Article

TOURISM ECONOMICS
The Business and Finance of Tourism and Recreation

Tourism Economics
2017, Vol. 23(1) 99–112

DOI: 10.5367/te.2015.0500
journals.sagepub.com/home/teu
SAGE

Tourism statistics: Correcting data inadequacy

Patricio Aroca
Universidad Adolfo Ibáñez, Chile

Juan Gabriel Brida
Universidad de la República, Uruguay

Serena Volo
Free University of Bolzano, Italy

Abstract
Tourism statistics are key sources of information for economic planners, tourism researchers, and operators. Still, several cases of data inadequacy and inaccuracy are reported in literature. The aim of this article is to propose a methodology useful to improve tourism statistics: a modified version of the Coarsened Exact Matching. The methodological steps herein proposed provide tourism statisticians and authorities with a tool to improve the reliability of available sample surveys. Data from a Chilean region are used to illustrate the method. This study contributes to the realm of tourism statistics literature in that it offers a new methodological approach to the creation of accurate and adequate tourism data.

Keywords
accommodations, accretion bias, attrition bias, Chile, sample weights, tourism planning

Introduction

Academic studies on tourism statistics emphasize the importance of collecting accurate and reliable statistics that can support tourism planning and forecasting (Aroca et al., 2013; Burkart and Medlik, 1981; De Cantis and Ferrante, 2013; Lickorish, 1997; Massieu, 2001; Meis, 2001; Pine, 1992; Volo, 2004; Volo and Pardew, 2013). Nevertheless, in many countries inconsistencies in tourism statistics are still quite frequent (Aroca et al., 2013; Guizzardi and Bernini, 2012; Volo and Giambalvo, 2008; Volo, 2010). In this study, a method to adjust data inconsistencies is presented

Corresponding author:
Serena Volo, Faculty of Economics and Management, Free University of Bolzano, Bolzano, Italy.
Email: serena.volo@unibz.it

and its benefits are shown when applied to accommodations' data. Thus, the aim of this article is to introduce tourism researchers to a methodology for reconstructing tourism databases using sample weights built with the Coarsened Exact Matching (CEM) and to show how to successfully use them in order to overcome some methodological issues of supply side tourism statistics.

Tourism supply data: Inconsistencies issues

Biases in data collection might lead to untrue representation of the studied phenomena. Particularly relevant in social sciences are the sampling and selection biases, from which tourism statistics and tourism studies are not exempted. Furthermore, incomplete data often derive from a change of the population of interest. Such changes may be due to the lack of initial inclusion or incomplete follow-up, mortality, or addition of new entities (Hofer and Hoffman, 2010). The loss of participants or entities over time due to transience, dropouts, withdrawals, and protocol deviations is known as attrition. The natural growth of the population often goes ignored during the process of collecting data, this problem is defined as gross-growth of the population, and it consists of ingrowth and accretion. The ingrowth relates to newly "grown" entities that were not initially present in the sampled population, while accretion refers to the "growth" of the sampled entities. Biases related to these two types of growth can arise when systematic differences occur in some of the outcome variables under study. Attrition, ingrowth, and accretion do occur in managerial statistics due to the lack of systematic updating of enterprises' directories and often they lead to biased results. Approaches to detect and correct these biases have been used in past literature, but so far tourism researchers have paid little attention to this matter.

The issue of incomplete tourism databases has been recently investigated in the studies by Aroca et al. (2013) and by Fontana and Pistone (2010). The latter describes a method used to complete the official statistics data on Italian tourism flows focusing on the imputation of missing values. Their proposed methodology removes the effect of nonrespondent accommodation establishments. The former by Aroca et al. (2013), however, proposes a technique for correcting attrition in tourism databases and corrected the misrepresented population of accommodation establishments using sample weights. The use of sample weights is effective to correct potential biases that might result from the nonrepresentativeness of the sample (Boudreau and Yan, 2010), but other techniques are available and may lead to better results in accounting for accretion, ingrowth, and attrition in databases. Some of the methods documented in medical and psychological literature are: selection modeling with a probit model for attrition and a regression model for the outcome, maximum likelihood methods, and the use of multiple imputations for missing data (McCoy et al., 2009; McGuigan et al., 1997;Schafer and Graham, 2002).

The method herein proposed consists of a sequential reconstruction of a tourism database with the calculation of sample weight using the CEM. The next section describes the CEM in detail, while the overall reconstruction of the datasets and an application of CEM is provided in the section "Reconstruction of a tourism dataset: The case of a Chilean region."

The CEM

Heitjan and Rubin (1991) refer to coarse data as a general type of incomplete data that arise from observing not the exact value of the data but a subset of the sample space. Their definition covers several incomplete-data problems including rounded, heaped, censored, and missing data (Kim and Hong, 2012). The CEM is a particular member of the matching methods known as Monotonic

Table 1. Comparison of the methodologies.

	Mahalanobis	Propensity score
D_{ij} is the distance between individuals i and j	$D_{ij} = (X_i–X_j)'\Sigma^{-1}(X_i - X_j)$, where X_i and X_j are the vectors of covariates of i and j, and Σ is the variance covariance matrix of X in the full control group (if we are interested in the average effect of the treatment on the treated), or in the pooled treatment and full control groups (when we analyze the average treatment effect).	$D_{ij} = \lvert e_i - e_j \rvert$, where e_i and e_j are the propensity scores for individuals i and j. The propensity score is defined as the probability of receiving the treatment given the observed covariates.
Advantages	Very good performance with few covariates.	(1) Propensity scores are balancing scores: at each value of e_i, the distribution of X defining the propensity score is the same in the treatment and control groups. (2) If given X, the treatment assignment is ignorable, also it is given the propensity score e_i.
Disadvantages	The distance does not work well when X is of high dimension. It may lead to many individuals to not being matched (which may result in biased results). Also, it has problems when covariates are not normally distributed.	If the treatment and control groups do not have substantial overlap (in terms of covariates), substantial errors may be introduced.
References	Imai et al. (2008), Gu and Rosenbaum (1993), Rosenbaum and Rubin (1985), Rubin (1979), Stuart (2010).	Abadie and Imbens (2011), Rosenbaum and Rubin (1983), Rubin and Thomas (1992, 1996), Stuart (2010).

Imbalance Bounding. Matching methods are useful tools for applied researchers and have been in use since the 1950s (Stuart, 2010) and they are used to appropriately select data when designing an observational study. A crucial step in the design of a matching method is that of defining the distance between the two individuals/entities under study. Several approaches to distance measurement have been implemented and while exact matching is ideal, in practice it is quite difficult to achieve. Until recently the Mahalanobis matching, the propensity score, and the linear propensity score were extensively used.

There exists evidence that CEM is in many respects superior to other common matching methods (e.g. the propensity score matching) and Iacus et al. (2011) offer some results that demonstrate the potential of CEM over other matching methods in terms of inference. It has to be noted that CEM is also successfully used as a method for policy evaluation, but in this article it is used in its functionality as matching procedure, which has been proved to be of much value in the case of small areas estimations (Puchner, 2015). Table 1 presents the main characteristics of two methods commonly used to measure the distance between entities under study. Stuart (2010) presents further details and comparisons on these methods.

CEM is a recent advance used to do exact matching on broad ranges of variables and it exploits categories rather than continuous measures (Iacus et al., 2011; Stuart, 2010). This allows to overcome the common problem of many individual/entities not being matched. CEM is widely

used in program evaluations where it is common to create control groups, based on specific covariates, in order to estimate the effects of the programs. Valid inference requires a method to randomly allocate beneficiaries to intervention or control groups. When there is an imbalance in background covariates between treated individuals and nonexposed individuals, CEM is an extended method that aims to correct this imbalance. As a member of the *Monotonic Imbalance Bounding* methods, CEM implies that the balance between the group that receives the treatment and the control group is chosen by the researcher before the analysis, in contrast to other methods where the balance is computed ex post and it is adjusted by re-estimations. The detailed description of this methodology and the formal proofs of its properties can be found in the work of Iacus et al. (2011).

Given an observational database, CEM creates a matched subsample. The methodology consists of two steps. First, a matched subsample is created by the CEM procedure and then, the new subsample is used to carry out the analysis. However, before the creation of the matched subsample, CEM requires the specification of two sets of variables: the treatment variables and the matching variables. The first set defines whether or not an individual received the treatment specified in the study. The second group includes those variables on which we want treatment and control groups to be similar after the matching process.

Once the treatment and matching variables are defined, the first step of CEM consists in coarsening each variable so that substantively indistinguishable values are grouped and assigned the same numerical value.

Let $X = (X1, X2, \ldots, Xk)$ denote a k-dimensional data set, where each column Xj includes the observed values of pretreatment variable j for the n sample observations. After recoding each variable, CEM creates clusters, each one composed by the same coarsened values of X.

Let us denote by s a generic cluster, by T^s the treated units in cluster s and by m_T^s. the number of treated units in cluster s. In the same way, for the control units, C^s and m_C^s. are defined. Then, the number of treated and control units are $m_T = U_{s \in S} m_T^s$. and $m_C = U_{s \in S} m_C^s$, respectively.

Finally, CEM assigns the following weight to each matched unit i:

$$w_i = \begin{cases} \mathbf{0}, \ i \in T^S \\ \dfrac{m_{Ti}^S + m_{Ci}^S}{m_{Ti}^S}, i \in C^S \end{cases}$$

If a unit is unmatched, it receives a weight of zero.

Finally, by using the computed weights a representative sample is created, and the main series are re-estimated. The description of the implementation of the methodology in different statistical platforms can be found in Firestone (2015).

Reconstruction of a tourism dataset: The case of a Chilean region

Tourism statistics in Chile

In order to illustrate the proposed method applied to tourism data, sample survey statistics from a region in Chile were used. During the last three decades, Chile presented a high economic performance that was followed by a significant growth in its tourism sector. In 2010, tourism contribution to the Gross Domestic Product was 3.23% and income from tourism (foreign exchange receipts) reached US$2316 million (Servicio Nacional de Turismo, 2011). Moreover, the number of international tourists doubled passing from 1.412 million in 2002 to 3.070 million in 2011 (INE, 2011). The tourism sample survey used is that of the Antofagasta region (Figure 1). This region,

Figure 1. Map of the Antofagasta region and its communes.
Source: Mapoteca, Biblioteca del Congreso Nacional de Chile.

Table 2. Average annual number of supplies of accommodation in Antofagasta, in the databases of INE-EMAT and SERNATUR.

	2003	2004	2005	2006	2007	2008	2009	2010	2011
SERNATUR	167	178	184	193	191	207	207	229	249
INE-EMAT	116	110	132	130	125	117	116	114	130

Source: INE-EMAT and SERNATUR.

placed in the north of Chile, is the second main destination of the country. It accounts for 15% of the total arrivals (national and international), whereas the region of Santiago and its surroundings registers 26% of arrivals. However, in terms of domestic tourism Antofagasta's arrivals equate those of Santiago (INE, 2008; INE, 2011).

As Aroca et al. (2013) already noted, sample surveys collected in the region of Antofagasta exhibit inconsistency, particularly those measuring the number and the capacity of suppliers of accommodations. There are several reasons for this. In Chile, there are three institutions responsible for tourism statistics: (1) the Central Bank's Department of National Accounts, (2) the National Statistical Institute (INE, Instituto Nacional de Estadísticas), and (3) the Chilean Official Tourism Destination Organization (SERNATUR, Servicio Nacional de Turismo). These institutions, albeit at different levels, are all responsible for tourism data collection. Particularly, the first institution—the Central Bank—collects essential tourism data to elaborate the National accounts and these are not object of the present study.

The second institution—INE—measures supply and demand of tourism accommodation through the Monthly Survey of Tourist Accommodation Facilities (named EMAT, Encuesta Mensual de Alojamiento Turístico) and this database for its nature and original objectives is a sample, but it is used as a census of tourism enterprises. However, this database is not regularly updated and does not take into account the natural life cycle of firms with their attrition, ingrowth, and accretion, creating therefore a distortion in the data. Thus, the resulting tourism statistics on arrivals and overnight stays are unreliable. It is the misuse that leads to a distorted representation of tourism activities, with consequences for policy-making decisions (Aroca et al., 2013). The third institution—SERNATUR—collects tourism statistics for management and policymaking.

Tourism data inconsistencies are showed in table 2, where the two sources—SERNATUR and INE-EMAT—are compared with regard to average number of accommodation's suppliers in Antofagasta (2003–2011). The INE-EMAT series shows a somewhat flat trend with few cyclical fluctuations, whereas the series of SERNATUR shows an almost continuous growth in the number of accommodations. The difference lies in the accretion, ingrowth, and attrition caused by the inability of INE-EMAT to regularly update the directory. Additionally, it is worth noticing that some accommodation suppliers are eliminated from the databases due to compliance to privacy regulations.

Application of the CEM to Chilean tourism data

Three tourism destinations in the region of Antofagasta (region II of Chile) have been considered for the purpose of this study. The region of Antofagasta is made up of three provinces and a total of nine communes (a commune is the smallest administrative district of Chile). Owing to data

Table 3. Methodological steps of the study.

Phase 1	Correction of tourism accommodation firms' directories Primary and secondary sources were used to reconstruct the pre-2010 directories of suppliers of accommodation in the three studied destinations: Antofagasta, Calama, and San Pedro de Atacama. Databases for the period 2003–2009 were reconstructed. Quality control was performed.
Phase 1.1: Secondary sources	Secondary sources consisted of two data sets: (i) directories of existing accommodation provided by each municipality and by SERNATUR for the year 2010 and (ii) other nontourism-specific data sources including, among others, tax records, websites, and phone directories.
Phase 1.2: Primary data	Primary data were collected through field visits, phone calls, and personal interviews. These aimed at identifying new suppliers of accommodation and confirming information on preexisting suppliers and ensured data reliability by capturing the changes in capacity and ownership of the suppliers of accommodation.
Phase 2	Sample weights computation using CEM Sample weights are computed using the first part of CEM method
Phase 2.1: Specification of variables	Treatment variables: indicate if the accommodation supplier was included in the INE-EMAT survey. Exact matching of variables: commune, the type of accommodation (hotel, apart hotel, etc.) and the number of rooms.
Phase 2.2: Coarse variables	Recoding in a way that substantively indistinguishable values are grouped and assigned the same numerical value X.
Phase 2.3: Clusters creation	Creation of clusters each one composed by the same coarsened values of X. For a generic cluster s: - the treated units in cluster s can be denoted as T^s and m_T^s is the number of treated units in cluster s. - control units in cluster s can be denoted as C^s and m_C^s is the number of control units in cluster s.
Phase 2.4: Assignment of weight	Weights are assigned to each matched unit as follows: $$w_i = \begin{cases} \mathbf{0}, \ i \in T^S \\ \dfrac{m_{Ti}^S + m_{Ci}^S}{m_{Ti}^S}, i \in C^S \end{cases}$$ If a unit is unmatched, it receives a weight of zero. This means, the weighed sample INE-EMAT is used to re-estimate the main series and the observations in the census that are not in the INE-EMAT were dropped.
Phase 3	Application to sample survey and reestimation By using the computed weights the new representative sample is created, and the main tourism series are re-estimated (number of accommodation suppliers, rooms, arrivals, and overnight stays).

Note: CEM: Coarsened Exact Matching.

availability, the three destinations used for the aim of this study are an aggregate of eight communes and do not necessarily overlap with the administrative structure of the provinces. However, as it can be seen from Figure 1 the chosen communes do have similar geographical characteristics. The three destinations considered are:

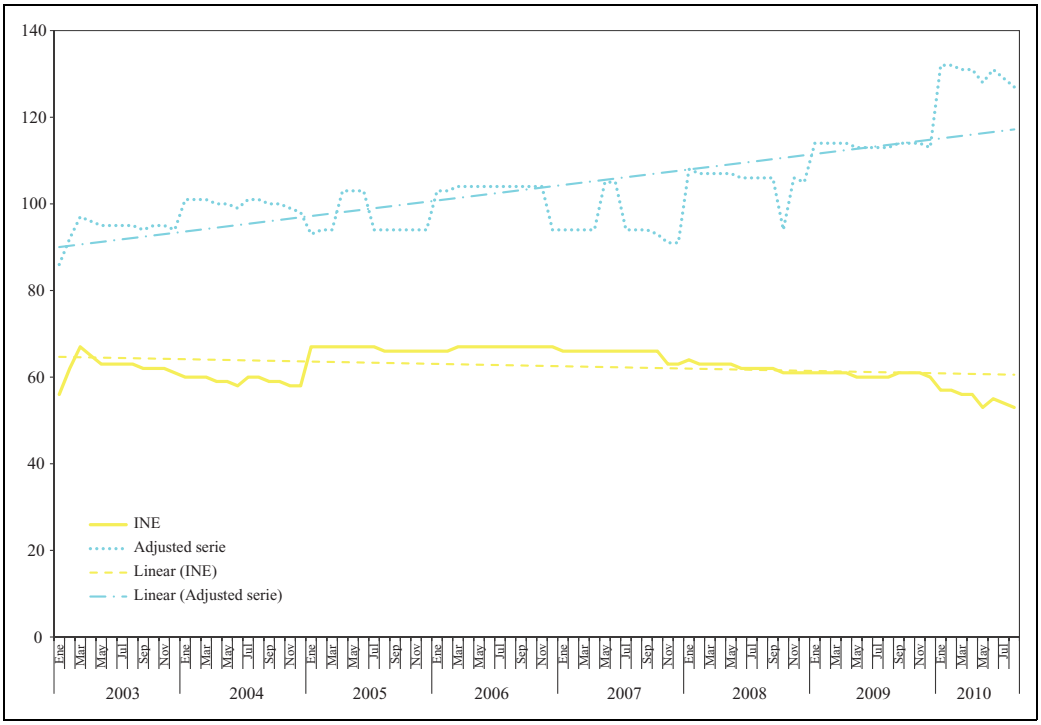

Figure 2. Number of accommodation suppliers by month, Anotofagasta, 2003–2010.

- Antofagasta, which includes the municipalities of: Antofagasta, Mejillones, Taltal, and Tocopilla, all having a coastline;
- San Pedro de Atacama, the innermost region at the border with Bolivia and Argentina; and
- Calama, which includes the municipalities of Calama, Ollagüe, and María Elena, located in the northern part of the region of Antofagasta.

In the tourism databases of the region of Antofagasta inadvertent omissions are present as some suppliers of newly created accommodation or changes in their sizes have not been recorded in time on existing registries (thus ignoring ingrowth and accretion of the dataset) while others are incorrectly present because the date of cessation of business activities is either not known or has not been accurately recorded (thus ignoring the attrition).

Aroca et al. (2013) have already introduced a methodology to correct tourism data distortions caused by attrition in nonrandom samples. In their work, sample weights to overcome statistical inaccuracy were created and applied to obtain valid estimates of population parameters. However, the CEM "is faster, is easier to use and understand, requires fewer assumptions, is more easily automated, and possesses more attractive statistical properties for many applications than do existing matching methods" (Blackwellet al., 2009, p. 524) and will be here applied and discussed.

The method herein used to correct the database (considering attrition, ingrowth, and accretion) consists of several sequential steps.

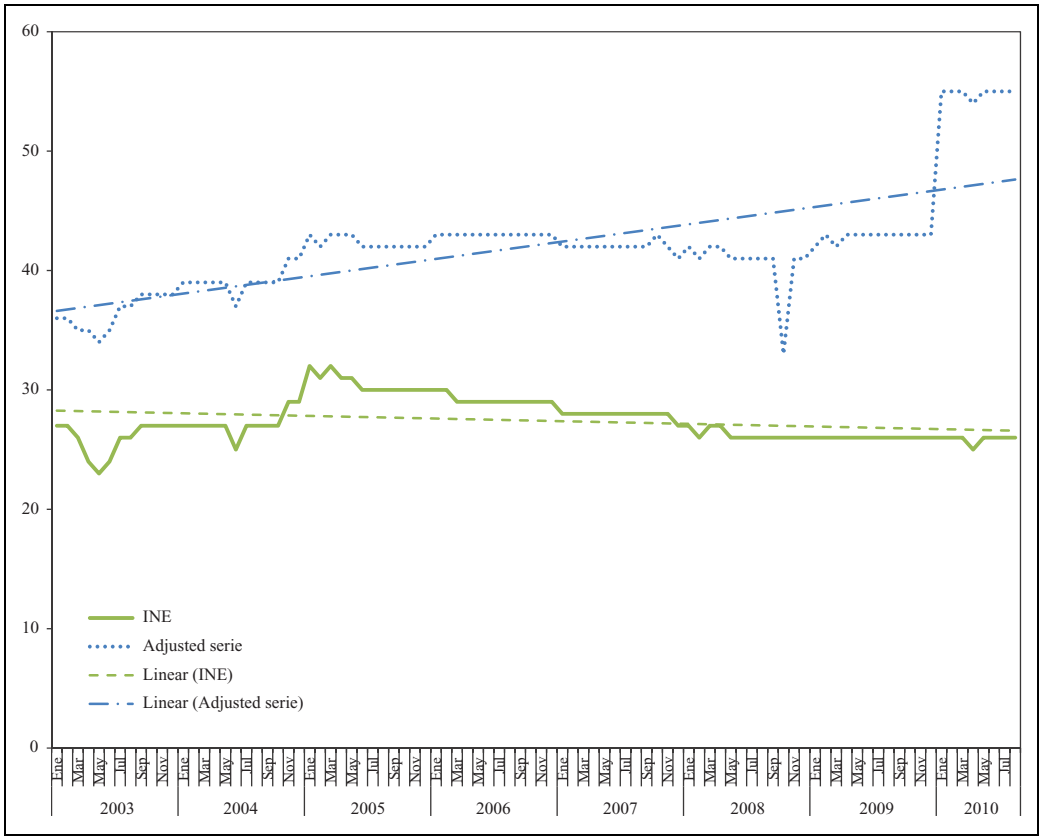

Figure 3. Number of accommodation suppliers by month, Calama, 2003–2010.

1) A census of all tourism accommodation in the studied communes (Antofagasta, Calama, and San Pedro de Atacama) was performed in 2010, and the directories for each of the years in the period 2003–2009 were reconstructed using the 2010 census data. That is, business establishments that were proved to have existed in previous years were added to the respective directories.
2) The sample weights were calculated for each year under investigation using the first part of the CEM method. In our application, the control units are those in the census, while the treated units are those in the INE-EMAT. For the purpose of this study, for each missing establishment a similar one included in the census was identified. Similarity was here defined in precise terms, which does not imply that the two units had to be identical, but close in their characteristics.
3) Using these weights—applied to the EMAT survey results—the most commonly used tourism statistics (number of suppliers of accommodation and rooms, arrivals, overnight stays) were re-estimated.

The methodology used in the case of the Antofagasta region comprises several phases that are presented in table 3.

Figure 4. Number of accommodation suppliers by month, San Pedro de Atacama, 2003–2010.

It has to be noted that this study adapts the CEM and thus to each missing establishment in the census it was possible to find a corresponding "similar" match. This methodological choice was possible due to the accentuated similarities of tourism enterprises in this region. One could speculate that if a different type of establishment would initiate its business during the time of the census, then no similar match could have been found. This is not the case of the current study, thus no unmatched observations are present. That is each business has a "clone." The main contribution of this study resides on the procedure to build the weights that allow to expand the results of the nonrandom sample to the whole population.

Empirical evidence

In this section the newly recalculated tourism statistics series for the period 2003–2010 and for each of the studied communes are presented to empirically show the effect of attrition, ingrowth, and accretion on tourism data.

The first recalculated series is the number of tourism accommodation suppliers. In Figures 2 to 4 the original and new series for each destination are compared. Clearly, the number of accommodation suppliers is significantly different before and after the re-estimation, and it becomes evident how tourism activities were underrepresented in the original series.

By looking at the figures, it is clear that all tourism activities represented in the data were similarly underrepresented in the original uncorrected series, and that, once corrected, the data

reflect levels of tourism activities that are more consistent with tourist arrivals, whether corrected or not.

The following monthly time series were recalculated on the basis of the new database: number of rooms, international tourism arrivals, occupancy rates, and number of employees permanently and temporarily working in the tourism sector. The recalculated series show substantial differences with the original time series showing the effect of accretion, ingrowth, and attrition on tourism statistics. Figures A1 to A12 in the appendix show the recalculated time series.

Conclusion

This study aimed at presenting a methodology to reconstruct tourism databases using sample weights built with the CEM, and showing how to successfully use them to overcome some methodological issues of supply-side tourism statistics.

The article outlined the issues related to statistical inaccuracies and focused mainly on the evaluation of tourism databases completeness and accuracy, showing that in many countries currently available statistics do exhibit inconsistencies that may lead to underrepresentation of supply and demand. Particular attention was devoted to the under-investigated issues of accretion, ingrowth, and attrition, and to the potential biases that they cause to tourism statistics.

A methodology to re-estimate a destination's tourism statistics was proposed and the subsequential steps presented. An innovative approach to calculate sample weights adapting the CEM was illustrated. The approach's ability to assist in correcting data distortion was discussed with an empirical application to a Chilean tourism dataset.

Tourism data in the region of Antofagasta are inaccurate enough to significantly degrade the tourism planning function. Correcting them—with the method herein proposed—allows to re-align the valence of the destination tourism industry and facilitates sound international comparisons. Tourism time series can be improved by sample weights—calculated with the CEM—and applied to a nonrandom sample of suppliers of tourism accommodations. A significant difference between the directory of suppliers of accommodation surveyed by the INE and the actual total number of suppliers of accommodation constituted the starting point of the empirical application. The difference—due to a misuse of the data—is substantial and has a negative effect on the perception of the evolution of regional tourism. The database accretion, ingrowth, and attrition were accounted for, thanks to the following three steps: (1) correction of tourism accommodation firms' directories; (2) sample weights computation using CEM; and (3) application to sample survey and re-estimation.

The results confirmed the disparities in levels of suppliers of accommodations and activities—up to twice that of the published estimates—as well as significant disparities between the pre- and post-weighing time-series trends.

Chilean tourism officials can apply the methodology to update their directory, rebuild the tourism time series of other regions, and those of the country as a whole.

More importantly, the methodology described here and applied to the case of Antofagasta, can be adapted to other destinations for which the population data required to construct the sample weights are available. It is conjectured that there are many such areas and that there are many opportunities to improve tourism statistics upon which economic and policy planning is based. From a methodological point of view, this research proposes and demonstrates the validity of a methodology to correct a common problem among accommodation statistics. The article also

contributes to the methodological advancements in that it presents the CEM and its benefits to better calculate sample weights and adjust data susceptible to attrition, ingrowth, and accretion biases.

This study contributes to the research stream on statistical imputation of missing value in tourism (e.g. Aroca et al. 2013; Fontana and Pistone, 2010) with relevance for a broader range of applications where tourism statistics are questionable or demonstrably inaccurate. Thereby it represents a tool for tourism destinations in order to improve tourism statistics. While the methodology offers ways to correct some data inaccuracy, considerable future research is necessary to address the following issues:

- ease of application, for example reducing the field research needed to update the tourism directories;
- evaluate the goodness of fit of the method, for example by comparing different matching methods using tourism data;
- test the model in different destinations as to verify easiness of replicability.

Still, the method described has a great potential that is waiting to be fully exploited in order to account for accretion, ingrowth, and attrition in tourism data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Patricio Aroca acknowledges the support of COES CONICYT/FONDAP/15130009.

Supplemental material

The appendix is available at: http://journals.sagepub.com/doi/suppl/10.5367/te.2015.0500.

References

Abadie A and Imbens GW (2011) Bias corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics* 29: 1–11.

Aroca P, Brida JG and Volo S (2013) Applying weights to correct distortions in a non-random sample: an application to Chilean tourism time series data. *Tourism Economics* 19(2):453–472.

Blackwell M, Iacus SM, King G, et al. (2009) CEM: Coarsened Exact Matching in stata. *The Stata Journal* 9(4): 524–546.

Boudreau C and Yan M (2010) *Construction and Use of Sampling Weights for the International Tobacco Control (ITC) Germany Survey, ITC Germany Survey Waves 1 (2007) and 2 (2009). Technical Report*. Available at: Http://Itc.Media-Doc.Com/Files/Report_Publications/Technical_Report/Nl_W13_Techreport_July62010rev.Pdf (accessed 2 May 2015).

Burkart A and Medlik S (1981) *Tourism, Past, Present and Future*. 2nd ed. London: Butterworth Heinemann.

De Cantis S and Ferrante M (2013) The implementation and main results of the TLS design in a survey on incoming tourism in Sicily. In: Oliveri AM and De Cantis S (eds) *Analysing Local Tourism*. Maidenhead: McGraw-Hill Education, pp. 255–268.

Firestone R (2015) *Evaluating Program Effectiveness: Key Concepts and How to Use Coarsened Exact Matching* Washington, DC: PSI. Available at: http://www.psi.org/publication/evaluating-program-effectiveness-key-concepts-and-how-to-use-coarsened-exact-matching/ (accessed 5 May 2016).

Fontana R and Pistone G (2010) Anticipating Italian census tourism data before their official release: a first solution and its implementation to Piemonte, Italy. *International Journal of Tourism Research* 12(5): 472–480.

Gu X and Rosenbaum PR (1993) Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal Computational and Graphical Statistics* 2(4): 405–420.

Guizzardi A and Bernini C (2012) Measuring underreporting in accommodation statistics: evidence from Italy. *Current Issues in Tourism* 15(6): 597–602.

Heitjan DF and Rubin DB (1991) Ignorability and coarse data. *Annals of Statistics* 19(4): 2244–2253.

Hofer SM and Hoffman L (2010) Statistical analysis with incomplete data: a developmental perspective. In: Little TD, Bovaird JA and Card NA (eds) *Modeling Contextual Effects in Longitudinal Studies*. Mahwah: Lawrence Erlbaum Associates, pp. 13–32.

Iacus SM, King G and Porro G (2011) Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* 106(493): 345–361.

Imai K, King G and Stuart EA (2008) Misunderstandings among experimentalists and observationalists in causal inference. *Journal of the Royal Statistical Society Series A* 171(2): 481–502.

INE (2008) Metodología encuesta mensual de establecimientos de alojamiento turístico. Available at: http://www.ine.cl/canales/chile_estadistico/estadisticas_economicas/turismo/metodo/turismometodologia.pdf (accessed 5 May 2016).

INE (2011) Informe anual de turismo. Available at: http://www.ine.cl/canales/menu/publicaciones/calendario_de_publicaciones/pdf/turismo_2011.pdf (accessed 5 May 2016).

Kim JK and Hong M (2012) Imputation for statistical inference with coarse data. *The Canadian Journal of Statistics* 40(3): 604–618.

Lickorish LJ (1997) Travel statistics-the slow move forward. *Tourism Management* 18(8): 491–497.

Massieu A (2001) A system of tourism statistics (STS) Scope and Content. In: Lennon JJ (ed) *Tourism Statistics*. London: Continuum, pp. 3–13.

McCoy TP, Ip EH, Blocker JN, et al. (2009) Attrition bias in a US internet survey of alcohol use among college freshmen. *Journal Studies Alcohol Drugs* 70(4): 606–614.

McGuigan KA, Ellickson PL, Hays RD, et al. (1997) Adjusting for attrition in school-based samples: bias, precision, and cost trade-offs of three methods. *Evaluation Review* 21(5): 554–567.

Meis S (2001) Towards comparative studies in tourism satellite accounts. In: Lennon JJ (ed) *Tourism Statistics*. London: Continuum, pp. 14–23.

Pine RJ (1992) Towards a useful measure of tourism activity at individual country level. *Tourism Management* 13(1): 91–94.

Puchner V (2015) *Evaluation of Statistical Matching and Selected SAE Methods: Using Micro Census and EU-SILC Data* Wiesbaden: Springer Spektrum.

Rosenbaum PR and Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Rosenbaum PR and Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39(1): 33–38.

Rubin DB and Thomas N (1992) Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics* 20: 1079–1093.

Rubin DB and Thomas N (1996) Matching using estimated propensity scores, relating theory to practice. *Biometrics* 52: 249–264.

Rubin DB (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74: 318–328.

Schafer JL and Graham JW (2002) Missing data: our view of the state of the art. *Psychological Methods* 7: 147–177.

Servicio Nacional deTurismo (2011) Estimación PIB turístico año 2010 y su evolución desde el año2003. Proyecto Cuenta Satélite de Turismo. Available at: http://www.sernatur.cl/documentos/?did=111 (accessed 5 May 2016).

Stuart EA (2010) Matching methods for causal inference: a review and a look forward. *Statistical science* 25(1): 1–21.

Volo S (2004) The role of roots in the perception of a destination: an exploratory study on sicily. *Journal of Hospitality & Leisure Marketing* 11(2–3): 19–29.

Volo S (2010) Seasonality in Sicilian tourism demand – an exploratory study. *Tourism Economics* 16(4): 1073–1080.

Volo S and Giambalvo O (2008) Tourism statistics: methodological imperatives and difficulties: the case of residential tourism in island communities. *Current Issues in Tourism* 11(4): 369–380.

Volo S and Pardew D (2013) The Costa Concordia and similar tragic events: the mathematics and psychology of the loss and restoration of travellers' trust. *Current Issues in Tourism* 16(2): 197–202.