

Full Length Article

Estimating multicountry tourism flows by transport mode

Carlos Llano^{a,*}, Juan Pardo^b, Santiago Pérez-Balsalobre^c, Julián Pérez^b^a Departamento de Análisis Económico: Teoría Económica e Historia Económica & Instituto "L.R.Klein"-UAM & CEPREDE, Facultad de CC.EE y EE, Módulo E-1, Universidad Autónoma de Madrid, campus Cantoblanco, 28049 Madrid, Spain^b Departamento de Economía Aplicada & Instituto "L.R.Klein"-UAM & CEPREDE, Facultad de CC.EE y EE, Módulo E-XIV, Universidad Autónoma de Madrid, Campus Cantoblanco, 28049 Madrid, Spain^c Departamento de Economía Aplicada & L.R. Instituto "L.R.Klein"-UAM & CEPREDE, Facultad de CC.EE y EE, Módulo E-XIV, Universidad Autónoma de Madrid, Campus Cantoblanco, 28049 Madrid, Spain

ARTICLE INFO

Article history:

Received 8 February 2023

Received in revised form 4 October 2023

Accepted 13 October 2023

Available online 4 November 2023

Associate editor: yang Yang

JEL codes:

R12,

L8,

L83

Keywords:

International tourism

Domestic tourism

UNWTO

Transport mode

Gravity model

ABSTRACT

This article describes the methodology used to obtain a harmonized database of country-to-country flows of tourists and visitors built upon the UNWTO original database. Focusing on 74 countries for the period 1995–2018 and starting with data on outbound tourism, we develop a methodology using other alternative indicators from the UNWTO which fills the gaps related to two main indicators (Foreign Visitors by country of Residence and Foreign Tourists by country of Residence). Subsequently, intra-national domestic tourism and the transport mode specific flows are obtained. The paper finally contrasts the robustness of the results obtained with the gravity equation for all flows, thereby confirming the solvency of the process and the final data retrieved.

© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

Introduction

Transport connectivity is a key factor explaining the competitiveness of touristic sites, by enhancing accessibility and improving consumer experience (Zamparini et al., 2022). Transportation is also the largest source of GHG emissions from tourism (Dickinson et al., 2011; Lee et al., 2022; Peeters & Dubois, 2010; Scott & Gossling, 2022), being the transition towards the cleaner transportation options a key driver of the contribution of the tourist sector to the global action against the climate change.

Although such ideas are commonly accepted, it is still very hard to develop empirical analysis using cross-country tourist flows considering the transport-mode dimension of the trips. Indeed, most of the analysis conducted till now focus on specific countries

* Corresponding author.

E-mail addresses: carlos.llano@uam.es (C. Llano), juan.pardo@uam.es (J. Pardo), santiago.perez@uam.es (S. Pérez-Balsalobre), julian.perez@uam.es (J. Pérez).

or cities (Hough & Hassanien, 2010; Kelly et al., 2007). The scarcity of empirical analysis including the transport-mode dimension in a global setup in the field of tourism, contrast with the growing literature on transport of goods, where data is more abundant and consolidated (Cherniwchan et al., 2017; Cristea et al., 2013; Ortiz et al., 2021). Moreover, the transport mode dimension is also relevant in the domestic displacements (Masiero & Zoltan, 2013), which are more frequent than the international ones (Carril-Caccia et al., 2022), and where the use of private car, one of the most polluting modes, is more prevalent.

In this line of thoughts, sustainability is becoming a new paradigm in tourism research (Buckley, 2012; Higham et al., 2016; Miller & Torres-Delgado, 2023; Prillwitz & Barr, 2011; Rehman et al., 2023). In a recent and insightful review, Sun et al., 2022, discuss whether tourism increases carbon emissions or not, revising 81 scientific articles between 2013 and 2021. Their results indicate a low consensus on the tourism-emissions nexus, with contradictory results. Beyond this debate, this study is very clear on the incompleteness of most of the analysis and give clear directions for future research. None of the 81 articles revised included international aviation emissions in their modelling, even though this is a major driver of tourism emissions (Lenzen et al., 2018). For the tourism variable, the majority use the inbound travel, using either tourism receipts or arrivals as parameters. Domestic tourism is largely ignored even though it contributes to 72 % of total global tourism expenditure in 2019. Airfare paid by international travel to national carriers is counted but emissions from international flights are excluded.

The main conclusion of this review is that most of current tourism-environment analysis conducted to date are characterized by omissions that make results questionable. Such omissions relate mainly to the absence of the imports made by the tourism sector (i.e., transportation) and claim for the incorporation of emissions from international air transport and shipping. Moreover, the explicit inclusion of the domestic tourism, which is heavily dependent on road transportation is also a must (Martín-Cejas & Sánchez, 2010). Furthermore, adding the various levels of carbon intensity, depending on average travel distance, transport mode, and the luxury of services (Becken & Simmons, 2002) is also critical. Such conclusions are in line with the ones of another classic reference to the topic (Hunter & Shaw, 2007), which claimed for the need to incorporating different modes of transport and the domestic tourism activities, remarking that, “perhaps the greatest need, however, is to collect ‘real world’ primary data”.

In line with this last reference, we firmly believe that most of these omissions are data-driven. To the best of our knowledge, no database exists offering detailed information about domestic and inter-country flows of tourism, spanning all countries and years and applying a homogeneous definition of the tourist category and the transport mode used in the displacement. Thus, several key issues related to the tourism-transport-environment nexus remain secluded, at least if the aim is to give solid answers to global questions. Such link cannot be answered using local or even national datasets, which are unable to capture critical dimensions such as the potential substitutability of different transport modes in the mobility of international and domestic tourists.

The United Nations World Tourist Organization (herein after, UNWTO), however, offers different databases including a large set of indicators related to inbound and outbound tourism, covering a large panel of countries and years. It also publishes additional datasets about domestic tourism, and certain dimensions related to the transport mode used. These datasets have been widely used in the empirical analysis of tourism (Gil-Pareja et al., 2007; Vietze, 2012; McKercher & Prideaux, 2014; Pratt & Tolkach, 2018; McKercher & Mak, 2019; Khalid et al., 2020; Provenzano, 2020; Rosselló et al., 2020; Gössling et al., 2021 just to cite some recent and salient papers).

In some of these articles, the indicators published by the UNWTO or the criteria adopted by some of the reporting countries are criticized (Frechtling & Hara, 2016; Pratt & Tolkach, 2018; Yang et al., 2019). In line with them, in this article we want to remark that the UNWTO dataset suffers from major drawbacks that can be summarized as follows: i) It does not offer a unique common indicator regarding the bilateral movement of tourists shared by all countries over a stable time window. In some cases, countries report “visitors” while others report “tourists”, the latter being defined as a visitor that stays overnight in the host country. ii) Moreover, when reporting the number of tourists or visitors, these are classified according to their *nationality* or their country of *residence*, two concepts that are highly correlated, but are not the same. iii) On top of that, the information provided for the bilateral international tourism is measured in aggregate terms, using the number of people as measurement unit, and adopting as reference the information about the *outbound tourism* of each country. Meanwhile, the information about the type of transportation, motive of the trip, expenses, etc., corresponds to *inbound tourism*, using the number of trips as a measurement unit, without reporting the travelers' country of origin. iv) Finally, the complementary data on domestic tourism is more sparse and include a larger number of gaps.

The aim of this paper is to develop a database employing homogeneous information about country-to-country flows of people, containing tourists as well as visitors, constructed upon the main indicators provided by the UNWTO original database. Moreover, such a dataset incorporates the intra-national tourism (domestic) and identifies the most likely transport mode used in the displacements of tourists, both, within and between countries.

We commence with the data on outbound tourism from the UNWTO, focusing on 74 countries and the years 2010–2018 within the period 1995–2018. Using the observable data already published by the original UNWTO dataset, without any other information or behavioral assumptions, we estimate 34 alternative models aiming to calculate and predict two main indicators, namely, Foreign Visitors by country of Residence (VFR) and Foreign Tourists by country of Residence (TFR). As a methodological option, the approach is justified by the desire to produce a database that is based, as far as possible, on survey data, allowing the output data produced to be modeled afterwards with the standard empirical quantitative approaches used in the field of tourism and travel decisions (i.e., a gravity equation).

As a second step, we add the original information on domestic tourism from the UNWTO and then complete the missing values by means of linear interpolation and forecasting. In a third step, the international and domestic flows for tourists

(overnight stays) and visitors, are split by five transport modes, using the transport-mode structure reported by the UNWTO for the inbound international tourism and the domestic movements. On purpose, no additional indicators out of the UNWTO are used when splitting the transport modes. Thus, the resulting dataset can be confronted with alternative sources (i.e., passenger movements, transport infrastructures, etc.) or modeled using causal frameworks.

We illustrate the contribution of our approach by focusing on the aggregate tourism flows across a selection of 32 European countries (the EU27, Switzerland, Liechtenstein, Iceland, Norway, and the UK). The methodology employed produces a dataset with only 10 % empty cells, compared to the 80 % of the original UNWTO dataset (more details are given in the Appendix). This data completion exercise fills in those flows of the cells for which the UNWTO original dataset provides survey-based indicators.

The article provides several graphical and econometric checks and finally contrasts the robustness of the results obtained with the gravity equation. Further robustness checks are provided using other quantitative methods (i) “multiple imputation” techniques; ii) the STATA *xtselvar* command for panel data model evaluation (Ugarte-Ruiz, 2020a, 2020b).

Our approach has been presented as an exercise which extracts as much information as possible from the dispersed data contained in the original UNWTO, to achieve a more accurate dataset. The latter is essential for analyzing the international and intra-national flows of tourism with the detail required to address relevant current issues (e.g., the substitution of domestic versus international tourism in times of additional mobility constraints such as the COVID19 pandemic, transport mode competition affecting traveling decisions worldwide, or the environmental implications of transport mode mix, given their different carbon footprints).

The process followed suggests that the information related to tourists and visitors is solid enough for the purpose of constructing a standard set of indicators across most countries. The results obtained verify a solid pattern of expansion of intra-national and inter-national tourism flows worldwide. The latter also indicate the presence of a clear *home bias* in favor of the domestic market. The econometric analysis also validates the likelihood of the final dataset, obtaining an outstanding performance of the gravity equation.

The structure of the rest of the paper is as follows: section two reviews the literature; section three describes the data; section four reports the methodology used to fill the gaps in the current UNWTO dataset; Section five reports the results obtained, comparing them with alternative indicators and testing them using the gravity equation. A final section concludes. A complementary online Appendix is also provided, where we offer additional information about the data used, and the results obtained, discussing in detail the solidity of the methods used, and comparing them with two alternative quantitative approaches.

Review of the literature

The aim of this section is to briefly contextualize our empirical analysis by summarizing three different strands of the literature: i) the relevance of the UNWTO dataset in empirical analysis of tourism; ii) the link between transportation and tourism; iii) the literature on forecasting tourist flows, and other techniques to handle missing data.

International tourist flows using the UNWTO dataset

As commented, the UNWTO dataset in international (inbound and outbound) and domestic tourism is probably the most solid dataset for conducting worldwide empirical analysis on tourism flows (Gil-Pareja et al., 2007; Gössling et al., 2021; McKercher & Mak, 2019; McKercher & Prideaux, 2014; Pratt & Tolkach, 2018; Rosselló et al., 2020; Vietze, 2012). Moreover, some of these articles adopt the gravity equation as the preferred quantitative framework to model bilateral flows. This section aims to summarize the literature and to contextualize our own contribution.

For example, Yang et al. (2019), analyses the obstacle of cultural distance in international tourist flows, based on data from UNWTO, commenting on two issues worth considering with respect to their research data: the first issue is that the countries of destination do not report the same countries of origin. For example, the US reports flows from >200 different origins while Argentina only reports 6. The second question, which is what motivates our research, is that they detect a significant heterogeneity in the definition when compiling tourist arrivals in each country, having to prioritize some indicators over others, due to the lack of homogeneity.

Similarly, Pratt and Tolkach (2018), uses the UNWTO from a more critical perspective, discussing the effect of the politicization of certain statistical categories. This article highlights three areas where tourism statistics appear misleading, serving industry and government interests.

McKercher and Mak (2019), uses the UNWTO arrival and departure data for 2016 to analyses the impact of distance on international tourism demand. Provenzano (2020), analyses the relationship between immigration and tourism through complex-network analysis and gravity models. The analysis is based on the UNWTO database, completed with official data from the statistical institutes of the countries analyzed, and concentrates on the indicator of visitors who spend at least one night outside their country of origin. Similarly, Khalid et al. (2020), analyze the effect of different financial and economic crises on international tourism flows through a gravity model of 200 countries from 1995 to 2010, using UNWTO visitor arrivals as the preferred indicator.

Finally, Shao et al. (2023), when investigating the relationship of economic globalization and international tourism within the network theory, use the UNWTO inbound and outbound tourist flows for a panel data for 47 countries from 1995 to 2018. When doing so, no hints are reported about the specific indicator used.

Transport and tourism

We now focus on the literature connecting tourism with transportation, with a clear focus on the transport mode choice of the international and domestic displacements.

To this regard, Kelly et al. (2007), outlines an approach for examining tourist-destination travel mode choices and forecasting the resulting environmental impact of those preferences. Using the tourism destination of Whistler, British Columbia, the article initially describes a discrete-choice experiment used to estimate tourist mode-choice behavior under different transportation-planning scenarios. Then, they plug these results in an energy-use model to obtain GHG emissions.

Complementary, Khadaroo and Seetanahb (2008), analyze the role of transport infrastructure in international tourism, using a gravity model and the UNWTO dataset, and considering a sample of 156 destinations from each one of the G-7 countries.

More recently, Masiero and Zoltan (2013), proposes that movement patterns and transportation mode choices are linked. Similarly, Thrane (2015) analyses the long-distance transportation mode choices of tourists. Their analysis focus on the domestic tourism market in Norway, and the Norwegians' winter vacation trip.

Finally, Zamparini et al. (2022), analysis the determinants of green mobility choices taken by young adults during their holidays. Using a survey conducted for Italian and Spanish University students, they show the intrinsic relationship between the transport mode choices at home and at tourism destinations. Such result is interesting to our goals, giving support to our extrapolation of one's country transport mode choices for the aggregate inbound flows (outbound) to each of its bilateral inflows (outflows).

Forecasting and missing data imputation techniques

To revise the literature on forecasting and missing data treatment, we recommend the curated collection in Annals of Tourism Research (Song et al., 2019), where we find all kind of examples: from time series analysis, spatial econometric approaches, structural equations to artificial intelligence models (Peng et al., 2014; Wan & Song, 2018; Yang & Zhang, 2019). Usually, these techniques are described as alternatives or complements to causal models or, as in our approach, as useful tools for dealing with missing data.

In line with the later, a key paradigm in the treatment of missing data is "multiple imputation" (MI), increasingly used in health and social science (Rubin, 1976; Rubin, 1996; Reiter & Raghunathan, 2007; Akande et al., 2017; White et al., 2011). In MI, the analyst creates multiple, completed datasets by replacing the missing values using predictions from the observed data. As commented by Akande et al. (2017), a wide range of routines are available such as "AMELIA II" in R (Honaker et al., 2011), proc. "MI" in SAS (Yang, 2011) or "MI" and "ICE" (MICE) in Stata (Royston et al., 2009; Wagstaff & Harel, 2011; Aloisio et al., 2014; Comulada, 2015).

The application of MI in tourism is scarce (Chujai et al., 2020; El Esawey, 2020). Finally, a word is devoted to a recent command (*xtselvar*) developed for Stata, able to compare the performance of alternative variables, specially defined for panel data frameworks (Ugarte-Ruiz, 2020a, 2020b). Both, the MI and the *xtselvar* are used in this article as robustness checks, giving strong support to our own approach.

Data

As commented, we adopt the UNWTO datasets on domestic and international tourism flows as the best official source for a global analysis of tourist flows. As commented in the Appendix, eight indicators exist (TFN, TFR, VFN, VFR, THSN, THSR, TCEN, TCER), although no country covers all of them. Our empirical analysis focuses on the period 1995–2018 of 74 countries. Regarding the international flows, we use the "outbound tourism" dataset. In addition, we also use two other relevant sections of the original UNWTO dataset: i) The information regarding intra-national (domestic) flows, which covers the same countries and years, but uses the number of trips rather than arrivals as unit of measurement. ii) We also use the information related to the transport mode used in the displacement. In this case, the information corresponds to inbound tourism, it is reported as "number of trips", and it is aggregated for each destination country, without reporting the country of origin. For practical reasons, it is assumed that the three concepts of travelers, trips, and arrivals are aligned.

Methodology

In this section, we describe the methodology followed in the estimation of our own dataset. First, we describe our choice regarding the two target indicators. Next, we describe the methodology applied to: i) the international origin-destination flows (using data on outbound tourism); ii) the domestic tourist flows; iii) the transport mode utilized (using data about domestic and inbound trips).

Our two target indicators are TFR (tourists with overnight stays) and VFR (visitors). We opt for "residency" rather than "nationality" deeming that the former is more convenient for further research. We also want to remark that TFR and VFR are the most general indicators.

In sum, our methodology aims at obtaining the most comprehensive panel of bilateral VFR and TFR flows possible of for the 74 countries included in the sample. Such methodology is synthesized three steps, with different sub-steps:

- Step 1. Treatment of International bilateral visitors and tourists.
- Step 2. Treatment of Domestic Tourism.
- Step 3. Disaggregation of international and domestic flows by transport mode.

Step 1: treatment of international bilateral visitors and tourists

This section describes the method used for estimating inter-country flows. We estimate several econometric models using the period 1995–2017, omitting the year 2018 just in case is needed for evaluating out-of-sample forecasting capacity of each model. The final version of the database will use the entire period for the final estimation and forecast.

The first step is a simplifying assumption, which consists in presuming that for any foreign visitor its country of residence is the one of nationality. This permits a reduction in the range of eight indicators (VFR_{ijt} , TFN_{ijt} , TFR_{ijt} , VFN_{ijt} , $THSN_{ijt}$, $THSR_{ijt}$, $TCEN_{ijt}$, $TCER_{ijt}$) to four (VFR_{ijt} , TFR_{ijt} , $THSR_{ijt}$, $TCER_{ijt}$), something that allows obtaining larger samples when pairs of countries share the same indicators in each year. Such decision was taken after testing how countries opt alternatively for the dimensions of “residence” or “nationality”, but rarely for both (the original UNWTO dataset does not report countries having the VFR (or the TFR) together with some other indicators in the same year. This is the case of: i) VFR_{ijt} vs TFN_{ijt} ; ii) VFR_{ijt} vs $TCEN_{ijt}$; iii) VFR_{ijt} vs $THSN_{ijt}$; iv) TFR_{ijt} vs $TCEN_{ijt}$).

Next, focusing on two target indicators of bilateral flows, we define Eqs. (1) and (2), able to explain all the different combinations considered:

$$\ln VFR_{ijt} = \beta_0 \ln (\text{Alternative indicator})_{ijt} + t + \varepsilon_{ijt} \quad (1)$$

VFR_{ijt} is the total number of visitors from a country of residence i to country of destination j in year t . ε_{ijt} denotes the classical disturbance term, while t corresponds to a trend dummy that starts with value 1 in 1995 and monotonically increases in one unit every year until the end of the period. Regarding the element “Alternative indicator”, corresponds to the other three indicators that might be available within the UNWTO original dataset for the same destination country (j) in year t : TFR_{ijt} , $THSR_{ijt}$, $TCER_{ijt}$. We want to remark that each model has been tested with the eight original indicators, that is TFN_{ijt} , TFR_{ijt} , VFN_{ijt} , $THSN_{ijt}$, $THSR_{ijt}$, $TCEN_{ijt}$, $TCER_{ijt}$, although the final reduction to the four selected ones offers the best results.

$$\ln TFR_{ijt} = \beta_0 \ln (\text{Alternative indicator})_{ijt} + t + \varepsilon_{ijt} \quad (2)$$

Similarly, TFR_{ijt} is the total number of tourists from country of residence i to country of destination j in year t . In this case, the element “Alternative indicator”, corresponds to any of the other three indicators for the same destination country (j) in year t : VFR_{ijt} , $THSR_{ijt}$, $TCER_{ijt}$.

Our assumption is that, when TFR_{ijt} (or VFR_{ijt}) is not available in the original dataset, an alternative flow between the same triad ijt is a good predictor of the target one, since it corresponds to a relevant share of the desired (but not observed) target indicator. For example, if the number of visitors received by Italy with origin in Germany in 2010 is not observable, a good predictor would be the number of overnight stays of tourists with residence (or nationality) in Germany staying in hotels in Italy in 2010.

Each of these two equations has been estimated using an OLS pooled regression as well as fixed effects (FE) and random effects (RE) panel data estimators. All these methods are valid knowing that the samples used do not include any zero flows. Depending on the quality of the results obtained and the forecasting accuracy, just six models are selected, using a complementary treatment for the values still missing. We also wish to remark that our approach will not produce a positive flow where there is no actual information related to any of the eight original indicators. So, even after applying our approach, any pair of countries with no flows in the original UNWTO will remain as a missing value. By contrast, any pair of countries with a positive value of these eight indicators will finally have a (reported or predicted) value for VFR_{ijt} and TFR_{ijt} .

In some cases, the previous procedure generates predictions where $VFR_{ijt} < TFR_{ijt}$, something that should be discarded since, by definition, the number of visitors is greater than the number of tourists. In these cases, a final correction is applied by means of Eqs. (3) and (4) for every observation where the raw prediction TFR_{ijt} is larger than the observed or predicted (VFR_{ijt}). To do so, we obtain a ratio between VFR_{ijt}/TFR_{ijt} by means of Eq. (3):

$$\left(VFR_{ijt}/TFR_{ijt} \right) = \beta_0 + t + \varepsilon_{ijt} \quad (3)$$

This Eq. (3) is estimated using OLS pooled regression with the trend dummy t , considering solely a sub-sample that satisfies the following conditions: VFR_{ijt} and TFR_{ijt} are not missing values, and $VFR_{ijt} > TFR_{ijt}$. Thus, combining the previous ratio with TFR_{ijt} it is possible to obtain a VFR_{ijt} higher than the former.

Then, Eq. (4) describes how a predicted VFR can be obtained by combining the previous ratio and the observed value for TFR.

$$VFR_{ijt} = \left(VFR_{ijt}/TFR_{ijt} \right) * TFR_{ijt} \quad (4)$$

Note that we use TFR as the base for obtaining VFR, given the better information available for this variable, and the most straightforward interpretation of its values.

Step 2: adding intra-national tourism flows

At this second stage, we incorporate the information included in the original UNWTO dataset regarding the intra-national (domestic) flows, covering the same sample of countries and years.

Where the data exists, the UNWTO dataset provides two indicators that are assumed to be equivalent to the two target ones described for the international flows: Visitors and Tourists (stay overnight). More explanations are given in the Appendix.

Starting with the original and highly sparse data on domestic tourism, it is necessary to point out that in this case, there are no additional indicators able to allow predictions without the use of external sources. Bearing this in mind, we adopt the following approach:

First, a linear interpolation process is applied to country-year specific flows where missing values exist for certain years, but with observations available before and after.

Next, we estimate Eqs. (5) and (6) to model the available domestic flows (including the interpolated values) alongside some relevant variables able to explain the intra-national domestic tourism.

$$\ln Vis_dom_{it} = \beta_0 + \beta_1 \ln GDP_{it} + \beta_2 \ln Pop_{it} + \beta_3 \ln Dist_int_i + \beta_4 \ln Area_i + \beta_5 \ln Citynum_{it} + \beta_6 \ln Pop_Dens_{it} + \beta_7 \ln GDP_Dens_{it} + \beta_8 Landlocked_i + \varepsilon_{it} \quad (5)$$

where Vis_dom_{it} is the total number of visitors within a country i in each year, ε_{it} denotes the classical disturbance term. After trying different alternatives, our preferred specification does not include a t trend dummy. Regarding the explanatory variables, we have considered the following:

$\ln(GDP_{it})$: logarithm of the GDP in current prices. Source: CEPII. Gravity dataset.

$\ln(Pop_{it})$: logarithm of the Population. Source: CEPII. Gravity dataset.

$\ln(Dist_int_i)$: logarithm of the internal distance within each country. Source: CEPII. Gravity dataset.

$\ln(Citynum_{it})$: logarithm of the number of big cities within the country, as reported by Henderson's database. Source: CEPII. Gravity dataset.

$\ln(Pop_dens_{it})$: logarithm of the population density in the country each year, computed as the share between the Population and the Area (Source: CEPII. Gravity dataset.)

$\ln(GDP_dens_{it})$: logarithm of the GDP density in the country each year, computed as the share between the GDP and the area. This variable complements the previous one, by measuring the higher tendency to observe intra-national domestic tourism in countries possessing amore unequal distribution of economic activity, with core-periphery structures and big metropolitan areas from which people wants to evade periodically.

$Landlocked_{it}$: Countries without coasts are less likely to generate internal tourism flows. We expect less variability in the climate within these countries, with Sun & Sand tourism not being an option for the country. For the EU27 + UK + EEE sample, small countries such as Switzerland and Austria are good examples.

Similarly, Eq. (6) defines an equivalent model for Tourists.

$$\ln Tour_dom_{it} = \beta_0 + \beta_1 \ln Gdp_{it} + \beta_2 \ln Pop_{it} + \beta_3 \ln Dist_int_i + \beta_4 \ln Area_i + \beta_5 \ln Citynum_{it} + \beta_6 \ln Pop_Dens_{it} + \beta_7 \ln Gdp_Dens_{it} + \beta_8 Landlocked_i + \varepsilon_{it} \quad (6)$$

where $Tour_dom_{it}$ is the total number of domestic tourists staying overnight within a country i each year. The rest of the variables are the same as in Eq. (5).

Based on these two models, we obtain the corresponding predictions, which allow completing the domestic flows for the countries with no observations in the sample.

Step 3: split international and intra-national flows by transport mode

At this third stage, we take advantage of additional rich information provided by the original UNWTO dataset, corresponding to the transport mode used in the displacements, both for domestic visitors and tourists and international ones. This information relates to inbound tourism and is reported as "number of trips" (thousands) and does not offer the double dimension of the transport-mode structure by country of origin (nor residence or nationality).

Aware of these limitations and knowing that the final purpose is to obtain the most likely structure of transport modes used for splitting every international and domestic flow, we proceed in the following way:

For the domestic flows, we apply Eqs. (7)–(12) which implies using three alternative mode-structures in cascade for each type of flow (visitors and tourists).

$$Vis_dom_{imt} = Vis_dom_{it} \times \frac{Trips_dom_{imt}}{\sum_{m=1}^m Trips_dom_{imt}} \text{ if } \sum_{m=1}^m Trips_dom_{imt} > 0 \quad (7)$$

$$Vis_dom_{imt} = Vis_dom_{it} \times \frac{Trips_dom_{im}}{\sum_{m=1}^m Trips_dom_{im}} \text{ If } \left\{ \begin{array}{l} \sum_{m=1}^m Trips_dom_{imt} = 0 \\ \sum_{m=1}^m Trips_dom_{im} > 0 \end{array} \right\} \quad (8)$$

$$Vis_dom_{imt} = Vis_dom_{it} \times \frac{Trips_dom_m}{\sum_{m=1}^m Trips_dom_m} \text{ If } \left\{ \begin{array}{l} \sum_{m=1}^m Trips_dom_{imt} = 0 \\ \sum_{m=1}^m Trips_dom_{im} = 0 \end{array} \right\} \quad (9)$$

*For simplicity, we use “0” for missing values.

Starting with domestic visitors, the transport share applied to the aggregate flow for each country “i” in the year “t” (Vis_dom_{it}) starts with Eq. (7), taking non-missing information regarding the transport mode structure reported in the original UNWTO. For each Eqs. (7)–(9), a suffix “m” denotes each of the five transport modes included in the original UNWTO, which we label as follows: *r* = road; *t* = train; *s* = ship; *a* = aircraft; *ot* = other mode by land.

When the original UNWTO does not report information about trips for this country “i” in a year “t”, but does it for other years in the sample, we apply the average transport mode structure observed for “i” considering the non-missing values for the entire period (Eq. (8)).

When Eq. (8) is also not applicable; since the original UNWTO does not report any information regarding the domestic trips transport structure for “i” along the period, we apply Eq. (9), which considers the on-average transport mode structure in the whole period (2010–2018) and countries for domestic tourism.

$$Tour_dom_{imt} = Tour_dom_{it} \times \frac{Trips_dom_{imt}}{\sum_{m=1}^m Trips_dom_{imt}} \text{ if } \sum_{m=1}^m Trips_dom_{imt} > 0 \quad (10)$$

$$Tour_dom_{imt} = Tour_dom_{it} \times \frac{Trips_dom_{im}}{\sum_{m=1}^m Trips_dom_{im}} \text{ If } \left\{ \begin{array}{l} \sum_{m=1}^m Trips_dom_{imt} = 0 \\ \sum_{m=1}^m Trips_dom_{im} > 0 \end{array} \right\} \quad (11)$$

$$Tour_dom_{imt} = Tour_dom_{it} \times \frac{Trips_dom_m}{\sum_{m=1}^m Trips_dom_m} \text{ If } \left\{ \begin{array}{l} \sum_{m=1}^m Trips_dom_{imt} = 0 \\ \sum_{m=1}^m Trips_dom_{im} = 0 \end{array} \right\} \quad (12)$$

*For simplicity, we use “0” for missing values.

Likewise, Eqs. (10)–(12) describe the equivalent procedure applied in cascade for splitting the aggregate flows of domestic tourism by transport mode.

The procedure applied to **international flows** is similar but adds the difficulty of transforming the monadic indicators on inbound transport mode for the origin and the destination countries into a dyadic variable for each pair of countries. We assume that the transport mode structure used for the trips arriving to a country (reported in the inbound statistics) is the same as the one for the trips leaving that country to a foreign destination (not reported). Hence, the transport mode structures of the two countries involved in each flow are interacted and then used to split the corresponding flow estimated in the first step of our methodology. The specific procedure is described in Eqs. (13)–(18) which again implies the use of three alternative transport-mode structures in cascade for each type of flow (visitors and tourists).

$$VFR_{ijmt} = VFR_{ijt} \times \frac{(Triplnt_{imt} + Triplnt_{jmt})}{\sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt})} \text{ if } \sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt}) > 0 \quad (13)$$

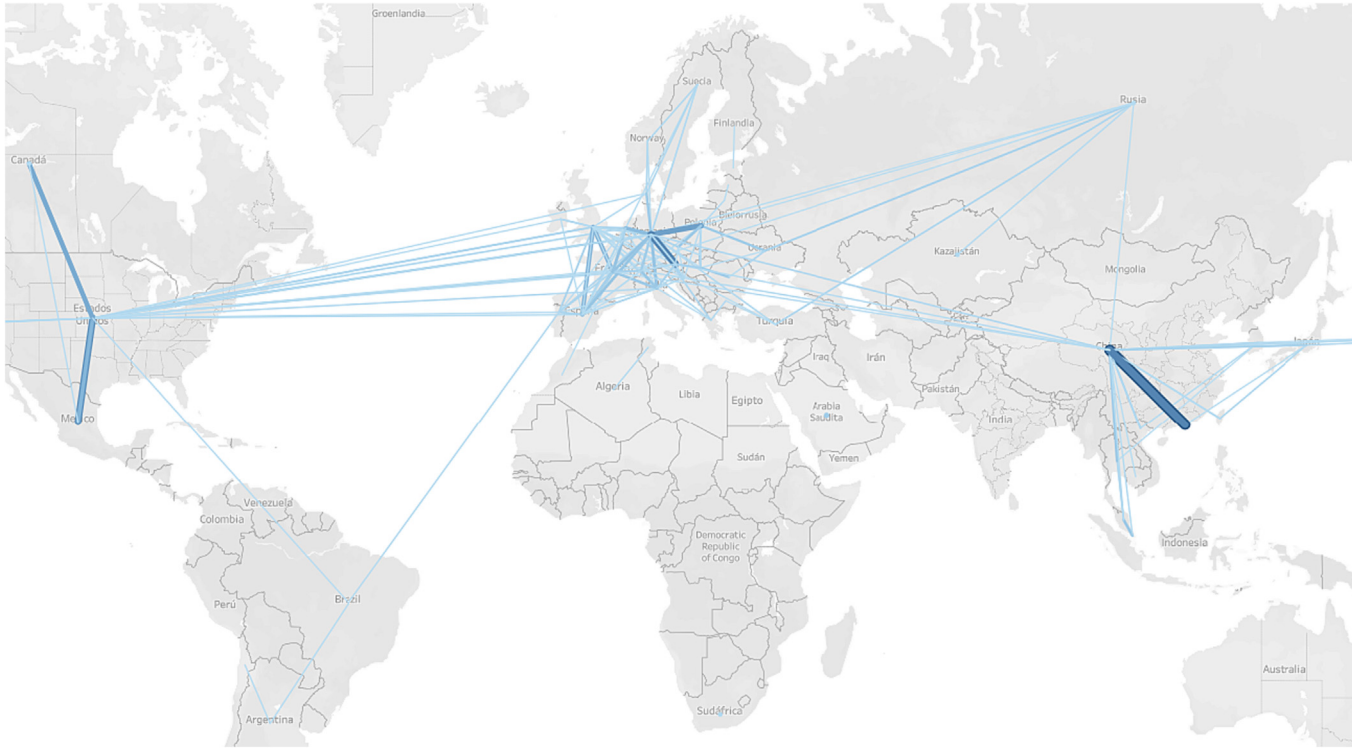


Fig. 1. . The main bilateral flows for VFR_{ijt} (flows >2 million) in 2018

Source: Own elaboration. Please, visit the following public site to interactively compare flows:

https://public.tableau.com/app/profile/author6764/viz/tourism_internacional_test_16511462531930/Dashboard1

$$\hat{VFR}_{ijmt} = VFR_{ijt} \times \frac{(Triplnt_{im} + Triplnt_{jm})}{\sum_{m=1}^m (Triplnt_{im} + Triplnt_{jm})} \text{ If } \begin{cases} \sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt}) = 0 \\ \sum_{m=1}^m (Triplnt_{im} + Triplnt_{jm}) > 0 \end{cases} \quad (14)$$

$$\hat{VFR}_{ijmt} = VFR_{ijt} \times \frac{(Triplnt_m + Triplnt_m)}{\sum_{m=1}^m (Triplnt_{mt} + Triplnt_m)} \text{ If } \begin{cases} \sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt}) = 0 \\ \sum_{m=1}^m (Triplnt_{im} + Triplnt_{jm}) = 0 \end{cases} \quad (15)$$

*For simplicity, we use “0” for missing values.

Starting with aggregate international visitors (VFR_{ijt}), the transport share applied to the aggregate flow from a country “i” to country “j” in year “t” starts with Eq. (13), taking non-missing information regarding the transport mode structure reported in the original UNWTO. As in the case of domestic flows, the Eq. (15) is applied when Eq. (14) is not possible and the latter when Eq. (13) has insufficient data.

$$\hat{TFR}_{ijmt} = TFR_{ijt} \times \frac{(Triplnt_{imt} + Triplnt_{jmt})}{\sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt})} \text{ if } \sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt}) > 0 \quad (16)$$

$$\hat{TFR}_{ijmt} = TFR_{ijt} \times \frac{(Triplnt_{im} + Triplnt_{jm})}{\sum_{m=1}^m (Triplnt_{im} + Triplnt_{jm})} \text{ If } \begin{cases} \sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt}) = 0 \\ \sum_{m=1}^m (Triplnt_{im} + Triplnt_{jm}) > 0 \end{cases} \quad (17)$$

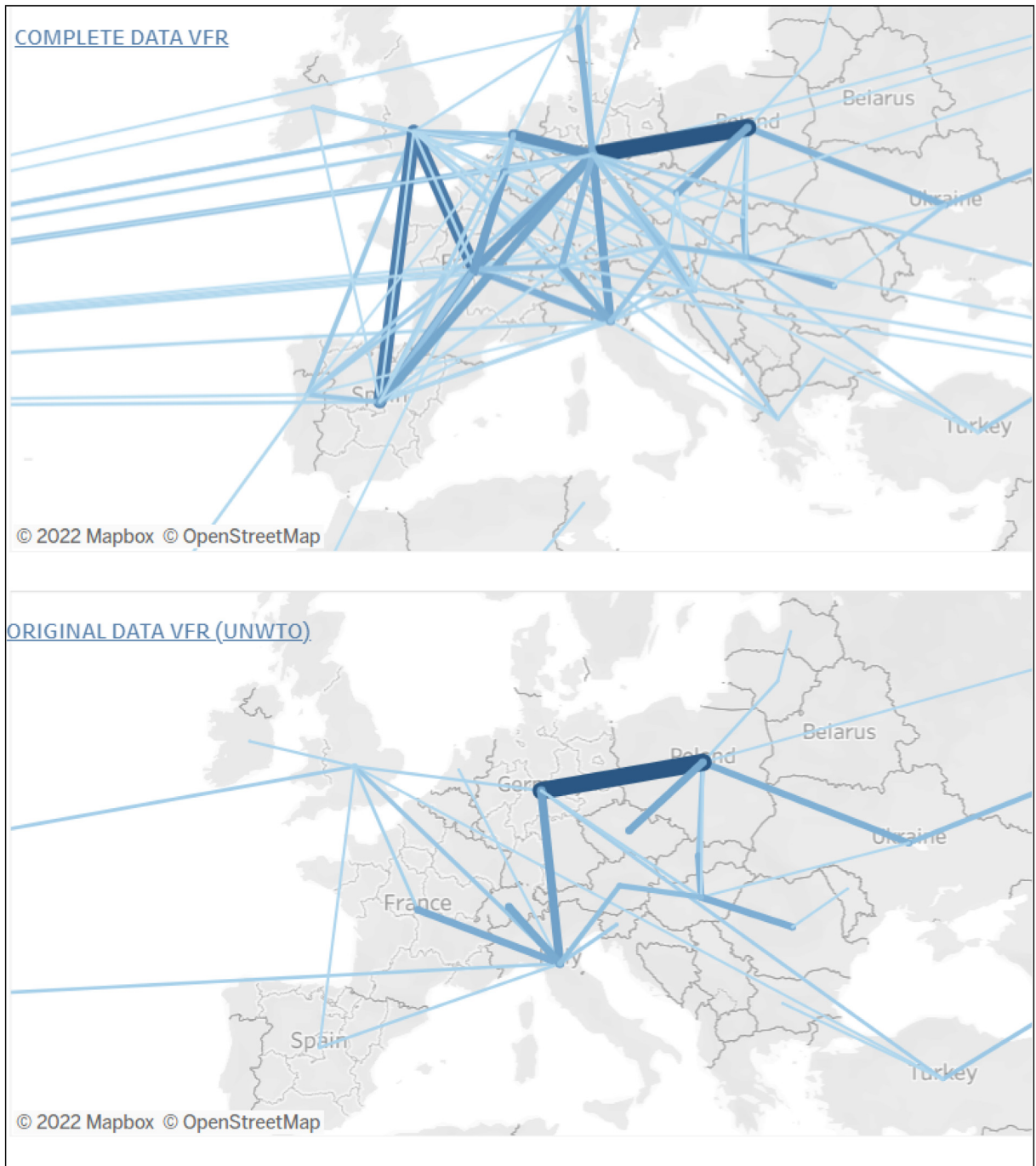


Fig. 2. . The main bilateral flows in Europe (flows >1 million) in 2018:

“original UNWTO VFR_{ijt} ” versus “complete VFR_{ijt} flows”

Source: Own elaboration. Please, visit the following public site to interactively compare flows:

https://public.tableau.com/app/profile/author6764/viz/tourism_internacional_test_16511462531930/Dashboard1

$$\hat{TFR}_{ijmt} = TFR_{ijt} \times \frac{(Triplnt_m + Triplnt_m)}{\sum_{m=1}^m (Triplnt_{mt} + Triplnt_m)} \text{ If } \left\{ \begin{array}{l} \sum_{m=1}^m (Triplnt_{imt} + Triplnt_{jmt}) = 0 \\ \sum_{m=1}^m (Triplnt_{im} + Triplnt_{jm}) > 0 \end{array} \right\} \quad (18)$$

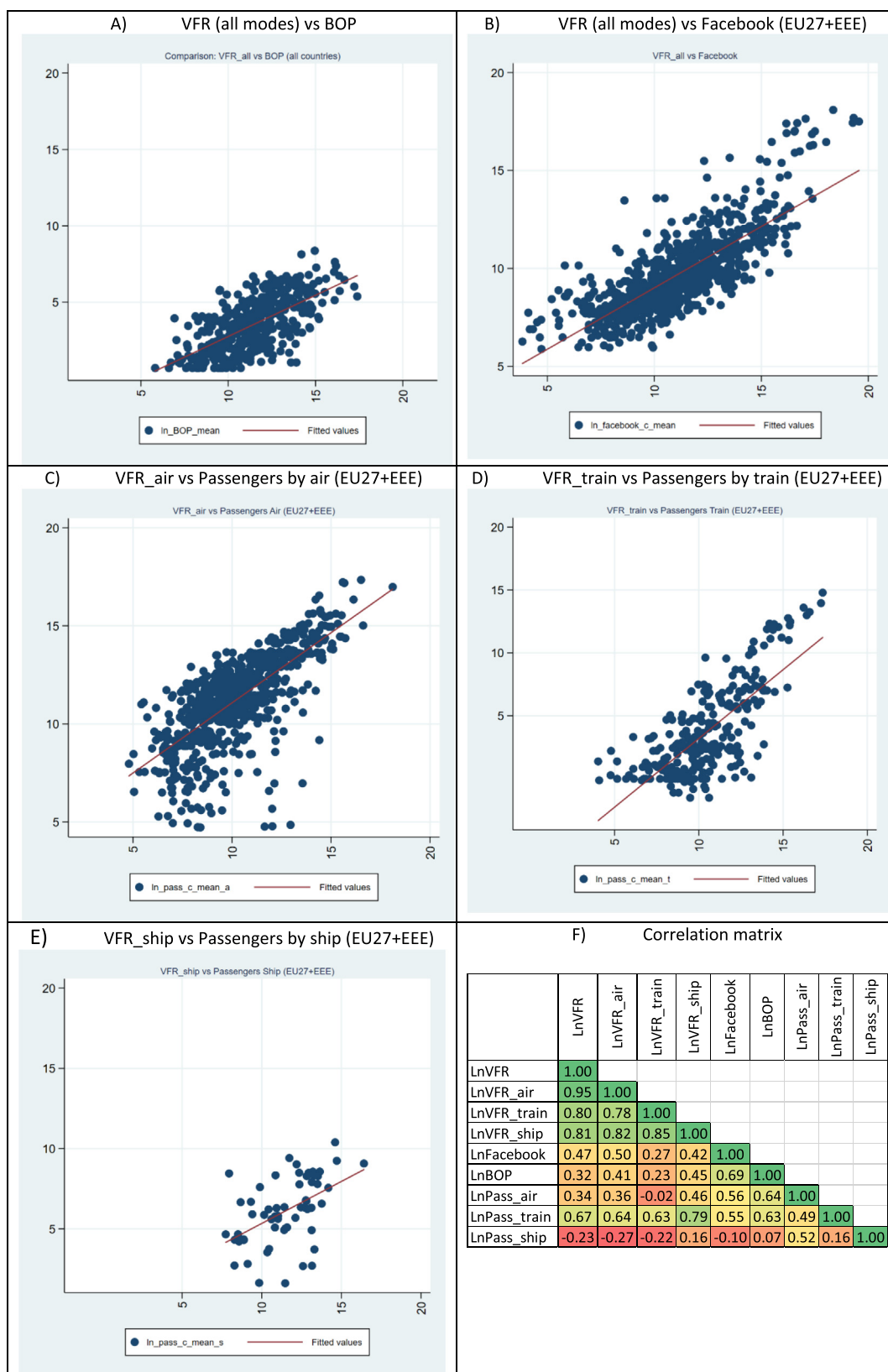


Fig. 3. . Comparison between the (logs) of VFR_{ij} and key bilateral indicators. Ave.2010–18.

Source: Own elaboration based on our own dataset based on UNWTO; WTO for BOPs; ESPON-IRIE for flows of passengers; for Facebook (Bailey et al., 2018).

*For simplicity, we use “0” for missing values.

Similarly, Eqs. (16)–(18) describe the equivalent procedure applied in cascade for splitting the aggregate international flows of tourists by transport mode.

Before describing the results obtained, and proceed with the evaluation of the flows, it is convenient to test the solidity of the methodology described before to produce the dataset. For brevity, we take this analysis to the online appendix (Section 10). Thus, the interested reader will find three specific analysis: the first one, focused on the core of the method (step 1), where we evaluate the forecasting capacity of all the models, using the international bilateral flows of tourists and visitors for the period 1995–2017, and five measures of forecast quality (MAPE, MAAPE, MAE, RMSE and the U’ Theil. Then, we provide the results obtained in two robustness analysis, one using multiple imputation techniques, and the other, using the new Stata command *xtselvar* for panel data selection.

Results

Once that the methodology of data estimation has been described, we evaluate the results obtained, first by means of a descriptive analysis, and then, using different specifications of the gravity equation..

To start with, Fig. 1 maps the main international VFR flows for 2018, while Fig. 2 focuses on European flows over the same period. The former shows how the most important flow is between China with Hong Kong and between the United States, Canada and Mexico. In Europe, the most relevant connections are between Germany, Spain, France, and the United Kingdom. Complementary analysis are provided in the online Appendix.

Comparison with alternative indicators

In this section we compare the flows obtained with other relevant indicators available. Although none of them are optimal, they provide interesting points of reference. Fig. 3 shows six panels where we compare the average bilateral flows for the period 2010–2018, expressed in logarithms: **Panel A** compares the complete bilateral VFR variable (people) with the monetary flows obtained from the WTO balance of payments (BOP) datasets. The monetary flows used in the comparison are exports (millions of dollars) of the “passenger all modes transport sector” (EBOPS code: SCA). This analysis only uses international flows, since BOP do not include domestic transactions. **Panel B** compares the same VFR indicator with an interesting measure of social relatedness computed using Facebook individual connections worldwide (Bailey et al., 2018). The Facebook index has the virtue of including domestic and international linkages, while a drawback, is cross-section. **Panels C, D and E** compare the bilateral transport-mode specific flows of VFR (air, train and ship) with three corresponding indicators on passenger flows recently estimated within the ESPON-IRIE project, constructed upon Eurostat official data. The analysis considers both domestic and international flows and refers to the 32 European countries included in such project (EU27 + EEE). Finally, **Panel F** reports the correlation matrix corresponding to these indicators.

As expected, the graphs and the correlations are not perfect but clearly positive. In the case of **Panel A**, aggregate VFR flows and the BOP are very aligned with a reasonable correlation of 0.32. Note that there is a strong conceptual jump from measuring how much tourists of a country *i* goes to country *j*, and how much expenses these travelers do in transportation firms of country *i*, in firms of country *j*, or in firms of a third country different than *i* and *j*. To this regard, it is important to understand that the BOP of a country *i* reports the monetary flows of their national firms offering transport services to foreigners, but such international services can take place in the country *i*, or in any other country in the world. An example: British Airways in 2000, being a global firm with residence in the UK, offers transport services to a British tourist traveling to Spain and to a Spanish tourist traveling to UK. The same might happen if such displacements are done by Iberia, a Spanish firm. In addition, another non-Spanish and non-British firm (i.e., Lufthansa) can transport such tourists, generating a German export of transport service to Spain and UK. In conclusion, the number of tourists traveling between Spain and UK might be related with the trade flows registered in the BOPs of Spain and UK, mainly if the Spanish and British transport firms account for the largest part of these bilateral trips. However, such figure of tourists can also be related with the service exports of Germany, if Lufthansa accounts for a strong share of the flights between these two countries.

Noticeably, **Panel B** also shows how the average VFR flow is also highly correlated (0.47) with the Facebook SCI, and even higher with the VFR_{air}. This can be interpreted as how the physical mobility of tourist is associated with the social connections created before (i.e., family) or after such displacements (i.e., friends made in the destination spot). Note that, in contrast to **Panel A**, this analysis includes the domestic and international flows, and the domestic Facebook index is obtained by aggregating the sub-national relations computed (almost) at the city level (NUTS 3 in Eurostat terminology).

Moreover, **Panel C** shows a high positive correlation between the average bilateral VFR indicator for air and a unique indicator on passenger flows by Air used. Although in principle the concepts of “visitor” and “passengers” might be very similar, they are usually register by different statistical means. Passenger statistics from Eurostat correspond to the records reported by each airport and might include multi-scale trips. The correlation coefficient is reasonably high for the different nature of these indicators (0.36). **Panel D** also shows a clear positive relation between the average bilateral VFR by train and the alternative indicator for passengers by train, built upon official Eurostat statistics. In this case the correlation coefficient is higher (0.63). Finally, the analysis provided for ship (LnVFR_{ship} vs LnPass_{ship}) in **Panel E** is less relevant, giving the small share of this mode, which also result in the lowest positive correlation (0.16). Unfortunately, there is no equivalent indicator for our VFR by road.

Table 1

The gravity model using the final complete flows (domestic and international tourism), 1995–2018. Countries: 74. PPML

Variables	(1) TFR_1	(2) TFR_2	(3) TFR_3	(4) VFR_1	(5) VFR_2	(6) VFR_3
Intra	5.853*** (0.713)	3.328*** (0.328)	3.325*** (0.297)	6.343*** (0.739)	3.654*** (0.346)	3.658*** (0.328)
LnDist	−0.428*** (0.156)	−1.143*** (0.128)	−1.151*** (0.117)	−0.388** (0.159)	−1.165*** (0.132)	−1.172*** (0.126)
Contig	3.990*** (0.572)	1.854*** (0.209)	1.854*** (0.192)	4.126*** (0.580)	1.916*** (0.210)	1.914*** (0.198)
Comlang_off	−0.887 (0.556)	−0.124 (0.249)	−0.142 (0.257)	−0.894 (0.560)	−0.215 (0.237)	−0.217 (0.249)
Landlocked	−0.242 (0.564)	0.385** (0.160)	0.394** (0.154)	−0.236 (0.576)	0.482** (0.156)	0.494*** (0.151)
Landlocked_both	−0.590 (0.434)			−0.755* (0.413)		
Colony	−1.615*** (0.616)	−0.125 (0.282)	−0.128 (0.273)	−1.437** (0.644)	−0.0725 (0.270)	−0.0937 (0.260)
EU	−0.942*** (0.304)	0.759*** (0.218)	0.775*** (0.221)	−0.867*** (0.334)	0.823*** (0.219)	0.858*** (0.221)
Island_both	−0.507 (0.435)	−1.702*** (0.469)	−1.669*** (0.462)	−0.472 (0.454)	−1.622*** (0.460)	−1.628*** (0.474)
Island	−0.366* (0.196)			−0.405** (0.201)		
Comrelig	−1.574*** (0.554)	0.486 (0.355)	0.432 (0.339)	−1.708*** (0.576)	0.481 (0.358)	0.416 (0.348)
Constant	16.15*** (1.299)	22.30*** (0.883)	22.13*** (0.820)	15.99*** (1.322)	22.60*** (0.923)	22.49*** (0.894)
Observations	72,445	72,445	72,356	71,990	71,990	71,898
Pseudo R2	0.755	0.968	0.987	0.758	0.967	0.990
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	Yes	Yes	No	Yes	Yes
Country-Year FE	No	No	Yes	No	No	Yes

Source: Own elaboration. Note1: Robust standard errors in parentheses.

*** p < 0.01.

** p < 0.05.

* p < 0.1.

To the best of our knowledge, no equivalent comparison has been published till now. Obviously, each of this analysis deserves further investigation using the appropriate econometric tools.

Validation of the final dataset using the gravity equation

As a further check, the complete dataset is tested using the gravity equation, one of the most well-known models used to explain bilateral flows of any kind (goods, services, people, capital), with many applications to tourism (De la Mata, 2014; Gil-Pareja et al., 2007; Khalid et al., 2020; Li et al., 2021; Liu et al., 2010; Llano, 2013; Morley et al., 2014; Provenzano, 2020; Song et al., 2019; Vietze, 2012; Witt & Witt, 1995). To do so, we define a generic Eq. (26) that will be fed with all the aggregate and transport-mode specific flows reported before.

$$\ln VFR_{fijt} = \beta_0 + \beta_1 \ln Gdp_{it} + \beta_2 \ln Gdp_{jt} + \beta_3 \ln Dist_{ij} + \beta_4 \ln Intra_{ij} + \beta_5 \ln Contig_{ij} + \beta_6 \ln Comlangoff_{ij} + \beta_7 \ln Landlocked_{ij} + \beta_8 \ln Landlocked_both_{ij} + \beta_9 \ln Colony_{ij} + \beta_{10} \ln EU_{ij} + \beta_{11} \ln Island_{ij} + \beta_{12} \ln Island_both_{ij} + \beta_{13} \ln Comrelig_{ij} + \delta_{it} + \mu_{jt} + \theta_i + \sigma_j + \alpha_t + \varepsilon_{ijt} \quad (26)$$

where VFR_{fijt} refers to the total number of visitors between countries ij and within i in year t , while the elements $\delta_{it}, \mu_{jt}, \theta_i, \sigma_j, \alpha_t$, corresponds to the “origin*time”, “destination*time”, “origin”, “destination”, and “time” fixed effects recommended by the state of the art of gravity equation (Anderson & Van Wincoop, 2003; Head & Mayer, 2014). Finally, ε_{ijt} denotes the disturbance term. All the explanatory variables have been described before, except “Intra”, which controls for the different nature of intra-national and inter-national flows: it is a dummy variable that takes value 1 for domestic flows and 0 otherwise; and “Contiguity”, which controls for the expected higher intensity of flows with the neighboring countries. In this case, the variable is also defined as a dummy that takes value 1 for neighboring countries and 0 otherwise. “Landlocked” and “Island” variables take values 1 when at least one of the destinations is landlocked or an island, while those named at the end with “_both” is a dummy variable that takes values 1 when both the origin and destination are landlocked or are they islands.

Table 1 reports the results when the gravity model is fed with the final complete flows, including international and domestic visitors and tourists. Each model is introduced with and without origin & destination fixed effects to show how the later absorbs

Table 2

Gravity model with final complete flows (domestic + international) by transport mode. 1995–2018. 74 countries. PPML.

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	TFR_r	TFR_s	TFR_t	TFR_a	TFR_ot	VFR_r	VFR_s	VFR_t	VFR_a	VFR_ot
Intra	3.489*** (0.252)	0.458 (0.354)	3.276*** (0.330)	2.650*** (0.358)	8.291*** (1.772)	3.811*** (0.264)	0.816** (0.387)	3.630*** (0.345)	2.954*** (0.373)	9.220*** (1.780)
LnDist	−1.375*** (0.106)	−1.388*** (0.136)	−1.350*** (0.140)	−0.633*** (0.125)	0.436 (0.586)	−1.410*** (0.109)	−1.421*** (0.150)	−1.371*** (0.145)	−0.644*** (0.130)	0.625 (0.530)
Contig	1.732*** (0.180)	1.675*** (0.239)	1.847*** (0.266)	2.050*** (0.204)	2.923*** (1.258)	1.801*** (0.175)	1.680*** (0.264)	1.878*** (0.265)	2.093*** (0.213)	3.410*** (1.307)
Comlang_off	−0.191 (0.397)	−0.0694 (0.406)	−1.167** (0.489)	0.307 (0.202)	2.238*** (0.501)	−0.301 (0.371)	−0.173 (0.409)	−1.169** (0.469)	0.211 (0.203)	2.336*** (0.490)
Landlocked	0.633*** (0.169)	0.243 (0.235)	−0.157 (0.381)	0.212* (0.126)	−1.785*** (0.432)	0.734*** (0.161)	0.360 (0.258)	−0.0130 (0.380)	0.229* (0.131)	−1.822*** (0.460)
Colony	−0.475* (0.259)	0.275 (0.333)	−0.107 (0.217)	0.574** (0.228)	1.345*** (0.437)	−0.359 (0.232)	0.230 (0.381)	−0.0807 (0.209)	0.628** (0.245)	1.342*** (0.443)
EU	0.964*** (0.293)	1.218*** (0.287)	0.328 (0.490)	0.291 (0.201)	0.981* (0.553)	1.015*** (0.270)	1.232*** (0.282)	0.397 (0.476)	0.402* (0.208)	1.313** (0.595)
Island_both	−0.850* (0.438)	−0.200 (0.731)	0.467 (1.101)	−1.781*** (0.474)	1.563* (0.849)	−0.784* (0.417)	−0.122 (0.797)	0.522 (1.132)	−1.700*** (0.469)	1.861** (0.744)
Comrelig	−0.0816 (0.493)	0.707 (0.553)	0.336 (0.636)	0.600** (0.297)	4.021*** (1.538)	−0.116 (0.479)	0.867 (0.622)	0.379 (0.629)	0.619* (0.320)	4.583*** (1.685)
Constant	22.80*** (0.725)	21.35*** (0.901)	21.77*** (0.935)	17.32*** (0.918)	3.837 (4.884)	23.24*** (0.751)	21.77*** (0.998)	21.96*** (0.972)	17.64*** (0.956)	2.417 (4.615)
Observations	72,347	72,241	72,290	72,320	72,208	71,889	71,783	71,832	71,862	71,749
Pseudo R2	0.990	0.942	0.980	0.939	0.984	0.993	0.948	0.985	0.950	0.987
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Source: Own elaboration. Note1: Robust standard errors in parentheses.

*** p < 0.01.

** p < 0.05.

* p < 0.1.

the information corresponding to certain variables. These specifications use the whole sample of the 74 countries for the period 1995–2018 using the PPML estimator (Santos Silva & Tenreiro, 2006; Santos-Silva & Tenreiro, 2011). The performance of all the models is outstanding, obtaining a high R^2 of 0.75 for the model with no fixed effects of origin or destination. The coefficient for the log of distance, the GDPs and the Contiguity are in line with the literature. The positive and significant coefficient for the “Intra” also coincides with the result expected, highlighting the presence of an apparent *home bias* in the flows, that is, a higher value of domestic tourism compared to the international ones (Anderson & van Wincoop, 2003; Carril-Caccia et al., 2022; Head & Mayer, 2014; Llano, 2013; McCallum, 1995). The variables *Comrelig*, *Colony*, are significant in models (1) and (4) but they are not when introducing the territorial fixed effects. The sign of the EU variable also shifts when introducing these multilateral resistance terms.

Finally, Table 2 reports the results obtained for equivalent regressions (PPML) but using the transport mode-specific flows for TFR and VFR, including domestic and international flows. All models include territorial and time fixed effects and perform very well, with R^2 above 0.9.

The elasticities obtained for the log of distance are negative and significant in almost all cases. The negative coefficients are higher for road and train than for the rest, which is reasonable considering the propensity to use the former in shorter distances. Indeed, the trips in “other-land” (buses, taxis, car rentals...) become non-significant or even positive. The *Contiguity dummy* is positive and significant for some specific flows, but not for all modes. Regarding the *Intra*, we find positive and significant coefficients for some flows, such as VFR (road, train, ship and others) and TFR (road, train and others). Again, this result is in line with economic intuition which suggests that, on average, after controlling for all non-observable factors related to the origins/destinations/time, the intra-national visitors are more intense than the inter-national ones for trips by land (road, train, others) and less likely for ship transport.

Complementary results for OLS estimations are reported in Tables A.12–A.13 in the Appendix.

Conclusions

The statistical scenario for international and domestic tourism is far from ideal, characterized by sparse and heterogenous indicators reported by each country. Even after the intense effort of data collection carried out by the UNWTO and other institutions (i.e., Eurostat), the information available is full of gaps and discontinuities.

This article describes the methodology used to obtain a potentially complete, harmonized database of country-to-country flows of tourists and visitors built upon the UNWTO original database. The resulting dataset covers 74 countries and the period 1995–2018. By means of 34 alternative models, we estimate two main indicators (VFR and TFR) using the observable data already

published by the original UNWTO dataset, omitting the use of any other information or behavioral assumptions. Our final dataset also includes domestic tourism and splits the flows into five transport modes.

Once the dataset has been mapped, it is confronted with key alternative indicators such as the balance of payments (services associated to passengers' transportation), the Facebook social connectivity index, or the flows of passengers by air, train or ship in Europe. Then, we model the flows using the gravity equation.

Said analysis has led to some relevant conclusions: i) from a methodological viewpoint: the process followed suggests that the information related to TFR, VFR, TCER and THSR are robust enough upon which to build a consolidated dataset of visitors and tourists across most countries. ii) Regarding the results obtained, we confirm a clear *home bias* favoring intra-national tourist flows. Within the EU27, the main flows are those between Germany-Poland, Germany-France, Germany-Italy, UK-Spain, France-Spain, France-Italy... Globally, the flows between China and Hong-Kong are on another scale with a major portion reported as using "other transport mode" (land). Also, the flows between the US and Mexico and Canada are of great intensity.

Once that this structural new dataset has been estimated for the pre-COVID19 period, further extensions are expected with respect to the post-pandemic one. Obviously, such development will face the additional challenges of new statistical gaps in statistical series and radical shifts in the mobility patterns due to external restrictions and changes in individual preferences. In our view, the two new approaches reported in the robustness analysis of this paper defines interesting pathways for overcoming such limitations, combining the UNWTO structural datasets published for these singular years with new indicators recently published (i.e., Google destinations; Google and Apple mobility indexes, Oxford University national Indexes on restrictiveness measures, etc.). Moreover, as commented in the main body of this analysis, further extensions are also in the pipeline, with respect to the estimation of equivalent datasets at the region-to-region level (NUTS 2) for the whole Europe (EU27 + EEE) or the estimation of the corresponding GHG emission patterns associated with the transport-mix obtained for each country and region.

CRedit authorship contribution statement

Carlos Llano: Conceptualization, Methodology, Software: Stata do file development, Writing- Original and final document; Coordination. **Juan Pardo:** Data curation, Software: complementary Stata development, Tableau Public Software development, Writing specific sections; Visualization. **Santiago Pérez-Balsalobre:** Software: stata do file development, Gravity equation leadership; Data curation for the gravity equation. **Julián Pérez:** Methodology; Forecasting leadership; Supervision.

Data availability

The dataset produced is visible in a public site in the format of an interactive dashboard. The aggregate flows are available at ESPON webpage. The detailed transport mode flows will be also available.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This article has been developed in the context of the ESPON IRIE Project: <https://www.espon.eu/interregional-relations-europe-new-project-espon>, developed by a big consortium during 2020-2022. The dataset developed in this article was used as a first step of the further extensions developed, at the region-to-region level by the Institute of Geography and Spatial Organization Polish Academy of Sciences (IGSO PAS). We want to acknowledge the collaboration received from all partners in the ESPON-IRIE Project, specially from the project officer in ESPON EGTC, Nicolas Rosignol; Xabier Velasco, manager of the project, at NASUVINSA, and other colleagues from the consortium, especially the ones from IGSOPAS. This paper was also developed in the context of another research project: the H2019/HUM-5761 INNOJOBMAD-CM Program from the Autonomous Community of Madrid (Comunidad de Madrid).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.annals.2023.103672>.

References

- Akande, O., Li, F., & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71(2), 162–170. <https://doi.org/10.1080/00031305.2016.1277158>.
- Aloisio, K. M., Micali, N., Swanson, S. A., Field, A., & Horton, N. J. (2014). Analysis of partially observed clustered data using generalized estimating equations and multiple imputation. *Stata Journal*, 14, 863–883.
- Anderson, J. E., & Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1), 170–192.
- Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3), 259–280.

- Becken, S., & Simmons, D. G. (2002). Understanding energy consumption patterns of tourist attractions and activities in New Zealand. *Tourism Management*, 23(4), 343–354. [https://doi.org/10.1016/S0261-5177\(01\)00091-7](https://doi.org/10.1016/S0261-5177(01)00091-7).
- Buckley, R. (2012). Sustainable tourism: Research and reality. *Annals of Tourism Research*, 39(2), 528–546.
- Carril-Caccia, F., Martín Martín, J. M., & Sáez-Fernández, F. J. (2022). How important are borders for tourism? The case of Europe. *Tourism Economics*, 0(0). <https://doi.org/10.1177/13548166221132452>.
- Cherniwchan, J., Copeland, B. R., & Taylor, M. S. (2017). Trade and the environment: New methods, measurements, and results. *Annual Review of Economics*, 9, 59–85.
- Chujai, P., Singthongchai, J., Yasaga, S., Surattara, N., & Buranakutti, K. (2020). The tourist attractions recommender system for Bangkok Thailand. *International Journal of Computer Theory and Engineering*, 12(1), 22–27 2020.
- Comulada, W. S. (2015). Model specification and bootstrapping for multiply imputed data: An application to count models for the frequency of alcohol use. *Stata Journal*, 15, 833–844.
- Cristea, A., Hummels, D., Puzello, L., & Avetisya, M. (2013). Trade and the greenhouse gas emissions from international freight transport. *Journal of Environmental Economics and Management*, 65, 153–173.
- De la Mata (2014). Does trade creation by social and business networks hold in services? *Applied Economics*, 46(13), 1509–1525.
- Dickinson, J. E., Lumsdon, L. M., & Robbins, D. (2011). Slow travel: Issues for tourism and climate change. *Journal of Sustainable Tourism*, 19(3), 281–300. <https://doi.org/10.1080/09669582.2010.524704>.
- El Sawey, M. (2020). Using spatio-temporal data for estimating missing cycling counts: A multiple imputation approach. *Transportmetrica A: Transport Science*, 16(1), 5–22. <https://doi.org/10.1080/23249935.2018.1440262>.
- Frechtling, D. C., & Hara, T. (2016). State of the world's tourism statistics and what to do about it. *Tourism Economics*, 22(5), 995–1013.
- Gil-Pareja, S., Llorca-Vivero, R., & Martínez-Serrano, J. A. (2007). The impact of embassies and consulates on tourism. *Tourism Management*, 28(2), 355–360.
- Gössling, S., Scott, D., & Hall, C. M. (2021). Pandemics, tourism and global change: A rapid assessment of COVID-19. *Journal of Sustainable Tourism*, 29(1), 1–20. <https://doi.org/10.1080/09669582.2020.1758708>.
- Head, K., & Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. Chapter 3. *Handbook of International Economics*. vol. 4. (pp. 131–195) (from Elsevier).
- Higham, J., Cohen, S. A., Cavaliere, C. T., Reis, A., & Finkler, W. (2016). Climate change, tourist air travel and radical emissions reduction. *Journal of Cleaner Production*, 111, 336–347.
- Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45, 1–47.
- Hough, G., & Hassanien, A. (2010). Transport choice behaviour of Chinese and Australian tourists in Scotland. *Research in Transportation Economics*, 26(1), 54–65.
- Hunter, C., & Shaw, J. (2007). The ecological footprint as a key indicator of sustainable tourism. *Tourism Management*, 28(1), 46–57.
- Kelly, J., Haider, W., & Williams, P. W. (2007). A behavioral assessment of tourism transportation options for reducing energy consumption and greenhouse gases. *Journal of Travel Research*, 45(3), 297–309. <https://doi.org/10.1177/0047287506292700>.
- Khadaroo, J., & Seetanah, B. (2008). The role of transport infrastructure in international tourism development: A gravity model approach. *Tourism Management*, 29, 831–840.
- Khalid, U., Okafor, L. E., & Shafiullah, M. (2020). The effects of economic and financial crises on international tourist flows: A cross-country analysis. *Journal of Travel Research*, 59(2), 315–334. <https://doi.org/10.1177/0047287519834360>.
- Lee, C.-C., Olasehinde-Williams, G. O., & Ibikunle, J. A. (2022). An asymmetric examination of the environmental effect of tourism in China. *Tourism Economics*, 28(7), 1872–1887. <https://doi.org/10.1177/13548166211021173>.
- Lenzen, M., Sun, Y. Y., Faturay, F., Ting, Y. P., Geschke, A., & Malik, A. (2018). The carbon footprint of global tourism. *Nature Climate Change*, 8(6), 522–528. <https://doi.org/10.1038/s41558-018-0141-x>.
- Li, X., Gong, J., Gao, B., & Yuan, P. (2021). Impacts of COVID-19 on tourists' destination preferences: Evidence from China. *Annals of Tourism Research*, 90, Article 103258.
- Liu, X., Whalley, J., & Xin, X. (2010). Non-tradable goods and the border effect puzzle. *Economic Modelling*, 27(10), 909–914.
- la Mata, D., & Llano, C. (2013). Social networks and trade of services: Modelling interregional flows with spatial and network autocorrelation effects. *Journal of Geographical Systems*, 15(3), 319–367.
- Martín-Cejas, R. R., & Sánchez, P. P. R. (2010). Ecological footprint analysis of road transport related to tourism activity: The case for Lanzarote Island. *Tourism Management*, 31(1), 98–103.
- Masiero, L., & Zoltan, J. (2013). Tourists intra-destination visits and transport mode: A bivariate probit model. *Annals of Tourism Research*, 43, 529–546.
- McCallum, J. (1995). National borders matter: Canada-US. Regional trade patterns. *The American Economic Review*, 85, 615–623.
- McKercher, B., & Mak, B. (2019). The impact of distance on international tourism demand. *Tourism Management Perspectives*, 31, 340–347. <https://doi.org/10.1016/j.tmp.2019.07.004>.
- McKercher, B., & Prideaux, B. (2014). Academic myths of tourism. *Annals of Tourism Research*, 46, 16–28. <https://doi.org/10.1016/j.annals.2014.02.003>.
- Miller, G., & Torres-Delgado, A. (2023). Measuring sustainable tourism: A state of the art review of sustainable tourism indicators. *Journal of Sustainable Tourism*. <https://doi.org/10.1080/09669582.2023.2213859>.
- Morley, C., Rosselló, J., & Santana-Gallego, M. (2014). Gravity models for tourism demand: Theory and use. *Annals of Tourism Research*, 48, 1–10.
- Ortiz, A. M. D., Outhwaite, C. L., Dalin, C., & Newbold, T. (2021). A review of the interactions between biodiversity, agriculture, climate change, and international trade: Research and policy priorities. *One Earth*, 4(1), 88–101.
- Peeters, P., & Dubois, G. (2010). Tourism travel under climate change mitigation constraints. *Journal of Transport Geography*, 18(3), 447–457.
- Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management*, 45, 181–193.
- Pratt, S., & Tolkach, D. (2018). The politics of tourism statistics. *International Journal of Tourism Research*, 20, 299–307. <https://doi.org/10.1002/jtr.2181>.
- Prillwitz, J., & Barr, S. (2011). Moving towards sustainability? Mobility styles, attitudes and individual travel behaviour. *Journal of Transport Geography*, 19(6), 1590–1600.
- Provenzano, D. (2020). The migration–tourism nexus in the EU28. *Tourism Economics*, 26(8), 1374–1393. <https://doi.org/10.1177/1354816620909994>.
- Rehman, F. U., Islam, M. M., & Miao, Q. (2023). Environmental sustainability via green transportation: A case of the top 10 energy transition nations. *Transport Policy*, 137, 32–44.
- Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462–1471.
- Rosselló, J., Becken, S., & Santana-Gallego, M. (2020). The effects of natural disasters on international tourism: A global analysis. *Tourism Management*, 79. <https://doi.org/10.1016/j.tourman.2020.104080>.
- Royston, P., Carlin, J. B., & White, I. R. (2009). Multiple imputation of missing values: New features for mim. *Stata Journal*, 9, 252–264.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473–489.
- Santos Silva, J. M. C., & Tenreiro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4), 641–658 November 2006.
- Santos-Silva, J., & Tenreiro, S. (2011). Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economic Letters*, 112, 220–222.
- Scott, D., & Gössling, S. (2022). A review of research into tourism and climate change - launching the annals of tourism research curated collection on tourism and climate change. *Annals of Tourism Research*, 95 art. no. 103409.
- Shao, Y., Huo, T., Yang, Y., & Li, Z. (2023). Does economic globalization shape the international tourism structure? A cross-National Panel Data Estimation. *Journal of Travel Research*, 62(5), 1121–1139. <https://doi.org/10.1177/0047287522119128>.
- Song, H., Qiu, R. T. R., & Park, J. (2019). A review of research on tourism demand forecasting. *Annals of Tourism Research*, 75, 338–362.
- Sun, Y. Y., Gössling, S., & Zhou, W. (2022). Does tourism increase or decrease carbon emissions? A systematic review. *Annals of Tourism Research*, 97, Article 103502.
- Thrane, C. (2015). Examining tourists' long-distance transportation mode choices using a multinomial logit regression model. *Tourism Management Perspectives*, 15, 115–121.

- Ugarte-Ruiz, A. (2020a). *XTSELVAR & XTSELMOD: Selection of Variables and Specification in a Panel Data Framework*. Virtual Conference Stata, USA Meeting. July 30–31, 2020.
- Ugarte-Ruiz, A. (2020b). XTSEL: Stata module for selection of variables and specification in a panel-data framework. *Statistical software components* S458816. Boston College Department of Economics revised 11 Nov 2022.
- Vietze, C. (2012). Cultural effects on inbound tourism into the USA: A gravity approach. *Tourism Economics*, 18(1), 121–138. <https://doi.org/10.5367/te.2012.0100>
- Wagstaff, D. A., & Harel, O. (2011). A closer examination of three small-sample approximations to the multiple imputation degrees of freedom. *Stata Journal*, 11, 403–419.
- Wan, S. K., & Song, H. (2018). Forecasting turning points in tourism growth. *Annals of Tourism Research*, 72, 156–167.
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377–399.
- Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11, 447–475.
- Yang, Y. (2011). Multiple imputation using Sas software. *Journal of Statistical Software*, 45.
- Yang, Y., Liu, H., & Li, X. (R.). (2019). The world is flatter? Examining the relationship between cultural distance and international tourist flows. *Journal of Travel Research*, 58(2), 224–240. <https://doi.org/10.1177/0047287517748780>.
- Yang, Y., & Zhang, H. (2019). Spatial-temporal forecasting of tourism demand. *Annals of Tourism Research*, 75, 106–119.
- Zamparini, L., Domenech, A., Miravet, D., & Gutierrez, A. (2022). Green mobility at home, green mobility at tourism destinations: A cross-country study of transport modal choices of educated young adults. *Journal of Transport Geography*, 103. <https://doi.org/10.1016/j.jtrangeo.2022.103412> art. no. 103412.

Carlos Llano is specialized in international trade with contributions in the use of the gravity equation and spatial econometrics.

Juan Pardo is specialized in the field of trade of services and tourism.

Santiago Pérez-Balsalobre is mainly dedicated to the analysis of interregional trade and flows of knowledge (complexity).

Julián Pérez has a wide experience in the field of economic modelling and forecasting, with a focus in impact evaluation, territorial analysis, and energy.