliacs@esi.us.es

Jose M. del Castillo was born in Seville, Spain, in 1965. He received the Ph.D. degree in Industrial Engineering from the University of Seville, in 1994. He was a Postdoctoral Researcher at the Institute of Transportation Studies at the University of California-Berkeley, in 1995. He is currently a Professor with the Department of Transportation Engineering, University of Seville. His research topics are applied statistics, transportation modelling, heuristic logistics and traffic flow.

Email: delcastillo@us.es

Abstract: Transport studies require, as a preliminary step, conducting a survey process to a sample of the universe of users of the transportation system. The statistical reliability of the data determines the goodness of the results and conclusions that can be inferred from the analyses and models generated. In this communication a methodology, based on the techniques of "bootstrapping", to the robust statistical estimate of freight transport matrices is presented; this allows to generate the confidence intervals of travel between origin-destination pairs defined by each cell of the OD matrix derived from a freight transport survey.

This result is of interest in defining the dimensions of certainty for matrix cells and subsequent adjustment by techniques based on aggregate data (i.e. traffic counts, cordon line matrices, paths, etc.).

The techniques of "bootstrapping" originated in the 70's, although widely used during the 90's, have not been fully exploited in the field of freight transport studies. To address this study a data set from a statistically reliable freight transport study conducted in Spain at the level of multi-province regions has been used.

1. INTRODUCTION

Origin-destination (OD) trip tables are required in most transportation applications to represent the spatial distribution of transport demand. The procedures to construct these tables are mainly based on available information collected by a transport survey. The level of the comprehensiveness and quality of the survey determines the confidence and reliability of the data captured. Incomplete and/or inaccurate data have negative consequences in characterising transport mobility and will invalidate subsequent stages (i.e. modelling, estimations, forecasting). As a complement to survey-based data-capturing techniques, other pieces of information, that might be easily available, quick or inexpensive, can help to improve the reliability of the eventually inferred OD trip table (i.e. link volumes, trips between macro-zones, cordons and screen-line counts, vehicle speeds, path travel times, path flows). To assess the quality of OD trip table estimates versus survey-captured tables, a large amount of statistical measures can be used to quantify the accuracy of the data observed (that is, of the pieces of information available).

The construction of freight transport matrices of a given region to be analysed feeds on the data collected in a process of surveying a sample of agents (users) of the transportation system. There are several techniques to perform freight data collection, of which the most commonly used can be classified into two families, based on the disaggregation level of the agents: a)

Individual agent level. In this case a sample of companies (prone to receive/ship goods) is chosen from the whole economic frame. This sample must be statistically representative of the economic distribution functions, which depend on many variables. b) Specific economic sector level. A sample is chosen from among the sector universe in the region. The sample must also be chosen to be statistically representative of the sector distribution according to variables associated with the item. Obviously, the level of aggregation of the sector variables affects the explanatory power of the collected data in relation to reality. This case is broadly used for the specific sector of transportation agents (i.e. freight transport companies and registered freight vehicles), though the data captured are limited (mostly origin-destination, product and load).

Of these techniques, one of the most widely used is based on surveying samples of registered freight vehicles distributed according to their registration plates. Once the studied region is discretised into transport areas by aggregating census districts, municipalities or counties, the sample size proves to be a function of the total number of vehicles distributed among the zones and according to the registered population; this ensures the high statistical reliability of information collected on a zonal level.

By this sampling technique, and for each zone, the number of freight vehicles registered therein, the vehicle type histogram can be easily obtained. The choice of the vehicle types to be surveyed is made through a process of random draws without replacement from the universe in each area, so that it reproduces the histogram. From the practical and professional standpoint, the sample and the universe generally are related through sampling rate (weight) coefficients. The weighting process (expansion) does not guarantee that the expanded data follow the same patterns as reality and the "representativeness" of the expanded data matrix, in relation to the real unknown matrix, is questionable. For a more precise characterization of the expanded matrix there are numerous techniques to refine this “representativeness”, of which confidence intervals are the most practical.

This piece of work describes a model that estimates the level of confidence of data captured for each OD pair and can be easily extended to its aggregated magnitudes by origin and destination. This objective is addressed by using the statistical technique of bootstrapping to evaluate the uncertainties in each OD pair estimate, which is used to infer confidence intervals for OD matrices retrieved from transport surveys. Preliminary results are obtained from applying the developed methodology to a selected case of freight transport at a national scale in Spain.

This paper is organised as follows: Next section justifies the interest of confidence intervals for the definition of constraints that should be verified during the adjustment of the OD matrix; it introduces a concise state of the art in the derivation of confidence intervals for each OD trip matrix cell, and a review of analytical methods and empirical techniques devoted to replicated bootstrap and its implementation for the inference of confidence intervals is also included. The case study section shows the results derived from an actual practical application; this allows a glimpse of the interest of the methodology presented. The final section ends up with major conclusions and further research lines to be followed.

2. CONFIDENCE INTERVALS FOR OD MATRICES

2.1 Problem definition

For a given study area divided into transport zones where agents can travel from each origin (ranging from 1 to n_o) to all destinations (from 1 to n_d), $\Upsilon = \left[\Upsilon_{ij} \right]$ denotes the OD trip matrix, where Υ_{ij} stands for the number of freight vehicle trips from origin zone *i* to destination zone *j*, and $\Upsilon = \sum_{i=1}^{n_o} \sum_{j=1}^{n_d} \Upsilon_{ij}$ the total number of trips within the study region. Obtaining matrix Υ requires the observation of all trips made in the area, by both the freight vehicle registered population and passers-by; this is an impossible task to tackle. Instead, a surveying process can be accomplished a number of times *E*, on samples taken from the population of vehicles from transport system which travel in the area, providing a series of matrices $\mathbf{T}^1, \mathbf{T}^2, ..., \mathbf{T}^E$ which represent a stochastic series where the total number of trips T^e is distributed among the $n_o \times n_d$ cells ($C = n_o \cdot n_d$ categories) according to a multinomial probability distribution of parameters $\boldsymbol{\pi} = \left[\pi_{ij} \right]$:

$$\mathrm{P}\left[T_{11} = T_{11}^e, ..., T_{n_o n_d} = T_{n_o n_d}^e \left| T^e, \pi_{11}, ..., \pi_{n_o n_d} \right. \right] = T^e !(\pi_{11})^{T_{11}^e} \cdot ... \cdot (\pi_{n_o n_d})^{T_{n_o n_d}^e} / T_{11}^e ! ... \cdot T_{n_o n_d}^e ! \quad (1)$$

where π_{ij} stands for the probability of detecting T_{ij}^e trips in pair *i-j*, and where $\sum_{i=1}^{n_0} \sum_{j=1}^{n_d} T_{ij}^e = T^e$, and $\sum_{i=1}^{n_0} \sum_{j=1}^{n_d} \pi_{ij} = 1$.

For a sufficiently high number *E* of samples, T^e may be approximated by a normal distribution. This approach is of low interest because of the impracticability and budget restrictions on conducting multiple repeated studies to obtain more than just one matrix. Instead, one can accept the

hypothesis that a single array $\mathbf{T} \equiv \mathbf{T}^1$, with a total travel $T \equiv T^1$, statistically characterizes the said series.

The generation of a large number of samples $\left\{\hat{\mathbf{T}}^m, \forall m = 1,...,M\right\}$, replicated by random samples from matrix $\mathbf{T}$, allows to estimate the parameters of the distribution (1) as:

$$\left\{ \hat{\pi}_{ij} = \frac{E\left[\hat{T}_{ij}^m,\ m = 1,...,M\right]}{\hat{T}^1 \equiv \hat{T}^2 \equiv ... \equiv \hat{T}^M} = \frac{T_{ij}^1}{T^1} \equiv \frac{T_{ij}}{T} \equiv p_{ij}^1 \equiv p_{ij},\ i = 1,...,n_0;\ j = 1,...,n_d \right\} \quad (2)$$

accepting T^1 and p_{ij}^1 as unbiased estimates of mean μ_T of the total number of trips and the probabilities of the number of cell trips (maximum likelihood estimator), respectively. Under these assumptions, expression (1) is particularised as: $\mathrm{P}\left[\mathbf{T}^* = \mathbf{T} \middle| T, \boldsymbol{p}\right] \equiv \mathrm{P}\left[T_{11}^* = T_{11},...,T_{n_o n_d}^* = T_{n_o n_d} \middle| T\ , p_{11},..., p_{n_o n_d}\right]$, which stands for the probability distribution function of all possible matrices $\mathbf{T}^*$ with parameters T and $\hat{\boldsymbol{\pi}} = \left\{p_{ij}\right\}$.

2.2 Analytical confidence intervals

When performing a statistical inference from a sample, the reliability of this has a decisive influence. Although there are several indexes to quantify this reliability, the confidence interval is the most widely used and accepted methodology. If *s* represents the parameter of interest, its classical confidence interval is defined as $P(s_l < s < s_u) = 1 - \alpha$ (replacing the equal sign in inequality $\geq$ in case of discrete variables), where $(s_l,\ s_u)$ represents the range within which the true value of *s* can be found with a probability of $(1-\alpha)100\ \%$.

In case of a matrix **T**, the confidence intervals are given by either $(L_{ij} \leq T_{ij} \leq U_{ij})$ or $(p_{ij}^l \leq p_{ij} \leq p_{ij}^u)$, where p_{ij} stands for trip proportion $(p_{ij} = \frac{T_{ij}}{T = \sum_{ij} T_{ij}})$.

There are other techniques, such as the hypothesis test, to perform statistical inference based on statistical distributions; but as a general rule, confidence intervals are more informative and preferred than hypothesis tests when both are available (Burdick and Graybill,1992).

For certain distributions, the expressions of the confidence intervals are well defined at analytical or numerical level. In case of the multinomial distribution there are different methods proposed in the literature, mainly

depending on the desired confidence level, the length of the interval, or a combination of both identified by the confidence index, the size of the sample and the matrix covariance of the probabilities. All these methods are grouped into two large families: a) analytical ones, based on approximate approaches, b) empirical methods, based on successive extractions.

2.3 Empirical confidence intervals

Bootstrap is a technique of replicating samples by extraction, presented in 1979 (Efron, 1979; Efron and Tibshirani, 1993), used to estimate a distribution from which to extract several parameters of interest (i.e, mean, variance). The assumptions made by this technique are minimal and limited to the distribution, followed by the estimator of the draws, and reliably reflect the properties of the estimator of the starting sample.

This technique involves random draws, with replacement, of subsets from the input data. The extractions are performed in such a way that each data item is represented identically in the random extraction scheme. Its characteristics differ from the Monte-Carlo method in connection with the sampling process. There are other variations of randomised replicating, such as the jackknife method, but analyses carried out up to day do not support the superiority of one over the other (Severiano et al., 2011).

With the aim of simulating a process of replicating trip matrices, a random number *m* of matrix samples $\mathbf{T}^*$ with n_o rows and n_d columns are extracted. The sum of cell elements T^* coincides with the total number of trips *T* of the starting data matrix. Each replicate sample $\mathbf{T}^* = \left[T_{ij}^*, i = 1,...,n_0; j = 1,...,n_d \right]$ is obtained in T random draws, with replacement, from the original data set $\mathbf{T} = \left[T_{ij}, i = 1,...,n_0; j = 1,...,n_d \right]$. To obtain the bootstrap confidence interval, for each pairwise cell of the *m* extractions, the percentile method for an intended coverage of $1-2\alpha$ is obtained directly from the distribution percentiles α and $1-\alpha$. Therefore, to obtain the 95% confidence interval lower and upper limits, the $0.025 \cdot m$ and $0.925 \cdot m$ values are computed from the bootstrap ordered indexes, as *m* extractions are available. Using multiple extractions, following Efron's bootstrap technique, a generic empirical statistics parameter estimator $\hat{\theta}$ of a statistics parameter θ, and confidence interval for θ can be constructed as summarised in the following pseudo-algorithms:

- *Estimate of statistics parameter* $\hat{\theta}$:
  - For the initial data set $(T_{11},...,T_{n_o n_d})$, estimate the multinomial proportions, from (2), and assume the hypothesis that these ratios correspond to the "true" population proportions.

– Generate *M* samples $\mathbf{T}^{*m} = \left[T_{ij}^{*m}, i = 1,...,n_0; j = 1,...,n_d \right]$ of size $N \equiv T = \sum_{i=1}^{n_o} \sum_{j=1}^{n_d} T_{ij}$ from the multinomial distribution of parameters $\hat{\boldsymbol{\pi}}$.

– Estimate the parameter set $\hat{\theta}$ from the M drawn samples related to each *ij*-th matrix cell:

$$\hat{\theta}^* = \left\{ \hat{\theta}_{ij}^*, i = 1,2,...,n_o; j = 1,2,...,n_d \right\} = \left\{ \left(\hat{\theta}_{ij}^m, m = 1,2,...,M \right), i = 1,2,...,n_o; j = 1,2,...,n_d \right\}$$

- *Estimate of cell standard error and mean:*

$$\hat{\sigma}_{ij} = \left\{ \sum_{m=1}^{M} \left[\hat{\theta}_{ij}^m - \bar{\theta}_{ij}^m \right] / (M-1) \right\}^{1/2}, \ \bar{\theta}_{ij} = \left(\sum_{m=1}^{M} \hat{\theta}_{ij}^m \right) / M.$$

- *Construction of a confidence interval for parameter θ_{ij} based on bootstrap percentiles:*

  – For each *ij*-th matrix cell, with all *M* bootstrap samples, histograms are constructed from $\hat{\theta}_{ij}^m$.

  – Compute percentiles $\hat{\theta}_{ij}^{\alpha/2} = \hat{F}_{ij}^{-1}(\alpha/2)$ and $\hat{\theta}_{ij}^{1-\alpha/2} = \hat{F}_{ij}^{-1}(1-\alpha/2)$, where $\hat{F}_{ij}(\hat{\theta}_{ij})$ is the empirical distribution.

  – Compute confidence intervals directly from the percentiles of the empirical distribution $\hat{F}_{ij}(\hat{\theta}_{ij}) : \left[\hat{F}_{ij}^{-1}(\alpha/2), \hat{F}_{ij}^{-1}(1-\alpha/2) \right]$, where $\hat{F}_{ij}^{-1}(\cdot)$ stands for the percentile of the bootstrap empirical distribution constructed by sorting the bootstrap estimators in ascending order.

2.4 OD matrix estimation approaches

The O-D matrix is the keystone piece of information fundamental input to most transportation systems analysis methods. This matrix evinces the volume of traffic between all origins and destinations in the transportation network. The O-D matrix is difficult and often costly to obtain by direct methods such as carrying a home-based survey; consequently, indirect or synthetic techniques that seek to infer this matrix based on indirect measures such as license plate surveys (Van der Zijpp, 1996), automatic vehicle identification (AVI) systems (Dixon and Rilett, 2000; Kwon and Varaiya, 2005) and cell phones (Caceres et al., 2011) are widely used.

The problem of OD inference, estimation and prediction has been dealt with during the last two and a half decades (Cascetta, 1984; Ben-Akiva, 1987; Cascetta et al.1993). In most of the published literature, OD estimation is based on historical demand information provided by a prior matrix and

additional information such as link count data and other more recent traffic surveillance technologies. The objective of this problem is simulating an OD matrix close to a prior or possibly outdated matrix and which, when assigned to the network model of the transport system, reproduces the observed magnitudes with a controlled error. Beside the hypothesis assumed and the approaches followed, there are factors that make it hard to be certain of the quality and reliability of the OD matrix estimated. To obtain a complete OD matrix by direct measurements describing the transport demand within a given region is an unfeasible task because of budget, manpower and time limitations. Therefore, OD matrices have customarily been estimated using different methodologies. The alternative most used over the past twenty-five years and with the largest amount of documented work in the literature is a mixed analytical-empirical method which uses traffic counts as measurements of link flows in a network model in order to adjust an existing matrix derived from a survey. The prior matrix can be regarded as an observation (a good approximation) of the “true” OD matrix to be estimated. In methods based on this approach, the prior OD matrix is iteratively “adjusted” or “changed” to reproduce the observed traffic counts when assigned to the transportation network. The most widespread adjustment methodology is based on obtaining trip matrices, expressed in equivalent vehicles, that replicate as closely as possible the volume observed when matrices are assigned to a reliable transport network model by an assignment code.

Estimating the unknown OD matrix using a limited observed/measured sample data from the traffic system is generally an underspecified problem; the number of OD unknown variables to be estimated is usually greater than the number of observations from the system. Therefore a quite large number of feasible solutions can be obtained for the OD matrix estimate problem. In consequence, additional pieces of information have to be incorporated to draw a unique solution. Supplementary hypothesis have to be set such as a metric relating observed and modelled magnitudes such as (i) measured link volumes, (ii) travel times, (iii) speeds, (iv) trajectories and path choices, (v) either full or partial prior OD matrices, among others. In summary, the OD trip matrix estimation goal is to infer the closest OD matrix to a prior matrix, such that when loaded to the transportation network model reproduces the observed measured data as closely as possible. Numerous metrics have been proposed in the literature: (i) Euclidean and non-euclidean least squares, (ii) maximum entropy (see Kapur, 1989 for a comprehensive review), (iii) stochastic methods, (iv) heuristic and metaheuristic methods, among other mixed approaches. As a consequence wide variations in the OD estimates are confronted.

In general one can affirm that the different methods of estimating OD trip matrices based on traffic counts, developed in the literature, have the following generic form (Yang at al.1992):

$$\begin{aligned} \underset{\mathbf{v},\,\mathbf{T}}{\text{Minimize}} \quad & \alpha F_1\left(\mathbf{T}, \bar{\mathbf{T}}\right) + \beta F_2\left(\mathbf{v}, \bar{\mathbf{v}}\right) \\ s.t. \quad & \mathbf{v} = Assign(\mathbf{T}) \\ & \alpha + \beta = 1 \\ & 0 \le (\alpha,\ \beta) \end{aligned} \tag{3}$$

where functions F_1 and F_2 are two metrics that measure distance between the estimated OD matrix $\mathbf{T}$, and the prior matrix, $\bar{\mathbf{T}}$, and between the estimated and the observed volumes in network links, v and $\bar{\mathbf{v}}$ respectively.

The proposed formulation follows the basics of scheme (3); however, to control the distortion of the prior matrix a set of bounded variable constraints (for each matrix cell) are prescribed. This manner of proceeding is intended to keep the variation of the information contained in the adjusted matrix compared to the prior matrix within a range considered to be feasible. Regarding the adjustment problem, the necessary volume data are inferred from data collected on traffic counts on certain links. The formulation proposed to adjust the prior OD matrix includes the Euclidean distance between estimated and observed volume data and the distance between the prior and estimated matrices; in addition, a set of variable bounds and functional constraints which define admissible ranges for individual OD pairs, zone productions and attractions, and total number of trips are included. These bounds are defined by the confidence intervals inferred by the bootstrap technique. Then a modified mathematical formulation from (3) results in the programming approach proposed in this investigation by incorporating the following constraints, as follows:

$$L_{ij} \le T_{ij} \le U_{ij}; \quad L_i^O \le \sum_{j \in D} T_{ij} \le U_i^O; \quad L_j^D \le \sum_{i \in O} T_{ij} \le U_j^D \tag{4}$$

where the necessary mathematical conventions to formulate the new OD matrix adjustment approach are summarised: $i \in$ O: origin zones (n_o); $j \in$ D : destination zones (n_d); U_{ij}, L_{ij} : upper and lower bounds for *(i, j)* OD pair; U_i^O, L_i^O: upper and lower bounds for trips generated by zone *i*; U_j^D, L_j^D: upper and lower bounds for trips attracted by zone *j*; $\bar{\mathbf{v}}$: observed travel demand through links; α, β: weights factor associated with the volume on links and OD matrix cells, respectively; v: volume on links; T_{ij}: inter-province travel demand (trips) from origin *i* to destination *j*. In addition to the above

dimensions established to control the distortion of the information contained in the matrices, and in order to preserve the basic structure of such information, one can set a series of maximum increments and decrements for those pairs of the prior matrix where no information is available (Doblas and Benitez, 2005; Caceres et al., 2011).

Then, a modified mathematical formulation from (3) results in the bi-level programming approach proposed in this investigation, formulated as follows:

$$
\begin{array}{ll}
\underline{\text{Upper Level}} & \underline{\text{Lower Level}} \\
\underset{T_{ij}}{\text{Min}} \quad \alpha F_1\left(\mathbf{T}, \bar{\mathbf{T}}\right) + \beta F_2\left(\mathbf{v}, \bar{\mathbf{v}}\right) & \underset{v_a}{\text{Min}} \quad \sum_{a \in A} \int_0^{v_a} s_a(v)\, dv \\
s.t. \quad \mathbf{v} = Assign(\mathbf{T}) & s.t.\, v_a = \sum_{i \in I} \sum_{j \in J} \sum_{k \in K_{ij}} \delta_{ak}\, h_k, \;\; \forall\, a \in A \\
\alpha + \beta = 1 & \sum_{k \in K_{ij}} h_k = T_{ij}, \qquad \forall\, i \in O, j \in D \\
0 \le (\alpha,\ \beta) \le 1 & h_k \ge 0 \forall\, k \in K_{ij}, i \in O, j \in D \\
L_{ij} \le T_{ij} \le U_{ij} \qquad \forall\, i \in O, j \in D & \\
L_i^O \le \sum_{j \in D} T_{ij} \le U_i^O \qquad \forall\, i \in O & \\
L_j^D \le \sum_{i \in O} T_{ij} \le U_j^D \qquad \forall\, j \in D & \\
L^R \le \sum_{i \in R_O} \sum_{j \in R_D} T_{ij} \le U^R \quad \forall\, i \in R_O,\, j \in R_D &
\end{array}
\tag{4}
$$

where the necessary mathematical conventions, to formulate the new OD matrix adjustment bi-level approach, are summarised.

Indices and sets

$i \in O$: origin zones (n_o); $j \in D$: destination zones (n_d); $a \in A$: network links; $k \in K_{ij}$: routes or paths from origin *i* to destination *j.*

Constants

δ_{ak}: 1 if link *a* belongs to path *k*, 0 otherwise; U_{ij}, L_{ij}: upper and lower bounds for (*i, j*) OD pair; U_i^O, L_i^O: upper and lower bounds for trips generated by zone *i*; U_j^D, L_j^D: upper and lower bounds for trips attracted by zone *j*; U^R, L^R:upper and lower bounds for total network trips; $\bar{\mathbf{v}} = \{\bar{v}_a,\ \forall a \in A\}$: observed travel demand through links $a \in A$ (*observed volume*); α, β: weights factor associated with the volume on links and OD matrix cells, respectively.

Functions

$s_a(v_a)$: performance (volume-delay or cost) function of link $a \in A$.

Variables

$\mathbf{v} = \{v_a,\ \forall a \in A\}$: volume on link *a*; h_k: flow on path *k*; T_{ij}: inter-province travel demand (trips) from origin *i* to destination *j,* (note that T_{ij} is variable for the global OD adjustment process, but constant for every assignment stage).

In addition, $\sum_{i \in R_O} \sum_{j \in R_D} T_{ij}$ stands for inter-macrozonal trips between pairs *i-j*, where origin *i* and destination *j* belong to macrozones *R*o and *Rd*, respectively; similarly $\bar{T}_{ij}$ represents the same quantity referred to the prior matrix $\bar{\mathbf{T}}$. As a general notation, bounds L_{ij} and U_{ij} (both with and without upper indexes) are identified with the endpoints of the uncertainty intervals inferred in formulation (4).

The lower level program stated in (4), known as Beckmann's transformation, is the basic model for obtaining those volumes v_a on all network links satisfying the *user-equilibrium* conditions for a given fixed demand T_{ij} (Sheffi, 1985).

3. CASE STUDY

A real case study has been carried out to demonstrate the application of the methodology and the importance of incorporating confidence interval information in mobility OD matrices.

As a first stage, starting from the origin-destination matrix (prior non-elevated matrix) retrieved from the non-elevated data provided by a transport survey, a bootstrap generating program estimates confidence intervals for each origin-destination matrix cell. This outcome defines the intervals where cell trips are allowed to fluctuate under a similar confidence level.

The second stage adjusts the prior matrix under a bi-level optimization scheme. The macroscopic assignment arrangement uses a commercial network tool to derive traffic flow on links of the modelled transport network. The upper level is an optimization scheme, which minimizes the deviation between modelled and measured traffic flows (all vehicles and trucks) on selected links. The information provided by the confidence intervals is incorporated as constraints in the optimization scheme.

The case analysed is the Spain Road Freight National Survey EPTMC (Fomento, 2008), on a sample captured of a continuous basis during 52/53 weeks every year. The study population consists of heavy goods vehicles registered in Spain, authorised to transport goods by road, with operations in the territory and abroad. The observation unit is vehicle-week (i.e. transport operations performed by selected vehicles during one week). This includes all operations that start in the reference week, although they may finish

afterwards. Data captured provide information on the characteristics of the vehicle, goods transported, origin, destination and distance of the operation. Transport operations relate to the movement of goods, which do not necessarily coincide with the movement of vehicles. Goods transported are grouped into ten classes; 0: agricultural products and live animals, 1: food and fodder, 2: solid mineral fuels, 3: petroleum products, 4: minerals and waste to recast, 5: iron products, 6: mineral raw or manufactured and construction materials, 7: fertilizers, 8: chemicals, 9: machinery, vehicles, manufactured objects and special transactions. Goods transported are quantified in gross tons (goods, packaging and container). The raw data of the survey present information at the origins and destinations at the province level, and are statistically representative at the regional level, but not significant at province level. The raw province disaggregated level is used for the application of the techniques presented hereinafter.

The sample design is based on a stratified random sampling with vehicle-week as the sampling unit. Samples were selected independently for each week of the year, at the rate of 1,000 vehicles per week, stratified by type of service (public / private) and type of vehicle. The selection of sampling units in each stratum is performed using a systematic sampling with random start upon the vehicle registration regional record. To expand the captured data, a stratified expansion estimator is used to correct incidences during the survey. The estimates are calculated in each stratum, yielding the total population as the sum of the estimates of each of them. The response rate for 2008 was 71.7%. The valid sample size surveyed amounts to 37,305 vehicles. The number of valid sample transport operations is 529,229, disaggregated into a) intra-municipal: 168,291, b) intra-regional: 302,825, c) inter-regional: 50,104, and d) international: 8,009.

3.1 OD matrix confidence intervals estimation

The simulations carried out comply with the empirical procedure introduced in section 3.3. The computer program was coded in Matlab. The simulated multinomial sample replication was generated by the subroutine MNRND. All simulation studies were performed on a 12 core Intel Xeon E5645 personal computer using parallel computing. To provide a reliable confidence interval, a large sample size is desirable. In this case a size of 10,000 bootstrap samples was used. These simulations consist of the following steps:

i. For the initial data set estimate the multinomial proportions p_{ij} and assume the hypothesis these ratios correspond to the “true” population proportions.

ii. Extract 10,000 multinomial samples from the survey matrix.

iii. Obtain confidence intervals for each cell sample on the 95% level, based on the drawn subset corresponding to each cell.

iv. Assess the average length of full, left and right halves of intervals as the mean of the difference between the upper and lower limits of each interval $(U_{ij} - L_{ij})$, the difference between the mean value and the lower limit $(T_{ij} - L_{ij})$, and the difference of the upper value and the mean value $(U_{ij} - T_{ij})$, respectively.

v. Weight (expand) each cell confidence interval according to the cell sampling rate.

Confidence interval lengths inferred versus trip nominal values for all OD matrix cells are depicted in Figure 1(a). The solid curve is the regression curve, obtained by a least-squares fit, with expression $U_{ij} - L_{ij} = e^a \cdot \bar{T}_{ij}^b$ where parameters $a = 0.2083,\ b = 0.4825$ with a t-statistics of 29 and 770 respectively.

The coefficient of determination of this adjustment, $Adj.R^2 = 0.996$, is sufficiently high to ensure the goodness of fit.

Figure 1(b) reflects the histogram of confidence interval lengths for OD cell trips. It is easy to notice the large number of null trip cells, a recursive behavior in most transport survey studies.

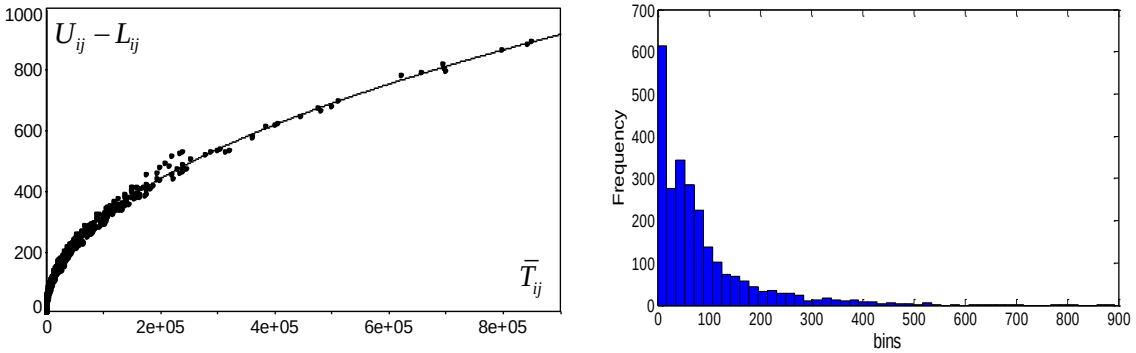


Fig. 1. Confidence interval length $U_{ij} - L_{ij}$ a) versus cell trips b) histogram.

3.2 Adjusting mobility matrices

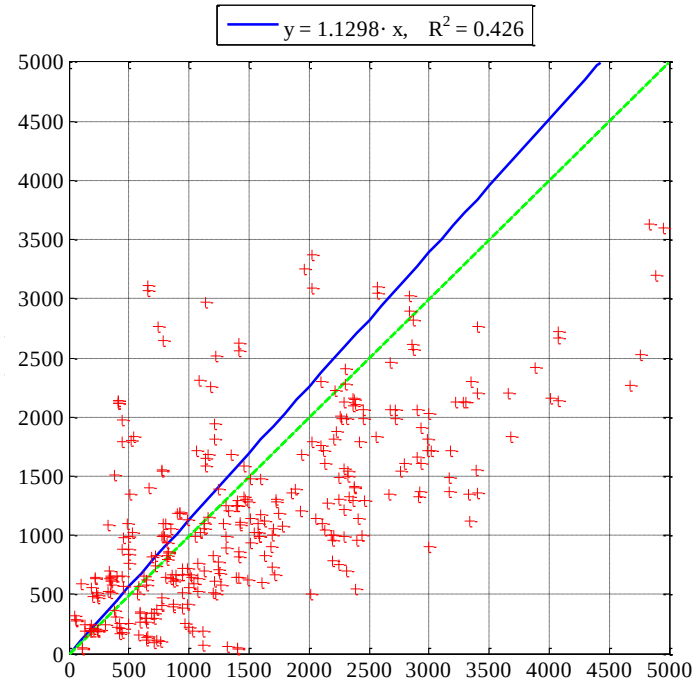
Figure 2 shows a summary of the results achieved in the assignment process of the prior matrix $\bar{\mathbf{T}} \equiv \mathbf{T}^1$ and the adjusted one $\mathbf{T}$ using the network model. In the case of the prior matrix the determination coefficient between observed and modelled volumes is $R^2 = 0.426$ (Figure 2a), while the assignment of the

adjusted matrix gives rise to a value of 0.604 for the same coefficient, (Figure 2b). The assessment of the methodology in terms of distortion of the information contained in the adjusted matrix in relation to the prior one, provides a high correlation value due to the bound constraints imposed (Figure 3). The determination coefficient between both matrices are $R^2 = 0.949$ for the inter-province case and $R^2 = 0.981$ for the inter-regional one.

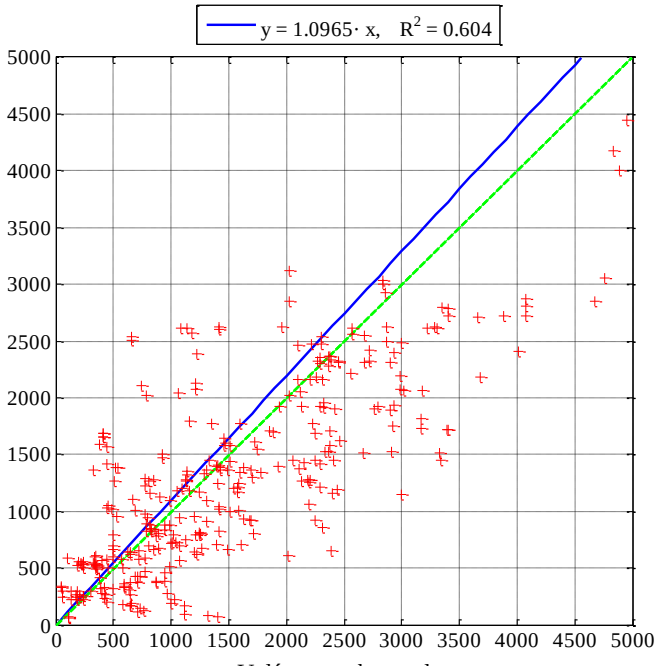
The solid straight lines are the linear regression lines, obtained by a least-squares fit..

The control in the OD estimation, containing the level of distortion between both prior and undated matrices, utilising the information incorporated by the cell confidence interval, guarantees reliability and brings a certain degree of soundness to the final results regarding the OD matrix obtained.

It is trivial and stated (Doblas and Benitez, 2005) that relaxing the constraints derived from the cell confidence intervals would both (i) increase the determination coefficient between observed and modelled volumes (unconstraint optimization yields better optimum values of the objective function than constraint optimization) and (ii) would deteriorate the correlation between prior and adjusted OD matrices; therefore a comparison in this terms does not offer valuable information worth to be analysed.

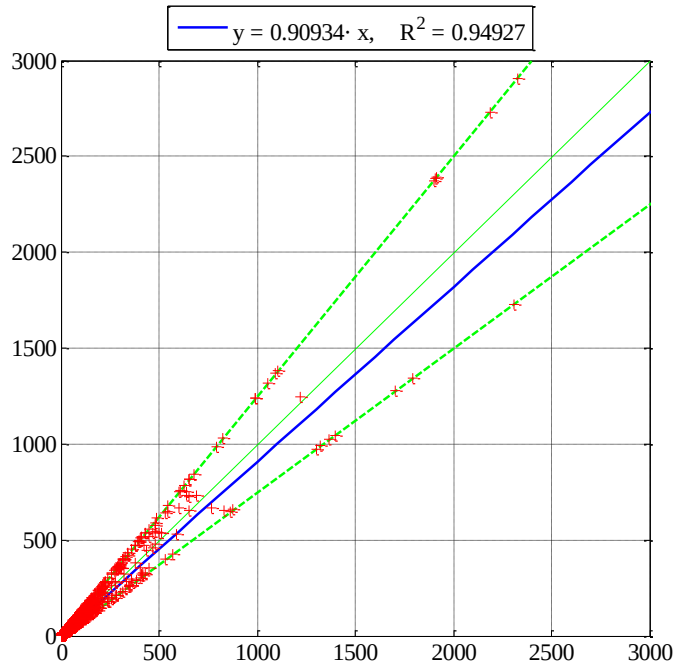


(a)

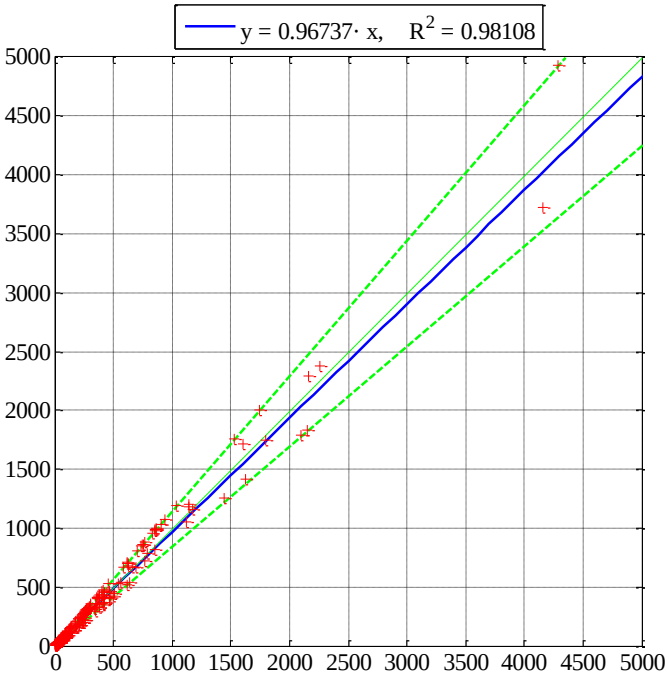


(b)

FIGURE 2 Relationship between measured (x-axis) and modelled volumes (y-axis) (in vehicles) using the (a) prior matrix or (b) the adjusted matrix.



(a)



(b)

FIGURE 3 Correlation between prior (x-axis) and adjusted (y-axis) OD matrices (in vehicles) using (a) inter-province matrix or (b) inter-regional matrix.

4. CONCLUSIONS

A general methodology for the development, treatment and incorporation of additional information sources to the problem of OD matrix estimation, based on the definition of confidence intervals for the trip matrix cells, is presented.

This approach is based on the definition of confidence intervals for the matrix cells extracted by a travel survey. The approach has been applied to the real case of the wide annual inter-province freight transport in Spain.

The experimental validation of the proposed models has shown evidence that the bootstrap technique is an alternative that may be considered for the determination of confidence intervals of the volume of trips between OD pairs. This allows defining an acceptable measure of the magnitudes to be imposed in the process of adjusting OD matrices. The consequences of this finding are significant, particularly for the generation of OD matrices that conform to that collected by a survey, diminishing the level of uncertainty involved in this extremely underspecified problem. To ensure the widespread professional application of this technique it will be necessary to further perform validation on large scale real cases in order to outline the degree of robustness, efficiency and numerical stability of outcomes.

ACKNOWLEDGEMENTS

The authors would like to thank the FEDER of European Union for financial support via project “G-GI3000IDII” of the “FEDER Operational Programme for Andalusia 2007-2013", and the Spanish Ministry of Economy and Science for the partial subsidy granted under the national R&D program, Project No. TRA2012-36930.

The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein, and do not necessarily reflect the official views or policy of the Ministry of Public Works of Spain (Fomento, 2008), owner of the data employed. The kind help provided by the Deputy General Directorate of Economic Studies and Statistics is kindly acknowledged.

REFERENCES

BEN-AKIVA, M. (1987), Methods to combine different data sources and estimate origin-destination matrices. In *Transportation and Traffic Theory*, Gartner, N. and N. Wilson (Eds). Elsevier Science Publishing, 459–481.

BURDICK, R.K. and GRAYBILL F.A. (1992), Confidence Intervals on Variance Components. New York: Marcel Dekker.

CACERES, N., ROMERO, L.M. and BENITEZ, F.G. (2011), Inferring origin–destination trip matrices from aggregate volumes on groups of links: a case study using volumes inferred from mobile phone data. Journal of Advanced Transportation, DOI:101002/1tr.187.

CASCETTA, E. (1984), Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. Transportation Research, 18B (4/5), 288–299.

CASCETTA, E., INAUDI, D. and MARQUIS, G. (1993), Dynamic estimators of origin-destination matrices using traffic counts. Transportation Science, 27(4), 363–373.

DIXON, M.P. and RILETT L.R. (2000), Real-time origin-destination estimation using automatic vehicle identification data. Proc. 79th Transportation Research Board Meeting, Washington, D.C.

DOBLAS, J. and BENITEZ F.G. (2005), An approach for estimating and updating origin-destination matrices based on traffic counts preserving prior structure. Transportation Research, B 39, 565-591.

EFRON, B. (1979), Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1–26.

EFRON, B. and TIBSHIRANI R.J. (1993), An introduction to the bootstrap. Boca Raton: Chapman & Hall/CRC.

FOMENTO, M. (2008), Encuesta de movilidad de las personas residentes en España. Movilia 2006/2007. Dirección General de Programación Económica, Ministerio de Fomento. http://www.fomento.gob.es.

KAPUR, J.N. (1989), Maximum-entropy models in science and engineering. John Wiley & Sons, New Delhi.

KWON, J. and VARAIYA, P. (2005), Real-time estimation of O-D matrices with partial trajectories from electronic toll collection tag data. In Transportation Research Record: Journal of the Transportation Research Board, No. 1923, 119–126.

SEVERIANO, A., CARRICO J.A., Robinson D.A., Ramirez M. and Pinto F.R.(2011), Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. Plos One, Vol. 6(5), e19539. doi:10.1371.

SHEFFI, Y. (1985), Urban transportation networks: equilibrium analysis with mathematical programming methods. Prentice-Hall, Inc.: Englewood Cliffs, NJ.

VAN DER ZIJPP, N. (1996) Dynamic origin-destination matrix estimation on motorway networks. PhD thesis, Transportation Planning and Traffic Engineering Subsection of the Faculty of Civil Engineering of Delft University of Technology.

YANG, H., SASAKI T., IIDA Y., and ASAKURA Y. (1992), Estimation of origin-destination matrices from link traffic counts on congested networks. Transport Research, 26(6), 417-434.