A strategic tourism knowledge base for socio-economic and environmental data analytics: The role of Big Data analysis

Sotirios Varelas

Neapolis University Pafos, Cyprus

E-mail: s.varelas@nup.ac.cy
ORCID: 0000-0003-4580-3256

Evangelia Kopanaki

University of Piraeus, Greece

E-mail: evik@unipi.gr
ORCID: 0000-0002-5527-3757

Nikolaos Georgopoulos

University of Piraeus, Greece

E-mail: ngeorgop@unipi.gr
ORCID: 0000-0003-3369-5542

Correspondence to:
Sotirios Varelas
Neapolis University Pafos
Department of Business
Administration
2 Danais Avenue
8042 Paphos, Cyprus
Tel. + 357 26 84 36 00

Abstract: This research studies the application of modern practices and technologies for the collection of a large volume and variety of data, in order to develop a research knowledge base for data mining and analysis in the tourism sector and especially on Cruise and Ship Lines Passengers. Emphasis was given to the application of appropriate methods of data analysis and processing, to produce tangible results for the benefit of sustainable tourism development. Current research focuses on structuring a data warehouse for the collected information in order to apply online analytical processing techniques on the stored data, as well as data mining and data visualization. A holistic approach is proposed, along with a new model for analyzing the impact of tourism activity in general—and cruises in particular—on local society. The results will be utilized as a strategic tool for decision-making by those involved in the tourism sector of cruise areas, with ways to maximize the benefits of tourism, such as increasing overnight stays and, more broadly, passenger consumption, and ways to reduce the environmental impact of visitors and passengers in the ecosystem of cruise areas.

Keywords: strategy, tourism, knowledge base, Big Data

1. Introduction

This research sets out to give important and extensive data on the local economy and ecosystem of cruise areas globally, via the "Development of a Strategic Tourism Knowledge Base for Socio-Economic and Environmental Data analytics on Cruise and Ship Lines Passengers". The objective of this

study is to propose a holistic approach, along with a new model for analyzing the impact of tourism activity in general—and cruises in particular—on local society. This model combines methodologies for examining the economic, social and environmental impacts of tourism activity, whose synthesis is an object of study in the contemporary research. Nonetheless, the research is considered to be innovative in so far as it aims to employ real-time data via social media along with practices taken from Big Data analytics.

Until now, research has focused on the criteria with which cruise companies and travel agencies choose their base ports (Tomigová, Mendes and Pereira, 2016; Wall-Reinius, Ioannides and Zampoukos, 2017) using limited case studies, due to the lack of a database of socio-economic selection criteria of the cruise travelling public. These approaches usually provide limited analysis results, leaving out any reference to the environmental impact of visitors. Given that tourism affects socio-economically and environmentally many sectors of the society and the economy of Piraeus (Apostolopoulos, Leivadi and Yiannakis, 2013), the data that need to be recorded, especially those concerning the critical category of cruise and ship lines passengers, are many and varied. Observations and records of different and detailed data are significant components of a tourist information unit and of modern destinations (Koutsouris 2009; Spilanis and Vayanni, 2014). However, such a large quantity of data has certain characteristics regarding its collection and processing. Information and Communication Technologies (ICT) provide different tools and platforms that may support the dissemination of information to all interested parties of a tourism ecosystem (Chiappa and Baggio 2015; Gretzel et al., 2015). At the same time, the use of a multidimensional data analytics model can enhance the competitiveness and innovation capabilities of a modern tourist destination, through knowledge creation, application and exploitation (Fuchs, Höpken and Lexhagen, 2014).

2. The role of Big Data

A large and growing body of literature has focused on the provision of tourism socio-economic and environmental effects at multiple domains, such as the society and the regional development (Apostolopoulos, Leivadi and Yiannakis, 2013; Cupul-Magaña and Rodríguez-Troncoso, 2017). Many longitudinal studies determined tourism results for economic investments, based on an evaluation of the environmental resources of a destination (Ritchie, Crouch, 2000; Cupul-Magaña and Rodríguez-Troncoso, 2017; Graymore, Sipe and Rickson, 2010; Miller, 2001; Nakajima and Ortega, 2016; Papageorgiou and Brotherton, 1999). Other studies refer to the use of a multidimensional data analytics model that may enhance the capacity and competitiveness capabilities of a modern tourist destination, through knowledge creation, application and exploitation (Fuchs, Höpken and Lexhagen, 2014). Observations and records of different and detailed data are significant components of a tourist information unit and of modern destinations. A major theoretical issue that has dominated the field of tourism studies for many years concerns the role of destination development, emphasizing on the requirements of both the environment and the society. Contemporary research often follows holistic approaches, taking into account economic, environmental, social, cultural, institutional and managerial indicators, combining them with national, regional or urban destinations factors (Liu, 2012; Wong, Tang and van Horen, 2006; Zhao et al., 2013). Heterogeneous data analysis is nowadays a trend that is usually part of

Big Data analysis and data mining. Big Data contributes as it adds the 4 "V"s: Volume, Velocity, Variety and Value, from which variety introduces the heterogeneous of the data while volume and velocity are referring to the algorithms and data mining techniques that must be applied in order to get the value our of them. In our research, data are designed to be collected both from automated and manual procedures and are stored in a complex hybrid database structure (SQL and noSQL DBMSs) in order to extract results by directly applying machine learning techniques on them. Considering the relevant research areas deriving from the aforementioned analysis, Srinivasan and Arunasalam (2013) describe Big Data and their potential research directions, while Hu, Wen, Chua and Li (2014) present research guidelines for Big Data systems. Cevher, Becker and Schmidt (2014) present an approach explaining the simplicity of the algorithms that Big Data should have, while Slavakis, Kim, Mateos and Giannakis (2014) discus fundamental contributions to Big Data theory and practice. An overall presentation of the Big Data analysis major research is presented in: Costa and Santos, 2017.

3. Suggested methodology

This section includes a critical analysis of methodological approaches and suggests a combination of methods that can be used for developing a knowledge base that combines economic, social and environmental data, combined with Big Data analysis. It proposes a new model for approaching primary and secondary data sources like social media for data mining and Big Data analysis. Therefore, the proposed research model is based on the synthesis of suggested methodological approaches presented through the methodology and will be implemented at some future date applied, at a future stage for the collection and analysis of real empirical data. The creation of a Research Knowledge Base for the Collection and Analysis of Socio-Economic and Environmental Data for cruises and ship lines passengers is an innovative research process with regard to the methodological approach as well as the way of organizing, collecting and extracting data through the use of technologies (Miah et al., 2017). In particular, the methodological approach considers holistic exploiting through Multi-Criteria Analysis of the Support Subsystems of Satellite Tourism Accounts and indicators to be mentioned in the tourist ecological footprint (Frechtling, 2010). This hybrid approach and incorporation of indicators of qualitative and quantitative characteristics, as well as environmental parameters, is a modern subject for contemporary research organizations and an effort to reach reliable sustainability indicators. Therefore, data management, its association with secondary sources and data from social media, as well as the analysis and extraction methods described, provide an innovative knowledge base that several international organizations in the field of Tourism seek to develop. The type of data that will be collected and stored in the Tourism Knowledge Base needs to be accurately specified. Therefore, an extensive analysis and description of varied socio-economic and environmental data needs to be conducted. For research purpose, primary and secondary data should be collected. Primary data will arise from basic and supporting subsystems of Tourism Satellite Accounts. A large-scale sample survey is going to be conducted in Cruise and Ship Lines areas, specifically on ship lines passengers and cruise visitors. Conceptual development regarding indicators' interconnections should be conducted prior to data collection. Concerning the calculation and recording of the

ecological footprint (e.g. CO2-eq transmission per square metre or CO2-eq per visitor, air pollution, noise pollution and solid waste), secondary data will be acquired from cooperating institutions. Finally, last methodological step is to extract and collect information from Social Media (Aggarwal, 2011), based on a combination of key-words, tagging and location, in order to efficiently collect further information about the area of Piraeus.

Regarding the methodology of data analysis, an innovative approach of basic and supporting subsystems of Tourism Satellite Accounts is designed to be adopted, using basic and supporting subsystems concerning the measurement of the ecological footprint of tourism activity (TEF). According to this method, five dimensions of data are to be analyzed: quantitative (costs, overnight stays, etc.), qualitative (demographics, nationality, cause of travel, accommodation), temporal (time), location (place) and environmental (ecological footprint). Practical association will be attempted between indicators regarding tourism basic products and tourism basic activities, which are used by cruise and ship lines passengers, combining them in an innovative way, with sustainable indicators and ecological footprint indicators (consumption of energy, particulates, etc.), based on primary and secondary sources. The following methodology approach should be compatible with the principles of the World Travel and Tourism Council (WTTC) and Special TSA country reports. Through this methodology the development of a holistic strategic framework will be attempted, which will result in multi-criteria analysis of Tourism Satellite Accounts and indicators of supporting subsystems, regarding the ecological footprint of tourism (Jones and Munday, 2007). As an indication, normalized indicators will be estimated, regarding cruise and ship lines passengers, through indicators such as Foreign Consumption, Balance of Services, Gross Domestic Product, etc. As for the ecological footprint, an estimation of the actual activity of cruise and ship lines passengers transportation will take place, using the model of Touristic Ecological Footprint (TEF), focusing on the category of transportation (Hunter, 2002).

Knowledge mining in metric relations regarding tourism is not considered as a new approach. The idea of posing fuzzy questions on structured databases may give answers that are not associated to simple statistics. This may lead to conclusions related to forecasts (predictive models) or related to data characterization (descriptive model). Although such methods of data analysis are common, a different algorithm/ method of analysis may be used depending on the complexity and particularity of each topic under investigation. From the analysis of Internet Marketing regarding tourism (Olmeda and Sheldon, 2002) to tourism proposals based on cooperative location algorithms (Zheng et al., 2010), knowledge mining has been extensively used for many years in tourism. Based on the techniques of knowledge mining, we will use predictive and descriptive models, in order to analyze the collected data. This technique goes a step further than simple statistic analysis. At the same time, the use of non-relational databases and specialized technologies can provide us with visualized results (Elastic, 2018; Kibana, 2018). Apart from the in-depth analysis of the collected data, we will attempt to analyze data collected from Social Media, where visitors/ tourists are free to express their views and share their tourism experiences (Xiang and Gretzel, 2010). Using the above technique, the analysis of data will be able to extract knowledge about tourist views on locations. Despite the fact that the velocity of the expected data is not considered to be very high, still, procedures of knowledge management will be followed.
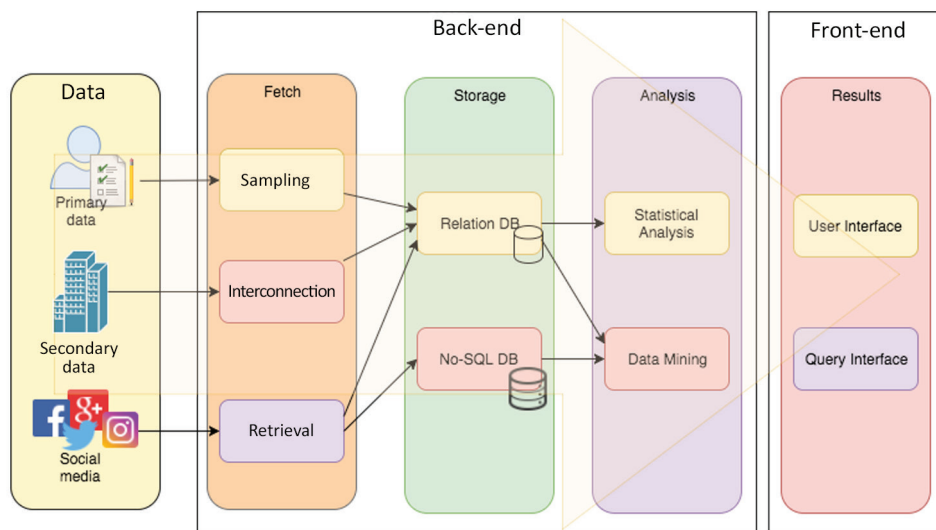
Figure 1. Knowledge base architecture

S o u r c e: Research team processing.

The creation of a Research Knowledge Base of Socio-Economic and Environmental Data, for the cruises and ship lines passengers, provides dynamic and timely research results regarding the social and economic impact of cruises and ship lines passengers, in relation to the environmental impact they bring. The Development of a Strategic Tourism Knowledge Base for Socio-Economic and Environmental Data analytics on Cruise and Ship Lines Passengers provides an important scientific contribution to the country's research effort to create reliable measurement systems, in the form of Satellite Tourism Accounts (Smeral, 2006). Furthermore, it will provide an advanced methodology addressing the sustainable development of destinations and the mapping and combination of quantitative, qualitative, social and environmental indicators. Consequently, the research results, the developed infrastructure and the acquired know-how will have a significant economic and social impact in the wider cruise areas.

4. Conclusion

The proposed research model will provide answers to questions regarding the exploitation of the large volume of visitors who enter the area of cruise marinas, such as how far much overnight stays and tourist consumption increase and social and environmental impacts decrease. In addition, the Information System and the Multicriteria Analysis of Socio-Economic and Environmental Data model will provide a research tool for collecting, creating, and organizing social, economic and environmental data that can be applied and adapted to other destinations, even across the territory, laying the foundations for a Sustainable Tourism Observatory at a national level. The dynamic and continuous results of the research will provide an important tool for strategy development and taking of critical decisions making in

the economy hosting the cruise. The research model in question sheds light not only on the complex nature issue of the impact on the three levels under examination (economic, social, environmental), but also on how the use of Big Data analysis can contribute to both real-time data and new approaches. The current model will be the object of future research at a passenger cruise marina, in which the impacts will be measured in combination employing the logic of a data warehouse through the use of Big Data analytics.

References

Aggarwal, C. C. (2011). An introduction to social network data analytics. In: C. C. Aggarwal (ed.). *Social network data analytics* (pp. 1–15). Boston: Springer. ISBN 9781441984616.

Apostolopoulos, Y., Leivadi, S., Yiannakis, A. (eds.). (2013). *The sociology of tourism: Theoretical and empirical investigations*. Vol. 1. London and New York: Routledge. ISBN 9780415271653.

Cevher, V., Becker, S., Schmidt, M. (2014). Convex optimization for Big Data: Scalable, randomized, and parallel algorithms for Big Data analytics. *IEEE Signal Processing Magazine*, *31*(5), 32–43.

Chiappa, G. D., Baggio, R. (2015). Knowledge transfer in smart tourism destinations: Analyzing the effects of a network structure. *Journal of Destination Marketing and Management*, *4*(3), 145–150.

Costa, C., Santos, M. Y. (2017). Big Data: State-of-the-art concepts, techniques, technologies, modeling approaches and research challenges. *IAENG International Journal of Computer Science*, *44*(3), 285–301.

Cupul-Magaña, A. L., Rodríguez-Troncoso, A. P. (2017). Tourist carrying capacity at Islas Marietas National Park: An essential tool to protect the coral community. *Applied Geography*, *88*, 15–23.

Dodds, R. (2007). Sustainable tourism and policy implementation: Lessons from the case of Calviá, Spain. *Current Issues in Tourism*, *10*(4), 296–322.

Elastic. (2018). Open source search, analytics [online, accessed: 2018-01-05]. Retrieved from: http://elastic.co.

Frechtling, D. C. (2010). The tourism satellite account: A Primer. *Annals of Tourism Research*, *37*(1), 136–153.

Fuchs, M., Höpken, W., Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations – A case from Sweden. *Journal of Destination Marketing and Management*, *3*(4), 198–209.

Graymore, M. L., Sipe, N. G., Rickson, R. E. (2010). Sustaining human carrying capacity: A tool for regional sustainability assessment. *Ecological Economics*, *69*(3), 459–468.

Gretzel, U., Werthner, H., Koo, C., Lamsfus, C. (2015). Conceptual foundations for understanding smart tourism ecosystems. *Computers in Human Behavior*, *50*, 558–563.

HSA. (2016). *Data for Arrivals at Hotels at Piraeus Destination* [online, accessed: 2018-01-05]. Piraeus: Hellenic Statistic Association. Retrieved from Internet: https://www.statistics.gr/el/statistics/-/publication/STO12/2016.

Hu, H., Wen, Y., Chua, T.-S., Li, X. (2014). Toward scalable systems for Big Data analytics: A technology tutorial. *IEEE Access*, *2*, 652–687.

Hunter, C. (2002). Sustainable tourism and the touristic ecological footprint. *Environment, Development and Sustainability*, *4*(1), 7–20.

ISI − Integrated Spatial Investments. (2017). *Integrated Urban Development Plan: Evaluation of the Integrated Spatial Investment Tool*. Piraeus: Municipality of Piraeus.

Jones, C., Munday, M. (2007). Exploring the environmental consequences of tourism: A satellite account approach. *Journal of Travel Research*, *46*, 164–172.

Kibana. (2018). Explore, visualize, discover data [online, accessed: 2018-01-05]. Mountain view: Elastic. Retrieved from: http://elastic.co/products/kibana.

Koutsouris, A. (2009). Social learning and sustainable tourism development; local quality conventions in tourism: A Greek case study. *Journal of Sustainable Tourism*, *17*(5), 567–581.

Liu, H. (2012). Comprehensive carrying capacity of the urban agglomeration in the Yangtze River Delta, China. *Habitat International*, *36*(4), 462–470.

Lozano-Oyola, M., Blancas, F. J., González, M., Caballero, R. (2012). Sustainable tourism indicators as planning tools in cultural destinations. *Ecological Indicators*, *18*, 659–675.

Miah, S. J., Vu, H. Q., Gammack, J., McGrath, M. (2017). A Big Data analytics method for tourist behaviour analysis. *Information and Management*, *54*, 771–785.

Miller, G. (2001). The development of indicators for sustainable tourism: Results of a Delphi survey of tourism researchers. *Tourism Management*, *22*(4), 351–362.

Nakajima, E., Ortega, A. (2016). Carrying capacity using energy and a new calculation of the ecological footprint. *Journal of Ecological Indicators*, *60*, 1200–1207.

Olmeda, I., Sheldon, P. (2002). Data mining techniques and applications for tourism Internet marketing. *Journal of Travel and Tourism Marketing*, *11*(2–3), 1–20.

Papadopoulos, S. I. (1989). Strategy development and implementation of tourism marketing plans: Part 2. *European Journal of Marketing*, *23*(3), 37–47.

Papageorgiou, K., Brotherton, I. (1999). A management planning framework based on ecological, perceptual and economic carrying capacity: The case study of Vikos-Aoos National Park, Greece. *Journal of Environmental Management*, *56*(4), 271–284.

Ritchie, J. R. B., Crouch, G. I. (2000). Editorial: The competitive destination: A sustainability perspective. *Tourism Management*, *21*(1), 1–7.

Slavakis, K., Kim, S.-J., Mateos, G., Giannakis, G. B.(2014). Stochastic approximation vis-à-vis online learning for Big Data analytics. *IEEE Signal Processing Magazine*, *31*(6), 124–129.

Smeral, E. (2006). Tourism satellite accounts: A critical assessment. *Journal of Travel Research*, *45*(1), 92–98.

Spilanis, I., Vayanni, H. (2004). Sustainable tourism: Utopia or necessity? The role of new forms of tourism in the Aegean Islands. In: B. Bramwell (ed.). *Coastal mass tourism: Diversification and sustainable development in southern Europe* (pp. 269–291). Bristol: Channel View Publications. ISBN 9781873150689.

Srinivasan, U., Arunasalam, B. (2013). Leveraging Big Data analytics to reduce healthcare costs. *IT Professional*, *15*(6), 21–28.

Tomigová, K., Mendes, J., Pereira, L. N. (2016). The attractiveness of Portugal as a tourist destination: The perspective of Czech tour operators. *Journal of Travel and Tourism Marketing*, *33*(2), 197–210.

Turner, R. K., Pearce, D., Bateman, I. (1994). *Environmental economics: An elementary introduction*. New York and London: Harvester Wheatsheaf. ISBN 0745010830.

Wall-Reinius, S., Ioannides, D., Zampoukos, K. (2017). Does geography matter in all-inclusive resort tourism? An investigation of the marketing approach of major Scandinavian tour operators. *Tourism Geographies*, 1–19. DOI: 10.1080/14616688.2017.1375975

Wong, S. W., Tang, B., van Horen, B. (2006). Strategic urban management in China: A case study of Guangzhou Development District. *Habitat International*, *30*(3), 645–667.

Xiang, Z., Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management*, *31*(2), 179–188.

Zhao, S., Feng, K., Gui, S., Cai, H., Jin, W., Wu, C. (2013). The emergy ecological footprint for small fish farm in China. *Ecological Indicators*, *29*, 62–67.

Zheng, V. W., Zheng, Y., Xie, X., Yang, Q. (2010). Collaborative location and activity recommendations with GPS history data. In: *Proceedings of the 19th International Conference on World Wide Web* (pp. 1029–1038). New York: ACM. ISBN 9781605587998. DOI: 10.1145/1772690.1772795.

Strategiczna baza danych społeczno-ekonomicznych i środowiskowych w turystyce – rola analizy zbiorów typu Big Data

Abstrakt: W artykule przedstawiono zastosowanie nowoczesnych narzędzi i technologii umożliwiających gromadzenie dużej ilości różnorodnych danych w celu tworzenia ich zbiorów, pozwalających na eksplorację i analizę naukową danych dotyczących sektora usług turystycznych, a zwłaszcza problemów linii i statków wycieczkowych oraz ich pasażerów. Szczególny nacisk położono na zastosowanie odpowiednich metod analizy i przetwarzania danych do uzyskiwania konkretnych wyników, które będą wspierać zrównoważony rozwój turystyki. Przepro-

wadzone badania koncentrują się na architekturze hurtowni danych, w których gromadzone są pozyskane informacje, umożliwiającej analityczne przetwarzanie zgromadzonych danych (OLAP), a także analizę zestawów danych (data mining) i ich wizualizację. Wyniki badań zostaną wykorzystane jako strategiczne narzędzie podejmowania decyzji w sektorze usług turystycznych, w szczególności w obszarze dotyczącym rejsów wycieczkowych, oraz maksymalizacji korzyści, takich jak zwiększenie liczby noclegów i szerzej – konsumpcji pasażerów, umożliwią także zmniejszenie wpływu odwiedzających i pasażerów statków wycieczkowych na ekosystemy odwiedzanych obszarów.

Słowa kluczowe: strategia, turystyka, bazy wiedzy, Big Data