Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 111 (2014) 1155 – 1164

Procedia
Social and Behavioral Sciences

EWGT2013 – 16th Meeting of the EURO Working Group on Transportation

Estimating of bootstrap confidence intervals for freight transport matrices

F.G. Benitez*,a, L. M. Romero[b], N. Caceres[a], J.M. del Castillo[a]

[a]*Transportation Engineering, Faculty of Engineering, University of Sevilla, E-41092 Sevilla, Spain*
[b]*Transportation Engineering, AICIA, E-41092 Sevilla, Spain*

Abstract

Freight transport studies require, as a preliminary step, a survey to be conducted on a sample of the universe of agents, vehicles and/or companies of the transportation system. The statistical reliability of the data determines the goodness of the outcomes and conclusions that can be inferred from the analyses and models generated.

The methodology contained herein, based on bootstrapping techniques, allows us to generate the confidence intervals of origin-destination pairs defined by each cell of the matrix derived from a freight transport survey. To address this study a data set from a statistically reliable freight transport study conducted in Spain at the level of multi-province inter-regions has been used.

Selection and/or peer-review under responsibility of Scientific Committee

Keywords: Origin-destination matrix; freight transport survey; matrix estimation; bootstrapping.

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

* Corresponding author: Tel.: +34-95.448.7315; fax: +34-95.448.7316.
E-mail address: benitez@esi.us.es

1877-0428
Selection and/or peer-review under responsibility of Scientific Committee
doi:10.1016/j.sbspro.2014.01.150

consequences in characterizing 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 analyzed, OD matrices, 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 is chosen from the whole economic frame. This sample must be statistically representative of the economic distribution functions, which depend on many variables (all cited above) or some of them (i.e. selected samples as a function of the spatial distribution of economic density).

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 (i.e. those mentioned above particularized for every single individual sector member, or some aggregated more representative ones such as the number of sector members). 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 discretized 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 z, denoting by V_z 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. Elements of the original sample that fail, as a result of any cause external to the survey for instance, are replaced by other elements with similar characteristics (i.e. the same type) in order to preserve the sampling distribution. From the practical and professional standpoint, the sample and the universe generally are related through sampling rate (weight) coefficients. For the present case, they are defined by $k_z^t = V_z^t / v_z^t$, where V_z^t and v_z^t stand for the number of existing vehicle types and respondents of type *t* in zone z, respectively.

The weighting process (expansion) does not guarantee that the expanded data follow the same patterns as reality and, while an analysis to compare certain statistical parameters of certain variables (i.e. vehicle type distribution, or other socioeconomic variables) may be carried out, it is a fact that the expanded data are severely affected by significant errors that are difficult to characterize. Therefore the "*representativeness*" of the expanded data matrix, in relation to the real unknown matrix, is questionable (or at least limited).

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 paper 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. Section 2

discusses the basic concern on deriving confidence intervals for OD matrices retrieved from transport surveys, continuing with a review of seminal previous works. Section 3 describes preliminary results obtained from applying the developed methodology to a selected case of freight transport at a national scale in Spain.

2. Problem definition and formulation

2.1. Introduction

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 transport system vehicles which travel in the area, yielding a series of matrices $\mathbf{T}^1, \mathbf{T}^2, ..., \mathbf{T}^E$. These matrices represent a stochastic series in which 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 $\boldsymbol{\pi} = \left[\pi_{ij} \right]$:

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

where π_{ij} is 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. 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 $\hat{\mathbf{T}}^m, \forall m = 1, ..., M$, replicated by random samples from matrix $\mathbf{T}$, allows us 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 the 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 particularized 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 $\boldsymbol{T}$ and $\hat{\boldsymbol{\pi}} = \ p_{ij}$.

2.2. Analytical and empirical confidence intervals for OD matrices

The index most widely used to quantify the reliability of statistical inference from a sample is the confidence interval. For a matrix T, the confidence intervals are given by either $(L_{ij} \le T_{ij} \le U_{ij})$ or $(p_{ij}^l \le p_{ij} \le p_{ij}^u)$, where p_{ij} stands for trip proportion $(p_{ij} = T_{ij} / T)$.

For certain distributions, the expressions of the confidence intervals are well defined at an analytical or numerical level. In the case of the multinomial distribution, different methods are 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, and b) empirical methods, based on successive extractions. For the multinomial distribution, the direct problem of determining the confidence interval has been addressed by several authors in the last century (Quesenberry & Hurst, 1964; Goodman, 1965; Bailey, 1980; Glaz & Johnson, 1984; Fitzpatrick & Scott, 1987; Sison & Glaz, 1995; Hou, Chiang & Tai, 2003; Wang, 2008). The objective of this set of methods is the determination of simultaneous confidence intervals, which handle multiple parameters for the entire sample. These intervals are simultaneously defined for each of the variables involved and present the same level of confidence. Some methods developed so far (Sison & Glaz, 1995) assume the same interval length for all proportions, p_{ij}, though when there are cells in which a number of elements dominate over others, the intervals predicted prove to be unreliable (May & Johnson, 1997). In other methods, when the sample size is large, the Central Limit Theorem hypothesis has been assumed (Roussas, 1973; Snedecor & Cochran, 1980; Meyer, 1986; Canavos, 1988; Walpole, 1992; Agresti & Caffo, 2000; Casella & Berger, 2002). Simulation studies carried out more than a decade ago (May & Johnson, 1997) provided results on methods developed in the late 60s and 80s (Goodman, 1965; Quesenberry & Hurst, 1964; Fitzpatrick & Scott, 1987) which confirm significant limitations for these analytical confidence intervals, such as the large length of the intervals or the limiting value of the number of elements in each cell and matrix size. To circumvent the previous drawbacks, empirical approaches have been gaining acceptance as practical techniques.

Bootstrapping is a technique of replicating samples by extraction, presented in the late 70s (Efron, 1979; Efron & Tibshirani, 1993), and used to estimate a distribution from which to extract several statistics of interest (i.e. mean, variance). In a broad range, the bootstrap methodology consists of estimating a statistical characteristic of the unknown population by simulating the characteristic when the true population is replaced by an estimated one. 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. The main advantage of bootstrapping is the easy way to perform the replicating. Several variations have been proposed to calculate confidence intervals in bootstrapping, such as the method of percentiles and others that correct possible bias when the number of draws is limited, but no definitive conclusions can, generically, be drawn (Efron, 1979; Weinberg, Carroll & Cohen, 1984). In the last fifteen years several studies have been conducted to investigate the performance of bootstrap methods versus analytical methods in terms of constructing confidence intervals for a single multinomial population with *r* categories. The review by Jhun and Jeong (2000) contains valuable conclusions on the applicability of diverse confidence interval analytical close expressions. They proved the superiority of the bootstrap method versus those analytical ones, although the analysis was limited to a small number of categories ($r \in 3,4,5$) and sample size (population $N \leq 200$). A more recent work (Morales, Pardo & Santamaria, 2004) has reached similar conclusions under a similar assumption of categories and sample sizes.

There are two main families of approaches to obtain bootstrap confidence intervals: based on the standard normal table, and based on bootstrap empirical percentiles (the pivot method). In the first one the intervals are calculated by using the estimated standard error and multiplying it with the corresponding percentile points of a normalized normal distribution. In the second one the distribution used is the empirical distribution constructed by sorting the bootstrap estimators in ascending order; for each of the *M* extractions (pairwise cell of the replicated matrix sample set), 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. In order to correct the bias in these empirically calculated intervals, several bootstrap confidence

interval methods have been constructed; the bias-corrected (BC) percentile and accelerated bias-corrected (BCa) percentile methods are among those worth citing (Diciccio & Romano, 1988). Most methods are described in Efron & Tibshirani (1993). In this research, taking into account the error drawn during the surveying process, such a level of correctness/accuracy is not needed for this type of OD matrix application, and for the sake of conciseness attention is concentrated on the bootstrap percentile method.

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 summarized 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: $\hat{\theta}^* = \hat{\theta}_{ij}^*, i = 1,2,...,n_o; j = 1,2,...,n_d = \hat{\theta}_{ij}^m, m = 1,2,...,M, i = 1,2,...,n_o; j = 1,2,...,n_d$.
- *Estimate of cell standard error and mean* $\hat{\sigma}_{ij} = \left\{ \sum_{m=1}^{M} \left[\hat{\theta}_{ij}^m - \overline{\theta}_{ij}^m \right] / (M-1) \right\}^{1/2}$, $\overline{\theta}_{ij} = \left(\sum_{m=1}^{M} \hat{\theta}_{ij}^m \right) / M$.
- *Construction of a confidence interval for parameter* θ_{ij} *based on bootstrap percentiles.*
  - For each *ij*-th matrix cell, in 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.

2.3. OD matrix estimation approaches

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, Inaudi & Marquis, 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, 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. 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, Sasaki & Asakura, 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 the 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 with 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:

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

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}: interprovincial 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 & Benitez, 2005; Caceres, Romero & Benitez, 2011).

3. Empirical analysis

The purpose of this analysis was to quantify the uncertainties in the data collected by a transport survey, and incorporate this piece of information into an origin–destination matrix updating a bi-level scheme on link flows to infer a more reliable demand matrix.

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

3.1. Procedure

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-analyzing tool (i.e. Emme INRO 2012, TransCAD Caliper 2010) 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 on selected links. The information provided by the confidence intervals is incorporated as constraints in the optimization scheme.

3.2. The real case

The case analyzed 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, authorized 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 presented information at the origins and destinations at the provincial level, and are statistically representative at the regional level, but not significant at provincial level. The raw provincial disaggregated level is used for the application of the techniques presented hereinafter.

3.3. The sample

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) intraregional: 302,825, c) interregional: 50,104, and d) international: 8,009.

3.4. Estimate of OD matrix confidence interval by bootstrap

The simulations carried out comply with the empirical procedure introduced in section 2.2. 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 that 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.

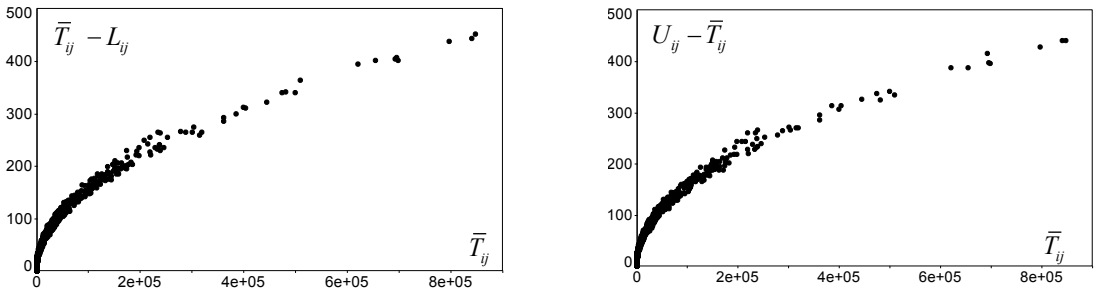

Fig. 1. Left and right half confidence interval versus cell trips

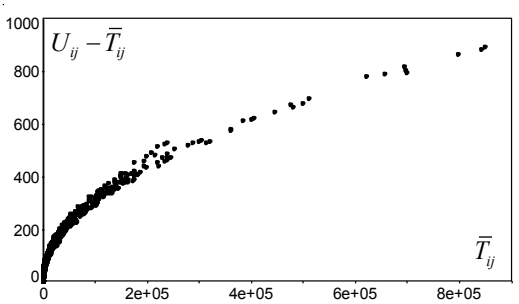

Fig. 2. Confidence interval length versus cell trips

The inferred left and right half confidence intervals versus trip nominal values for all OD matrix cells are depicted in Figure 1. Full confidence interval lengths versus cell trip mean values are shown in Figure 2. The solid curves are the regression models, obtained by a least-squares fit, with expression $Length_{ij} = e^{a} \cdot \bar{T}_{ij}^{b}$ of which the parameters and statistical values are reported in Table 1.

Table 1. Parameters of the fitting regression models to the confidence intervals.

Fitting curve	*Parameter*	*mean*	*variance*	*statistics t*	*95% confidence interval*		*Adj.* R^2
					lower limit	upper limit	
$\bar{T}_{ij}^{m} - L_{ij}$	a	-0.5032	0.0075	-66.8626	-0.5180	-0.4885	0.9960
	b	0.4841	0.0007	738.8437	0.4828	0.4854	
$U_{ij} - \bar{T}_{ij}^{m}$	a	-0.4659	0.0075	-62.1562	-0.4806	-0.4512	0.9960
	b	0.4808	0.0007	735.9120	0.4795	0.4821	
$U_{ij} - L_{ij}$	a	0.2083	0.0072	28.9490	0.1941	0.2224	0.9964
	b	0.4825	0.0006	769.8773	0.4813	0.4837	

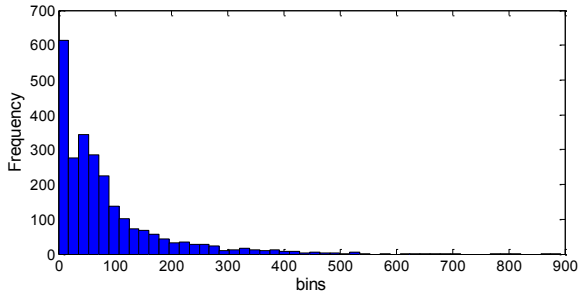

Fig. 3. Histogram of confidence interval length $U_{ij} - L_{ij}$

The determination coefficients of these adjustments, $Adj.\ R^2$, are sufficiently high to ensure the goodness of fit. Figure 3 reflects the histogram of confidence interval lengths for OD cell trips. It is easy to see the large number of null trip cells, a recursive behavior in most transport survey studies.

4. Conclusions

A general methodology for estimating confidence intervals for OD matrix cells, extracted from a travel survey, is presented. The approach has been applied to the real case of the extensive annual interprovincial freight transport in Spain. This allows us to estimate a measure of the magnitudes to be imposed in the process of adjusting OD matrices. The consequences of this finding are significant for the generation of OD matrices that contend with real uncertainty in data collected by a survey, diminishing the level of uncertainty involved in this extremely underspecified problem.

The procedure to calculate confidence intervals with a large level of certainty is simple. The OD matrix updated bootstrap methodology, where the sample-related uncertainties are incorporated as restrictions in the optimization problem, is useful for inferring a more reliable trip distribution matrix. The product of this methodology should help to increase the confidence in the updated matrices for developing more reliable demand forecasting models that are founded on a more robust data structure.

Acknowledgements

This research is funded by the Public Road Agency of the Andalusian Regional Government (AOP-JA, Spain Project G-GI3000/IDII) and EU FEDER Funds. 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 (Ministerio de Fomento), owner of the data employed.

References

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician,* 54(4), 280–288.

Bailey, B. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformations of cell frequencies. *Technometrics,* 22(4), 583–589.

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

Caceres, N., Romero, L.M., & 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 Trasportation*, DOI:101002/1tr.187.

Caliper Corporation (2010). TransCAD. Release 2010.

Canavos, G. (1988). *Probabilidad y Estadística: Aplicaciones y Métodos*. Mc- Graw Hill, Madrid.

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., & Marquis, G. (1993). Dynamic estimators of origin-destination matrices using traffic counts. *Transportation Science*, 27(4), 363–373.

Casella, G., & Berger, R. (2002). *Statistical Inference*. 2 Edn, Duxbury, United States of America.

Diciccio T. J., & Romano J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society*, B 50(3), 338-354.

Doblas, J., & 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. *The Annals of Statistics*, 7(1),1-26.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Boca Raton: Chapman & Hall/CRC.

Fitzpatrick, S., & Scott, A. (1987). Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82(399), 875–878.

Fomento, M. (2008). *Encuesta Permanente de Transportes de Mercancías por Carretera (EPTMC)*. Subdirección General de Estudios Económicos y Estadísticas. Ministerio de Fomento.

Glaz, J., & Johnson, B. (1984). Probability for multivariate distribution with dependence structures. *Journal of the American Statistical Association*, 79, 411–436.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics,* 7(2), 247–254.

Hou, C., Chiang, J., & Tai, J.J. (2003). A family of simultaneous confidence intervals for multinomial proportions. *Computational Statistics & Data Analysis*, 43, 29-45.

INRO (2012). Emme. Released 4.0.1.

Jhun, M., & Jeong, H.C. (2000). Applications of bootstrap methods for categorical data analysis. *Computational Statistics & Data Analysis,* 35, 83–91.

May, W.L., & Johnson, W.D. (1997). Properties of simultaneous confidence intervals for multinomial proportions. *Communication in Statistics- Simulation and Computation*, 26, 495-518.

Meyer, P. (1986). *Probabilidad y Aplicaciones Estadísticas*. Addison-Wesley Iberoamericana, Mexico.

Morales, D., Pardo,L., & Santamaria, L. (2004). Bootstrap confidence regions in multinomial sampling. *Applied Mathematics and Computation*, 155, 295–315.

Quesenberry, C. P., & Hurst, D. C. (1964). Large-sample simultaneous confidence intervals for multinomial proportions. *Technometrics* 6(2), 191–195.

Roussas, G. (1973). *A first Course in Mathematical Statistics*. 2 Edn, Addison-Wesley, Massachusetts.

Sison, C. P., & Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association,* 90(429), 366–369.

Snedecor, G., & Cochran, W. (1980). *Statistical Methods*. 7 Edn, The Iowa State University Press: Ames, Iowa.

Walpole, R. E., & Myers R. H. (1993). Probability and Statistics for Engineers and Scientists. 5th Edn. New York: MacMillan.

Wang, H. (2008). Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis,* 99, 896 – 911.

Weinberg, S. L., Carroll, J. D., & Cohen, H. S. (1984). Confidence regions for indscal using the jackknife and bootstrap techniques. *Psychometrika*, 49(4), 475-491.

Yang, H., Sasaki, T., Iida, Y., & Asakura, Y. (1992). Estimation of origin-destination matrices from link traffic counts on congested networks. *Transport Research*, Part B, 26(6), 417-434.