The Use of New Data Analysis Techniques in Tourism: A Bibliometric Analysis in Data Mining, Big Data and Structural Equations Models

Jesús Palomo[a],
Cristina Figueroa-Domecq[a],
Mª Dolores Flecha-Barrio[a], and
Mónica Segovia-Pérez [a]

[a]The Faculty of Social Sciences and Law
Rey Juan Carlos University, Spain
jesus.palomo@urjc.es
cristina.figueroa@urjc.es
mariadolores.flecha@urjc.es
monica.segovia@urjc.es

Abstract

Tourism is an information intensive sector and consequently decision-making entails managing and analysing an increasing quantity of information. Tourists generate incredible amounts of information, before, during and after their trips. Consequently, all kinds of tourism organizations are working to try to adapt new data techniques and methodologies that will allow the compilation of information, its connection, and its analysis. For these reasons, research in tourism has increasingly focused on data mining, big data and structural equations modelling techniques. The aim of this paper is the analysis of the use over time that the tourism research area is making of these new methodologies through a bibliometric analysis. It covers publications between 1994 and 2015. Results confirm the increasing importance of these techniques in the tourism research area.

Keywords: Data mining; Big Data, Structural Equations Modelling; Tourism; Bibliometric Analysis

1 Introduction

The last years have shown how, though the Internet usage by travellers has stabilized, social media and mobile systems have dramatically changed the way tourists relate to the tourism industry (Xiang, Magnini & Fesenmaier, 2015), changing the dynamics of online communication. The current environment informs that travel companies should innovate continuously in order to learn and rapidly adapt to the new requirements of tourists, to manage the distribution of their products quickly and securely, and to improve productivity (Miles, 2002). Furthermore, tourists generate incredible amounts of information, before, during and after their trips. Consequently, tourism is an economic activity in which decision-making entails managing an increasing quantity of information without economic, political, social, or physical boundaries on both the supply and the demand side (Lemmetyinen & Go, 2009; Shaw & Williams, 2009).

For this reason, tourism requires tools such as information and communication technologies (ICTs) to enable secure, rapid, and inexpensive transfer of information

(Buhalis & Licata, 2002) at the same time it needs specialized methodologies that will allow the correct analysis of all the data generated, in order to create knowledge (Fuchs, Höpken & Lexhagen, 2014). The usage of new analytical techniques is reshaping the decision-systems used by practitioners but also by researchers. Some of the most important techniques or methodologies for a complex and advanced data analysis are Data-Mining (DM); Big Data (BD); and Structural Equation Modelling (SEM).

2 Research Objectives

The aim of this paper is the analysis of the usage that research on tourism has made of new methodologies, with the main focus on data mining, big data and structural equations models. This will be done through a Bibliometric Analysis.

3 Methodology

The bibliometric research procedures described in, e.g. Bordons et al. (2003), Figueroa et al. (2015), and Ensslin et al. (2015), settle the basis for the bibliometric methodology. The selected database for the analysis is SCOPUS, a bibliographic database containing abstracts and citations for academic journal articles. Only articles have been included in the analysis, since this selection confirm the quality of the published papers in two levels.

A proper selection of the keywords to define the screening process of the articles defines the accuracy of the research. Consequently, 20 articles on tourism, big data, data minning and structural equation were selected at random and analysed. The following keywords were identified and booleanly combined: tourism; hotel; data mining or data-mining; big data; structural equation. In July 2016, a total of 785 articles related to data mining, big data and structural equations were initially identified up to year 2015. The coding process then started with the identification of the most important variables for the analysis (year of publication; names, and number of authors; name of the institutional affiliation and the country; language; impact factor of the journal and number of citations received by the article).

Quantitative analysis was then performed (see e.g. Cavero et al., 2014; Abramo et al., 2013 for related bibliometric analysis methods). A statistical multivariate analysis of all the characteristics of the documents has been performed through the powerful statistical software R.

4 Results

The first paper in the tourism area including the concepts of DM, BD or SEM was written in 1994, but it is not until 2006, with 24 articles, that the amount of articles seems to increase at a steady growth and this type of techniques become settled. Since 1994 a total of 785 articles have been published, 17,58% of these papers in 2014 and 21,4% in 2015.

In order to assess whether or not the scientific production fits Price's law (Price, 1963) of exponential scientific growth, different regression models have been tested

to obtain the model that best fits this bibliometric analysis. This law focus on the evolution of publications over time. The following plot (Figure 1) shows that the production follows an exponential function; hence, the Price's Law is verified (Figure 1). The results of the adjustment are as follows: Analysis model: exponential. R2= 0.9762276; Residual standard error: 0.2575 on 19 degrees of freedom; Multiple R-squared: 0.9774, Adjusted R-squared: 0.9762; F-statistic: 822.3 on 1 and 19 DF, p-value: < 2.2e-16.

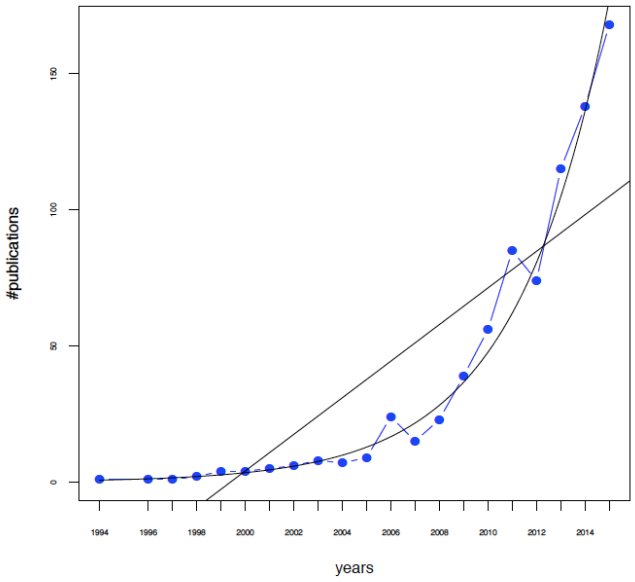

Fig 1. Price's Law

Bradford law analysis (or the Scatter of journals) indicates that a total of 258 journals have published 785 papers. Bradford's law of scattering was formulated as (Bradford, 1948): "if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject, and several groups or zones containing the same number of articles as the nucleus, where the number of periodicals in the nucleus and succeeding zones will be [1: n: n2]."

The average number of papers per journal is 3.043, which seems quite diversified among journals. Nevertheless, as it can be seen in Table 1, there are important journal specialized in the publication of papers related to DM, BD and SEM. The most important journal publishing is Tourism Management (72 articles and 9,2% of all the articles published), followed by International Journal of Hospitality Management (49 articles and 6,2% of all the articles published), and they are both included in Q1 from a research quality point of view.

Table 1. Most important journal publishing in the area of DM, BD and SEM.

	Journal	Number of articles published	% Total articles
1	Tourism Management	72	9,2%
2	International Journal of Hospitality Management	49	6,2%
3	Journal of Hospitality and Tourism Research	35	4,5%
4	Journal of Travel Research	28	3,6%
5	International Journal of Contemporary Hospitality Management	27	3,4%
6	Journal of Travel and Tourism Marketing	27	3,4%
7	Asia Pacific Journal of Tourism Research	26	3,3%
8	Tourism Analysis	22	2,8%
9	Annals of Tourism Research	19	2,4%
10	International Journal of Tourism Research	16	2,0%
11	Expert Systems with Applications	11	1,4%
12	Journal of Hospitality Marketing and Management	11	1,4%
13	Journal of Hospitality and Tourism Technology	10	1,3%
14	Journal of Sustainable Tourism	10	1,3%

When testing the original Bradford's law (Bradford, 1948; Avramescu, 1980; Rao, 1998), the relationship of each zone [6: 40: 212] does not fit into the original Bradford's distribution. So it is proposed to use the Leimkuhler Model with the following results with r0=8.197611 ~ 8 and k= 5.042796

- The nucleus 1 is formed by 8 journals; there are 286 articles in-group 1 (the nucleus).
- Group 2 is formed by 41 journals; There are 245 articles in group 2
- Group 3 is formed by 09 journals; There are 254 articles in group 3

Now, the relationship of each zone [8: 8*5.043: 8*5.043^2], which follows closely [8: 40.32: 203.21] the distribution of journals across the 3 different zones and Bradford's law. The percentage of error is Error = (251.7807 - 258)/ 258 = -0.02410598 = -2.4%. Under the adjusted model, the expected number of articles per journal in zone 1 is 35.75, in zone 2 is 5.98, in zone 3 is 1.22

The journals in the nucleus, consequently, the journal more interested in these research methodologies, are: Tourism Management; International Journal of Hospitality Management; Journal of Hospitality and Tourism Research, Journal of Travel Research, International Journal of Contemporary Hospitality Management, Journal of Travel and Tourism Marketing, Asia Pacific Journal of Tourism Research, Tourism Analysis.

Finally, another important variable to analysis is related with the authorship. In terms of authorships, these are the top 10 authors with 8 or more authorships: Assaker G. (8 articles); Gursoy D. (8 articles); Nunkoo R. (8 articles); Song H. (8 articles); Uysal M.

(8 articles); Ramkissoon H. (11 articles); Han H. (14 articles); Li G. (15 articles); Karatepe O.M. (18 articles); Law R. (19 articles). These results confirm a high degree of expertise of certain authors, that have not been found in other research areas in tourism (Figueroa-Domecq et al., 2015).

5 Conclusion

The usage of DM, BD and SEM techniques are becoming more important according to the results of the performed bibliometric analysis in the tourism research area. The first article published in SCOPUS that related tourism and DM, BD or SEM is found in 1994, but it is not until 2006 that these techniques seem to settle; actually, in 2015, 21,4% of the articles in the area are published. Nevertheless, there are still important bibliometric analysis to perform in this area: identification of most relevant topics using these techniques, as well as the collaboration among institutions in this area.

6 References

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). Gender differences in research collaboration. *Journal of Infometrics,* 7(4), 811–822.

Avramescu, A. (1980). Theoretical foundation of Bradford's Law. *International Forum for Information and Documentation,* 5,15-22.

Bordons, M., Morillo, F., Fernandez, M., & Gomez, I. (2003). One step further in the production of bibliometric indicators at the micro level: Differences by gender and professional category of scientists. *Scientometrics,* 57(2), 159–173.

Bradford, S. C. (1948). *Documentation*. London. UK: Crosby, Lockwood.

Buhalis, D. & Licata, M. C. (2002). The future of tourism intermediaries. *Tourism Management,* 23(3), 207-220.

Cavero, J. M., Vela, B., & Caceres, P. (2014). Computer science research: More production, less productivity. *Scientometrics,* 98(3), 2103–2111.

Ensslin, L., Dutra, A., Ensslin, S. R., Chaves, L. C., & Dezem, V. (2015). Research process for selecting a theoretical framework and bibliometric analysis of a theme: Illustration for the management of customer service in a bank. *Modern Economy,* 6(06), 782- 796.

Figueroa-Domecq, C., Pritchard, A., Segovia-Pérez, M., Morgan, N. & Villacé, T. (2015). Tourism gender research: A critical accounting. *Annals of Tourism Research,* 52, 87–103.

Lemmetyinen, A. & Go, F.M. (2009). The key capabilities required for managing tourism business networks. *Tourism Management,* 30 (1), 31-40.

Miles, I. (2002): Service innovation: towards a tertiarization of innovation studies. In Gadrey, J. & Gallouj, F. (Eds.), *Productivity, innovation and knowledge in services*. Cheltenham: Edward Elgar Publishing.

Palomo, J. & Montalvo, S. (2011). An international platform for teaching support based on breaking news. *Arbor, Ciencia, Pensamiento y Cultura,* 187, 249–253.

Price, D. J. (1963). *Little science, big science*. New York, US: Columbia University.

Rao, I. R. (1998). An analysis of Bradford multipliers and a model to explain law of scattering. *Scientometrics,* 41(1-2), 93–100.

Shaw, G. & Williams, A.M. (2009). Knowledge transfer and management in tourism: an emerging research agenda. *Tourism Management,* 30(3), 325-335

Xiang, Z., Magnini, V. P., & Fesenmaier, D. R. (2015). Information technology and consumer behavior in travel and tourism: Insights from travel planning using the internet. *Journal of Retailing and Consumer Services,* 22, 244-249.