Tourism Management 68 (2018) 187–197

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

Using big data from Customer Relationship Management information systems to determine the client profile in the hotel sector

Pilar Talón-Ballestero[a], Lydia González-Serrano[a,*], Cristina Soguero-Ruiz[c], Sergio Muñoz-Romero[b,c], José Luis Rojo-Álvarez[b,c]

[a] Department of Business Economics, Rey Juan Carlos University, Camino del Molino s/n, 28943 Fuenlabrada, Madrid, Spain
[b] Center for Computational Simulation, Universidad Politécnica de Madrid, Boadilla, 28223 Madrid, Spain
[c] Department of Signal Theory and Communications and Telematic Systems and Computation, Rey Juan Carlos University, Camino del Molino s/n, 28943 Fuenlabrada, Madrid, Spain

ARTICLE INFO

Keywords:
Big data
Hospitality industry
Customer relationship management
Client profile
Bootstrap resampling
Hotel chains

ABSTRACT

Client knowledge remains a key strategic point in hospitality management. However, the role that can be played by large amounts of available information in the Customer Relationship Management (CRM) systems, when addressed by using emerging Big Data techniques for efficient client profiling, is still in its early stages. In this work, we addressed the client profile of the data in a CRM system of an international hotel chain, by using Big Data technology and Bootstrap resampling techniques for Proportion Tests. Strong consistency was found on the most representative feature of repeaters being *traveling without children*. Profiles were more similar for British and German clients, and their main differences with Spanish clients were in the *stay duration* and in *age*. For a vacation chain, these results suggest further analysis on the target orientation towards new market segments. Big Data technologies can be extremely useful for analyzing indoor data available in CRM information systems from hospitality industry.

1. Introduction

Customer knowledge is vital for the hospitality industry, and it plays a crucial role in improving the offer with better quality services (i.e., more adapted and customized), the relationship with customers, and the approach of marketing strategies (Adomavicius & Tuzhilin, 2001; Min, Min & Emam, 2002). All of them result in better customer satisfaction that increases the loyalty and ensures repeating customers, as well as higher profitability (Tseng & Wu, 2014). Over the last several years, this information has been mainly managed in many hotels by proactively gathering and recording customer preferences into the so-called Customer Relationship Management (CRM) systems (Sarmaniotis, Assimakopoulos, & Papaioannou, 2013). CRMs have become a key strategy for improving customer satisfaction and retention, especially in hotels (Padilla-Meléndez & Garrido-Moreno, 2013), and they are remarkably beneficial to those organizations by generating large amounts of valuable information about their customers (Chadha, 2015; Kotler, 2002; Nguyen, Sherif, & Newby, 2007).

Nevertheless, it has been recently pointed out (Dursun & Caber, 2016) that even advanced analysis techniques, such as data mining, are not yet being adequately used in the hotel industry for the purpose of effectively profiling the customers by using the comprehensive data that are routinely collected with hotel CRM systems. A large amount of information is available nowadays in hotel companies, either internal and structured (from the Property Management and the CRM systems), or external and unstructured (such as opinion platforms, social networks, or geolocalization, among many others). This brings the need to consider powerful tools available from Big Data technologies, which have already been successfully used in other fields such as bioinformatics, healthcare, or finance (George, Haas, & Pentland, 2014), to name just a few.

Big Data technologies are providing unprecedented opportunities for statistical inference on massive analysis, but they also bring new challenges to be addressed, especially when compared to the analysis of carefully collected smaller data sets. In Sivarajah, Kamal, Irani, and Weerakkody (2017), a systematic and illustrative review is presented on the state-of-art analysis of the literature on Big Data techniques and Big Data Analytics, which highlights the key challenges in terms of different data types, data processing, and data management. As pointed therein, descriptive statistics are the simplest form of Big Data analytic

* Corresponding author.
E-mail addresses: pilar.talon@urjc.es (P. Talón-Ballestero), lydia.gonzalez@urjc.es (L. González-Serrano), cristina.soguero@urjc.es (C. Soguero-Ruiz), sergio.munoz@urjc.es (S. Muñoz-Romero), joseluis.rojo@urjc.es (J.L. Rojo-Álvarez).

https://doi.org/10.1016/j.tourman.2018.03.017
Received 11 August 2017; Received in revised form 12 March 2018; Accepted 20 March 2018
Available online 28 March 2018
0261-5177/

methods, and they involve the summarization and description of knowledge and patterns by using simple statistical tests, such as mean, median, mode, variance, or proportions. When scrutinizing the usefulness of Big Data technologies in a new application field, it is necessary to establish well the behavior and scope of basic statistics, before going into more sophisticated analytics such as data mining or advanced machine learning.

In the present work, our main practical objective was to determine the client profile in an international hotel chain by exploiting the overall information in its CRM system. For this purpose, we identified the relevant variable groups, and we analyzed their practical meaning by using Big Data analytics on proportion tests from ratios between repeaters and first-timers. The use of robust and reliable proportion tests in this scenario has been tackled by using Bootstrap resampling techniques, which provides us with clear cut-off tests for decision making even in massive analysis conditions. It was possible for us to obtain two types of implications, namely, those related with the application of Big Data techniques to CRM exploitation, and those related with the results derived from the specific application to this hotel chain.

The scheme of the paper is as follows. In the next section, we introduce the relevance of CRM systems and their applications in the hotel sector, the basics on Big Data techniques and their scope in current analytics for hotel clients, and some relevant studies dealing with client profiling in terms of their repeating behavior. Then, we present the theoretical foundations of the methods to be used in our client profiling study, consisting of proportion tests and on bootstrap resampling. We then present the results on a real database from a large-scale hotel chain. Discussion is established on our and others' results, and finally, concise conclusions are drawn.

2. Literature review

Our main objective in the present work was to determine the repeater client profile by exploiting CRM systems in hotel chains and using Big Data technologies. For this purpose, we start by presenting a review of the state of art focused on the three main topics developed here, and the CRM system concept is first scrutinized. Recent applications on Big Data technologies are then summarized, both of them in the hotel industry, and finally, we present recent studies analyzing the repeaters versus first-timers profiles.

2.1. CRM in the hospitality industry

In the last several years, CRM has grown in relevance both in the operational and in the strategic points of view. Two of the major reasons for this are the increasing market competitiveness and the lower cost for client retention than for new client recruiting (Petrick, 2004; Yoo & Bai, 2013). Hence, CRM has become a key strategy for personalizing the client experience and for increasing their satisfaction.

The present work deals with CRM systems. A CRM system is a "firm tool that is technology-based for developing and leveraging consumer knowledge to nurture, maintain, and strengthen profitable relationships with consumers" (Elfving & Lemoine, 2012). According to Buttle (2004), a CRM system is a crucial part of a global CRM strategy. Soltani and Navimipour (2016) stated that CRM systems provide the infrastructure that facilitates the construction of long-term relationships with customers. Some examples of the functionality of CRM systems are sales force automation, data warehousing, data mining, decision support, and reporting tools (Hendricks, Singhal & Startman, 2007; Katz, 2002; Soltani & Navimipour, 2016).

For the hospitality sector, several studies consider CRM as one of the best strategies for improving a company's results and for ensuring long-term survival (Abu Kasim & Minai, 2009; Keramati, Mehrabi, & Mojir, 2010; Kim & Choi, 2010; Sigala, 2005; Wu & Li, 2011). Accordingly, CRM systems are nowadays a fundamental tool in the hotel sector, especially when properly implemented, due to the large amount of data that hotels integrate from their clients. These data could be turned into useful knowledge (Chadha, 2015; Dev & Olsen, 2000; Kotler, 2002; Lin & Su, 2003; Nasution & Mavondo, 2008; Nguyen et al., 2007), and the implementation of CRM systems allows us to identify the host behavioral patterns and to retain them in the long term (Chadha, 2015; Papastathopoulou, Avlonitis, & Panagopoulos, 2007; Verdugo, Oviedo-Garcia, & Roldan, 2009).

It is evident that customer loyalty and profitability are correlated (Payne & Frow, 2005). Therefore, one of the main assumptions of CRM systems is that satisfying and creating long-term relationships with profitable customers enhances the business success of the company (Wu & Lu, 2012). However, the role that large amounts of information currently available in CRM systems can play in efficient client profiling has not been studied enough yet, even for simple and well known statistical descriptions. In addition, there is evidence that advanced analysis techniques are not yet being properly used in the hotel industry to effectively profile customers from comprehensive data collected via hotel CRM systems (Dursun & Caber, 2016). Hotels are not fully exploiting the potential of CRM systems, but there is strong interest and ongoing work towards their successful implementation (Padilla-Melendez & Garrido-Moreno, 2013). This way, both the CRM-effort efficiency and a company's competitiveness could be dramatically increased.

2.2. Big data in the hospitality industry

Big Data is drastically changing the hotel sector management and the client-to-business relationship, by making the decision-making process from large amounts of data easier (Fox & Do, 2013). Nowadays, the technological bases of both the tourism organizations and the hoteliers make relevant that marketers and managers improve their access to data intelligence to make the best use of it (Peter, 2014). These professionals have invested heavily in recent years in organizing strong scientific teams and including statisticians and database experts who are well equipped to build and analyze the contents of their Data Warehouses (Ramos et al., 2017). Though human analysis is often required, today Big Data can enhance the decision making and increase the organizational output from five possible approaches, namely, descriptive analytics, inquisitive analytics, predictive analytics, prescriptive analytics, and preemptive analytics. Most Big Data analytics are descriptive and exploratory in nature, but even simple descriptive statistics allow businesses to discover simple and clear patterns that become extremely useful for decisions.

The hospitality industry has become an information intense sector, where large data volumes have been stored with practical applications which are not so widespread. With the arrival of Big Data, it is possible to manage such data to achieve the objectives and to transform the information into knowledge (Xiang, Schwartz, Gerdes & Uysal, 2015). Data are stored in very different formats, and their analysis becomes a complex task due to their heterogeneity, going from structured data in conventional databases (from Property Management Systems and CRM systems) to semi-structured and unstructured data. Furthermore, the available information systems often can include meta-search generated data, e.g., Tripadvisor, Kayak, Trivago, or social networks such as Facebook, Twitter, or LinkedIn (Ramos et al., 2017; Santana-Cerdeña, Ramos, & Bobur, 2014).

The hotel industry is starting to use Big Data technologies mainly in product sales, social media and online behavior of customers, as well as offline data retrieval and analysis (Zhang, Shu & Wang, 2015). Some examples are tourist's location (Hjorth, 2012; Silva & Mateus, 2003; Vu, Li, Law, & Ye, 2015), blogs (Litvin, Goldsmith, & Pan, 2008; Tseng, Wu, Morrison, Zhang, & Chen, 2015), photography (Balomenou & Garrod, 2014), internet behavior (Rong, Vu, Law, & Li, 2012), search engines (Pan & Li, 2011), and Online Travel Agencies (Ramos et al., 2017), to cite just a few. There is a growing interest in the hospitality field to exploit user-generated data and gain insight into research problems that

Table 1
Variable description.

Group of variables	Variables	Categories	Type	Examples
Demographic	Age	1	Numerical	
	Gender	3	Categorical	Female, male, unknown
	Civil status	3	Categorical	Single, with partner, unknown
	Country of residence	> 30	Categorical	Spain, Germany
	State	> 150	Categorical	Berlin, Madrid
Behavioral	Preferred destination	10	Categorical	
	Motivation	> 150	Categorical	Leisure, business
	Family holidays	3	Categorical	Yes, no, unknown
	Traveling with a partner	3	Categorical	Yes, no, unknown
	Traveling with children	3	Categorical	Yes, no, unknown
	How did client know about the hotel chain?	> 30	Categorical	Internet, friends, others
Generated by the hotel chain	Repeaters vs. first timers	2	Binary	Yes or no
	Cluster 1 done by the hotel		Categorical	
	Cluster 2 done by the hotel		Categorical	
	Cluster 3 done by the hotel		Categorical	
Transactional Data	How was the customer registered?	> 40	Categorical	Cardex, call center
	Entity in the last visit	8	Categorical	Agency, B2B
	Type of price applied to the customer	3	Categorical	Promotion

have not yet been well understood using conventional methods (Yang, Pan, & Song, 2014; Ye, Law, & Gu, 2009). The most used Big Data method in this setting is Text Analytics for information retrieval, and it usually involves machine learning, statistical analysis, and computational linguistics (Özköse, Arı, & Gencer, 2015). The main functionalities of Big Data and CRM systems in hotel management focus on Revenue Management and on Marketing, and their use includes forecasting, pricing, and bench marking (Haynes, 2016; Noone, 2016; Pan & Yang, 2016; Ramos et al., 2017; Song & Liu, 2017).

However, the growing interest in Big Data for dealing with external and unstructured data from clients has brought attention to the fact that large hotel chains can also access a huge amount of internal and structured data through their CRM systems. These systems may not often be deeply exploited for client knowledge purposes, even though simple statistical analysis in Big Data can yield significant and powerful pattern descriptions. Whereas some other studies have paid attention to Big Data approaches with advanced analytics and to its usefulness on internal data generated by hotel clients (Lee, Hwang, Jo, & Kim, 2016), the present study aims to gain insight specifically on the client profiling and restricting ourselves to simple Big Data analytics tools, such as proportion tests.

2.3. Client profiling in first-timers vs repeaters

Researchers have pointed out the relevance of understanding the differences between first-timers and repeaters in the hospitality and tourism industry (Anwar & Sohail, 2004; Morais & Lin, 2010; Oppermann, 1997; Petrick, 2005). Marketing researchers suggested that understanding the differences between first-timers and repeaters can provide us with an excellent basis for market segmentation (Formica & Uysal, 1998; Ryu & Han, 2011). Several studies have identified both tourist profiles, mostly focused on touristic destinations (Fallon & Schofield, 2003; Kim & Prideaux, 2005; Lau & Mckercher, 2004; Li, Cheng, Kim, Petrick, 2008; Morais & Lin, 2010; Oppermann, 1997; Tasci, 2016; Vu et al., 2015; Wang, 2004), festival and cultural events (Anwar & Sohail, 2004; Formica & Uysal, 1998), restaurants (Ryu & Han, 2011), cruises (Petrick, 2004), or whitewater rafting (Fluker & Turner, 2000). These differences are fundamental to developing effective business and marketing strategies, as well as to understand client motivation and to build theoretical knowledge on decision-making (Lau & Mckercher, 2004; Oppermann, 1997; Petrick, 2004). However, few academic references can be found in the hotel sector (Kim, Knutson, & Vogt, 2014).

It has been suggested (Chakravarti & Day, 1991) that different consumer profiles should be determined and identified in terms of different features. Some authors have pointed out the relevance of behavior variables, such as past buying patterns (Bayer, 2010; Kim, Jung, Suh, & Hwang, 2006; Wind & Lerner, 1979), and even more emphasis has been made on the relationship between the repetition habit of the client and the loyalty (Yoo & Bai, 2013). Moreover, repeaters can become efficient communication channels for relatives, friends, colleagues, and other potential consumers (Petrick, 2004).

On the other hand, Tasci (2016) shows that loyal consumers are different from others in terms of socio-demographic, psychographic, and behavioral features. Hence, demographic factors (such as education, gender, and age), as well as travel behavior variables (including purpose of travel), have been found to strongly influence consumer loyalty (Homburg & Giering, 2000; Skogland & Siguaw, 2004), and they pose notable differences between first-timers and repeaters (Lau & Mckercher, 2004; Li et al., 2008; McKercher, 2004; Oppermann, 1997). When socio-demographics were analyzed, significant differences were found based on age, spending patterns, length of stay, and nationality (Gitelson & Crompton, 1984; Li et al., 2008; McKercher, 2004).

Our study focuses on the profile of repeaters in an international chain. McKercher (2004) and Lau & Mckercher (2004) classified travelers to holiday destinations as either first-timers or repeaters. Many holiday destinations rely heavily on repeated visitations (Anwar & Sohail, 2004; Fallon & Schofield, 2003), hence, the results of this differentiation can be very relevant for this sector.

3. Statistical and data analytics methods

3.1. Big database

The database used in this study was assembled from a CRM system routinely used by a widely known international hotel company with international scope. The Information Technology unit of the company supervised the access to the information, and a business object was created upon the CRM system universe allowing us to scrutinize the stored features (variables) and to pre-filter the possibly useful ones. A Web Intelligence document was specifically created in the system to support this preliminary analysis.

The information of 4,935,806 different clients (those who stayed overnight at the hotels in the chain) was recorded in the CRM system during years 2013 and 2014. Table 1 describes the variables used in this work. A total of 18 variables as well as their corresponding categories were analyzed, including among them demographic variables, behavioral variables, and transactional data. In addition, the information generated by the hotel chain was also considered as relevant for the present study: guests grouped into repeaters (those clients who stayed

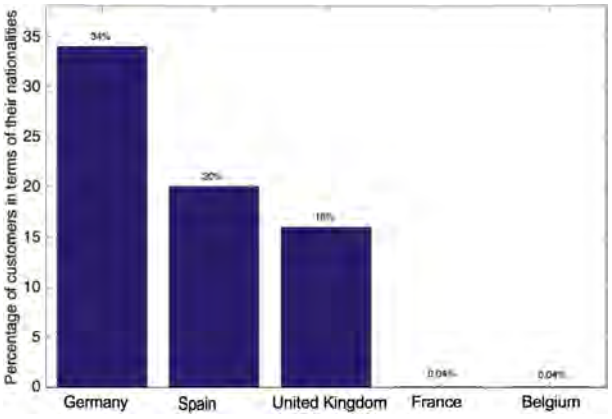

Fig. 1. Percentage of customers in terms of several of their nationalities.

more than once at the hotels in the chain) or first-timers and a client clustering used by the chain.

Some studies have indicated that travelers with different nationalities have different motivations for traveling (Kim & Lee, 2000; Kozak, 2002; Mok & Armstrong, 1998; Seddighi, Nuttall, & Theocharous, 2001) and that they may show differences in their behavioral characteristics (Kim & Prideaux, 2005). Based on that, we focused on the three nationalities with the largest number of clients in our chain, namely, Germany, Spain, and the United Kingdom (UK), which represented up to 70% of the total clients of this hotel chain (see Fig. 1).

3.2. Proportion tests and bootstrap resampling

In this study, we need to compare an overall high number of categories distributed among the features stored in the CRM system. We limited our study to the most usual feature types in the CRM system, i.e., the ones described either by dichotomous or by categorical values, possibly (and very often) in the presence of missing values. We worked with the whole set of categories in terms of two groups which are of our special interest, i.e., repeaters (Group 1) vs first-timers (Group 2).

As an example, think of a simple feature, such as *gender*, for which we have three possible values in the database, namely, female (value = 0), male (value = 2) and unknown (value = 3). In this case, we build a simple vector consisting of three elements, as described by the following. If we denote by $p_f(g_1)$ the female proportion in Group 1, and by $p_f(g_2)$ the female proportion of female in Group 2, we can define the Category Proportion Difference (CPD) as

$$\Delta p_f = p_f(g_1) - p_f(g_2) \quad (1)$$

Accordingly, if (for example) our analysis yielded $\Delta p_f > 0$, this would indicate that this value of the feature should be more present in Group 1 than in Group 2, and as far as Group 1 is repeater client, it could be read as *women are more prone to be repeater clients*.

We can build the same statistical description for male and for unknown values of the feature, namely, Δp_m and Δp_u. Note that these proportions are quite related among them, nevertheless, we are scrutinizing explicitly all the possible values in terms of dichotomic categories. Note also that in the case of unknown or missing data, we should read $\Delta p_u > 0$ as *unknown gender is more prone to be present in the database for repeater clients*. Hence, for this example simple feature, we have a three-dimensional vector of difference of proportions, which will be denoted as

$$\underline{\mathbf{v}}_{gender} = [\Delta p_f,\ \Delta p_m,\ \Delta p_u] \quad (2)$$

and it is called here the Feature Proportions Difference Vector (FPDV). The FPDV quantification and visualization provides us with a clear, exhaustive, and simple description of the trends in the values of this feature, and its extension to categorical features is straightforward.

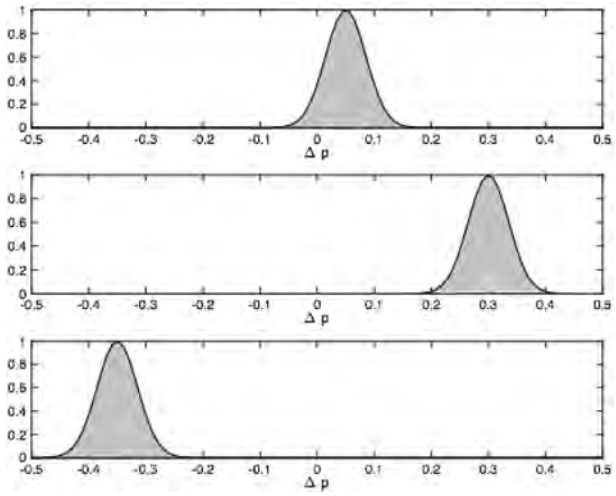

Fig. 2. Hypothesis test for the i-th CPD.

When the proportions of the i-th category are quite similar between groups, their CPD, denoted from now as Δp_i will be about zero. Nonzero values will deserve special attention, as they are indicative of differences between both groups. At this point, we need to establish a simple-to-use statistical test for determining whether a value of Δp_i is large enough to indicate a significant difference in this category between the compared groups. Therefore, the statistical test that we need has the following hypothesis:

1. *Null hypothesis*, H0: $\Delta p_i = 0$, and there is no proportion difference between groups for the i-th category.
2. *Alternative hypothesis*, H1: $\Delta p_i \neq 0$,and there is a proportion difference between groups for the i-th category. If $\Delta p > 0$ (if $\Delta p < 0$), then the proportion of the i-th category is significantly larger in Group 1 (in Group 2).

This hypothesis test is depicted in Fig. 2. The described CPD hypothesis test will be intensively used in large databases, as in the case of our CRM features analysis. The proportion test for analyzing p_i could be dealt with simple statistical tests, however, it is well known that proportion tests should be dealt with special caution (Sá, 2003), and furthermore, the definition of parametric tests for Δp_i can be sensitive in classical statistics, for instance in terms of the actual assumed distributions. For these reasons, we propose using a new method for estimating clear cut-off hypothesis tests for CPD by using bootstrap resampling, as described next in short.

The bootstrap resampling is a widely extended and common statistical technique which relies on random sampling with replacement (also known as the plug-in principle) to provide statistical tests. The plug-in principle says that the empirical distribution function of a statistic can be used as an approximation for the actual distribution function (Efron & Tibshirani, 1994). The rationale of the bootstrap resampling method is that if we want to make some inference of a population in terms of some decision statistic whose calculation is known, but its actual statistical distribution is not easy to obtain, we can resample the population data and make the inference on the resample.

Fig. 3 shows the general approach to bootstrap resampling. In our case, we want to provide a non-parametric hypothesis test for the CPD statistic in each category in the database. Given a population, we sample it with replacement several times B, which needs to be large enough to represent the tails of the statistical distributions (typically B = 50, 200, or 1000, depending on the application). These statistical copies of the original population are known as bootstrap resamples. The calculation of the decision statistic (in this case, the CPD for a category) is made now on each resample, hence yielding a new value for it,

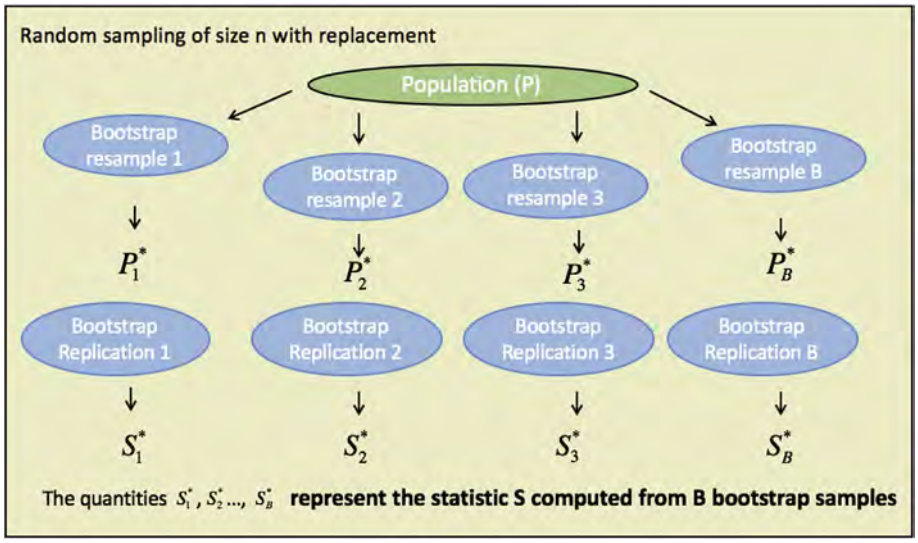

Fig. 3. Bootstrap resampling fundamentals for providing with a non-parametric hypothesis test for CPD in each category in the database.

known as *bootstrap replication*. The B replications of the statistics can be used to build a histogram, which stands for an estimation of the actual CPD probability density function, and then, simple methods, such as ordered statistics, can be used to determine the hypothesis test. Note in Fig. 3 and in the following that asterisk symbol (*) is used to identify the quantities that have been estimated throughout the bootstrap process from the original ones, these last often called the *empirical* statistics.

3.3. Big data and map-reduce implementation

The increasing amount of data that is generated, along with the need to extract useful knowledge from them, has created many problems in the last several years. These problems are mainly due to the nature of these available data, since the tools used so far were not applicable to these data sets. In particular, these problems affect different areas, such as hardware storage and accessibility of these data, database management, pre-processing, simple analysis, or automatic extraction of hidden patterns.

In order to address all these problems, a plethora of techniques have emerged that allow, in a much more efficient way, to extract value from data that were difficult to address, either by the large size of the available database, or by its low quality, or for being a union of diverse data sources with different nature. The framework encompassing this whole set of tools and solutions tends to be referred to as *Big Data*. In this article, we will use those tools that allow us to pre-process and analyze large volumes of available data, since their size make it unfeasible to analyze them using the classic tools. In particular, one of the tools that can do this is known as *map-reduce* (Jeffrey & Ghemawat, 2008, pp. 107–113), an algorithmic framework that takes advantage of both parallel hardware architecture and distributed file systems. In our area of knowledge, one of the Big Data challenges aimed to extract useful knowledge from large amounts of data is to create or adapt statistical or machine learning techniques to the parallelizable framework known as map-reduce. This map-reduce framework consists mainly of two basic operations, map and reduce. The map function is distributed and executed in parallel in the different computing nodes that constitute the hardware architecture. The purpose of this map function is to transform the available data format (possibly unstructured) into a structured data form, which is known as *key-value* and that allows us to solve the problem. The reduce function, also executed in parallel, takes as input the outputs of a set of map functions and summarizes all the key-value pairs belonging to the same key in a single key-value pair. Therefore, the obtained result returns as many key-value pairs as keys exist. Because the data has to be transformed into this rigid key-value structure, not all problems seem to be resolved efficiently with this map-reduce framework. In addition, most of the algorithms that allow advanced data analysis (as are those in the machine learning field) are not easily parallelizable (embarrassingly parallel), and they require great effort to be implemented in the map-reduce form.

In this paper, we propose to apply a combination of classical statistical techniques (described in Subsection 3.2) that are embarrassingly parallel in the map-reduce framework, in order to extract useful knowledge from the large available data collection. For its implementation, each bootstrap resampling block is distributed to each available parallel computing node, and the proportion test is applied as a map function. The union or sum of the outputs of each proportion test constitutes the reduce function, thus obtaining the desired advanced data analysis. Before this analysis, a necessary preprocessing is applied through map-reduce procedure in order to prepare the appropriate variables and the necessary information for the study.

4. Results

4.1. Some descriptive statistics

As previously described, data were recorded from 4,935,806 different clients in the CRM system during the years 2013 and 2014, and the goal of this work consisted of modeling the profile of first-timers and repeaters within this population, as it was found usual that many clients stayed in the chain hotel several times. Towards that end, we analyzed first-timers vs repeaters considering their nationality. The whole analysis on the complete data set took about 15 min on a MacBook Pro(R) (3.5 GHz, Intel Core i7), whereas it can be expected to be reduced to just a few seconds when using instead a graphic processing unit (GPU) or another computation server. For each of the selected nationalities (Germany, the UK, and Spain), the age distribution is represented in terms of the gender (men, women, and unknown) for first-timers and repeaters in Fig. 4. Table 2 also summarizes the mean and the standard deviation. In general, first-timer men were older than repeaters, and the age for British citizen was lower than for the other nationalities.

4.2. Advantages of big data hypothesis tests

Hypothesis tests based on probabilities are known to be sensitive in terms of statistical robustness. This can be checked in Fig. 5(a), which

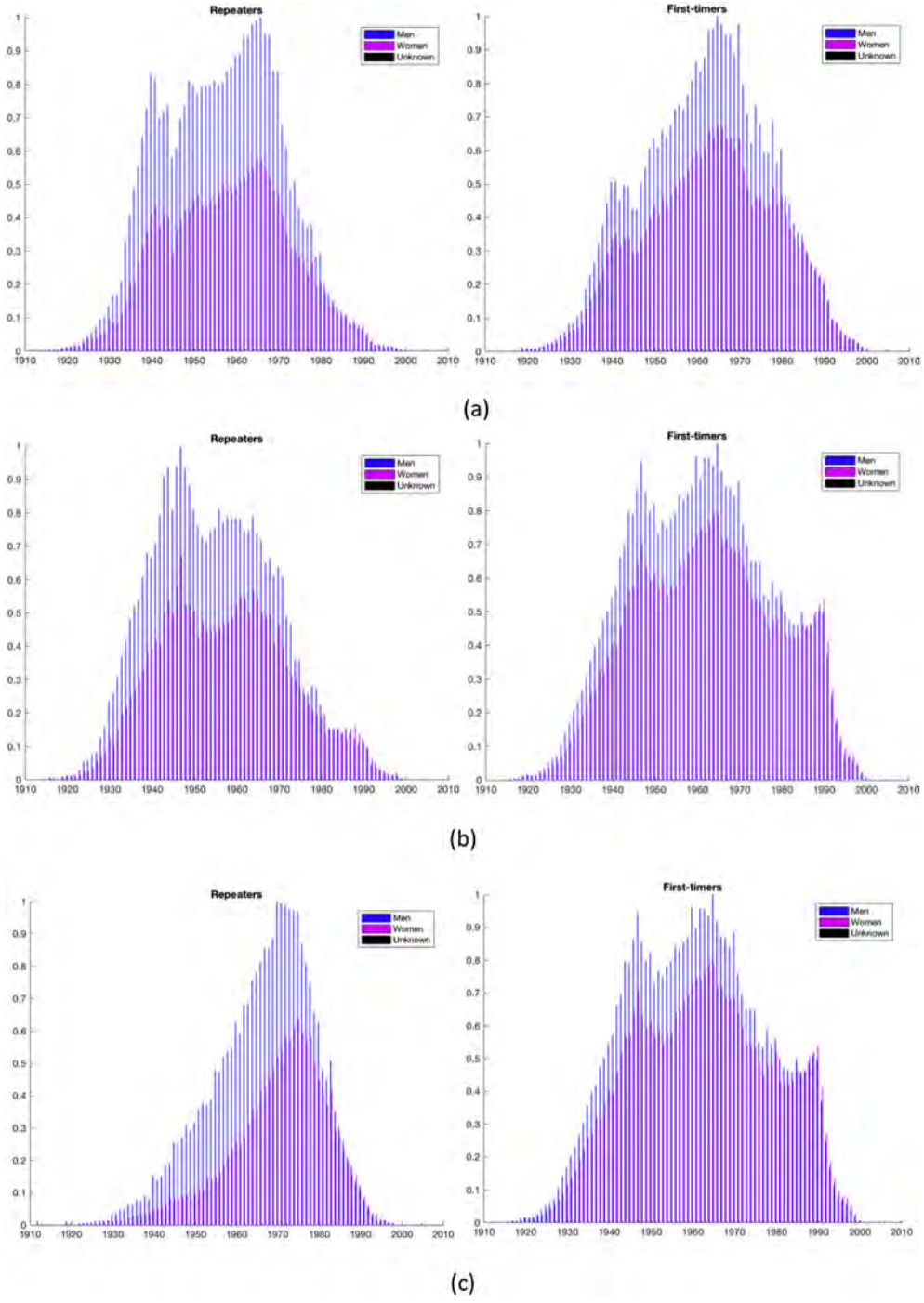

Fig. 4. Birth year distribution in terms of gender (men, women and unknown) for repeater and first-time visitors: (a) Germany; (b) UK; and (c) Spain.

represents the histograms estimated inside each map-reduce chunk for the CPD statistic of *gender* feature (*unknown* value is not represented in this example). Although the use of these statistical distributions for hypothesis testing separately in each chunk would show a trend to identify these variables as significant, there is still strong variability in these distributions, so that they do not represent a good-quality representation of the probability density function of the CPD statistics. This is surprising even given the large size of each chunk (200,000 cases).

However, Fig. 5(b) shows the aggregated estimation of the probability density functions from the accumulated averaged histograms after each chunk. Whereas the first 10 chunks still provide a quite noisy and fragmented density estimation, it soon becomes a smoother estimation, and it reaches convergence (highly similar estimations) after about half of the chunks have been included in the averaging. This holds even while the distribution estimation is capturing subtle but present multimodalities, which justifies the use of nonparametric statistical methods such as the bootstrap resampling. These observations make evident that even simple statistics are efficiently dealt with Big Data analytics (panel b) than with conventional approaches (panel a).

4.3. Chromosome and spiderweb results for proportion tests

We obtained a non-parametric hypothesis test for the CPD statistics in each category in the database and for each nationality. Hypothesis tests for the i-th CPD were calculated by using the bootstrap resampling

Table 2
Mean and standard deviation of birth year in terms of gender, for first-timers and repeaters. First (second, third) row refers to German (Spanish, British) customers.

Gender	Nationality	First-timers	Repeaters
Men	German	1962.18 +- 14.64	1956.80 +- 13.86
	Spanish	1966.19 +- 14.46	1966.65 +- 12.61
	British	1961.14 +- 16.26	1955.17 +- 14.55
Women	German	1963.19 +- 15.28	1958.16 +- 14.48
	Spanish	1968.32 +- 14.86	1970.31 +- 12.68
	British	1962.88 +- 16.69	1957.39 +- 15.22
Unknown	German	1960.88 +- 18.16	1956.02 +- 15.37
	Spanish	1949.85 +- 24.20	1964.12 +- 17.09
	British	1963.92 +- 17.56	1957.62 +- 20.33

technique for the proportion tests. Two different visualization techniques were used for the large number of categories scrutinized in the Big Database, namely, the so-called chromosome proportions plot and the spiderweb representation. The first one shows all the proportion rates for all the variables except for motivation due to visualization purposes (more than 150 categories) in three different nationalities (Germany, the UK, and Spain), as seen Fig. 6. All the values significantly higher (lower) than 0 indicate that a significant difference exists between the proportions of the i-th category for repeaters (first-timers). Red-dotted lines are used as a threshold to identify the most relevant features with a detailed and extensive description in terms of all the scrutinized categories, so that the higher its absolute value, the more relevant that feature is. We can see that interesting features often are viewed as clusters or interest regions. Alternatively, a joint visualization for defining the clients' profile based on just the most relevant features are represented by a spiderweb plot (see Fig. 7), which can provide the managers with a coarse and rough view of the profile. Both representations are complementary and show different details to the manager for analyzing the CRM features from a business point of view.

Several conclusions can be obtained after analyzing these two representations, almost at a glance. The clearest one (based on Fig. 6) is that the profiles of German and British repeaters are similar to each other and very different from that of the Spanish. In the case of German and British repeaters, two groups of customers are important: clients without children (category c85) and seniors with long stays (category c73). In both cases, it is significant that they are single (category c149) and travel in pairs (category c150). With respect to the first-timers profile, and in the case of the British, the favorite destination is Mallorca (category c167). Several fields with unknown value are significant, such as civil status (category c148), customer segment (category c39), customer group (category c71), or habitual residence (category c400). Therefore, more information is required from this group of clients in order to be able to better establish their specific

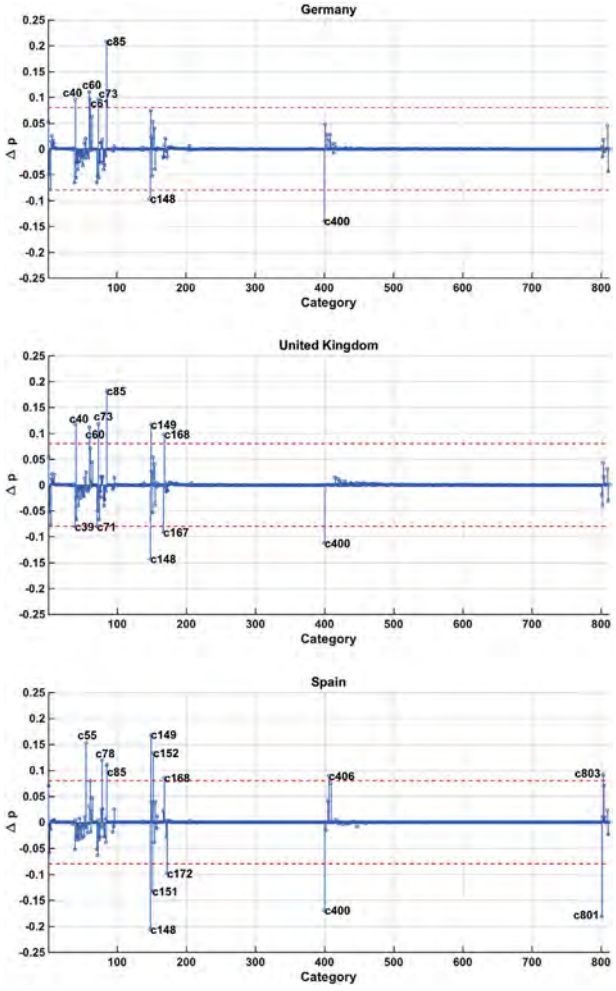

Fig. 6. Chromosome proportions plots for Germany, UK, and Spain.

profile. However, Spanish repeaters belong to the young-client escape group (category c78), traveling without children (category c85), and with high frequency (category c55). It is also significant that they are often single (category c149) and male (category c1). Among Spanish, the first-timers profile is characterized by traveling as a family (category c151) and booking through an agency (category c801).

5. Discussion and implications

The work in this paper allows us to scrutinize and present two

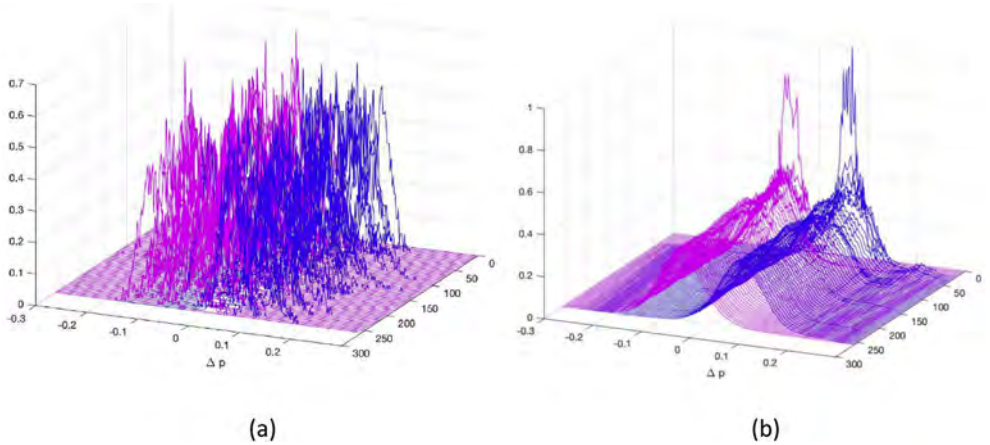

Fig. 5. Histograms yielding the use of hypothesis testing for CPD statistic in an example category (gender in Spain, men in blue, women in pink): (a) Normalized histograms estimated for each chunk in the map-reduce procedure; (b) Aggregated normalized histograms after each chunk in the map-reduce procedure. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

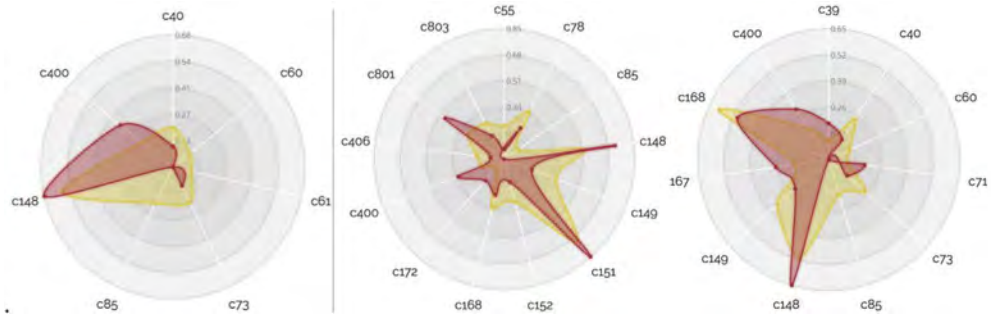

Fig. 7. Spiderweb representations for Germany (left), Spain (center) and UK (right). Repeaters and first-timers profiles are represented by red and yellow areas, respectively. These areas are defined by the $|\ \Delta pi\ |$ for the most significant categories above threshold.

different types of implications, namely, the methodological issues with respect to the application of the Big Data from CRM information, and those derived from the hotel chain study concerning the profile of the client. Both are summarized next after the discussion on the presented results.

5.1. Discussion on the Big Data case study

In our study, the repeaters profile is more similar in German and British guests than in Spanish guests. In the case of foreign repeaters, they are characterized by being senior citizens traveling for long periods without children. The importance of this segment in the tourism market has already been pointed out, and its behavior has become of great interest due to its size and potential growth (Chen, Liu, & Chang, 2013). Older consumers are presumed to have firmer brand loyalty than younger consumers (Henry, 2000). Also, German and British first-timers are significantly young clients on their holidays without children in our data, similar to the results by Gitelson and Crompton (1984), who noted that first-timers are more likely to be younger and single.

Spanish repeaters belong to the young-client escape group. These results differ for Spanish clients and match for German and British when compared with the results supported by Oppermann (1997), who demonstrated that first-timers tend to spend more money, but they stay for a shorter time than repeaters. Consistent findings were also observed by Wang (2004) and Lau & Mckercher (2004), however, contradictory findings regarding the length of stay were reported by Li et al. (2008), who concluded that first-timers are most likely to stay for longer periods, while repeaters are more likely to take weekend trips for visiting friends and relatives. The latter issue is easier for Spanish than for non-Spanish guests, since most of the chain establishments are located in Spain and this makes it more accessible to national customers to be able to move more frequently. This difference between the stay duration of Spanish and non-Spanish guests has been previously pointed out by Talón-Ballestero, González-Serrano, and Rodríguez-Antón (2016), who showed that the length of stay of foreign clients is longer in the Spanish hotel sector.

In addition, repeaters are single (especially for British and Spanish), in contrast with the ideas of Tasci (2016), who concluded that married clients were more loyal customers, and of Gitelson and Crompton (1984), who highlighted that the first-timers were mostly single. The fact of traveling without a family coincides in all three cases, which stands for a remarkable result, since according to Talón-Ballestero et al. (2016), up to 60% of the Spanish holiday market is families. Most of them are men and, in the case of the Spanish, they hired their last B2C visit. In making travel decisions, repeaters seem to rely more on their own experiences than on other information sources, hence, they spend much less time on planning (Li et al., 2008). On the other hand, Spanish first-timers have significantly purchased their last visit through an agency. According to Li et al. (2008), first-timers are more likely to rely on advice from travel professionals for making their travel decisions.

Women in the three nationalities are less often repeaters consistently in our data, which could be related to previous studies that have shown the different behavior in men and women in terms of their hotel preferences (Ariffin & Maghzi, 2012; Lutz & Ryan, 1993; McCleary, 1994; Sammons, Moreo, Benson & Demico, 1999).

Interestingly, first-timers are less reported in the CRM systems, for instance in terms of their civil status, among others. This brings into view the need for as complete as possible data recording in order to better identify customers and offer an adapted service, hence increasing their willingness to repeat (Padilla-Melendez & Garrido-Moreno, 2013).

5.2. Technical implications

Our proposal paves the way towards a systematic yet simple approach to exploit the data from CRM systems through Big Data in order to determine the client profile. This analysis has also shown the remarkable effect of map-reduce techniques for yielding consistency to probability based statistical tests, thus making it evident that even simple statistical descriptions are more efficiently tackled than conventional and lower-scale approaches. To our understanding, the most relevant technical contribution of this work is to show that Big Data, with simple statistical concepts and structured data available in the organizations, yields a better knowledge of the client.

The data structure of CRMs is a strategic asset for companies and their information is confidential, hence, CRM access for research purposes can be possible through collaboration agreements between companies and universities with shared interest to address Big Data exploitation. This is a common situation in Data Science for companies, and in our work, it has allowed us to obtain useful information from the existing data structures in the organization. Nevertheless, the developed methodology can be completely replicated in different CRM databases by any practitioner or data scientist. In addition, our results have provided useful information to the managers of the hotel chain with a view to reconsider the data structure of the CRM and its possible modification, and hence to improve its exploitation.

In this work, the clustering developed by the chain has been used as a set of variables, but no additional client grouping has been generated by our study, because our objective was to show the possibilities of Big Data to exploit the information available in the CRM system by univariate study of categorical variables. There is no doubt that other advanced multivariable analytics can highlight interesting cross-relationships among the business-relevant variables identified here by simple proportion tests.

A detailed statistical study of convergence from the theoretical point of view has not been carried out. But still, in the analyzed data, the asymptotic behavior of the distributions has been empirically observed (see Fig. 4) and this has been the expected one. A theoretical bootstrap convergence study with large data is a future and interesting work that exceeded our current objectives.

5.3. Hotel chain implications

Previous client classification studies based on statistical approaches have been used with varying success, but they will not always be able of

dealing with very large data sets. The use of Big Data techniques has allowed us to establish the repeaters vs first-timers profile in terms of many features and categories. Our results show that nationality, gender, age, length of stay, family conditions or selling channel, are strongly relevant for this profiling in the studied hotel chain, and special attention should be paid to their correct recording in the CRM system.

The more information recorded in the CRM systems, the easier it will be to generate detailed profiles with simple Big Data techniques. Hence, hotel managers will improve their client knowledge for increasing satisfaction and loyalty, ensuring the repeating visit and the increased profit (Tseng & Wu, 2014). From a marketing viewpoint, these results can help to identify clusters of clients for guiding supply of the company products and services in terms of their needs.

Moreover, it would be desirable to scrutinize the reorientation towards the growing female segment (40,7% of the analyzed sample) and to establish different actions in terms of the selling channel (b2c for repeaters and agencies for first-timers). In addition, repeaters mostly travel without children, which can support the current chain reorientation towards non-family vacation segments and “adults only” products.

One of the limitations of our study comes from the consideration of a single hotel chain, and thus, of a specific CRM system. As pointed out, access to these data requires collaboration agreements with companies. Future research may use this methodology in different companies because the method can be easily applied to other CRM systems by Data Scientists and experts.

Not only the conceptual approach, but also the practical guidance for hospitality managers, can be usefully improved by our results. As it has been observed, the correct management through the Big Data techniques of the large volume of data generated by the clients during their stay, which are regularly collected in the CRM system, allows a greater and better knowledge of their characteristics, which will improve the customer satisfaction, carry out personalized marketing campaigns, as well as give offers to selected customers to book an adequate room at the selected rate.

Future research lines can generalize the current results, for instance, by comparing the differences between first-timers vs. repeaters for different types of Hotel Chains, different purposes of visit (business clients), and in different nationalities.

6. Final conclusions

We concluded that the repeater profile in this chain corresponds to single, men, and traveling without children in the three scrutinized nationalities, however, there are differences among nationalities in terms of length of stay (larger in British and German than in Spanish) and age (senior in British and German vs. younger in Spanish). Moreover, due to the large number of tourists considered and the high volume of their handled information, the profile detected in this chain can be very useful not only to hotels, but also to tourist companies and destinations, in order to conveniently adapt their products and their marketing actions.

Overall, the great amount of available data from clients creates relevant opportunities for the hotel companies, which can turn into a strong competitive advantage. Further technical and more advanced tools will allow us to better exploit the best available information about the clients and their purchasing behavior. Our study has shown that even simple statistics as the proportion tests can be used for stating a solid large-scale information retrieval for client profiling, and it paves the way towards Big Data approaches yielding strong support for decision making of hotel managers.

Acknowledgements

This work has been partly supported by Spanish Projects PRINCIPIAS (TEC2013-48439-C4-1-R), FINALE (TEC2016-75161-C2-1-4), TEC2016-75361-R & ECO2016-75379-R from Spanish Government.

References

Adomavicius, G., & Tuzhilin, A. (2001, 03). Using data mining methods to build customer profiles. *Computer, 34*(3), 74–82. http://dx.doi.org/10.1109/2.901170.

Anwar, S. A., & Sohail, M. S. (2004, 04). Festival tourism in the United Arab Emirates: First-time versus repeat visitor perceptions. *Journal of Vacation Marketing, 10*(2), 161–170. http://dx.doi.org/10.1177/135676670401000206.

Ariffin, A. A., & Maghzi, A. (2012, 03). A preliminary study on customer expectations of hotel hospitality: Influences of personal and hotel factors. *International Journal of Hospitality Management, 31*(1), 191–198. http://dx.doi.org/10.1016/j.ijhm.2011.04.012.

Balomenou, N., & Garrod, B. (2014, 10). Using volunteer-employed photography to inform tourism planning decisions: A study of St David's peninsula, wales. *Tourism Management, 44*, 126–139. http://dx.doi.org/10.1016/j.tourman.2014.02.015.

Bayer, J. (2010, 09). Customer segmentation in the telecommunications industry. *The Journal of Database Marketing & Customer Strategy Management, 17*(3–4), 247–256. http://dx.doi.org/10.1057/dbm.2010.21.

Buttle, F. (2004). *Customer relationship management: Concepts and tools.* Amsterdam: Elsevier Butterworth-Heinemann.

Chadha, A. (2015, 03). Case study of hotel taj in the context of CRM and customer retention. *KCAJBMR Kuwait Chapter of Arabian Journal of Business and Management Review, 4*(7), 1–8. http://dx.doi.org/10.12816/0018976.

Chakravarti, D., & Day, G. S. (1991, 10). Market driven Strategy: Processes for creating value. *Journal of Marketing, 55*(4), 116. http://dx.doi.org/10.2307/1251961.

Chen, K., Liu, H., & Chang, F. (2013, 12). Essential customer service factors and the segmentation of older visitors within wellness tourism based on hot springs hotels. *International Journal of Hospitality Management, 35*, 122–132. http://dx.doi.org/10.1016/j.ijhm.2013.05.013.

Dev, C. S., & Olsen, M. D. (2000, 02). Marketing challenges for the next decade. *Cornell Hotel and Restaurant Administration Quarterly, 41*(1), 41–47. http://dx.doi.org/10.1177/001088040004100122.

Dursun, A., & Caber, M. (2016, 04). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives, 18*, 153–160. http://dx.doi.org/10.1016/j.tmp.2016.03.001.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap.* New York: Chapman & Hall.

Elfving, J., & Lemoine, K. (2012). *Exploring the concept of customer relationship management: Emphasizing social*Phd Master thesis. Supervisor: Karin Brunsson. Department of Business Studies, Uppsala University 2012-05-25. .

Fallon, P., & Schofield, P. (2003, 01). First-timer versus repeat visitor Satisfaction: The case of orlando, Florida. *Tourism Analysis, 8*(2), 205–210. http://dx.doi.org/10.3727/108354203774076742.

Fluker, M. R., & Turner, L. W. (2000, 05). Needs, motivations, and expectations of a commercial whitewater rafting experience. *Journal of Travel Research, 38*(4), 380–389. http://dx.doi.org/10.1177/004728750003800406.

Formica, S., & Uysal, M. (1998, 04). Market segmentation of an international cultural-historical event in Italy. *Journal of Travel Research, 36*(4), 16–24. http://dx.doi.org/10.1177/004728759803600402.

Fox, S., & Do, T. (2013, 09). Getting real about big Data: Applying critical realism to analyse big data hype. *International Journal of Managing Projects in Business, 6*(4), 739–760. http://dx.doi.org/10.1108/ijmpb-08-2012-0049.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal, 57*(2), 321–326.

Gitelson, R. J., & Crompton, J. L. (1984, 01). Insights into the repeat vacation phenomenon. *Annals of Tourism Research, 11*(2), 199–217. http://dx.doi.org/10.1016/0160-7383(84)90070-7.

Haynes, N. (2016). The evolution of competitor data collection in the hotel industry and its application to revenue management and pricing. *Journal of Revenue and Pricing Management, 15*(3–4), 258–263.

Hendricks, K. B., Singhal, V. R., & Stratman, J. K. (2007). The impact of enterprise systems on corporate performance: A study of ERP, SCM, and CRM system implementations. Journal. Henry, C. D. (2000, 07). Is customer loyalty a pernicious myth? Business Horizons, 43(4), 13–16. doi:10.1016/s0007-6813(00)00066-5.

Henry, C. D. (2000). Is customer loyalty a pernicious myth? *Business Horizons, 43*(4), http://dx.doi.org/10.1016/S0007-6813(00)00066-5 13–13.

Hjorth, L. (2012, 12). Relocating the mobile: A case study of locative media in seoul, South Korea. *Convergence: The International Journal of Research Into New Media Technologies, 19*(2), 237–249. http://dx.doi.org/10.1177/1354856512462360.

Homburg, C., & Giering, A. (2000). Personal characteristics as moderators of the relationship between customer satisfaction and loyalty?an empirical analysis. *Psychology and Marketing, 18*(1), 43–66. http://dx.doi.org/10.1002/1520-6793(200101)18:13.0.co;2-i.

Jeffrey, D., & Ghemawat, S. (2008). *MapReduce: Simplified data processing on large clusters.* Commun. ACM 51, 1 (January 2008) https://doi.org/10.1145/1327452.1327492.

Abu Kasim, N. A., & Minai, B. (2009). Linking CRM strategy, customer performance measures and performance in the hotel industry. *International Journal of Economics and Management, 3*(2), 297–316.

Katz, H. (2002). How to embrace CRM and make it succeed in an organization. *SYSPRO white paper.* Costa Mesa, CA: SYSPRO.

Keramati, A., Mehrabi, H., & Mojir, N. (2010, 10). A process-oriented perspective on customer relationship management and organizational performance: An empirical investigation. *Industrial Marketing Management, 39*(7), 1170–1185. http://dx.doi.org/10.1016/j.indmarman.2010.02.001.

Kim, B., & Choi, J. P. (2010, 06). Customer information Sharing: Strategic incentives and new implications. *Journal of Economics and Management Strategy, 19*(2), 403–433. http://dx.doi.org/10.1111/j.1530-9134.2010.00256.x.
Kim, S., Jung, T., Suh, E., & Hwang, H. (2006, 07). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications, 31*(1), 101–107. http://dx.doi.org/10.1016/j.eswa.2005.09.004.
Kim, M., Knutson, B. J., & Vogt, C. A. (2014). Posttrip behavioral differences between first-time and repeat guests: A two-phase study in a hospitality setting. *Journal of Hospitality Marketing & Management, 23*(7), 722–745. http://dx.doi.org/10.1080/19368623.2014.891960.
Kim, C., & Lee, S. (2000, 07). Understanding the cultural differences in tourist motivation between anglo-american and Japanese tourists. *Journal of Travel & Tourism Marketing, 9*(1–2), 153–170. http://dx.doi.org/10.1300/j073v09n01_09.
Kim, S. S., & Prideaux, B. (2005, 06). Marketing implications arising from a comparative study of international pleasure tourist motivations and other travel-related characteristics of visitors to Korea. *Tourism Management, 26*(3), 347–357. http://dx.doi.org/10.1016/j.tourman.2003.09.022.
Kotler, P. (2002). When to use CRM and when to forget it!. *Paper presented at the academy of marketing science* Sanibel Harbour Resort and Spa, 30 May.
Kozak, M. (2002, 06). Comparative analysis of tourist motivations by nationality and destinations. *Tourism Management, 23*(3), 221–232. http://dx.doi.org/10.1016/s0261-5177(01)00090-5.
Lau, A. L., & Mckercher, B. (2004, 02). Exploration versus acquisition: A comparison of first-time and repeat visitors. *Journal of Travel Research, 42*(3), 279–285. http://dx.doi.org/10.1177/0047287503257502.
Lee, S., Hwang, E., Jo, J. Y., & Kim, Y. (2016). Big data analysis with hadoop on personalized incentive model with statistical hotel customer data. *International Journal of Software Innovation, 4*(3), 1–21.
Li, X., Cheng, C., Kim, H., & Petrick, J. F. (2008, 04). A systematic comparison of first-time and repeat visitors via a two-phase online survey. *Tourism Management, 29*(2), 278–293. http://dx.doi.org/10.1016/j.tourman.2007.03.010.
Lin, Y., & Su, H. (2003, 08). Strategic analysis of customer relationship management—a field study on hotel enterprises. *Total Quality Management and Business Excellence, 14*(6), 715–731. http://dx.doi.org/10.1080/1478336032000053843.
Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008, 06). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29*(3), 458–468. http://dx.doi.org/10.1016/j.tourman.2007.05.011.
Lutz, J., & Ryan, C. (1993, 10). Hotels and the businesswoman. *Tourism Management, 14*(5), 349–356. http://dx.doi.org/10.1016/0261-5177(93)90003-4.
Mccleary, K. (1994, 04). Gender-based differences in business travelers' lodging preferences. *Cornell Hotel and Restaurant Administration Quarterly, 35*(2), 51–58. http://dx.doi.org/10.1016/0010-8804(94)90019-1.
Mckercher, B. (2004, 11). Understanding tourism Behavior: Examining the combined effects of prior visitation history and destination status. *Journal of Travel Research, 43*(2), 171–179. http://dx.doi.org/10.1177/0047287504268246.
Min, H., Min, H., & Emam, A. (2002, 11). A data mining approach to developing the profiles of hotel customers. *International Journal of Contemporary Hospitality Management, 14*(6), 274–285. http://dx.doi.org/10.1108/09596110210436814.
Mok, C., & Armstrong, R. W. (1998, 10). Expectations for hotel service quality: Do they differ from culture to culture? *Journal of Vacation Marketing, 4*(4), 381–391. http://dx.doi.org/10.1177/135676679800400406.
Morais, D. B., & Lin, C. (2010, 03). Why do first-time and repeat visitors patronize a destination? *Journal of Travel & Tourism Marketing, 27*(2), 193–210. http://dx.doi.org/10.1080/10548401003590443.
Nasution, H. N., & Mavondo, F. T. (2008, 04). Organisational capabilities: Antecedents and implications for customer value. *European Journal of Marketing, 42*(3/4), 477–501. http://dx.doi.org/10.1108/03090560810853020.
Nguyen, T. H., Sherif, J. S., & Newby, M. (2007, 05). Strategies for successful CRM implementation. *Information Management & Computer Security, 15*(2), 102–115. http://dx.doi.org/10.1108/09685220710748001.
Noone, B. M. (2016). Pricing for hotel revenue management: Evolution in an era of price transparency. *Journal of Revenue and Pricing Management, 15*(3–4), 264–269.
Oppermann, M. (1997, 05). First-time and repeat visitors to New Zealand. *Tourism Management, 18*(3), 177–181. http://dx.doi.org/10.1016/s0261-5177(96)00119-7.
Özköse, H., Arı, E. S., & Gencer, C. (2015). Yesterday, today and tomorrow of big data. *Procedia-Social and Behavioral Sciences, 195*, 1042–1050.
Padilla-Meléndez, A., & Garrido-Moreno, A. (2013, 06). Customer relationship management in hotels: Examining critical success factors. *Current Issues in Tourism, 17*(5), 387–396. http://dx.doi.org/10.1080/13683500.2013.805734.
Pan, B., & Li, X. R. (2011). The long tail of destination image and online marketing. *Annals of Tourism Research, 38*(1), 132–152.
Pan, B. I. N. G., & Yang, Y. A. N. G. (2016). Monitoring and forecasting tourist activities with Big Data. *Management Science in Hospitality and Tourism: Theory, Practice, and Applications, 43*.
Papastathopoulou, P., Avlonitis, G. J., & Panagopoulos, N. G. (2007, 04). Intraorganizational information and communication technology diffusion: Implications for industrial sellers and buyers. *Industrial Marketing Management, 36*(3), 322–336. http://dx.doi.org/10.1016/j.indmarman.2005.10.002.
Payne, A., & Frow, P. (2005, 10). A strategic framework for customer relationship management. *Journal of Marketing, 69*(4), 167–176. http://dx.doi.org/10.1509/jmkg.2005.69.4.167.
Peter, T. (2014). Use hotel data to drive growth. http://www.hotelnewsnow.com/Article/14553/Use-hotel-data-to-drive-growth, accessed 26/01/2016.
Petrick, J. F. (2004, 08). Are loyal visitors desired visitors? *Tourism Management, 25*(4), 463–470. http://dx.doi.org/10.1016/s0261-5177(03)00116-x.
Petrick, J. F. (2005, 10). Segmenting cruise passengers with price sensitivity. *Tourism Management, 26*(5), 753–762. http://dx.doi.org/10.1016/j.tourman.2004.03.015.
Ramos, C. M., Martins, D. J., Serra, F., Lam, R., Cardoso, P. J., Correia, M. B., et al. (2017). Framework for a hospitality big data Warehouse: The implementation of an efficient hospitality business intelligence system. *International Journal of Information Systems in the Service Sector, 9*(2), 27–45.
Rong, J., Vu, H. Q., Law, R., & Li, G. (2012, 08). A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining. *Tourism Management, 33*(4), 731–740. http://dx.doi.org/10.1016/j.tourman.2011.08.006.
Ryu, K., & Han, H. (2011, 09). New or repeat customers: How does physical environment influence their restaurant experience? *International Journal of Hospitality Management, 30*(3), 599–611. http://dx.doi.org/10.1016/j.ijhm.2010.11.004.
Sammons, G., Moreo, P., Benson, L. F., & Demicco, F. (1999, 05). Analysis of female business travelers' selection of lodging accommodations. *Journal of Travel & Tourism Marketing, 8*(1), 65–83. http://dx.doi.org/10.1300/j073v08n01_04.
Santana-Cerdeña, L., Ramos, S., & Bobur, S. (2014). Potential y Retos del Big Data en Turismo. *X Congreso de Turismo y Tecnologías de la Información y las Comunicaciones* (pp. 21–35). Universidad de Málaga, Facultad de Turismo.
Sarmaniotis, C., Assimakopoulos, C., & Papaioannou, E. (2013, 07). Successful implementation of CRM in luxury hotels: Determinants and measurements. *EuroMed Journal of Business, 8*(2), 134–153. http://dx.doi.org/10.1108/emjb-06-2013-0031.
Sá, J. P. Marques De (2003). *Applied Statistics: Using SPSS, STATISTICA, and MATLAB.* Berlin: Springer.
Seddighi, H., Nuttall, M., & Theocharous, A. (2001, 04). Does cultural background of tourists influence the destination choice? an empirical study with special reference to political instability. *Tourism Management, 22*(2), 181–191. http://dx.doi.org/10.1016/s0261-5177(00)00046-7.
Sigala, M. (2005, 09). Integrating customer relationship management in hotel operations: Managerial and operational implications. *International Journal of Hospitality Management, 24*(3), 391–413. http://dx.doi.org/10.1016/j.ijhm.2004.08.008.
Silva, A. P., & Mateus, G. R. (2003, 05). A location-based service application for a mobile computing environment. *SIMULATION, 79*(5–6), 343–360. http://dx.doi.org/10.1177/0037549703037151.
Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017, 01). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263–286. http://dx.doi.org/10.1016/j.jbusres.2016.08.001.
Skogland, I., & Siguaw, J. A. (2004, 08). Are your satisfied customers loyal? *Cornell Hotel and Restaurant Administration Quarterly, 45*(3), 221–234. http://dx.doi.org/10.1177/0010880404265231.
Soltani, Z., & Navimipour, N. J. (2016, 08). Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research. *Computers in Human Behavior, 61*, 667–688. http://dx.doi.org/10.1016/j.chb.2016.03.008.
Song, H., & Liu, H. (2017). Predicting tourist demand using big data. *Analytics in smart tourism design* (pp. 13–29). Springer International Publishing.
Talón-Ballestero, P., González-Serrano, L., & Rodríguez-Antón, J. M. (2016). *Fundamentos de Dirección Hotelera, Vol. I*Madrid: Síntesis978-84-9077-392-5.
Tasci, A. D. (2016). A quest for destination loyalty by profiling loyal travelers. *Journal of Destination Marketing & Management*. https://doi.org/10.1016/j.jdmm.2016.04.001.
Tseng, S., & Wu, P. (2014, 03). The impact of customer knowledge and customer relationship management on service quality. *International Journal of Quality and Service Sciences, 6*(1), 77–96. http://dx.doi.org/10.1108/ijqss-08-2012-0014.
Tseng, C., Wu, B., Morrison, A. M., Zhang, J., & Chen, Y. (2015, 02). Travel blogs on China as a destination image formation agent: A qualitative analysis using leximancer. *Tourism Management, 46*, 347–358. http://dx.doi.org/10.1016/j.tourman.2014.07.012.
Verdugo, C. M., Oviedo-Garcia, A. M., & Roldan, L. J. (2009). The employee-customer relationship quality: Antecedents and consequences in the hotel industry. *International Journal of Contemporary Hospitality Management, 21*(3), 251–274.
Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2015, 02). Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tourism Management, 46*, 222–232. http://dx.doi.org/10.1016/j.tourman.2014.07.003.
Wang, D. (2004, 02). Tourist behaviour and repeat Visitation to Hong Kong. *Tourism Geographies, 6*(1), 99–118. http://dx.doi.org/10.1080/1461668032000172355.
Wind, Y., & Lerner, D. (1979, 02). On the measurement of purchase Data: Surveys versus purchase diaries. *Journal of Marketing Research, 16*(1), 39. http://dx.doi.org/10.2307/3150872.
Wu, S., & Li, P. (2011, 06). The relationships between CRM, RQ, and CLV based on different hotel preferences. *International Journal of Hospitality Management, 30*(2), 262–271. http://dx.doi.org/10.1016/j.ijhm.2010.09.011.
Wu, S., & Lu, C. (2012, 03). The relationship between CRM, RM, and business performance: A study of the hotel industry in taiwan. *International Journal of Hospitality Management, 31*(1), 276–285. http://dx.doi.org/10.1016/j.ijhm.2011.06.012.
Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can Big Data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120–130.
Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research, 53*(4), 433–447.
Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management, 28*(1), 180–182.
Yoo, M., & Bai, B. (2013, 06). Customer loyalty marketing research: A comparative approach between hospitality and business journals. *International Journal of Hospitality Management, 33*, 166–177. http://dx.doi.org/10.1016/j.ijhm.2012.07.009.
Zhang, Y., Shu, S., Ji, Z., & Wang, Y. (2015, March). A study of the commercial application of big data of the international hotel group in China: Based on the case study of marriott international. *Big data computing service and applications* (pp. 412–417). .

Pilar Talón-Ballestero is a PhD holder in Advanced Marketing, Associate Professor in Business Economics Department and Director the university Master in Revenue Management of Rey Juan Carlos University. She is the Director and researcher in numerous studies and consultancy projects developed with private and public entities: Spanish Confederation of Hotels and Touristic Acommodations, Spanish Hotel Technological Institute, Iberostar, Nh and Palladium Hotels groups, etc. She has published numerous publications (articles and books), including some impact ones about hotel sector. She was hotel and travel agency manager. Her areas of specialization are revenue management, distribution, big data and gender in tourism.

Lydia González-Serrano is a PhD holder in Economics and Business Administration (Finance) and Associate Professor in BusinessEconomics Department of Rey Juan Carlos University. She is the Director and researcher in numerous research and consultancy projects developed with private and public entities: Institute for Women, Spanish Agency for International Cooperation and Development, Spanish Confederation of hotels and Touristic Acommodations, Spanish Technological Institute, etc. Her research activity has been reflected in numerous publications, including some impact ones. Her several books about hotel management have been published. Research lines are focused on two key issues: finance and risk analysis and hotel management.

Cristina Soguero-Ruiz received the Telecommunication Engineering degree, the B.Sc. degree in Business Administration and Management, and the M.Sc. degree in Biomedical Engineering from the University Rey Juan Carlos, Madrid, Spain, in 2011 and 2012. She got the Ph.D. degree in Machine Learning with Applications in Healthcare in 2015 in the Joint Doctoral Program in Multimedia and Communications in conjunction with University Rey Juan Carlos and University Carlos III. She was supported by FPU Spanish Research and Teaching Fellowship (granted in 2012). She won the Orange Foundation Best PhD Thesis Award by the Spanish Official College of Telecommunication Engineering.

Sergio Muñoz-Romero earned his PhD in Machine Learning at Universidad Carlos III de Madrid, where he also received the Telecommunication Engineering Degree. He has led pioneering projects where machine learning knowledge was successfully used to solve real Big Data problems. Currently, he is a researcher at Universidad Rey Juan Carlos. Since 2015, he has worked at Persei vivarium as Head of Data Science and Big Data. His present research interests are centered in machine learning algorithms and Statistical Learning Theory, mainly in dimensionality reduction and feature selection methods, and their applications to Big Data. The whole team have developed the same amount of work. The engineers (JLRA, CSR, and SMR) focused on the development of the analysis tools, programmed the big data methods, parsed the raw data and generated the results. Business management members (PTB and LGS) proposed the idea of the big data application on CRM and its linkage with hotel sector management. All the authors contributed to write the abstract, introduction, results and discussion-conclusion sections. PTB and LGS elaborated Section 2, and SMR, CSR, and JLRA elaborated Section 3.

José Luis Rojo-Álvarez received the Telecommunication Engineering Degree in 1996 from University of Vigo, Spain, and the PhD in Telecommunication Engineering in 2000 from the Polytechnic University of Madrid, Spain. Since 2016, he has been a Full Professor in the Department of Signal Theory and Communications, University Rey Juan Carlos, Madrid, Spain. He has published more than 100 indexed papers and more than 170 international conference communications. He has participated in more than 60 projects (with public and private fundings) and directed more than 10 of them. In 2016 he received the Rey Juan Carlos University Price to Talented Researcher.