ADJUSTING ORIGIN-DESTINATION MATRICES BASED ON TRAFFIC COUNTS AND BOOTSTRAPPING CONFIDENCE INTERVALS

F.G. Benitez[1*], Professor
L. Romero[2], PhD. Research Associate
N. Caceres[3], PhD. Research Associate
J.M. del Castillo[4], Professor

Transportation Engineering, Faculty of Engineering
University of Seville
Camino de los Descubrimientos, s/n, Seville 41092, Spain
Tel.: +34 954 488135, Fax: +34 954 487316
Email: benitez@esi.us.es[1], l_m_romero@esi.us.es[2], noeliacs@esi.us.es[3], delcastillo@us.es[4]

(*) Corresponding Author

Word count: 6705 + 3 Figures = 7455

ABSTRACT

Mobility studies require, as a preliminary step, conducting a survey to a sample of users of the transportation system. The statistical reliability of the data determines the goodness of the results and conclusions which can be inferred from the analyses and models generated. Due to the high economic costs of the collecting field stages, collected data are partially reused in either a disaggregated o aggregated manner. In the first case, the statistical reliability is not always guaranteed, affecting drastically the results to be derived from projections and estimates of future hypothetic scenarios.

In this paper we present a methodology, based on the techniques of "bootstrap", for the robust statistical estimation of the mobility matrices, and generate the confidence intervals of travel between origin-destination (OD) pairs defined by each matrix cell derived from a 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.).

To address this task we have counted with a statistically reliable data mobility study conducted in Spain at the level of regions. This paper presents the results derived from disaggregating date at interprovincial level, and an application to the posterior mobility matrix adjustment based on traffic counts data. The study results demonstrate the potential of the methodology developed and the usefulness of conclusions.

INTRODUCTION

Modeling transport demand based on transport data of different nature and captured by different procedures has been disseminated by research papers, spread out in specialized books and deployed in professional studies. Although there are a large variety of methods, most approaches follow the traditional construction of an origin–destination (OD) trip matrix estimate based on available information collected by a transport survey. Either estimating or adjusting an OD matrix that generates information (i.e. flows, speeds) that are most compatible with observed field data (i.e. link volumes, trips between zones, cordons and screen-line counts, vehicle speeds) is the main goal of most matrix estimation methods. The confidence on the results, derived from the modeling and its derived planning study, is based on the reliability of each of the prior stages of this type of studies and in particular on the first one: transport data captured during the survey process. Inaccurate OD estimates could have far reaching negative consequences including unrealistic mobility forecast patterns.

The construction of mobility matrices of a given region to be analyzed, OD matrices, feeds on the collected data in a process of surveying a sample of users of the transportation system. Clearly, the number of trips carried out in the region under study and their characteristics (i.e. spatial and temporal distribution, modes of transport used, purposes, stages, etc.) is a function not only of the population size, but of other endogenous factors to the existing transportation system (i.e. infrastructure, modes, services, accessibility, etc), endogenous to individuals within the resident population characterized by socioeconomic attributes (i.e. age, education, profession, resources, etc) and endogenous to the region through geo-socio-economics characteristics (i.e. jobs, shops, services, etc) mainly. Of these factors, population size is the most correlated attribute with the aggregate total number of trips generated in the region. For the highest level of disaggregation, the individual, geo-socio-economics characteristics of users and the level of service of the transport system are the factors that best explain the generation/attraction of trips (1).

Although there are several techniques to perform data collecting, the most customary used can be classified into two families, based on the disaggregation level of the population:

a) Individual level. In this case a sample of individuals is chosen from the population. This sample must be statistically representative of the population distribution functions.

b) Household level. A sample is chosen from among the household universe in the region. The sample must also be chosen to be statistically representative of the distribution of households according to variables associated with the item. Obviously, the level of aggregation of the household variables affect the explanatory power of the data collected in relation to reality.

c) Other higher levels of aggregation. They are less relevant as general methodologies, but are used to supplement the data captured by some of the previously cited methods, it is the case of surveys at specific trip generators or attractors and other places where non-residents and passers-by can be collected.

From these techniques, one of the most widely used is based on the research of mobility patterns of samples based on family units or households (2). Once the studied region is discretized into

transport areas by aggregating census districts, the sample size turns to be a function of the total number of households distributed among the zones and according to the resident population; this ensures a high statistical reliability of trip information collected on a zonal level. Census data provide insight into the distribution of households by family size, this is the reason this variable is aggregated in order to obtain in a direct manner the household histogram by family size. Other useful variables that characterize households in a more disaggregated way, available in the government and public administration, are the number of vehicles per household, the assessed value of housing, among others.

By this sampling technique, and for each area z, denoting by H_z the number of households in the area, the household histogram by family size can be easily obtained. The choice of those households 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. The elements of the original sample that fail, by any external cause to the survey for instance, are replaced by other substitutes with similar characteristics (i.e. same size) in order to preserve the sampling distribution. From the practical and professional standpoint, the sample and the universe generally are related through expansion coefficients. For the present case, they are defined by $k_z^t = \frac{H_z^t}{h_z^t}$, where H_z^t and h_z^t stand for the number of existing households and respondents of size *t* in zone *z*, respectively.

The expansion process 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. age distribution, or other socioeconomic variables) may be carried out, it is a fact that the expanded data are severely affected by significant errors of a difficult characterization. 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*”, where confidence intervals are the most practical. This process of “*representativeness*” can be approached from the perspective of inferring confidence intervals for each of the terms of the OD matrices; this requires following two different paths according to the reliability that is given to the expansion coefficients:

a) Obtaining confidence intervals from the OD matrices data sample (pre-expanded) and then affecting them with the expansion coefficients.

b) Obtaining confidence intervals from the expanded OD matrices.

The difference between these two cases is the length of the intervals inferred. Those obtained by the first procedure are more conservative (large intervals) that those derived from the second procedure.

This paper describes a model that estimates 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 bootstrap to evaluate the uncertainties in each pair of the OD matrix estimates. The model is attractive because of two aspects. First, it incorporates statistics features that improve the knowledge on the data yield by the survey. Second, it uses confidence intervals of all available information to define bounds for the feasible solution space where the OD matrix estimate is sought.

This paper is organized 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, based on previous variant b); 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.

PROBLEM DEFINITION AND FORMULATION

Introduction

For a given study area divided into $n_o + n_d$ transport zones where users 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 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.

To obtain matrix Υ requires the observation of all trips made in the area, both by the resident population and non-resident as 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 users who 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 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 $\sum_{i=1}^{n_0} \sum_{j=1}^{n_d} \pi_{ij} = 1$.

Reliable estimation of the parameter matrix $\boldsymbol{\pi}$ requires the availability of a sufficient high number E of samples, and in this case the total number of trips T^e follows a normal distribution $N(\mu_T, \sigma_T)$. This approach is of a low interest because of the impracticability and budget restrictions to conduct multiple repeated studies to obtain more than just one matrix $\mathbf{T}^1$. Therefore 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 series $\mathbf{T}^1, \mathbf{T}^2, ..., \mathbf{T}^E$.

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

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

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 (eq.1) is particularized as:

$$\begin{aligned} &\mathrm{P}\left[\mathbf{T}^{*} = \mathbf{T} \middle| T, \boldsymbol{p}\right] \equiv \mathrm{P}\left[T_{11}^{*} = T_{11}, ..., T_{n_o n_d}^{*} = T_{n_o n_d} \middle| T\ , p_{11}, ..., p_{n_o n_d}\right] \\ &= T\ !(p_{11})^{T_{11}} \cdot ... \cdot (p_{n_o n_d})^{T_{n_o n_d}} / T_{11}! \cdot ... \cdot T_{n_o n_d}! \end{aligned} \tag{2}$$

the probability distribution function of all possible matrices $\mathbf{T}^*$ with parameters $\boldsymbol{T}$ y $\hat{\boldsymbol{\pi}} = \left\{p_{ij}\right\}$,

where $\sum_{i=1}^{n_0}\sum_{j=1}^{n_d} T_{ij} = T$, and $\sum_{i=1}^{n_0}\sum_{j=1}^{n_d} p_{ij} = 1$.

Confidence intervals for OD matrices

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 the 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 (3).

For certain distributions, the expressions of the confidence intervals are well defined at analytical or numerical level. In the 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. The following is a brief state of the art for the case of the multinomial distribution.

Analytical methods

For the multinomial distribution, the problem of determining the appropriate sample size for a population has been addressed by several authors in the last half century (4-11). Similarly, the direct problem of determining the confidence interval has been treated, among others, by (12-18). Hou et al. (19) present a review and comparison of different confidence intervals defined in the last 5 decades. The objective of this set of method 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.

For a multinomial distribution function (eq. 2), the interest pursued is the construction of a set of simultaneous confidence intervals. Sison and Glaz (17) derive a practical and easy approach to use. Unfortunately the method works well when all proportions are similar for all *n* cells, there being a quasi-uniform dispersion in the number of elements in each cell. When the number of cells is large, the results tend to those predicted by methods based on the assumption of normality. Conversely, if there are cells in which the number of elements dominates over others, the intervals predicted turn to be unreliable (20).

For the case where the sample size is large enough, the Central Limit Theorem allows some simplifying hypothesis, yielding the results derived by some authors (21-26) and Agresti and Caffo (27) in particular. Leemis and Trivedi (28) define an algorithm to determine empirically the endpoints of the interval (L,U) for a confidence level $(1-\alpha)100\%$ for $x \equiv x_i$ such that $P(X \geq x | \pi = L) = \alpha / 2$, $P(X \leq x | \pi = U) = \alpha / 2$, in function of the empirical distribution function *F*.

Simulation studies carried out a decade ago (20) provided results on methods developed in the late 60s and 80s (12,13,16) 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. This is the reason empirical methods have been gaining ground and acceptance as useful techniques from the practical perspective.

Replicated empirical methods using bootstrap

Bootstrap is a technique of replicating samples by extraction, presented in 1979 (29-30), 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 randomized replicating, such as the *jackknife* method, but analyzes carried out up to day do not support the superiority of one over the other (31).

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 a 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. There are several methods to correct the bias in these empirically calculated intervals (30).

Using multiple extractions, following the *bootstrap* technique, the histogram can be built to derive the computation of percentiles. The following steps summarize the pseudo-algorithm:

Step 1. Obtain sample parameters $\hat{\pi}_i = \frac{n_i}{N}, \quad \forall i = 1,2,...,n$;

Step 2. Generate M samples of size N from a multinomial distribution of parameters $\hat{\pi}_i$;

Step 3. Estimate, for each simple *m*, parameters $\hat{\pi}_i^m \quad \forall i = 1,2,...,n$;

Step 4. For each *observation i*, in all M samples, the histogram is constructed from $\left\{\hat{\pi}_i^m;\ \forall m = 1,2,...,M\right\}$.

Step 5. Compute percentiles $\hat{\pi}_i^{\alpha/2} = Percentil\left[\frac{\alpha/2}{(k-1)}\right],\ \hat{\pi}_i^{1-\alpha/2} = Percentil\left[\frac{1-(\alpha/2)}{(k-1)}\right]$.

There are studies on real applications showing the superiority of this method versus those approximates introduced in the previous section (32, 33).

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 (34), automatic vehicle identification (AVI) systems (35-36) and cell phones (37) are widely used.

The problem of OD inference, estimation and prediction has been dealt with during the last two and a half decades (38-40). 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.

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 modeled 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 (41) 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.

Beside the hypothesis assumed and the approaches followed, there are other factors that

impede to be certain on the quality and reliability of the OD matrix estimated (42).

A high percentage of the effort in estimating OD matrices is concentrated in spending an important share of the budget on surveys to collect data to create information regarding mobility of people and goods. Even though, either a shortage on economic resources or technical expertise may end up collecting data of dubious quality, since respondents are chosen following bias extractions from the whole population. This is the reason data are not fully accessible to the public, and are mostly aggregated just to disclose trends. Banks and Reiter (43) show a broad sketch of the factors that affect the final accuracy of collected data.

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 based on a) empirical methods, such as conducting a survey on a sample of individuals, applying a trip distribution model, or using traffic counts as measurements of link flows in a network model in order to adjust an existing matrix; b) analytical methods, just as applying a trip model; or c) empirical-analytical methods, as any mixed approach of the two previous ones. The third alternative is the one that has mostly been used over the past twenty-five years and a considerable amount of work has been documented in the literature (44); a set of traffic counts and a prior O–D matrix are prerequisites. The prior matrix is typically assumed to come from a survey using a finite data set (instead of using the whole population). The survey data need to be corrected, expanded, and validated in order to achieve a representative and reliable prior matrix to be used in matrix estimation methods (45-46) So, this prior matrix can be regarded as an observation (a good approximation) of the “true” O–D matrix to be estimated. However, the fact that a prior matrix is regarded as a “good approximation” of the “true” O–D matrix does not imply that it can be used directly as a result.

A few of errors may arise during the processes of building, calibrating, and forecasting the prior matrix with models, such as sampling errors, measurement inaccuracies, transfer omissions, or aggregation errors. The prior matrix is therefore adjusted using traffic flows, which are one type of information that can be collected automatically on a subset of links in a network, not on all links (this would be impossible nowadays due to economic budgetary restrictions). In methods based on this third alternative, the prior O–D matrix is iteratively “adjusted” or “changed” to reproduce the observed traffic counts when assigned to the transportation network. The aforementioned errors can also be mitigated by adjusting the prior O–D matrix to satisfy the traffic counts. In this manner, one may obtain a “reasonable” estimate of the O–D matrix; hence, this alternative is the most widely used in practical applications.

In mobility studies a vast amount of information, which should be given only a certain degree of reliability, is handled. Facing all available information one can observe inconsistencies between some data, so it is necessary to conduct a thorough analysis of the possible causes of such inconsistencies. A clear example corresponds to the discrepancy observed between the volumes of vehicles measured in reality (flows in the network) and those modeled by OD matrices. The usual practice in professional studies is to attribute this discrepancy solely to an incorrect definition of the above matrices, so one just proceeds to modify the OD matrix extracted from the mobility survey, the so called prior matrix. However, that matrix has been obtained by expensive processes based on carrying thorough surveys and it provides a fair idea of the distribution and magnitude of trips. Thus any excessive distortion of the information overrides the budget and human efforts devoted to these surveying tasks.

The most widespread adjustment methodology is based on obtaining trip matrices,

expressed in equivalent vehicles, that replicate as closely as possible those observed volume when matrices are assigned to a reliable transport network model by an assignment code. In general one can affirm that the different methods to estimate OD trip matrices based on traffic counts, developed in the literature, have the following generic form (47):

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

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 estimated and observed volumes in network links, v and $\bar{\mathbf{v}}$ respectively. The most common accepted expressions for functions F_1 and F_2 are the maximum likelihood, generalized least squares and derivations of the principle of maximum entropy. Parameters α and β are the corresponding weight factors that reflect the relative confidence in the available data $\bar{\mathbf{T}}$ and $\bar{\mathbf{v}}$. Finally, pseudo function *Assign* represents the assignment process used (48) to model link volumes from the estimated matrix.

Adjusting OD matrices using confidence interval bound constraints

The proposed formulation follows the basics of scheme (eq. 3), however to control the distortion of the prior matrix a set of bounded variable constraints (for each matrix cell) and functional restrictions (relative to the information contained in the aggregate matrix, as it is the case of information from aggregations of cells, called macrozones, which represent specific conditions between macrozones) are prescribed. This manner of proceeding pursues to keep the variation of the information contained in the adjusted matrix compared to prior matrix within a range considered to be feasible. This last set of restrictions may be due, as is the case at hand, to the fact that the survey the prior matrix is obtained from was designed to derive statistically significant results at the level of macro zones (i.e. regional). Thus, even though an OD trip matrix at the zone level is available, care should be taken in using the survey data in the estimation process of incorporating such circumstances regarding inter macrozone magnitudes.

Based on the assumption set forth in the introduction of the problem definition, consider the study area partitioned into n_o+n_d traffic zones with trips from any origin to all destinations and an OD trip matrix denoted by $\mathbf{T}=[T_{ij}]$, its (i, j) element being the number of trips from origin i to destination j during a certain time period. The road network corresponding to the study area is abstracted into a graph model consisting of a set of *regular nodes* and a set of *directed links*. The service level associated with the links is given by link performance functions $s_a(v_a)$, which relate the travel time on each link to the flow across the link. Finally, the assignment of the OD matrix to the network model in order to obtain flow and travel time on each link is considered to be a deterministic (or stochastic) user-equilibrium procedure, whose behavioural principles are described by the two conditions usually attributed to J.G. Wardrop (48).

Regarding the adjustment problem, the necessary volume data are inferred from collected data on traffic counts on certain links. The formulation proposed to adjust the prior OD matrix includes euclidean distance between estimated and observed volume data and distance between prior and estimated matrices; in addition a set of variable bounds and functional constraints which define admissible ranges for individual OD pairs, zonal 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 (eq.3) results in the bi-level programming approach proposed in this investigation, formulated as follows:

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

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 links $a \in A$.

Variables

$\mathrm{v} = \{v_a,\ \forall a \in A\}$: volume on link a; h_k : flow on path k; T_{ij}: interprovincial travel demand (trips) from origin i to destination j, (note that it 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 $\overline{T}_{ij}$ represents the same quantity referred to the prior matrix $\overline{\mathbf{T}}$. 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 (eq.4).

The lower level program stated in (eq.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} (49).

In addition to the above dimensions established to control the distortion of the information contained in the matrices, one can set a series of maximum increments and decrements for those pairs of the prior matrix where no information is available (50,37).

CASE STUDY

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.

The case analyzed corresponds to the survey of long distance mobility for people living in Spain. The published version (51) contains information on total aggregated trips by regions, and therefore statistically significant at this level. The level of aggregation used for the observations is the household, regardless of any demographic feature. The spatial and temporal ranges correspond to the nation territory and the month before the conclusion of the survey. The scope includes all trips of those who travel a distance exceeding 50 km and those of shorter length including an overnight stay outside the town of origin. The raw data of the survey presented information at the origins and destinations at the provincial level, not statistically significant, and they are used for the application of the techniques presented hereinafter.

Estimate of OD matrix confidence interval by bootstrap

The simulations carried out comply with the empirical procedure introduced in the *Analytical methods* subsection and the less conservative variant (b) from expanded matrices, discussed in the *Introduction* section. These simulations consist of the following steps:

1. For the initial data set $(T_{11},...,T_{n_o n_d})$, estimate the multinomial proportions $(p_{11} = \frac{T_{11}}{T},..., p_{n_o n_d} = \frac{T_{n_o n_d}}{T})$ and assume the hypothesis that these ratios correspond to the "true" population proportions.
2. Extract 1,000 multinomial samples from the survey matrix.
3. Obtain confidence intervals for each cell sample, based on the drawn subset corresponding to each cell, with a nominal confidence level $(1-\alpha) = 0.95$.
4. Assess the average length of intervals as the difference between the upper and lower limits of each interval.

The computer program was coded in Matlab. The simulated multinomial samples were generated by the subroutine MNRND. All simulation studies were performed on a 12 core Intel Xeon E5645 personal computer using parallel computing.

Confidence interval lengths inferred versus trip nominal values for all OD matrix cells are depicted in Figure 1. The solid curve is the regression curve, obtained by a least-squares fit, with expression $U_{ij} - L_{ij} = e^{a} \cdot T_{ij}^{b}$ where parameters $a = -2.0995,\ b = 0.5009$ with a t-statistics of -349 and 1159 respectively.

The coefficient of determination of this adjustment, $R^2 = 0.998$, is sufficiently high to ensure the goodness of fit.

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 modeled volumes is $R^2 = 0.70$ (Figure 2a), while the assignment of the adjusted matrix gives rise to a value of 0.8 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 is $R^2 = 0.99$.

The solid straight line is the linear regression line, obtained by a least-squares fit, with expression $T_{ij} = -2422.74 + 0.9833\bar{T}_{ij}$.

The control in the OD estimation, containing the level of distortion between both prior and undated matrices, utilizing 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 (50) that relaxing the constraints derived from the cell confidence intervals would both (i) increase the determination coefficient between observed and modeled volumes (unconstraint optimization yields more 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 analyze.

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 interprovincial mobility 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 compel with real uncertainty in data 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.

UNKNOWLEDGMENTS

This research was funded by the Ministry of Public Works of Spain (*National R&D Program* Project P63/08 SIMETRIA); part of the theoretical developments of the matrix adjustment methodology has been framed within the developments of projects ENE2008-05552 and TRA2007-63564. 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 of policy of the Ministerio de Fomento (*Ministry of Public Works*), owner of the data facilitated.

REFERENCES

1. Oppenheim, N. *Urban travel demand modeling*. Wiley-Interscience, N.Y., 1995.
2. Wilson, F.R. *Journey to work-Modal split*. MacLaren and Sons Ltd, London, 1967.
3. Burdick, R. K. and F.A. Graybill. *Confidence Intervals on Variance Components*. New York: Marcel Dekker, 1992.
4. Hurtubise, R. Sample size and confidence intervals associated with a Monte Carlo simulation model possessing a multinomial output. *The American Statistician*, Vol. 12, 1969, pp. 71–77.
5. Angers, C. A graphical method to evaluate sample sizes for the multinomial distribution. *Technometrics*, Vol. 16(3), 1974, pp. 469–471.
6. Angers, C. Sample size estimation for multinomial populations. *The American Statistician*, Vol. 33, 1979, pp.163–164.
7. Angers, C. Large sample sizes for the estimation of multinomial frequencies from simulation studies. *Simulation*, Vol. 27, 1984, pp. 175–178.
8. Angers, C. Note on quick simultaneous confidence intervals for multinomial proportions. *The American Statistician*, Vol. 43, 1989, pp. 91.
9. Tortora, R. A note on sample size estimation for multinomial populations. *The American Statistician*, Vol. 32(3), 1978, pp. 100–102.
10. Thompson, S. Sample size for estimating multinomial proportions. *The American Statistician*, Vol. 41, 1987, pp. 42–46.
11. Bromaghin, J. Sample size determination for interval estimation of multinomial probabilities. *The American Statistician*, Vol. 47(3), 1993, pp. 203–206.
12. Quesenberry, C. P. and D.C. Hurst. Large-sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, Vol. 6(2), 1964, pp. 191–195.
13. Goodman, L. A. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, Vol. 7(2), 1965, pp. 247–254.
14. Bailey, B. Large sample simultaneous confidence intervals for the multinomial probabilities based on transformations of cell frequencies. *Technometrics*, Vol. 22(4), 1980, pp. 583–589.
15. Glaz, J. and B. Johnson. Probability for multivariate distribution with dependence structures. *Journal of the American Statistical Association*, Vol. 79, 1984, pp. 411–436.
16. Fitzpatrick, S. and A. Scott. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, Vol. 82(399), 1987, pp. 875–878.
17. Sison, C.P. and J. Glaz. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, Vol. 90(429), 1995, pp. 366–369.
18. Wang, H. Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. *Journal of Multivariate Analysis*, Vol. 99, 2008, pp. 896 – 911.
19. Hou, C., J. Chiang and J.J. Tai. A family of simultaneous confidence intervals for multinomial proportions. *Computational Statistics & Data Analysis*, Vol. 43, 2003, pp. 29-45.
20. May, W.L. and W.D. Johnson. Properties of simultaneous confidence intervals for multinomial proportions, *Communications in Statistics - Simulation and Computation*, Vol. 26, 1997, pp. 495-518.
21. Roussas, G. *A first Course in Mathematical Statistics*. 2 edn, Addison-Wesley, Massachusetts, 1973.

22. Snedecor, G. and W. Cochran. *Statistical Methods*, 7 edn, The Iowa State University Press: Ames, Iowa, 1980.
23. Meyer, P. *Probabilidad y Aplicaciones Estadísticas*, Addison-Wesley Iberoamericana, Mexico, 1986.
24. Canavos, G. *Probabilidad y Estadística: Aplicaciones y Métodos*, Mc- Graw Hill, Madrid, 1988.
25. Walpole, R. E. and R.H. Myers. *Probability and Statistics for Engineers and Scientists*. 5th edn. New York: MacMillan, 1993.
26. Casella, G. and R. Berger. *Statistical Inference*, 2 edn, Duxbury, United States of America, 2002.
27. Agresti, A. and B. Caffo. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, Vol. 54(4), 2000, pp. 280–288.
28. Leemis, L. and K.A. Trivedi. A comparison of approximate interval estimators for the binomial parameter. *The American Statistician*, Vol. 50(1), 1996, pp. 63–68.
29. Efron, B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, Vol. 7, 1979, pp. 1–26.
30. Efron, B. and R.J. Tibshirani. *An introduction to the bootstrap*. Boca Raton: Chapman & Hall/CRC, 1993.
31. Severiano, A., J.A. Carrico, D.A. Robinson, M. Ramirez and F.R. Pinto. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. *Plos One*, Vol. 6(5), 2011: e19539. doi:10.1371.
32. Gonzalez, D. Comparación de intervalos de confianza para la distribución multinomial. Master Thesis. Facultad de Ciencias. Universidad Nacional de Colombia, 2010.
33. Bishop, Y., S. Feinberg and P. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.
34. Van der Zijpp, N. 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, 1996.
35. Dixon, M.P. and L.R. Rilett. Real-time origin-destination estimation using automatic vehicle identification data. Proc. 79th Transportation Research Board Meeting, Washington, D.C., 2000.
36. Kwon, J. and P. Varaiya. 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*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 119–126.
37. Caceres, N., L.M. Romero and F.G. Benitez. 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*, 2011, doi:101002/1tr.187.
38. Cascetta, E. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator, *Transportation Research*, Vol. 18B, (4/5), 1984, pp. 288–299.
39. Ben-Akiva, M. 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, 1987, pp. 459–481.
40. Cascetta, E., D. Inaudi and G. Marquis. Dynamic estimators of origin-destination matrices using traffic counts. *Transportation Science*, Vol. 27(4), 1993, pp. 363–373.

41. Kapur, J.N. *Maximum-entropy models in science and engineering*. John Wiley & Sons, New Delhi, 1989.
42. Hugosson, M. B. Quantifying uncertainties in a national forecasting model. *Transportation Research*, Vol. 39A, 2005, pp. 531-547.
43. Banks, D. and J.P. Reiter. Confidentiality issues related to transportation use of census data for transportation planning: Preparing for the future. Proc. 84th Transportation Research Board Meeting, Washington, D.C., 2005.
44. Abrahamsson, T. Estimation of Origin-Destination Matrices Using Traffic Counts – A Literature Survey. Report IR-98021, Int. Institute for Applied Systems Analysis, 1998.
45. Robillard, P. Estimating the O–D matrix from observed link volumes. *Transportation Research*, Vol. 9, 1975, pp. 123–128.
46. Wilson, A.G. *Entropy in Urban and Regional Modelling*. Poin, Ltd., London, England, 1970.
47. Yang, H., T. Sasaki, Y. Iida, and Y. Asakura. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transport Research*, Vol. 26(6), 1992, pp. 417-434.
48. Patriksson, M. *Traffic assignment problem. models and methods*. VSP International Science, Utrecht, The Netherlands, 1994.
49. Sheffi, Y. *Urban transportation networks: equilibrium analysis with mathematical programming methods*. Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1985.
50. Doblas, J. and F.G. Benitez. An approach for estimating and updating origin-destination matrices based on traffic counts preserving prior structure. *Transportation Research*, Vol. B 39, pp. 565-591.
51. Fomento, M. *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, 2008.

LIST OF TABLES AND FIGURES

List of Tables:

None

List of Figures:

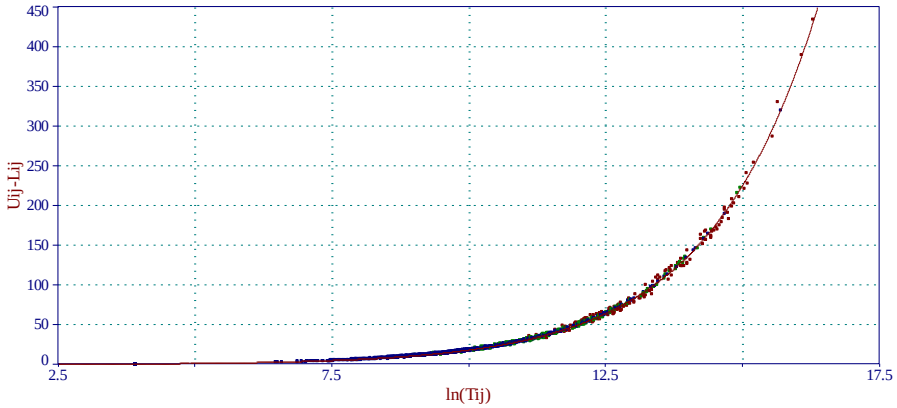
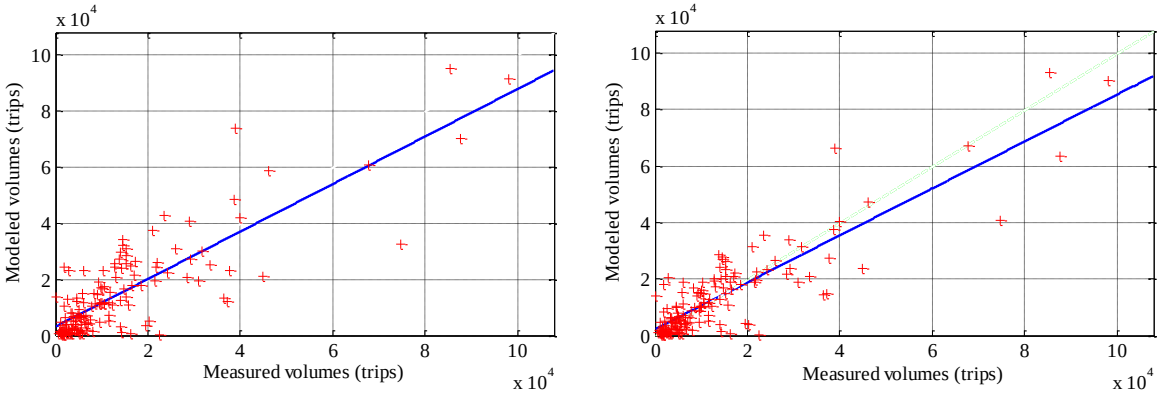


FIGURE 1 Confidence interval length vs. cell trips.



(a) Prior matrix assignment (b) Adjusted matrix assignment

FIGURE 2 Relationship between measured and modeled volumes using the prior matrix (a) and the adjusted matrix (b).

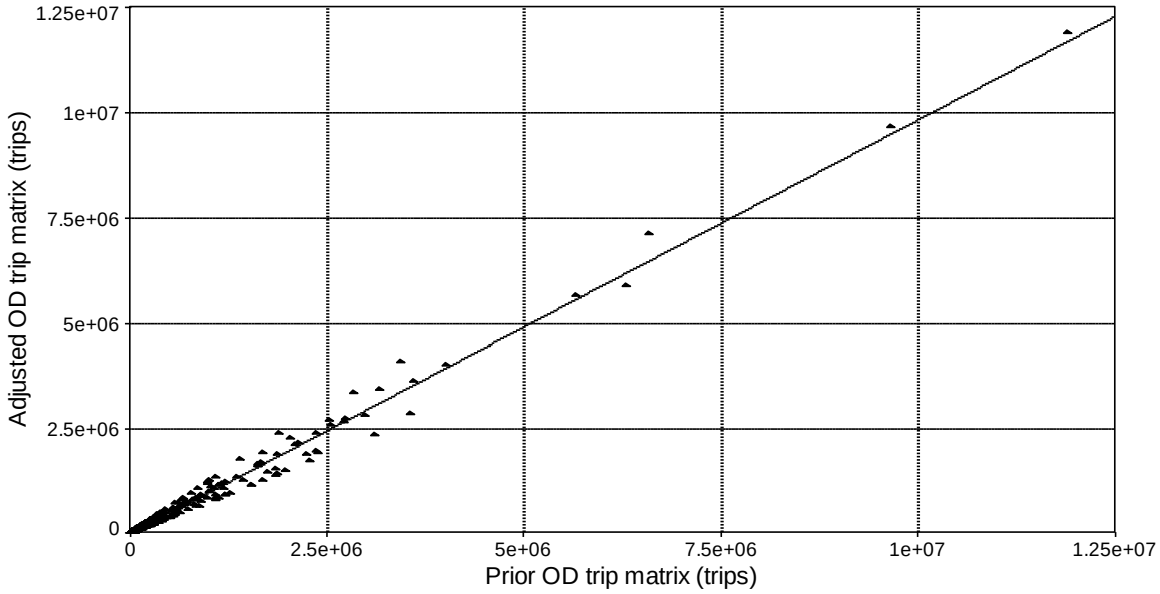


FIGURE 3 Correlation between prior and adjusted OD matrices.