engineering proceedings

MDPI

Proceeding Paper

Tourism and Big Data: Forecasting with Hierarchical and Sequential Cluster Analysis †

Miguel Ángel Ruiz Reina

Department of Theory and Economic History (Staff of Fundamentals), PhD Program in Economics and Business, University of Malaga, s/n, Plaza del Ejido, 29013 Málaga, Spain; ruizreina@uma.es

† Presented at the 7th International conference on Time Series and Forecasting, Gran Canaria, Spain, 19–21 July 2021.

Abstract: A new Big Data cluster method was developed to forecast the hotel accommodation market. The simulation and training of time series data are from January 2008 to December 2019 for the Spanish case. Applying the Hierarchical and Sequential Clustering Analysis method represents an improvement in forecasting modelling of the Big Data literature. The model is presented to obtain better explanatory and forecasting capacity than models used by Google data sources. Furthermore, the model allows knowledge of the tourists' search on the internet profiles before their hotel reservation. With the information obtained, stakeholders can make decisions efficiently. The Matrix U1 Theil was used to establish a dynamic forecasting comparison.

Keywords: Big Data; forecasting; Google Trends; cluster

Citation: Ruiz Reina, M.Á. Tourism and Big Data: Forecasting with Hierarchical and Sequential Cluster Analysis . *Eng. Proc.* **2021**, *5*, 14. https://doi.org/10.3390/engproc2021005014

Academic Editors: Ignacio Rojas, Fernando Rojas, Luis Javier Herrera and Hector Pomare

Published: 28 June 2021

1. Introduction

Big Data is a keyword in digitised markets. Technological development and the incorporation of analysis tools have meant a structural change for organisations, firms and institutions. The interpretation and visualisation of complex data are the core of Data Science [1,2]. Technology companies have the most precious asset in a digitised economic environment: information as a competitive advantage [3].

This new digital economy involves reducing information barriers in markets where intermediaries traditionally existed [4]. Consumers, through their searches on the internet, reveal their intentions. These intentions can be used as a predictive modelling tool for future demands of certain products. Hotel demand in a globalised market can be described through searches for potential consumers [5]. Researchers have paid attention to the selective secondary data sources of the internet network. This means a contribution to traditional analysis [6,7].

Methodologies currently applied have attempted to examine regularities in consumer behaviour data [8–10]. The difficulty lies in trying to explain quantitative and qualitative aspects in the modelling. In the field of time series with high dimensions and complex Big Data problems, attention has been paid to concepts such as "The Freedman's Paradox using an Info-Metrics perspective" [11] or "the power of Text in multidimensional contexts with high frequency" [12].

This article is interested in constructing a Hierarchical and Sequential Cluster Analysis (HSCA) for discrete time series. The analysis carried out focused on the decision-making mechanisms of economic agents for the demand for Hotel Accommodation in Spain (HADS). In particular, there are several generic words that consumers search for on the internet that reveal their intention of HADS. Google Trends (GT) provides an amount of information, which is used in this paper. A better understanding of previous searches can be translated into modelling inputs for structuring the forecasting of HADS.

The contribution of this paper is an improvement to current articles in the literature. The previous methodology has been proven to be an adequate input as a predictive tool,

but it lacks classification and hierarchy by topics. The inclusion of a cluster of keywords (124) will allow identifying and segmenting potential consumers. The GT search indexes are for keywords related to tourist interest to visit Spain, and "broad matching" has been used [13,14]. This modelling could be used on internet forecasting for the tourism industry and hospitality, among other fields. Once a volume of temporary searches is known, companies will adjust the offers to their consumers, and there will be a gain in efficiency in decision-making. This fact allows us to model consumer behaviours and to project the regularities of the online tourism market.

Periodicity is essential to reveal systematic behaviours. As we previously cited, a Big Data analysis's difficulty lies in combining qualitative and quantitative research while maintaining traditional modelling standards. We will build the predictions on discrete-time-series variables and seasonal variable dummies (sampling January 2008 to December 2019).

The HSCA method is compared with SARIMA models [15], ADRL + SEASONALITY model [5], Hierarchical Neural Networks (HNN) [16] and Singular Spectrum Analysis (SSA) [5,8]. As a model selection criterion for forecasting, we will use the Matrix U1 Theil decision matrix [5]. The results obtained from the HSCA methodology reveal improvements in predictive capacity about the other models.

The remainder of this investigation is as follows: Section 1.1 provides a review of the existing literature on the forecasting of Big Data applied to Tourism; in Section 2, the theoretical methodology is performed; in Section 3, data analysis of primary and secondary data sources is done; Section 4 is dedicated to discussing the empirical results obtained after applying the methods proposed. Finally, Section 5 is for the mains conclusions obtained and bibliographic references.

1.1. Literature Review

The grouping in time series occurs when we are interested in the collection into categories or clusters. Nowadays, the application is interesting for finance, economics, medicine, engineering, or computing [17–19]. Clustering approaches for time series are time series clustering by features [20–22], clustering models in time series [23–25], or dependency clustering models [26,27].

Regarding predictive modelling of the use of GT, it should be noted that it is relatively recent. The new datasets from Google resources are a disruptive change in the traditional analysis of HDAS worldwide. The model's predictive capacity evolution was determined by techniques previously developed by mathematicians and statisticians. The conventional scientific research was joined by technology development, meaning a breakthrough summarised in Big Data Technologies.

In the scientific literature published using GT in tourism, we would highlight studies with an extensive literature review [9,10], or new modelling and forecasting developments. These studies have found standard results in the forecasting techniques concerning other fields such as parametric and non-parametric techniques [8].

In recent years, authors have published papers with secondary databases from Google. In addition, Neural Networks, Machine Learning, Statistical Methods, and traditional Econometrics have been used as forecasting methods in the tourism sector. Recently, attention has been paid to the spurious relationship between GT Searches and tourism demand [14].

Hierarchical algorithm approaches for clusters have been applied to tourism but have always been used to cross-section data. In particular, secondary data obtained from the travel and tourism competitiveness index are analysed to create clusters. Subsequently, multidimensional scaling techniques are applied to detect the most and most minor influential determinants in tourist destinations' competitiveness [28].

Moreover, a causality method called Granger Causality and seasonality testing has recently been developed, supposing an improvement to Granger's traditional process of causality [5,29,30]. Furthermore, a new dimensionless model selection criterion has recently emerged called the Matrix U1 Theil. This new criterion is a comparative advantage

compared to usual forecasting criteria such as Root of the Mean Square Error, Mean Absolute Error, Theil inequality index, and Diebold–Mariano criterion [5].

2. Methods

This methodological section will develop a new cluster criterion named Hierarchical and Sequential Clustering Analysis (HSCA). This grouping methodology was designed to classify the amount of information existing on the internet network. HSCA will improve and overcome the limitations of keywords previously used in econometric modelling [5]. For this, some properties are cited for modelling with large volumes of data. The first property is Effectiveness and Replicability criteria; the use of HSCA can be replicated in other fields related to Big Data. A second property, identifying clusters with correlation and testing criteria, reveals the importance and causality in our explanatory variables' modelling. A third property, Noise Tolerance and Outliers Values working with large volumes of data, makes the usual theoretical assumptions to be relaxed in favour of accessible interpretation and usability of the model. Finally, a property, Parsimony Criterion, will determine the best model with the least number of explanatory variables.

In real Big Data applications, it is not easy to find a single algorithm that meets the properties described above. The diagram (Figure 1) represents the sequence from a universe of words related to a variable of interest to predict. The graph shows how the keywords initially relate to each cluster and the predicted variable.

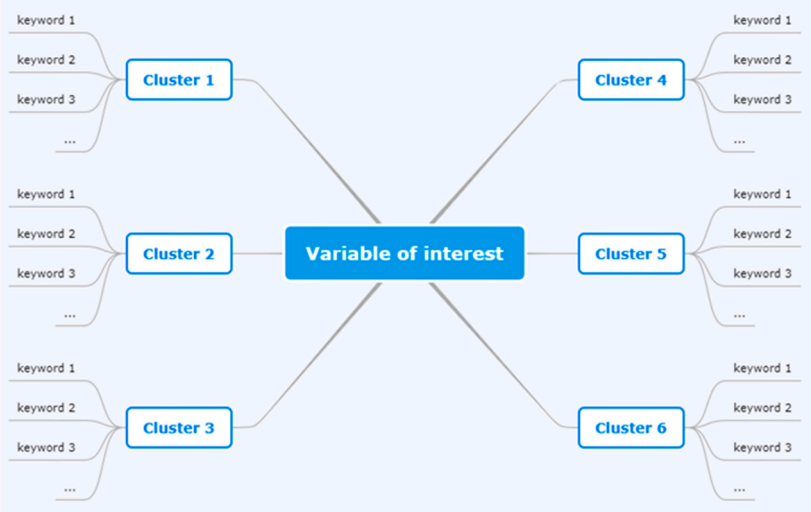

Figure 1. Clustering scheme for a predictive variable (Variable of interest). Own Elaboration.

2.1. Hierarchical and Sequential Clustering Analysis (HSCA)

In this subsection, we will describe the HSCA method. We could divide the methodology into the following sequential steps:

First step: Relevant explanatory variables ($keywords_t$) are selected for forecasting $\{keywords_{mt} \in \mathbb{R}^+;\ m = 1,2,3\ldots;\ t \in T = 1,2,3\ldots T\}$.

In our model, $keywords_t$ are words that future consumers search on the internet before their tourist demand, for instance, Google searches and "broad matching" such as "visit Spain", "rent a car in Spain", or "Weather in Spain" among others. The search words and clusters obtained from GT will be presented in the data section.

Second step: the words of the first step are organised by clusters (topics). $\{keywords_{mlt} \subset cluster_{lt};\ \forall\ (cluster_{1t}, cluster_{2t}\ldots, cluster_{lt});\ l = 1,2,3\ldots\}$.

Third step: auxiliary regressions (y_t and $(keyword_{m1t}, keyword_{m2t}, \ldots, keyword_{mlt})$ are expressed in natural logarithms) are performed for the same forecasting variable (y_t) classified by the cluster. The hierarchy of each group is determined by its R^2. The models present the same dependent variable, and the explanatory variables are different in each grouping.

$$y_t = f(cluster_{1t}) + \sum_{i=1}^{12} \alpha_i w_i + u_{1t} = \sum_{m=1}^{j} \beta_m keyword_{m1t} + \sum_{i=1}^{12} \alpha_i w_i + u_{1t} \quad (1)$$

$$y_t = f(cluster_{2t}) + \sum_{i=1}^{12} \lambda_i w_i + u_{2t} = \sum_{m=1}^{k} \delta_m keyword_{m2t} + \sum_{i=1}^{12} \lambda_i w_i + u_{2t} \quad (2)$$

$$y_t = f(cluster_{lt}) + \sum_{i=1}^{12} \tau_i w_i + u_{1t} = \sum_{m=1}^{o} \psi_m keyword_{mlt} + \sum_{i=1}^{12} \tau_i w_i + u_{lt} \quad (3)$$

where w_i ($for\ monthly\ data\ i = 1, 2, \ldots, 12$) is a deterministic seasonal dummy and uses the HAC covariance method [31].

$$\begin{array}{c} w_1 = -1,\ for\ others\ w_i = 0 \\ w_1 = -1,\ w_2 = 1\ for\ others\ w_i = 0; \\ w_1 = -1,\ w_3 = 1\ for\ others\ w_i = 0; \\ \vdots \\ w_1 = -1,\ w_{12} = 1\ for\ others\ w_i = 0 \end{array} \quad (4)$$

Once the regressions and tests of individual significance of the parameters were made, we determine the most relevant keywords within each cluster. The model selection criteria that verify the clustering procedure developed in this article are the usual ones from Akaike (AIC) and Hannan–Quinn [32]. For instance, to contrast any keyword, we define the null hypothesis as the statement that narrows the model and the alternative hypothesis as the broader one [32].

$$\begin{array}{l} y_t = f(cluster_{1t}) + \sum\limits_{i=1}^{12} \alpha_i w_i + u_{1t} = \sum\limits_{m=1}^{j} \beta_m keyword_{m1t} + \sum\limits_{i=1}^{12} \alpha_i w_i + u_{1t} \\ H_o : \beta_m = 0 \\ H_1 : \beta_m \neq 0 \end{array} \quad (5)$$

Fourth step: after the most relevant words of each cluster were selected, a final preliminary auxiliary regression is performed with the most pertinent explanatory variables of each group.

$$\begin{array}{l} y_t = f(\widehat{cluster}_{1t}) + f(\widehat{cluster}_{2t}) + \cdots + f(\widehat{cluster}_{lt}) + \sum\limits_{i=1}^{12} \vartheta_i w_i + \varepsilon_t = \\ = \sum\limits_{m=1}^{j} \gamma_1 \widehat{keyword}_{m1t} + \sum\limits_{m=1}^{k} \phi_1 \widehat{keyword}_{m2t} + \cdots + \sum\limits_{m=1}^{l} \omega_1 \widehat{keyword}_{mlt} + \sum\limits_{i=1}^{12} \vartheta_i w_i + \varepsilon_t \end{array} \quad (6)$$

The model is simplified under the parsimony criterion, seeking the fewest number of significant explanatory variables with explanatory capacity.

$$\begin{array}{l} y_t = f(\widehat{\widehat{cluster}}_{1t}) + f(\widehat{\widehat{cluster}}_{2t}) + \cdots + f(\widehat{\widehat{cluster}}_{lt}) + \sum\limits_{i=1}^{12} \vartheta_i w_i + \widehat{\widehat{\varepsilon}}_t = \\ = \sum\limits_{m=1}^{j} \gamma_1 \widehat{\widehat{keyword}}_{m1t} + \sum\limits_{m=1}^{k} \phi_1 \widehat{\widehat{keyword}}_{m2t} + \cdots + \sum\limits_{m=1}^{l} \omega_1 \widehat{\widehat{keyword}}_{mlt} + \sum\limits_{i=1}^{12} \vartheta_i w_i + \widehat{\widehat{\varepsilon}}_t \end{array} \quad (7)$$

The interpretation of coefficients are elasticities, and the dummy variables are semi-elasticities [33].

2.2. Comparison of Forecasting and Evaluation

Forecasting and control problems are closely linked. To forecast, we will define the following expression for our modelling as follows:

$$\begin{array}{l} E(y_{t+h}|x_t, w_i) = E(\sum\limits_{m=1}^{j} \gamma_1 \widehat{keyword}_{m1t+h} + \sum\limits_{m=1}^{k} \phi_1 \widehat{keyword}_{m2t+h} + \cdots \\ + \sum\limits_{m=1}^{l} \omega_1 \widehat{keyword}_{mlt+h} + \sum\limits_{i=1}^{12} \vartheta_i w_i) \end{array} \quad (8)$$

where h represents the time horizon, and the residuals of the forecasting are white noise $\left\{ E(\widehat{\varepsilon}_{t+h}|x_{t+h}, w_i) = 0;\ \mathrm{var}(\widehat{\varepsilon}_{t+h}|x_{t+h}, w_i) = \sigma^2_{\widehat{\varepsilon}};\ \mathrm{cov}(\widehat{\varepsilon}_{t+h}|x_{t+h}, w_i) = 0 \right\}$.

As a model selection criterion, we will base ourselves on the Matrix U1 Theil decision matrix. A dimensionless matrix is designed for the decision to select predictive models [5].

3. Data

Data were collected from Jan. 2008 to Dec. 2019. Therefore, we can differentiate two data sources, on the one hand, the official data sources from the INE (Spanish National Statistics Institute (Instituto Nacional de Estadística) https://ine.es/ (accessed on 24 June 2021).) for the predicted variable (HDAS), and the explanatory variables are obtained from Big Data secondary sources, in particular, from the GT tool.

HDAS presents some relevant characteristics in the time series analysis; it is worth noting the high seasonality and a growing trend throughout the period analysed (Figure 2).

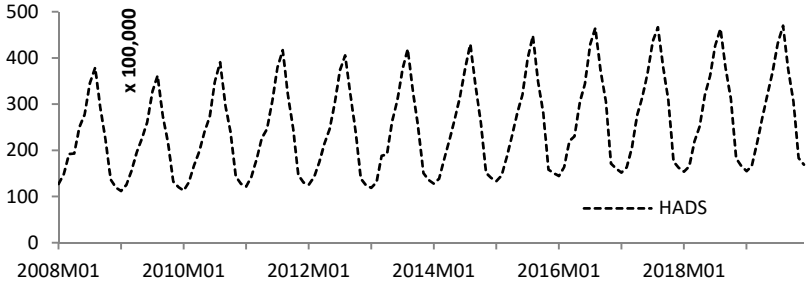

Figure 2. Number of HADS (January 2008 to December 2019). Data source: INE. Own Elaboration.

From a statistical point of view, it should be noted that the maximum values for each year occur in the summer season, the highest value in August 2019 with 46,998,612 hotel overnights in Spain, and the lowest value in January 2009 with 11,203,819. For the 144 observations analysed (Table 1), the existence of unit roots (ADF (p-value) = 0.85) and stationarity variance (KPSS (p-value) = 0.56) should be highlighted [34,35]. The KPSS (stationary variance) results allow us in our modelling to adjust dummy variables for the repetitive behaviours of the series (seasonality).

Table 1. Descriptive Statistics and Stationary Analysis of HADS (Jan. 2008 to Dec. 2019). Own Elaboration.

Mean	Maximum	Minimum	ADF (p-Value)	KPSS (p-Value)	Observations
24,989,874	46,998,612	11,203,819	0.85	0.56	144

The sample period includes 18,000 contemporary observations. From INE data, there are 144 for the variable to be predicted (HADS). The search terms related to planning a visit to Spain were collected from GT and are presented in Table A1 (see Appendix A). In this document, we worked with 17,856 observations of search variables contemporary to the HADS variable. The information is summarised in nine clusters with 124 search terms related to hotel tourism demand from January 2008 to December 2019 for tourists worldwide. All the keywords were searched using "broad match" and combination with other terms. e.g., entering "Spain Hotel", "Spain culture", and so on [13].

4. Results

In the following section of empirical results, we describe a training period between January 2008 and December 2018, with a testing sample to forecast 12 months in 2019. The applied methodology is previously mentioned in Section 3—Table 2 shows the most relevant keywords within each tourist interest cluster. Regarding the hierarchy, we can

indicate that all the keywords finally selected in each group are the most descriptive capacity. For example, finding all values between 0.95 and 0.99, highlighting the terms related to the "social" cluster, which shows that these search engines have a high explanatory capacity, highlighting "Airbnb", "Youtube", "English", "Tripadvisor", "Twiter". However, the differences between the clusters and their hierarchy are minimal. An aspect to highlight is that the dummy variables described for systematic seasonality were relevant for all models in all the sets.

Table 2. Summary of clusters and keywords (broad matching) relevant for HADS. Sample January 2008–December 2018. Own Elaboration.

Cluster	Relevant Keywords	R-Squared
Sports	sport	0.95
Laws	visa	0.97
Transport	car, flight	0.98
Seasonality	summer, winter	0.95
Social	Airbnb, Youtube, English, Tripadvisor, Twiter	0.99
Welfare	Android, Xiaomi	0.98
Searches	low-cost, Spain Tourism, visit Spain	0.98
Culture	alcohol, city breaks, monuments, architecture	0.97
Places	Beach, Canary Island, Alhambra, Plaza de España, Sagrada Familia	0.98

Once the main information clusters were selected to predict the variable of interest, we carried out final modelling for the set of variables in the groups to choose the best regressors to evaluate their predictive capacity. In our modelling, we expressed all the variables in natural logarithms, except the seasonal dummy variables, with the *p*-values in parentheses. We obtain the following result as follows:

$$\begin{aligned}\widehat{y}_t = \underset{(0.00)}{15.90} + \underset{(0.04)}{0.08}\,Airbnb_t + \underset{(0.00)}{0.06}\,Apple_t - \underset{(0.00)}{0.12}\,car_t + \underset{(0.00)}{0.03}\,city\_breaks_t + \underset{(0.00)}{0.07}\,flight_t \\ - \underset{(0.00)}{0.08}\,Samsung_t + \underset{(0.01)}{0.03}\,sport_t + \underset{(0.00)}{0.13}\,visa_t + \underset{(0.06)}{0.07}\,visit\_Spain_t + \sum_{i=1}^{12}\vartheta_i w_i\end{aligned} \tag{9}$$

$$\begin{aligned}\sum_{i=1}^{12}\vartheta_i w_i = \underset{(0.00)}{0.10}\,w_2 + \underset{(0.00)}{0.34}\,w_3 + \underset{(0.00)}{0.50}\,w_4 + \underset{(0.00)}{0.69}\,w_5 + \underset{(0.00)}{0.82}\,w_6 \\ + \underset{(0.00)}{1.05}\,w_7 + \underset{(0.00)}{1.16}\,w_8 + \underset{(0.00)}{0.90}\,w_9 + \underset{(0.00)}{0.69}\,w_{10} + \underset{(0.00)}{0.18}\,w_{11} + \underset{(0.00)}{0.10}\,w_{12}\end{aligned} \tag{10}$$

$$\begin{aligned}&T = 132;\ Sample:\ 2008M01\ 2018M12;\\ &Method:\ Least\ Squares\ HAC\ standard\ errors\ \&\ covariance\\ &(Bartlett\ kernel,\ Newey-West\ fixed = 5.0000);\\ &R^2 = 0.99;\ Adjusted\ R^2 = 0.99;\\ &AIC = -3.04;\ P-Value\ (Wald\ F-Statistic) = 0\end{aligned} \tag{11}$$

The final model selected presents a high explanatory capacity $R^2 = 0.99$. All the parameter interpretations are studied as the percentage increases of the regressors (1%). For instance, the variable "Airbnb" implies an increase of HADS of 8%; in the explanatory variables, the variables "flight" and "visit Spain" are interpreted as a 7% increase in HADS. It is interesting to mention that the variables "Car" (-0.12) have a negative sign and "flight" (0.07) represents a positive sign. The technological variables (Samsung, Apple), "sports", and "City Breaks" are relevant.

The prediction of the final HSCA model is compared to other models cited in the Introduction section. The comparative graph of the forecasting time series can be seen in Figure 3.

Table 3 below shows the comparison between the HSCA model and the other predictive models (ADRL + SEASONALITY, SARIMA, HNN, SSA) using Matrix U1 Theil (values more

significant than one will indicate better predictive capacity than HSCA; otherwise, we find values less than 1). The HSCA model shows the best predictive power in test h = 3 and h = 6. For a time horizon of h = 12, it would be below ADRL + SEASONALITY and HNN.

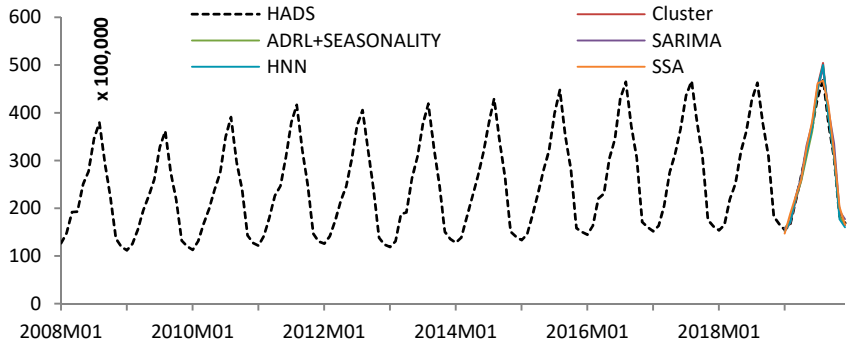

Figure 3. Out-sample forecast HADS h = 12 (January 2018 to December 2019). Own Elaboration.

Table 3. Summary of forecasting accuracy. Out-Sample training Jan. 2019–Dec. 2019. Own Elaboration.

	HSCA	ADRL + SEASONALITY	SARIMA	HNN	SSA
HSCA (h = 3)	1.00	0.39	0.36	0.43	0.15
HSCA (h = 6)	1.00	0.50	0.69	0.92	0.39
HSCA (h = 12)	1.00	1.14	0.86	1.13	0.79

5. Conclusions

In the present investigation, a grouping model was developed for hotel accommodation forecasting (HADS). The properties described in the methodological section were central to the research (Section 3). Databases from primary (INE) and secondary (GT) sources were studied. The HSCA model shows a forecasting and causality capacity. A total of 124 Keywords were analysed in a time series from January 2008 to December 2019 (18,000 observations, including HADS). We determined the primary search keywords by topic (Table 2). The hierarchy of each cluster was also fixed.

Furthermore, this research was compared with other models with high predictive capacity, such as ADRL + SEASONALITY: SARIMA, HNN and SSA. Analysing the Matrix U1 Theil results for time horizons $h = 3$, we found HSCA (coefficients less than 1) as the best model. For an annual time horizon, we discovered that ADRL + SEASONALITY (1.14) and HNN (1.13) performed better results than HSCA. Let us compare the causal explanatory capacity ($R^2 = 0.99$). We can say that HSCA is the best since it includes many more explanatory variables (search topics) than the rest of the models studied. With the information obtained from the HSCA model, it is possible to adjust tourist profiles based on their searches. Primary and secondary tourism industries can benefit from this knowledge of the global market.

We can deduce that previous studies' explanatory capacity was improved from this work, providing relevant and novel information to the scientific literature. Furthermore, this research is the basis for future empirical work related to stakeholders' Big Data field and decision-making. Currently, the most developed economies are focused on a digital environment. Both firms and consumers are expanding their activities on digital platforms, which makes it possible to measure market actions. Furthermore, the engineering of search engines such as Google comes from valuable information to improve the predictive capacity of the models. The results presented in this study refer to consumers' active search, but the data generated can generate predictive information for future tourism consumers. The impact on this type of study's economy supposes a paradigm shift in traditional tourism analysis studies.

The study was applied to the tourism field. However, this methodology can be applied to the finance, insurance or airline field, where decision-making is critical in competitive markets.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and analysed during the current study are available in the repository (https://drive.google.com/file/d/1T5_dQGxCLlG7IKMOPvgFUD9lsCe344fn/view?usp=sharing) (accessed on 23 June 2021).

Acknowledgments: The author wishes to acknowledge the support given by the University of Malaga. PhD. Program in Economics and Business, effective from 16 July 2013. Especially to Associate Professor Antonio Caparrós Ruiz from the University of Malaga for reviewing this work. Group of research: "SEJ157-INIDICADORES SOCIALES".

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Keywords and clusters correlated with HADS (broad matching). January 2008 to December 2019. Own Elaboration.

Sports	Laws	Transport	Seasonality	Social	Welfare	Searches	Culture	Places
Sport	Taxes	Transport	Weather	Spanish People	Hospitality	Trip Spain	Monuments	Beach
Football	Tax free	Flight	Winter	Mind	Environment	Visit Spain	Musueums	Mountain
Basketball	Laws	Train	Summer	vegan Spain	Relax	Spain Tourism	Congress	Island
Athletics Spain	Schengen	Roads	Autumn	English	Stress	Hotel Spain	Study	nature
Swimming Spain	Spain passport	Cruise ships	Spring	French	Life style	Apartment Spain	Disco	Mediterranean area
Volleyball Spain	Visa Spain	Helicopter Spain	Climate Change	Italian	Hospital	Best travel	Concert	Canary Island
Tennis Spain	Spain travel insurance	Bus Spain	Easter week Spain	German	Apple Spain	Resort	Food	Zoo Spain
Boxing Spain	Medical certificate Spain	Car Spain	Christmas Spain	Facebook	Android Spain	Ecotourism	Wine	Andalusia Spain
Soccer Spain	Spain driving license	Tolls Spain	-	Twitter	Samsung Spain	Family Trip	theme parks Spain	Catalonia Spain
Hockey on ice Spain	-	Motorhomes Spain	-	Tripadvisor	Xiaomi Spain	low cost	nightlife Spain	Alcázar de Toledo
Baseball Spain	-		-	Hotels.com Spain	Huawei Spain	Rural Spain	Spain architecture	Monasterio del Escorial
-	-	-	-	Booking.com Spain	-	Agriculture Spain	alcohol	Palacio Real
-	-	-	-	Wimdu	-	Fishing Spain	City Breaks	Muralla de Ávila
-	-	-	-	Kayak Spain	-	Livestock Spain	-	Alcázar de Segovia
-	-	-	-	Airbnb	-	Blitz	-	Valencian Community Spain
-	-	-	-	Instagram	-	-	-	Plaza de España
-	-	-	-	Youtube	-	-	-	Teatro Romano de Mérida
-	-	-	-	Terrorism	-	-	-	Acueducto de Segovia
-	-	-	-	Overtourism	-	-	-	Mezquita de Córdoba
-	-	-	-	Tourism Phobia	-	-	-	Sagrada Familia
-	-	-	-	Wifi Spain	-	-	-	La Giralda
-	-	-	-	3G, 4G and 5G Spain	-	-	-	La Alhambra and Tours

References

1. Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; Mullainathan, S. Human decisions and machine predictions. *Q. J. Econ.* **2018**, *133*, 237–283. [CrossRef]
2. Carrizosa, E.; Guerrero, V.; Morales, D.R. Visualising data as objects by DC (difference of convex) optimisation. *Math. Program.* **2018**, *169*, 119–140. [CrossRef]
3. Mikalef, P.; Pappas, I.O.; Krogstie, J.; Giannakos, M. Big data analytics capabilities: A systematic literature review and research agenda. *Inf. Syst. e-Bus. Manag.* **2018**. [CrossRef]
4. Palos-Sanchez, P.R.; Correia, M.B. The collaborative economy based analysis of demand: Study of airbnb case in Spain and Portugal. *J. Theor. Appl. Electron. Commer. Res.* **2018**. [CrossRef]
5. Ruiz-Reina, M.Á. Big Data: Does it really improve Forecasting techniques for Tourism Demand in Spain? In *International Conference on Time Series and Forecasting*; Godel Impresiones Digitales S.L.: Granada, Spain, 2019; pp. 694–706.
6. Song, H.; Li, G. Tourism demand modelling and forecasting—A review of recent research. *Tour. Manag.* **2008**, *29*, 203–220. [CrossRef]
7. Pan, B.; Wu, D.C.; Song, H. Forecasting hotel room demand using search engine data. *J. Hosp. Tour. Technol.* **2012**, *3*, 196–210. [CrossRef]
8. Wu, D.C.; Song, H.; Shen, S. New developments in tourism and hotel demand modeling and forecasting. *Int. J. Contemp. Hosp. Manag.* **2017**, *29*, 507–529. [CrossRef]
9. Mariani, M.; Baggio, R.; Fuchs, M.; Höepken, W. Business intelligence and big data in hospitality and tourism: A systematic literature review. *Int. J. Contemp. Hosp. Manag.* **2018**. [CrossRef]
10. Li, J.; Xu, L.; Tang, L.; Wang, S.; Li, L. Big data in tourism research: A literature review. *Tour. Manag.* **2018**, *68*, 301–323. [CrossRef]
11. Macedo, P. Freedman's Paradox: An Info-Metrics Perspective. In *International Conference on Time Series and Forecasting*; Godel Impresiones Digitales S.L.: Granada, Spain, 2019; pp. 665–676.
12. Gabrielyan, D.; Masso, J.; Uuskula, L. Powers of Text. In *International Conference on Time Series and Forecasting*; Godel Impresiones Digitales S.L.: Granada, Spain, 2019; pp. 677–693.
13. Choi, H.; Varian, H. Predicting the Present with Google Trends. *Econ. Rec.* **2012**, *88*, 2–9. [CrossRef]
14. Bokelmann, B.; Lessmann, S. Spurious patterns in Google Trends data—An analysis of the effects on tourism demand forecasting in Germany. *Tour. Manag.* **2019**, *75*, 1–12. [CrossRef]
15. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis: Forecasting and Control*, 4th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2013.
16. Athanasopoulos, G.; Hyndman, R.J.; Kourentzes, N.; Petropoulos, F. Forecasting with temporal hierarchies. *Eur. J. Oper. Res.* **2017**, *262*, 60–74. [CrossRef]
17. Liao, T.W. Clustering of time series data—A survey. *Pattern Recognit.* **2005**, *38*, 1857–1874. [CrossRef]
18. Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering—A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [CrossRef]
19. Caiado, J.; Maharaj, E.A.; D'Urso, P. Time-series clustering. In *Handbook of Cluster Analysis*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2015; pp. 241–264.
20. Kakizawa, Y.; Shumway, R.H.; Taniguchi, M. Discrimination and clustering for multivariate time series. *J. Am. Stat. Assoc.* **1998**, *93*, 328–340. [CrossRef]
21. Scotto, M.G.; Alonso, A.M.; Barbosa, S.M. Clustering time series of sea levels: Extreme value approach. *J. Waterw. Port Coast. Ocean Eng.* **2010**, *136*, 215–225. [CrossRef]
22. D'Urso, P.; Maharaj, E.A.; Alonso, A.M. Fuzzy clustering of time series using extremes. *Fuzzy Sets Syst.* **2017**, *318*, 56–79. [CrossRef]
23. Alonso, A.M.; Berrendero, J.R.; Hernández, A.; Justel, A. Time series clustering based on forecast densities. *Comput. Stat. Data Anal.* **2006**, *51*, 762–766. [CrossRef]
24. Scotto, M.G.; Barbosa, S.M.; Alonso, A.M. Model-based clustering of Baltic sea-level. *Appl. Ocean Res.* **2009**, *31*, 4–11. [CrossRef]
25. Vilar, J.A.; Alonso, A.M.; Vilar, J.M. Non-linear time series clustering based on non-parametric forecast densities. *Comput. Stat. Data Anal.* **2010**, *54*, 2850–2865. [CrossRef]
26. Alonso, A.M.; Peña, D. Clustering time series by linear dependency. *Stat. Comput.* **2019**, *29*, 655–676. [CrossRef]
27. Alonso, A.M.; Galeano, P.; Peña, D. A robust procedure to build dynamic factor models with cluster structure. *J. Econom.* **2020**, *216*, 3552. [CrossRef]
28. Chávez, J.C.N.; Torres, A.I.Z.; Torres, M.C. Hierarchical Cluster Analysis of Tourism for Mexico and the Asia-Pacific Economic Cooperation (APEC) Countries. *Rev. Tur. Anál.* **2016**, *27*, 235–255. [CrossRef]
29. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
30. Ruiz-Reina, M.Á. Forecasting using Big Data: The case of Spanish Tourism Demand. In *International Conference on Time Series and Forecasting*; Godel Impresiones Digitales S.L.: Granada, Spain, 2019; pp. 782–789.
31. Newey, W.K.; West, K.D. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* **1987**, *55*, 703–708. [CrossRef]
32. Greene, W.W.H. *Econometric Analysis*, 7th ed.; Prentice Hall: Hoboken, NJ, USA, 2012.

33. Peng, B.; Song, H.; Crouch, G.I.; Witt, S.F. A Meta-Analysis of International Tourism Demand Elasticities. *J. Travel Res.* **2015**, *54*, 611–633. [CrossRef]
34. Dickey, D.A.; Fuller, W.A. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica* **1981**, *49*, 1067–1072. [CrossRef]
35. Kwiatkowski, D.; Phillips, P.C.B.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root? *J. Econom.* **1992**, *54*, 159–178. [CrossRef]