Methods for the Redistribution of Tourist Expenditure in Spain: From EGATUR to Reallocation Matrices

Diego Frade-Amil, Manuel Oviedo de la Fuente, Salvador Naya, Javier Tarrío-Saavedra, Luisa Carpente, and Mario Francisco-Fernández

MODES, Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: diego.frade.amil@udc.es

DOI: *https://doi.org/10.17979/spu.23.c35*

Abstract: To estimate the expenditures made by foreign visitors in Spain, the National Statistics Institute (INE) produces the "Tourist Expenditure Survey" (EGATUR), which allocates all spending to the main destination region, even if it was not fully carried out there. Consequently, there arises a need to redistribute expenditure across the Autonomous Communities actually visited, for which we rely on supplementary information from transactions with foreign bank cards. Our proposal is to estimate a reallocation matrix whose elements allow the redistribution of expenditure from the main destination regions to all those effectively visited. Several approaches are explored, including constrained least squares, overnight-stay proportion matrices, and Data Envelopment Analysis (DEA), among others.

1 EGATUR: A Brief Introduction

The estimation of actual tourist expenditure represents a key challenge for the production of official statistics and the design of public policies. In Spain, the National Statistics Institute (INE) conducts the *Encuesta de Gasto Turístico* (EGATUR) (Instituto Nacional de Estadística (INE), 2025b), the official survey designed to measure the expenditure made by international visitors to Spain. EGATUR collects microdata on trips, destinations, country of residence, purpose of travel, duration (including overnight stays), and detailed expenditure categories (Instituto Nacional de Estadística (INE), 2025a). In its standard dissemination, the total expenditure of each trip is assigned to the main destination Autonomous Community (ACs) indicated by the traveler, based on different criteria (e.g., most important destination, longest stay, highest expenditure). While this convention simplifies data reporting, it may lead to an over- or under-representation of expenditure in the designated main destination, since travelers may also incur spending in other ACs during their stay (for instance, in multi-stop itineraries or day trips).

In this context, complementary methods are required to reallocate expenditure across the ACs actually visited. This is the objective of one of the research lines of the project "Ciencia e ingeniería de datos para la mejora de la función estadística oficial" (CIDMEFEO), funded by INE, in particular Line 9, entitled "Utilización de información de pagos con tarjetas bancarias para repartir el gasto de los turistas y excursionistas no residentes asignado a la CCAA de destino principal de sus visitas a España entre todas las CCAA que se han visitado." To address this problem, we combine EGATUR information with additional data from foreign payment-card transactions in order to estimate a reallocation matrix that redistributes expenditure from the main destination to the actual destinations.

Section 2 presents the methodological framework, describing the procedure and the elements involved. Section 3 introduces the different estimation methods proposed and the steps to identify the most suitable one. Finally, Section 4 provides some concluding remarks and outlines future lines of research.

2 Methodological Framework

Let $\mathbf{x}$ denote the expenditure vector in the main destination ACs extracted from EGATUR, and let $\mathbf{b}$ denote the vector built from foreign payment card transactions in the ACs. Because the expenditures recorded in $\mathbf{x}$ do not correspond to the *real* spatial distribution of spending, we estimate those real expenditures by first estimating a reallocation matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & \cdots & p_{1,n} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,n} \end{pmatrix},$$

whose elements p_{ij} take values in $[0, 1]$ and whose column sums are equal to 1. The rows correspond to *actual* destination ACs, whereas the columns correspond to *main* destination ACs, so each p_{ij} represents the proportion of expenditure assigned to the main destination AC j that was actually incurred in the actual destination AC i.

Once $\mathbf{P}$ is estimated, we compute the estimated real-destination expenditure vector $\mathbf{g}$ (we denote by $\hat{\mathbf{P}}$ and $\hat{\mathbf{g}}$ their estimates.) (Boyd and Vandenberghe, 2018):

$$\hat{\mathbf{g}} = \hat{\mathbf{P}}\mathbf{x}.$$

The following considerations apply to our estimation:

- The data span the years 2018–2023. For each year, observations are grouped into 14 countries (or country groups) and 19 ACs ($n = 19$) (Instituto Nacional de Estadística (INE), 2025b).
- Since card transaction data are confidential, access is restricted and the analysis is carried out in a secure environment provided by INE. To test the methods on personal computers, we generated a vector b from the total annual expenditure reported in EGATUR and distributed across the ACs according to the proportion of overnight stays in each AC. The overnight-stay proportions are also computed from EGATUR.
- Although in this case the total main-destination expenditure (i.e., $\sum_i x_i$) equals the total payment-card amount (by construction), in general these totals may differ. In such cases we work with *percentage* vectors $\mathbf{x}$ and $\mathbf{b}$. For readability we keep the same symbols, but it is understood that real-amount $\mathbf{x}$ would be used when getting the estimation of vector g.
- A matrix is estimated for each country and, in addition, an aggregate (global) matrix—hereafter referred to as the "global case".

Basically, the methods employed in the process of estimating $\mathbf{P}$ were constrained least squares and corrected versions of the overnight-stay proportion matrix. These methods are detailed below.

After obtaining $\hat{\mathbf{g}}$ with each technique, we compare the results against previously simulated scenarios to quantify the goodness of fit of those methods.

It should be noted that the processing and preparation of the data were carried out using R (R Core Team, 2024), whereas the formulation and solution of the models were implemented in Python (Python Software Foundation, 2025) with the Pyomo optimization modeling library (Bynum et al., 2021; Hart et al., 2011).

3 Estimation Methods

3.1 Proportion Matrices Based on Overnight Stays

Assuming travelers spend more where they stay longer, a first estimate of $\mathbf{P}$ is the matrix of overnight-stay proportions, denoted **perno**:

$$\hat{\mathbf{P}} = \mathbf{perno} = \begin{pmatrix} \text{perno}_{1,1} & \cdots & \text{perno}_{1,n} \\ \vdots & \ddots & \vdots \\ \text{perno}_{n,1} & \cdots & \text{perno}_{n,n} \end{pmatrix}.$$

This matrix is obtained by column-normalizing the overnight-stay matrix **Pe**, built from EGATUR:

$$\mathbf{Pe} = \begin{pmatrix} Pe_{1,1} & \cdots & Pe_{1,n} \\ \vdots & \ddots & \vdots \\ Pe_{n,1} & \cdots & Pe_{n,n} \end{pmatrix},$$

where Pe_{ij} is the number of overnight stays in actual destination AC i that have been assigned to main destination AC j.

This method serves as a first approximation to the estimation of real-destination tourist expenditure (Frade-Amil et al., 2024). However, because our baseline proposal leverages payment-card information for the estimation, we temporarily set **perno** aside and return to it for corrected variants below.

3.2 Least Squares with Equality and Inequality Constraints (LSEI)

Our first optimization approach solves the following least squares problem with equality and inequality constraints (LSEI):

$$\begin{aligned} \min \quad & \|\mathbf{Px} - \mathbf{b}\|^2 \\ \text{s.t.} \quad & p_{ij} \geqslant 0, \quad \forall i, j \in \{1, \ldots, n\}, \\ & p_{ij} \leqslant 1, \quad \forall i, j \in \{1, \ldots, n\}, \\ & \sum_{i=1}^{n} p_{ij} = 1, \quad \forall j \in \{1, \ldots, n\}. \end{aligned}$$

3.3 LSEI with Additional Constraints

Starting from the LSEI model, two new models can be obtained by adding two sets of constraints: distances between ACs and the number of overnight stays in the ACs for a given main destination AC.

Distance-Based Constraints

Let $\mathbf{D} = (D_{ij})$ be the symmetric matrix of distances between ACs[1]. For main destination j and two actually visited ACs i and k, we impose:

$$p_{ij} \geqslant p_{kj} \quad \text{if} \quad D_{ij} < D_{kj}.$$

If distances are equal, we impose equality:

$$p_{ij} = p_{kj} \quad \text{if} \quad D_{ij} = D_{kj}.$$

[1] The distances between Autonomous Communities (both between capitals and between geographic centers) were calculated using Distance.to (2024).

The LSEI model then becomes

$$\begin{aligned} \min \quad & \|\mathbf{Px} - \mathbf{b}\|^2 \\ \text{s.t.} \quad & p_{ij} \geqslant 0, \;\; p_{ij} \leqslant 1, \;\; \sum_i p_{ij} = 1, \quad \forall i, j, \\ & p_{ij} \geqslant p_{kj} \;\text{ if }\; D_{ij} < D_{kj}, \quad \forall i, j, k, \; i \neq k, \\ & p_{ij} = p_{kj} \;\text{ if }\; D_{ij} = D_{kj}, \quad \forall i, j, k, \; i \neq k. \end{aligned}$$

Overnight-Stay Constraints

Analogously, if $Pe_{ij} > Pe_{kj}$ for a given main destination j,

$$p_{ij} \geqslant p_{kj}.$$

When $Pe_{ij} = Pe_{kj}$, we impose $p_{ij} = p_{kj}$. Now, the constrained problem is the following one:

$$\begin{aligned} \min \quad & \|\mathbf{Px} - \mathbf{b}\|^2 \\ \text{s.t.} \quad & p_{ij} \geqslant 0, \;\; p_{ij} \leqslant 1, \;\; \sum_i p_{ij} = 1, \quad \forall i, j, \\ & p_{ij} \geqslant p_{kj} \;\text{ if }\; Pe_{ij} > Pe_{kj}, \quad \forall i, j, k, \; i \neq k, \\ & p_{ij} = p_{kj} \;\text{ if }\; Pe_{ij} = Pe_{kj}, \quad \forall i, j, k, \; i \neq k. \end{aligned}$$

3.4 Corrected Overnight-Stay Proportion Matrices

Although we initially set **perno** aside, it can serve as a basis for corrections that incorporate card information. We explore two corrections: one based on *discrepancies* and another adding *centrality measures*.

Discrepancy-Based Correction

Starting from **perno**, we compute an initial actual expenditure estimation $\hat{\mathbf{g}}_{\text{perno}} = \mathbf{perno} \cdot \mathbf{x}$ and define a row-wise correction ratio that aligns $\mathbf{g}_{\text{perno}}$ with $\mathbf{b}$. Let $\mathbf{r}$ denote the vector obtained as the elementwise ratio between $\mathbf{b}$ and $\mathbf{g}_{\text{perno}}$ (Boyd and Vandenberghe, 2018):

$$\mathbf{r} = (r_1, \ldots, r_n) = \left(\frac{b_1}{g_{\text{perno},1}}, \ldots, \frac{b_n}{g_{\text{perno},n}} \right) = \frac{\mathbf{b}}{\mathbf{g}_{\text{perno}}}.$$

Then, we build the diagonal matrix $\mathbf{R} = \text{diag}(r_1, \ldots, r_n)$. Applying a left multiplication and then renormalizing each column we obtain the following expression:

$$\hat{P}_{ij} = \frac{r_i \, \text{perno}_{ij}}{\sum_{i=1}^{n} r_i \, \text{perno}_{ij}}.$$

Centrality-Weighted Correction with Discrepancy

If we model the problem as a graph on ACs, we can compute a centrality measure for each node (e.g., PageRank, eigenvector), aggregated into $\mathbf{c} = (c_1, \ldots, c_n)$ (Bender and Williamson, 2010). Then, we define $\mathbf{C} = \text{diag}(c_1, \ldots, c_n)$ and apply it to **perno**, renormalizing by columns:

$$\hat{P}_{ij} = \frac{c_i \, \text{perno}_{ij}}{\sum_{i=1}^{n} c_i \, \text{perno}_{ij}}.$$

Then, we compute the estimation $\hat{\mathbf{g}}_{\text{perno\_cent}} = \hat{\mathbf{P}}\mathbf{x}$ and compute the vector $\mathbf{r}$ as above, applying a second row-wise correction with column renormalization:

$$\hat{P}'_{ij} = \frac{r_i \, P_{ij}}{\sum_{i=1}^{n} r_i \, P_{ij}}.$$

3.5 Method Comparisons

After obtaining results with the aforementioned methods, we compare them to simulated "real" scenarios. Using payment-card data (or synthetics derived from EGATUR), for each country of residence (and for the global case) we gather all available years (in percentages) and construct two bounding vectors: the elementwise minimum across years and the elementwise maximum. We then simulate vectors within the min–max range, adding the constraint that elements sum to 100% within a tolerance (e.g., 10^{-4}).

Each method's estimated real-destination vector is compared to each simulated scenario using appropriate metrics (e.g., RMSE) and/or decision criteria to identify the most suitable method.

4 Conclusions

In this work we have addressed the problem of estimating real tourist expenditure in Spain, arising from the limitation of EGATUR, which assigns all expenditure to the main destination community. To overcome this, we have proposed different redistribution methods based on the construction of a reallocation matrix. Among them, constrained least squares and corrected overnight-stay proportion matrices stand out, as they allow the incorporation of additional information and yield more coherent estimates.

At present, the validation of these methods with the real data described in Section 3 is underway. As future work, we plan to incorporate additional techniques such as Data Envelopment Analysis (DEA) (Cooper et al., 2007) to compare the performance of the different methods.

Acknowledgements

This work has been partially funded by the INE project "*Ciencia e ingeniería de datos para la mejora de la función estadística oficial*" (CIDMEFEO). It has also been supported by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2024/14) and by CITIC, as a research center accredited for excellence within the Galician University System and member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

Bibliography

E. A. Bender and S. G. Williamson. *Lists, Decisions and Graphs: With an Introduction to Probability*. 2010. Course notes, University of California, San Diego.

S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-47122-1.

M. L. Bynum, G. A. Hackebeil, W. E. Hart, C. D. Laird, B. L. Nicholson, J. D. Siirola, J.-P. Watson, and D. L. Woodruff. *Pyomo – Optimization Modeling in Python*, volume 67 of *Springer Optimization and Its Applications*. Springer, 3rd edition, 2021. ISBN 978-3-030-68935-1. doi: 10.1007/978-3-030-68935-1.

W. W. Cooper, L. M. Seiford, and K. Tone. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. Springer, New York, 2nd edition, 2007. ISBN 978-0-387-45283-8. doi: 10.1007/978-0-387-45283-8.

Distance.to. Calculador de distancias. *https://es.distance.to/*, 2024. Accessed November 2024.

D. Frade-Amil, M. O. de la Fuente, S. Naya, J. Tarrío-Saavedra, and M. Francisco-Fernández. The distribution of the tourist expenditure in spain among the visited autonomous communities: a first approach. In M. L. Rodríguez, T. V. Rodeiro, J. Pereira-Loureiro, and M. F. G. Penedo, editors, *Proceedings XoveTIC 2024: Impulsando el talento científico*, pages 161–168, 2024.

W. E. Hart, J.-P. Watson, and D. L. Woodruff. Pyomo: Modeling and solving mathematical programs in python. *Mathematical Programming Computation*, 3(3):219–260, 2011. doi: 10.1007/s12532-011-0026-8.

Instituto Nacional de Estadística (INE). *Estadística de Movimientos Turísticos en Frontera y Encuesta de Gasto Turístico (FRONTUR-EGATUR): Metodología*, March 2025a. URL *https://www.ine.es/daco/daco42/frontur/frontur_egatur_metodologia.pdf*.

Instituto Nacional de Estadística (INE). Encuesta de gasto turístico (egatur), 2025b. URL *https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177002&menu=ultiDatos&idp=1254735576863*.

Python Software Foundation. The python language reference, version 3.x. *https://docs.python.org/3/reference/*, 2025. Accessed January 2025.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL *https://www.R-project.org/*.