Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen - Bolzano

Fakultät für Wirtschaftswissenschaften | Facoltà di Economia | School of Economics and Management

BEMPS –

Bozen Economics & Management Paper Series

NO 22 / 2014

Tourism statistics: correcting data inadequacy using coarsened exact matching

Patricio Aroca, Juan Gabriel Brida,
Juan Sebastián Pereyra, Serena Volo

TOURISM STATISTICS: CORRECTING DATA INADEQUACY USING COARSENED EXACT MATCHING

Patricio Aroca[1], Juan Gabriel Brida[2], Juan Sebastián Pereyra[3] and Serena Volo[4]

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 paper is to describe Coarsened Exact Matching, a methodology useful to improve tourism statistics. This method provides 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.

[1] Business School, Universidad Adolfo Ibáñez, Viña del Mar, Chile. E-mail: patricio.aroca@uai.cl

[2] School of Economics and Management - Free University of Bolzano, Italy. E-mail: JuanGabriel.Brida@unibz.it

[3] School of Economics and Management - Free University of Bolzano, Italy. E-mail: JuanSebastian.PereyraBarreiro@unibz.it

[4] School of Economics and Management - Free University of Bolzano, Italy. E-mail: serena.volo@unibz.it

Keywords: attrition bias, accretion bias; sample weights; accommodations; tourism planning; Chile.

1.INTRODUCTION

Tourism statistics are collected by national statistical offices and national tourism organizations with the aim of assisting economists, public officials and tourism operators in their decision making. Academic studies on tourism statistics emphasize the importance of collecting accurate and reliable statistics that can support tourism planning and forecasting (Aroca, Brida & Volo, 2013; Burkart & Medlik, 1981; Lickorish, 1997; Massieu, 2001; Meis, 2001; Pine, 1992). Particularly, tourism statistics are needed by (a) governments to evaluate dimensions and significance of tourism, (b) destinations to predict tourism flows and manage tourism impacts; and (c) industry's decision-makers for strategic marketing purposes (Massieu, 2001; Volo & Giambalvo, 2008; Wöber, 2000). Therefore, the collection of tourism statistics is a main concern for many tourism organisations at both national and international level.

Issues of data reliability, the problem of respondent´s mobility, physical difficulties in data collection and the variety of data users and needs have been

acknowledged in tourism research (Edwards, 1991; Hannigan, 1994; Latham and Edwards, 2003; Parroco, Vaccina, De Cantis, & Ferrante, 2012; Ritchie, 1975). The recent effort undertaken by the United Nation World Tourism Organization (1994, 1995, 1998, 2000) to harmonize and standardize the collection of tourism data is certainly the most noteworthy action in the direction of validity and comparability of tourism statistics across countries. Nevertheless, in many countries inconsistencies in tourism statistics are still quite frequent (Aroca et al., 2013; Guizzardi & Bernini, 2012; Volo & Giambalvo, 2008).

Volume statistics, the most common of which being the measure of tourists´ arrivals, are usually the result of counting procedures either at a entereing the destination or at accommodation establishments. In measuring tourism, accommodation statistics are of particular relevance, but regrettably these often underestimate true domestic and international mobility (Guizzardi & Bernini, 2012; Latham & Edwards, 2003). In this study, a method to adjust data inconsistencies is presented and its benefits are shown when applied to accommodations´ data. Thus, the aim of this paper is to introduce tourism researchers to a methodology for reconstructing tourism databases using sample weights built with the Coarsened Exact Matching and to show how to successfully use them in order to overcome some methodological issues of supply-side tourism statistics.

The organization of this paper is as follows. In the second section, an overview of the inconsistencies and difficulties related to collection of accurate and valid statistics is presented with attention to the issue of attrition and accretion biases in data collection. Then, in section three the Coarsened Exact Matching (CEM) is presented and its ability to compute more reliable sample weights is discussed. A methodology to reconstruct a tourism database is proposed and its application to tourism supply statistics is illustrated in section four, where tourism statistical data from a Chilean region (Antofagasta) are used to test the method. A comparison between the original data and those obtained after computing the sample weights with the CEM algorithm is presented. The final section offers a conclusive discussion about the significance of the method, its benefits for policy makers and tourism statisticians.

2.TOURISM SUPPLY DATA: INCONSISTENCIES ISSUES

The relevance of defining and measuring tourism along with the issue of comparability with other industries have been at the centre of the scientific debate of those interested in tourism measurements and benchmarking (Lickorish, 1997; Smith, 1988). Due to the temporal and spatial nature of tourism, the collection of tourism statistics is somewhat replete with methodological challenges and impediments. Furthermore, the highly mobile nature of the sampled population – tourists – makes it difficult to ensure probabilistic sampling procedures (De Cantis & Ferrante, 2013; Latham & Edwards, 2003; Mendola & Milito, 2013). Consequently, several inconsistencies affect the creation or selection of tourism-relevant data resulting in misleading statistics (Volo & Giambalvo, 2008). Most of the issues documented arise from the lack of, or variations in, operational definitions of methodologically relevant constructs or from changes of the administrative sources of data. Recent advancements have been made in the international collection of tourism statistics and growing attention still pervades the efforts of the national statistical offices as to ensure data availability, accuracy and comparability.

The commonly used measures of tourists´ movements and the related variables of interests are: volume statistics, expenditures and tourists' profiles. The two basic procedures to record international movements comprise: data collection at international boarders and registration of arrivals at the accommodation establishments. These procedures, originally designed for purposes different than that of collecting tourism flows data, are most commonly used as they reduce the burden on respondents/tourists (Latham & Edwards, 2003; Shackeleford, 1980). Indeed, it is common practice around the word to collect statistics on tourism flows from the supply side, that is: statistical offices collect information on tourism flows from accommodation establishments. Data are then used to study variation of flows, estimate the relevance of tourism in the country and forecast future patterns. Thus, the relevance of collecting accurate data is paramount in order to make sound tourism planning decisions.

Despite the national statistical offices´ awareness of the importance of data accuracy and consistency and their continuous effort towards new approaches to data collection, in some countries unreported tourism is

witnessed (Aroca et al., 2013; Guizzardi & Bernini, 2012; Volo & Giambalvo, 2008).

Some of the most common problems that tourism statisticians have to face are those related to biases in data collection which 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 & Hoffman, 2010).

The loss of participants or entities over time due to transience, dropouts, withdrawals and protocol deviations is known as attrition. These selectivity problems are common in longitudinal, panel and multi-wave studies, which as a result account for increasing nonresponses throughout time, therefore threatening the studies´ external and internal validity (Frees, 2004), and in extreme cases, preventing researchers and practitioners from a full and accurate use of the datasets (Aroca et al., 2013). The existence of attrition in a database,

and particularly of an high attrition rate, may lead to an attrition bias. However, it is worth noticing that the bias happens only when there is a systematic difference in the outcome variables of interest between the entities that remained in the sample and those that dropped out (McCoy et al., 2009). Scientists in the fields of psychology, medicine and epidemiology have recently started to assess the impact of attrition biases with meta-analytical approaches (Tierney & Stewart, 2005), while in management and economics the issue remains quite marginal and therefore results and conclusions of many studies could be susceptible to attrition biases.

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. For their nature, accretion biases are frequently studied in natural sciences and the threat they pose to studies´ external validities

have also been acknowledged in longitudinal and panel studies in applied social research (Tebes et al., 1996).

Attrition, in-growth 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 non-respondent 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 in Chile using sample weights. The use of sample weights is effective to correct potential biases that might result from the non-representativeness of the sample (Boudreau & Yan, 2010), but other techniques are available and may

lead to better results in accounting for accretion, ingrowth and attrition in databases. For instance some of the methods documented in medical and psychological literature are: selection modelling 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 & Graham, 2002).

The present study seeks to provide tourism researchers with an alternative methodology to assist in dealing with incomplete data of a given database. The method herein proposed consists of a sequential reconstruction of a tourism database with the calculation of sample weight using the Coarsened Exact Matching (CEM). The next section describes the CEM in detail, while the overall reconstruction of the datasets and an application of CEM is provided in section four.

3. THE COARSENED EXACT MATCHING

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 Coarsened Exact Matching (CEM) is a particular member of the matching methods known as Monotonic 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 (for example 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. 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.

==========

PLEASE INSERT TABLE 1 APPROXIMATELY HERE

==========

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. 2009; 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 non-exposed 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, . . ., 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 denote by “s” a generic cluster, by “Ts” 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, “Cs” and “m_C^s” are defined. Then, the number of treated and control units are “ $m_T = \cup_{s \in S} m_T^s$ ” and “ $m_C = \cup_{s \in S} m_C^s$ ”, respectively.

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

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

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 Population Services International (2011).

4. RECONTRUCTION 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 (GDP) was 3.23% and income from tourism (foreign exchange receipts) reached US$ 2,316 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, 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).

==========

PLEASE INSERT FIGURE 1 APPROXIMATELY HERE

==========

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 the Central Bank's Department of National Accounts, the National Statistical Institute (INE) and Chilean Official Tourism Destination Organization (SERNATUR) are all responsible, albeit at different level, of tourism data collection. Particularly, the National Statistical Institute (INE) measures supply and demand of tourism accommodation through the Monthly Survey of Tourist Accommodation Facilities (named EMAT). However, the database of facilities to be surveyed is prepared by INE and while, for its nature and original objectives it is a sample, this database is then used as a census of tourism enterprises. The main issue is that the 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 accommodation data collection and producing unreliable tourism statistics on arrivals and overnight stays. As noticed in Aroca et al. (2013) this misuse leads to a misrepresentation of tourism activities, with consequences for policy-making decisions. As an example, table two shows the difference between the two data sources, SERNATUR and INE-EMAT, for the period 2003-2011 with regard to the average number of suppliers of accommodation in Antofagasta. 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.

==========

PLEASE INSERT TABLE 2 APPROXIMATELY HERE

==========

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). Due to data 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 one the chosen communes do have similar geographical characteristics. The three destinations considered are:

- 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 non-random 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" (Blackwell, Iacus, King & Porro, 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.

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 INE-EMAT survey, while the treated units are those in the census.

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 three.

==========

PLEASE INSERT TABLE 3 APPROXIMATELY HERE

==========

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 two, three and four 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 under-represented in the original series.

==========

PLEASE INSERT figures 2, 3 and 4 APPROXIMATELY HERE

==========

By looking at the figures, it is clear that all tourism activities represented in the data were similarly under-represented 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 re-calculated 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. The recalculated time series are available in appendix.

5.CONCLUSION

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

The paper 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 under-

representation 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 sub-sequential steps presented. An innovative approach to calculate sample weights – the Coarsened Exact Matching – was illustrated and its ability to assist in correcting data distortion 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 non-random 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 was accounted for thanks to the following three steps: (1) Correction of tourism accommodation firms´ directories; (2) Sample weights computation using Coarsen Exact Matching; 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 paper also contributes to the methodological advancements in that it presents the Coarsen Exact Matching 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, e.g. reducing the field research needed to update the tourism directories,
- evaluate the goodness of fit of the method, e.g. 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, in growth and attrition in tourism data.

TABLES AND FIGURES

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, 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 interesting 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).	$D_{ij}= \|e_i-e_j\|$, 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).	If the treatment and control groups do not have substantial overlap (in terms

	Also, it has problems when covariates are not normally distributed.	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).

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

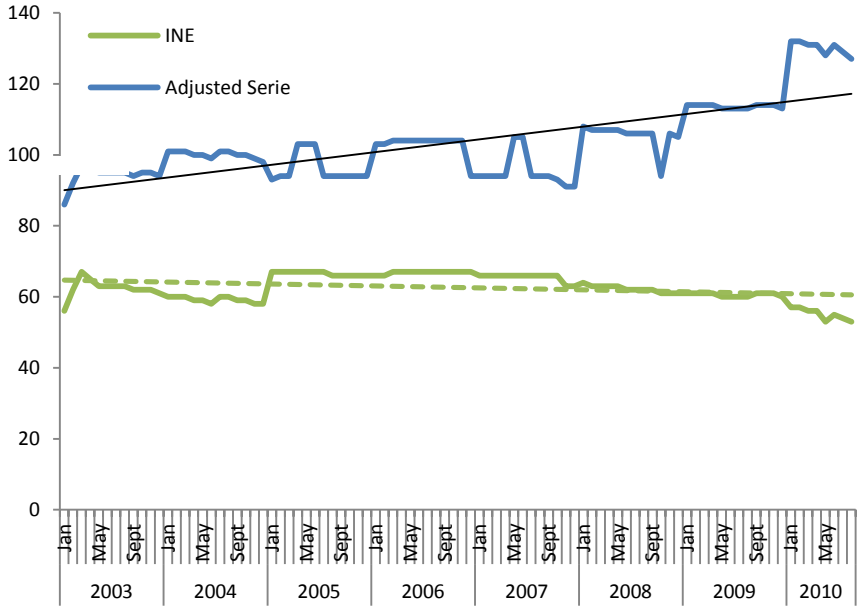
TABLE 3 Methodological steps of the study

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 datasets: (i) directories of existing accommodation provided by each municipality and by SERNATUR for the year 2010; and (ii) other non-tourism 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 pre-existing suppliers and ensured data reliability by capturing the changes in capacity and ownership of the suppliers of accommodation.	
Phase 2	***Sample weights computation using Coarsen Exact Matching*** 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 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. -the 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: $$\boldsymbol{w_i} = \begin{cases} \mathbf{0}, i \in T^S \\ \dfrac{m_{Ti}^S + m_{Ci}^S}{m_{Ti}^S}, i \in C^S \end{cases}$$ 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 re-estimation***	
	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).	

FIGURE 1 Map of the Antofagasta region and its communes.

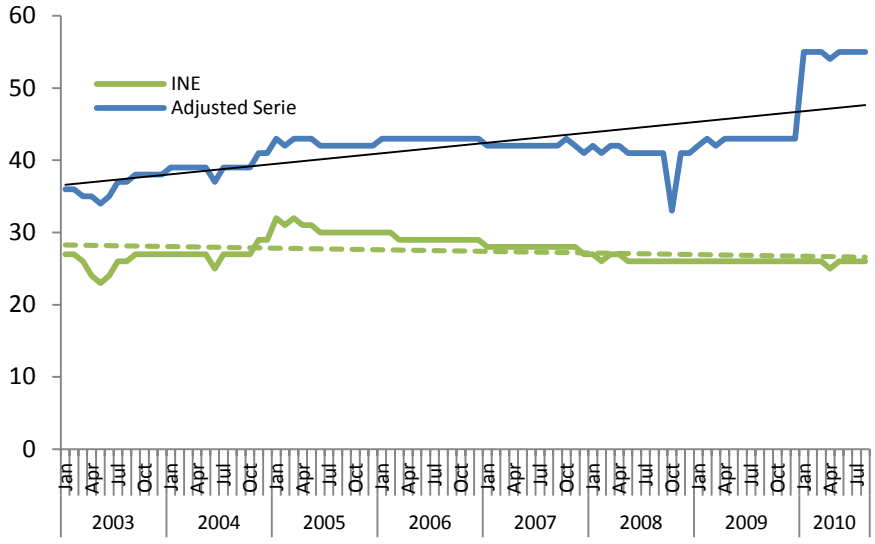
Source: http://www.manuelpena007.blogspot.it/2012_09_01_archive.html (last accessed 11 April 2014)

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



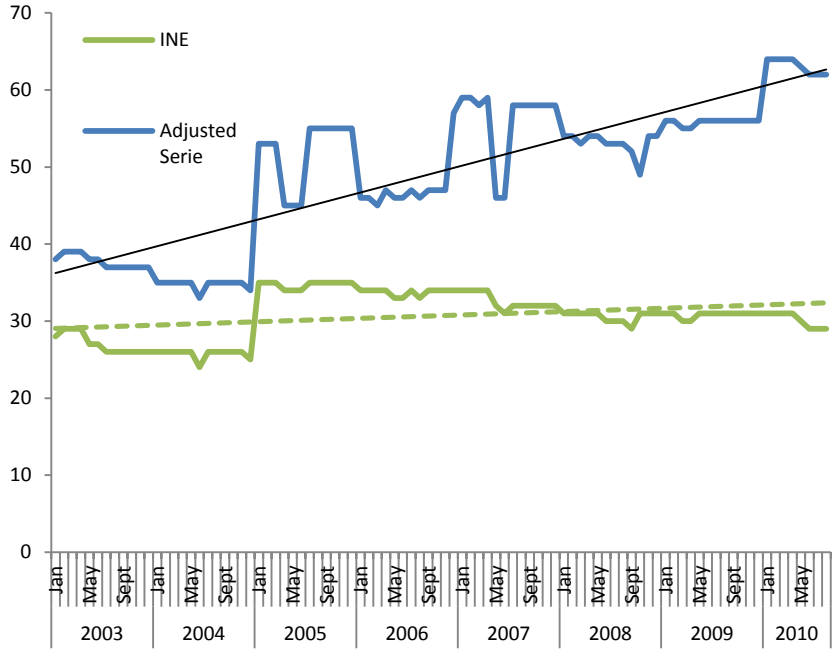
Source: Authors' elaboration from EAT-INE database.

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



Source: Authors' elaboration from EAT-INE database.

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



Source: Authors' elaboration from EAT-INE database.

REFERENCES

Abadie, A. & Imbens, G. W. (2011). Bias corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, 29, 1-11.

Aroca, P., Brida J.G., & 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), April 2013, 453-472.

Blackwell, M., Iacus, S.M., King, G. & Porro, G. (2009) cem: Coarsened Exact Matching in Stata, *The Stata Journal*, 9 (4), 524-546.

Boudreau, C., & 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 (http://itc.media-doc.com/files/Report_Publications/Technical_Report/nl_w13_techreport_july62010rev.pdf).

Burkart, A., & 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 A. M. Oliveri & S. De Cantis (Eds.), *Analysing Local Tourism* (255-268). United Kingdom: McGraw-Hill Education.

Edwards, E. (1991) The reliability of tourism statistics, *EIU Travel and Tourism Analyst,* 1: 62-75.

Fontana, R., & 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.

Frees, E.W. (2004). *Longitudinal and Panel Data. Analysis and Applications in the Social Sciences*, Cambridge University Press, Cambridge.

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

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

Hannigan, K. (1994), Developing European community tourism statistics. *Annals of Tourism Research*, 21(2), 415-417

Heitjan, D. F. & Rubin, D. B. (1991) Ignorability and Coarse Data. *Annals of Statistics.* 19 (4), 2244-2253

Hofer, S.M., & Hoffman, L. (2010). Statistical analysis with incomplete data: a developmental perspective, in T.D. Little, J.A. Bovaird, &

N.A. Card, (Eds.), *Modeling Contextual Effects in Longitudinal Studies.* Mahwah, NJ: Lawrence Erlbaum Associates.

Iacus, S.M., King, G., & 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. & Stuart, E. A. (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. (http://www.ine.cl/canales/chile_estadistico/estadisticas_economicas/turismo/metodo/turismometodologia.pdf).

INE (2011), Informe Anual de Turismo. (http://www.ine.cl/canales/menu/publicaciones/calendario_de_publicaciones/pdf/turismo_2011.pdf).

Kim, J. K. & Hong, M. (2012), Imputation for statistical inference with coarse data *The Canadian Journal of Statistics* 40(3), 604–618.

Latham J. and Edwards C. (2003) The Statistical measurements of Tourism, in C. Cooper, (Ed.),*Classic Reviews in Tourism* England: Channel View Publications.

Lickorish, L.J. (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 J.J. Lennon (Ed.) *Tourism Statistics* (3-13) London: Continuum.

McCoy, T.P., Ip, E.H., Blocker, J.N., Bloker, J.N., Champion, H., Rhodes, S. D., Wagoner, K. G., Mitra, A., & Wolfson, M. (2009) Attrition bias in a US internet survey of alcohol use among college freshmen. *Journal Studies Alcohol Drugs* 70 (4):606–614.

McGuigan, K. A., Ellickson, P. L., Hays, R. D., & Bell, R. M. (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 J.J. Lennon (Ed.) *Tourism Statistics* (14-23) London: Continuum.

Mendola, D. & Milito, A. M. (2013) Sampling in Local Tourism Quantification: Critical Issues and Field Experience. In A. M. Oliveri & S. De Cantis (Eds.), *Analysing Local Tourism* (53-64). United Kingdom: McGraw-Hill Education.

Parroco, A. M., Vaccina, F., De Cantis, S. & Ferrante, M. (2012). Multi-Destination Trips and Tourism Statistics: Empirical Evidences in Sicily. *Economics: The Open-Access, Open-Assessment E-Journal*,

Vol. 6, 2012-44. http://dx.doi.org/10.5018/economics-ejournal.ja.2012-44

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

Population Services International (2011), Use of Coarsened Exact Matching (CEM) in *Program Evaluation.* Manuscript PSI.

Ritchie, J. R. Brent (1975). Some Critical Aspects of Measurement Theory and Practice in Travel Research. *Journal of Travel Research,* 14 (1) 1–10.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39(1), 33–38.

Rubin, D. B. & Thomas, N. (1992). Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics* 20 1079–1093.

Rubin, D. B. & Thomas, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics* 52 249–264.

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

Schafer, J.L. & Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychological Methods*, 7, 147-177.

Servicio Nacional de Turismo (2011), Estimación PIB turístico año 2010 y su evolución desde el año 2003. Proyecto Cuenta Satélite de Turismo. (http://www.sernatur.cl/estudios-y-estadisticas?did=407).

Shackleford, P. (1980), Keeping tabs on tourism: A manager's guide to statistics. *International Journal of Tourism Management,* 1(3), 148-157.

Smith, S. L. J. (1988), Defining tourism a supply-side view. *Annals of Tourism Research,* 15(2), 179-190.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science*, 25(1), 1-21.

Tebes, J.K., Snow, D.L., Ayers, T. S. & Arthur, M. W. (1996). Panel Accretion and External Validity in Adolescent Substance Use Research, *Evaluation Review* 20, 470-484.

Tierney, J.F., Stewart, L.A. (2005) Investigating patient exclusion bias in meta-analysis. *International Journal of Epidemiology* 34, 79–87.

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.

Wöber, K.W. (2000) Standardizing City Tourism Statistics, *Annals of Tourism Research* 27, 1, 51-68.

World Tourism Organisation (1994) *Recommendation on Tourism Statistics*. Madrid: World Tourism Organisation.

World Tourism Organisation (1995) *Technical manual n. 2 the collection of Tourism Expenditure statistics*. Madrid: World Tourism Organisation.

World Tourism Organisation (1998) *A satellite account for tourism* (4th draft). Madrid: World Tourism Organisation.

World Tourism Organisation (2000) *Tourism Satellite Account (TSA): Methodological References.* Madrid: World Tourism Organisation.

APPENDIX

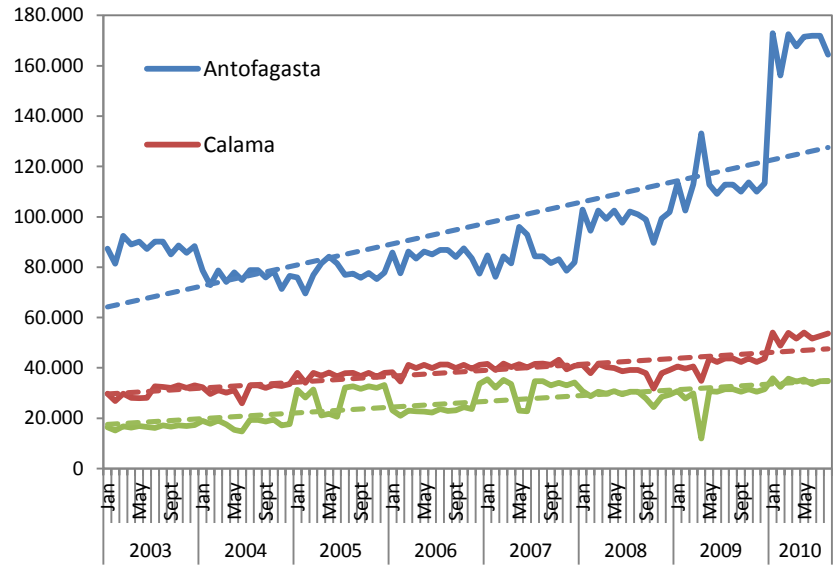


Figure 5: Number of rooms by month by commune, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

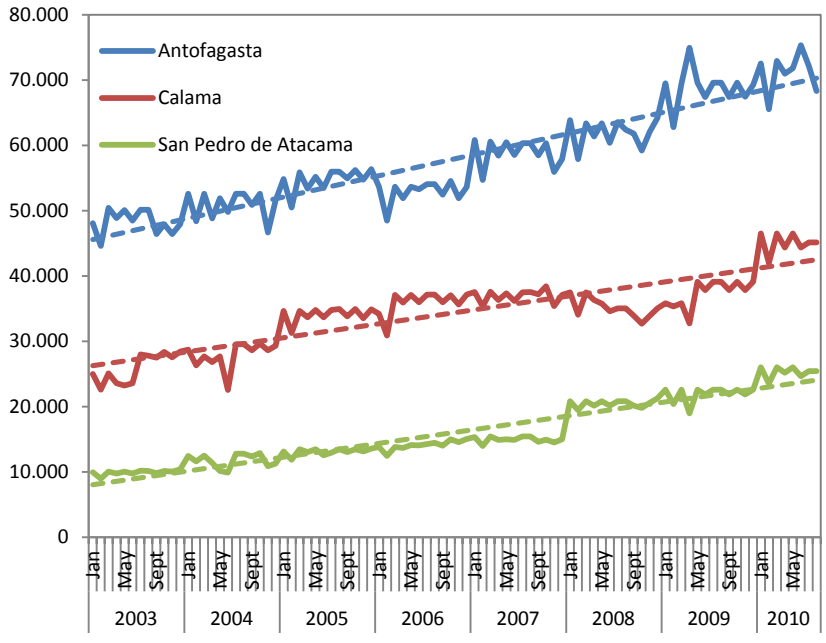


Figure 6: Number of rooms offer by hotels by commune, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

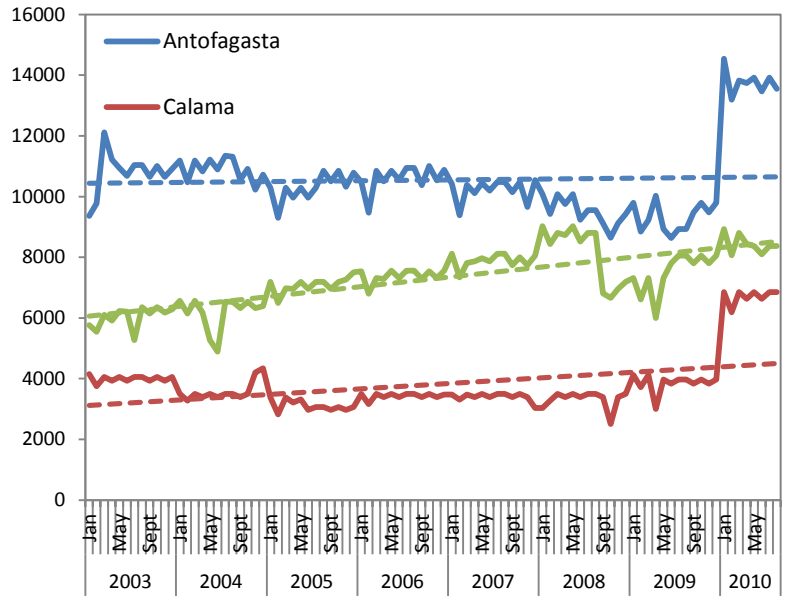


Figure 7: Number of rooms offer by residential and motels by commune, 2003-2010.

Source: Authors' elaboration from EAT-INE database.

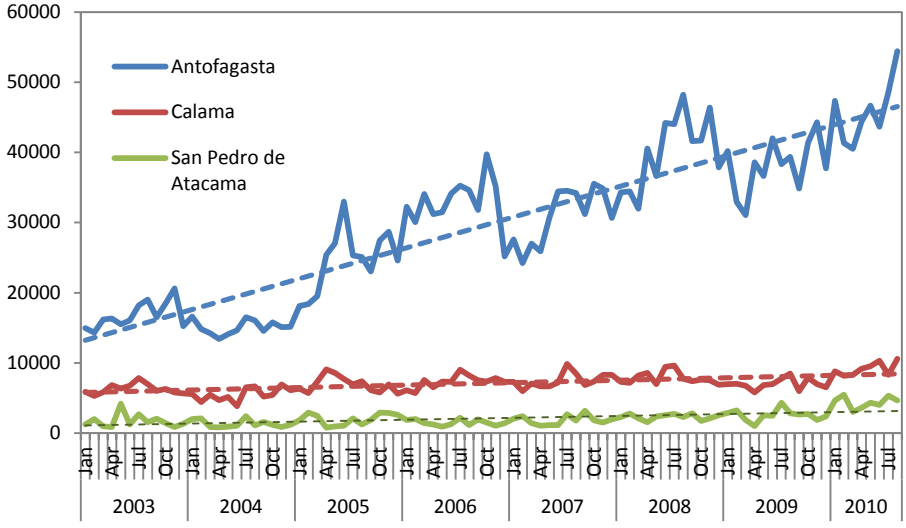


Figure 8: Number of national tourist arrivals by month by commune, 2003-2010.

Source: Authors' elaboration from EAT-INE database.

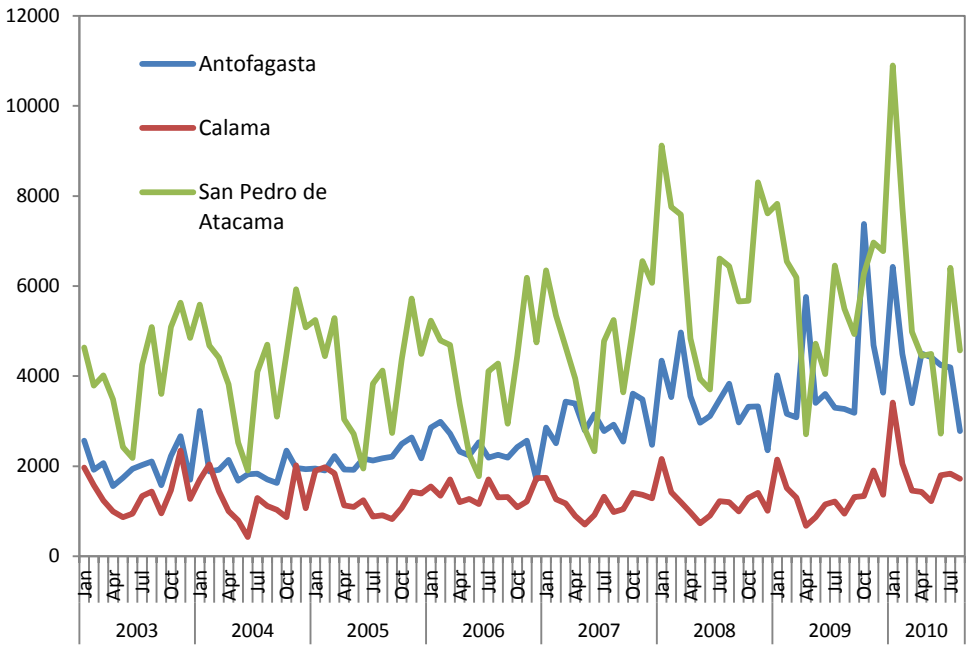


Figure 9: Number of international tourist arrivals by month, by commune, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

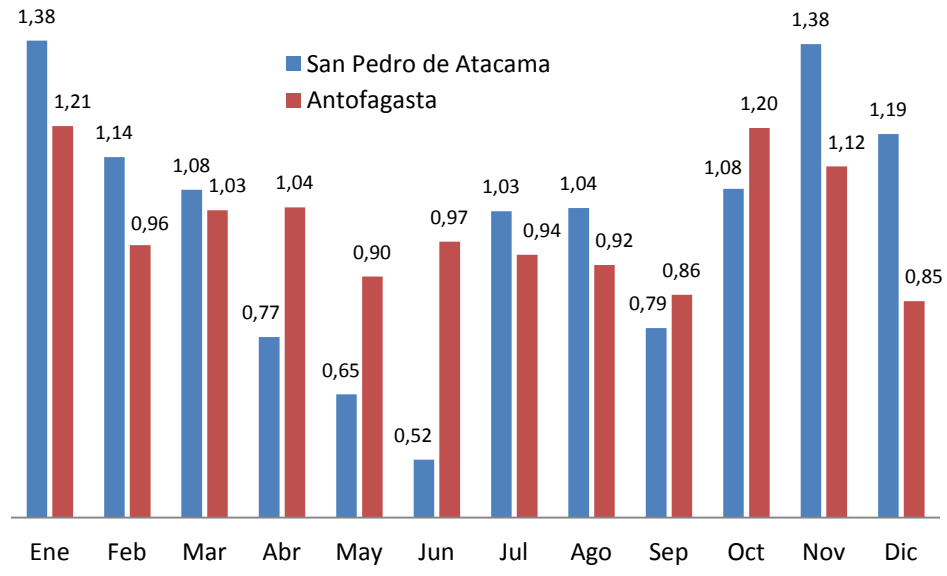


Figure 10: Seasonality of international tourist arrivals by commune.
Source: Authors' elaboration from EAT-INE database.

Figure 11: Occupancy rate in hotels, by commune, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

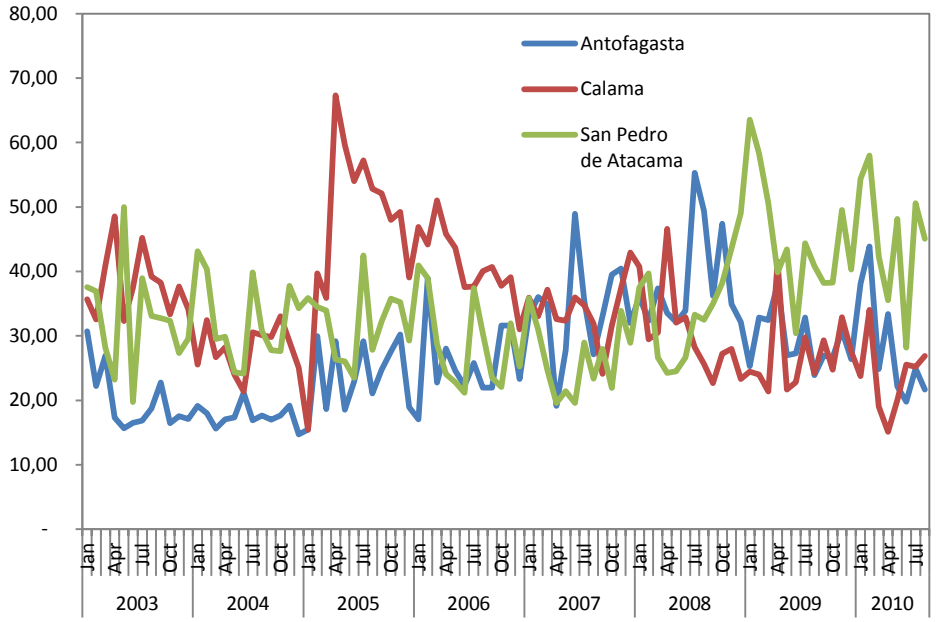


Figure 12: Occupancy rate in residenciales and motels, by commune, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

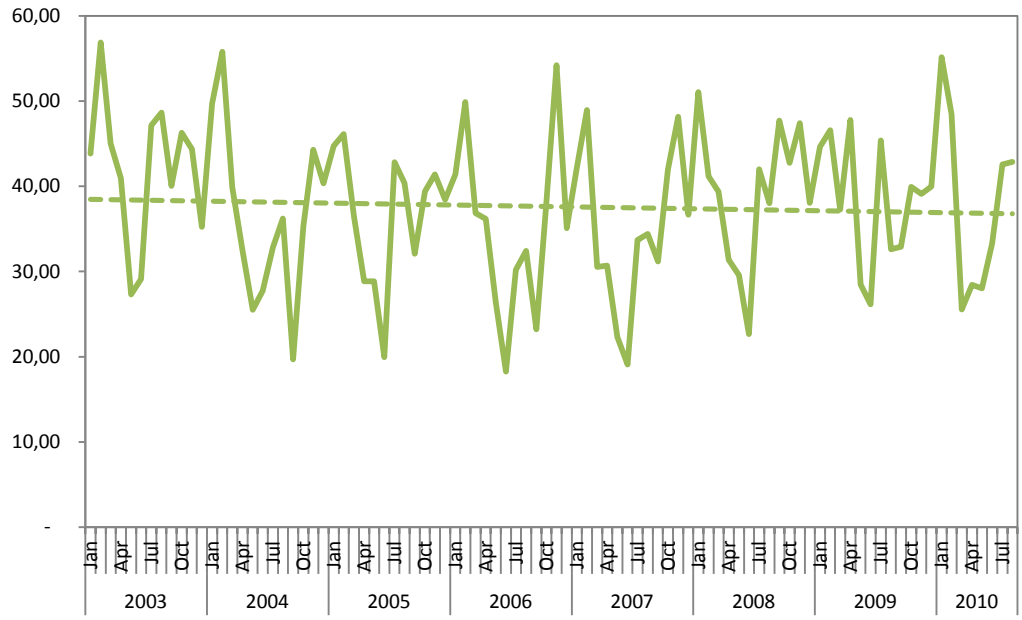


Figure 13: Occupancy rate in hotels, San Pedro de Atacama province, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

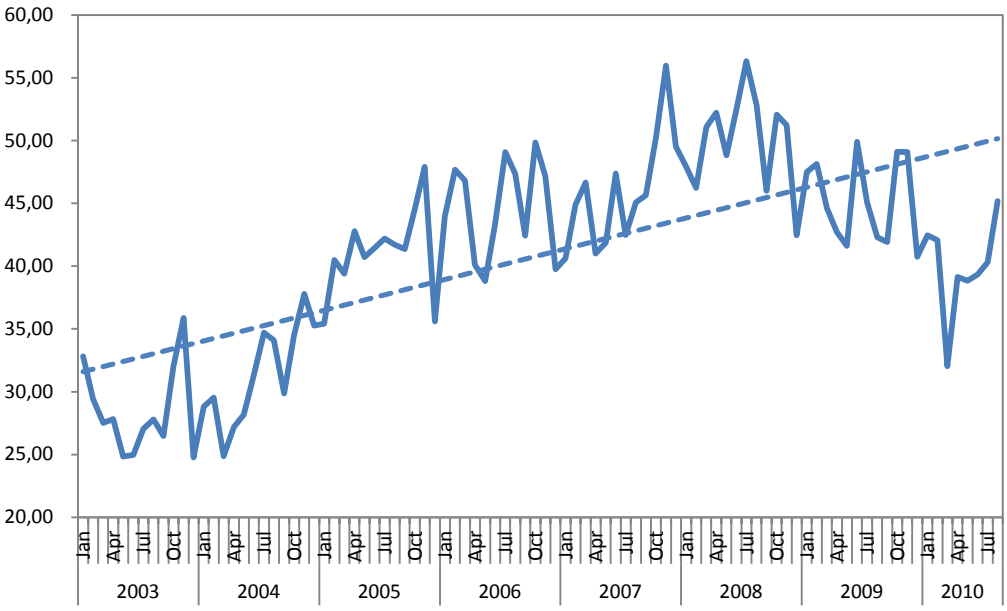


Figure 14: Occupancy rate in hotels, Antofagasta region, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

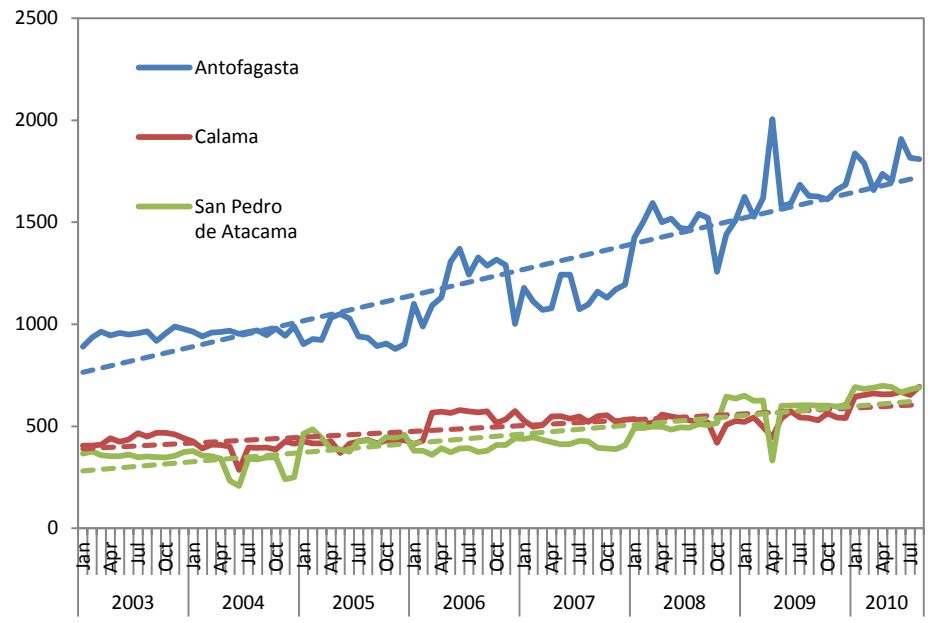


Figure 15: Number of workers in accommodation firms by month, by commune, 2003-2010.
Source: Authors' elaboration from EAT-INE database.

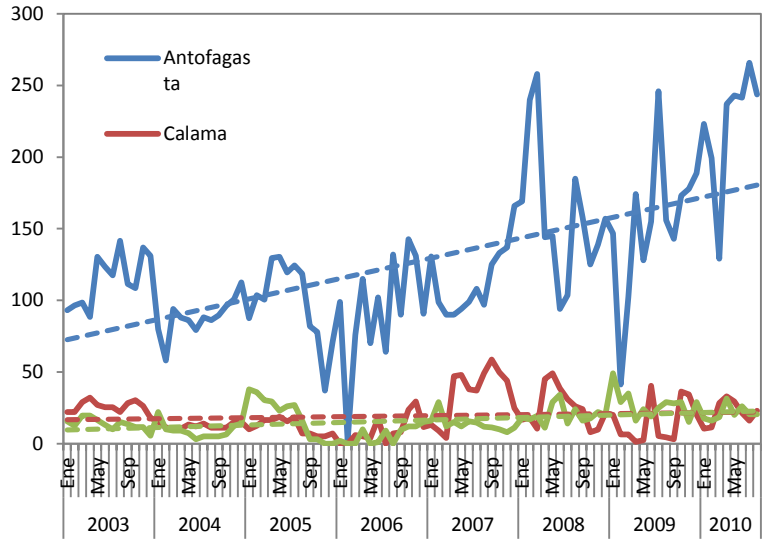


Figure 16: Temporary employment in the accommodation firms, by commune, 2003-2010.
Source: Authors' elaboration from EAT-INE database.