International Journal of Contemporary Management
Volume 19 (2020) Number 2, pp. 7–35
doi:10.4467/24498939IJCM.20.004.12669
www.ejournals.eu/ijcm

THE USE OF BIG DATA IN TOURISM SALES FORECASTING

*Magdalena Kachniewska** https://orcid.org/0000-0003-3163-0868

Abstract

Background. The explosion of big data (BD), automation, and machine learning have allowed contemporary businesses to better understand and predict human behavior. In scientific research big data have been widely used to study consumer journey and opinions. One of the tools enabling forecasting of sales volume is the Bass diffusion model, which universal nature has been proven in many applications in forecasting the sale of products belonging to various market segments. This article considers the use of BD as exogenous variables in the Bass model to predict the sales of tourist packages.

Research aims. The purpose of the research is to assess the impact of using big data on improving the accuracy of forecasts for the sale of tourist packages. The Generalized Bass Model (GBM) has been thus expanded to include big data, which means that exogenous variables include: (1) marketer-generated content (MGC) and (2) user-generated content (UGC), including volume of web search and blog posts.

Methodology. This article analyzes online news, blog posts and web search traffic volume related to tourist packages, and then integrates the information into the Bass model, treating it as part of the exogenous variables representing the marketing efforts of tour operators. It has been assumed that the volume of tour operators' web news is a proxy for content generated by marketers (MGC), while the volume of blog posts and web search traffic constitute user-generated content (UGC).

Key findings. The empirical analysis found that by incorporating big data into the Bass model provides more accurate prediction of tourist packages' sales volume. In addition, UGC (as an exogenous variable) is better at predicting sales volume than MGC. UGC is a fairly good tool explaining the level of interest and involvement of potential tourists. However, it has been shown that forecasting efficiency is different for blog posts and web search traffic volumes.

Keywords: Bass model, big data, sales prediction, tourism market, user-generated content.

JEL Codes: M31, M37, C55.

* Warsaw School of Economics (SGH), Warsaw, Poland. E-mail: magdalena.kachniewska@sgh.waw.pl.

INTRODUCTION

Due to the development of information and telecommunications technologies (ICT), data digitization and the growing propensity of buyers to use the Internet and mobile devices, a rapid increase in the amount of data collected in IT systems is observed. Big data is currently considered one of the key resources of enterprise data and a source of a breakthrough in management and marketing research, however the use of big data raises various management and methodological challenges.

At the same time, the diffusion processes of new products and services have become increasingly complex and multifaceted in recent years, as consumers are exposed to a wide range of influence, including oral communication, externalities, and social signals. Diffusion modeling, a field of marketing research that aims to understand the spread of innovation throughout the life cycle, has adapted to the description of these influences.

Information overload that consumers have to deal with everyday means that they do not have sufficient capacity to process and evaluate every alternative available to them. Instead of wasting time and money exploring all the possibilities, consumers usually make decisions based on trust in others. It is a simple and economical way to obtain information in the age of their multitude. The basic mechanism in this area is the word of mouth (WOM) mechanism, which is shaping public opinion with increasing strength.

The use of large transactional databases, and even more dispersed unstructured data, requires significant investments and overcoming methodological, technical and organizational challenges, as evidenced in the literature (e.g., Marr, 2016; Vriens & Kidd, 2014, p. 1; Bean, 2017). It is difficult to expect that tourism entrepreneurs, relatively poorly equipped with IT competences, will take up the challenge in the field of advanced big data analysis systems. However, it is worth using the already known forecasting tools to analyze this part of the data, which can be obtained at a relatively low cost, the more that the knowledge acquired this way can be effectively used to make current business decisions.

The Bass product diffusion model seems to be the best tool binding the issue of big data and diffusion innovation models: it describes the process of acquiring new products as an interaction between its

current and potential users, taking into account the fact that virtual space is becoming the most important interaction environment today. Therefore, this article is the proposal to enrich the well-known innovation diffusion model (Bass model) with marketer-generated content (MGC) and user-generated content (UGC) data that significantly affect the purchasing decision process on the tourism market.

This research is organized as follows: the next section (Background) depicts the literature review in big data and diffusion models (with special emphasis on the Bass diffusion model), section 3 (Method) describes the data (variables used in the model), section 4 (…) describes the procedure for data collection and the empirical analysis of the diffusion of tour operators offer. Finally, the effects of big data variables are validated. The concluding section presents the study's theoretical implications and suggested extensions.

BACKGROUND

Insight into the concept and possible applications of big data

Although big data (unstructured data) is the subject of discussions in scientific research and has numerous practical applications, the term "big data" is difficult to define. In general, it concerns the technology of collecting and analyzing data of large volume and complexity. The data comes from both traditional databases, e.g., functioning in enterprises, which contain the so-called internal data as well as from other sources (documents, e-mails, blogs, social media, various types of electronic sensors, and location devices, e.g., GPS). The data has both a defined and indefinite structure, which hinders their distribution and processing using the available IT infrastructure (architecture and analytical tools) and computational methods. This complexity and diversity of data creates a new challenge for data analysts.

When interpreting Big Data, reference should be made to new technological solutions regarding the processing of large volumes of data of a completely different nature (quantitative and qualitative) than before (McKinsey, 2011). One of the first definitions treats BD as large data, the number of which should be maximized in the analysis

process in order to extract information values (Cox and Ellsworth, 1997). META Group (an analytical and consulting company specializing in information technologies) based the definition of BD on the concept of three attributes (the "3V" model): *volume, variety, and velocity* (Doug, 2001). In 2012, two additional dimensions were introduced relating to BD: *variability* and *veracity* (Katal et al., 2013). Fan and Bifet (2012) define BD as data sets that cannot be managed using current exploration methods or software tools due to their large size and complexity. IBM emphasizes the fact that BD is a variety of data generated from various sources, at high speed and in large quantities (Zikopoulos et al., 2013). Big Data is not only a technological phenomenon. They are considered mainly as a source of economic *value*, hence the original set of data was finally expanded to 6Vs.

Undoubtedly, one cannot indicate a specific number of bytes defining big data, because the rate of their increase is too high (Cisco, 2016). In this study, the term BD will also mean data sets that are at the same time characterized by large volume, diversity, variability and complexity, and additionally, require the use of innovative technologies, tools, and IT methods to extract new and useful knowledge.

Buyers continuously generate both structured transaction data and unstructured data regarding their behavior (Erevelles et al., 2015, p. 2; Hofacker et al., 2016, p. 89). These data have great analytical potential, which is becoming more and more interesting for companies. The rapid development of IT enables building increasingly better business intelligence systems processing big data, identifying their relationships, dependencies or behavior patterns, and developing on the basis of algorithms for future marketing activities. Large data sets are treated as a new form of capital, and their growing resources and significance provoke us to talk about a "megatrend" or "big data revolution" (Erevelles et al., 2015, p. 897; Mayer-Schönberger & Cukier, 2013; Bosch, 2016, p. 57).

Big Data enthusiasts believe that this phenomenon leads to a civilizational breakthrough comparable to the invention of the Internet, steam engine, or print (Mayer-Schönberger & Cukier, 2013; Minelli et al., 2013). Based on the literature review, it can be concluded that this technology creates new opportunities for researchers working on BD. The large amount of unstructured data has a decisive impact on the quality of results in scientific research. In the field of business activity, BD means new perspectives for the company's development

(Hofacker et al., 2016; Saboo et al., 2016). The most important possibilities in this area include: gaining a competitive advantage on the market by creating tailored predictive models, where the client is offered products and services perfectly tailored as well as optimization of enterprise management (Zhan et al., 2017; Erevelles et al., 2015, p. 900–901; Wedel & Kannan, 2016, p. 97–102; Xu et al., 2016, p. 1564–1565). The BD analyzes enable focusing advertising activities, assigning sellers to customers' needs, tracking network movements or providing sales to the sales department information on the scale of interest in a particular product. Therefore, the future of computerization in enterprises and maintaining a favorable market position will be based on Big Data techniques.

From the perspective of this research one of the most important features od BD is variety, as the data come from many sources, occur in various formats and are saved in various forms (numbers, text, image, sound) and generated in a variety of ways, including:

- internal organization data (transaction, accounting, HR data);
- data from external sources (e.g., Internet user generated content);
- transactional data (Chang et al., 2006), generated from e-commerce, service providers, financial institutions;
- the so-called Deep Web data that are not indexed by standard search engines (He et al., 2007; Wei et al., 2010), including dynamic websites, private websites that can be accessed after logging in or restricted websites;
- data created by a large number of information nodes and connections between them (Aggarwal & Wang, 2010).

People using the Internet through various types of websites, portals, applications, on stationary and mobile devices are specific data generators. They leave a digital trail: sequences of websites visited, IP addresses, GPS geolocation data, data from cellular networks, electronic and cashless payments, blog entries, forums, social networks, shopping choices on websites auctions, and e-commerce (Paharia, 2013). Nowadays, sensors connected to the Internet and installed in objects that surround us (Internet-of-Things, IoT) are becoming an increasingly important source of data. Moreover, entries on a social networking site are, after all, texts written freely by its users (which are accompanied by searches, clicks, "likes" and commentaries as well as photos and shared music), so such data is

radically different from entries in the database of transactions made by credit card or registers of temperature measurements and pressure in an F1 car engine. Therefore, they require different analytical tools (Frizzo-Barker et al., 2016).

Of course, not all of the data types listed above are necessary for all purposes. Amazon (founded in the late 1990s as one of the first online bookstores) from the beginning had the idea of increasing book sales by recommending clients new and interesting items. These instructions were reviews, written and published by a team of high-class literary critics on Amazon.com. Positive reviews of the crew significantly increased the sale of specific books (Marcus, 2004). However, the high cost of this method meant that over time a much more efficient method was used: they began to scrutinize all the collected data for relationships not between customers and products but just the products. No hypotheses were made, only statistical relationships between all books were examined, regardless of the author and content. The resulting algorithm worked in real time, displaying recommendations to the customer on the store's website (Mayer-Schönberger & Cukier, 2013).

Another interesting example of the practical application of BD in marketing activities is the analysis of potential influenza epidemic centers. Cooper et al. (2005) showed that search activity and media coverage can be used to monitor and predict the outbreaks of the disease. The Google Flu Trends solution was created in 2009, when the fear of the H1N1 influenza epidemic was growing. It was intended to estimate the spread of the virus based on Google searches. A statistical algorithm was developed based on the correlation of certain phrases typed in the search engine by Internet users with "influenza" in a specific area: the expressions came from a huge mass of these unstructured data, while information about the actual incidents came from structured data of healthcare providers (Ginsberg et al., 2009). Data from 2007–2008 were studied, with approximately three billion queries each day. About 450,000 different mathematical models were then tested and the one best suited to historical data was determined, including 45 search phrases (Ginsberg et al., 2009). The Google Flu Trends tool had an enormous advantage over government agency information in that it forecast flu spread in virtually real time (a maximum delay of 1 day), taking into account the phrase "googled" by millions of American users at a given time. Government

information reports were available with a delay of about two weeks (Mayer-Schönberger & Cukier, 2013).

Another example is the use of data to optimize sales of Target discount store network. By analyzing huge amounts of customer purchase data, the contents of the customer's shopping basket in the first trimester of pregnancy (large quantities of odorless cosmetics, dietary supplements) were determined. Based on the collected data, Target analysts are able to assign pregnancy to any woman who meets the set purchase criteria and predict the term of delivery as well as send discount coupons appropriate for each stage of pregnancy. Nowadays, data correlation occurs to be more important than understanding buyer behavior.

Applications could be multiplied: from forecasting air ticket prices for the cheapest purchase (Mayer-Schönberger & Cukier, 2013), credit risk assessment and optimization of credit costs (Przanowski, 2014), making a diagnosis based on medical and genetic data (Minelli et al., 2013), fighting crime (Hurvitz et al., 2013), after indicating the properties most at risk of fire or choosing the optimal time to replace parts on city buses (Mayer-Schönberger & Cukier, 2013), sports (by using sensors data are collected from players during a game in order to improve their playing schemes), weather (more than 60 years of global weather analysis are used to predict the risk of future extreme events), logistics (smart trucks and smart spaces), agriculture (monitoring weather and soil conditions for optimum point of harvesting), manufacturing (industry 4.0) and energy and telecommunication (smart grid and virtual plants) (OSC, 2015). According to the most optimistic assumptions, the data-driven company is to "become smarter than the people who run it" (Minelli et al., 2013, p. XVII). In general, the main possibilities of using big data analysis in marketing include:

- market forecasting,
- profiling target buyers,
- quantifying the needs of buyers and their purchase motives,
- identifying how products are perceived, purchased, and used,
- designing innovation,
- introducing new products,
- optimization of marketing and production activities,
- relationship management (Vriens & Kidd, 2014; Bean, 2017; Kalkowski, 2016).

Big data are most often analyzed in order to identify accurately the characteristics of buyers, patterns of their shopping and functional behavior in time and space and their relationship with products, brands, companies, etc. Analytical systems are used to predict the dynamics of buyer behavior (total value of customers, size and frequency of their purchases, migration between suppliers, future purchasing choices) and to adapt operational marketing activities to them, achieving measurable financial effects.

The problem in practice is not only the quality of individual data, but also the ability to build data sets that have actual significance, among other useless data (Lee & Sohn, 2016). The analysis of the new product diffusion should take into account at least the data that are within the reach of potential buyers, i.e., data provided by the seller as marketer-generated content (MGC) and information shared by buyers, i.e., user-generated content (UGC).

Literature review on innovation diffusion and Bass model

Consumers today are exposed to many factors that include word of mouth (WOM) communication, external networks, and social signals. In this context, the definition proposed by Peres et al. (2010) seems appropriate: "Innovation diffusion is the process of the market penetration of new products and services that is driven by social influences, which include all interdependencies among consumers that affect various market players with or without their explicit knowledge" (p. 91).

The empirical foundations for the social approach to diffusion of innovation were created as early as in 1950: Ryan and Gross (1950) stated that social contacts and interpersonal communication are important factors that influence the adoption of new behaviors. From an economic point of view, the theory of innovation diffusion is based on the statement that innovation should be understood as a dynamic process dominated by social influences (Kiesling et al., 2012, p. 183). Mathematical modeling of innovation diffusion has enjoyed great interest among scientists since the 1960s. They used it to forecast sales of food products (Fourt & Woodlock, 1960), as well as the rate of imitation of technical change between two companies (Mansfield, 1961). Frank (2004) used a logistic diffusion model to forecast the wireless market in Finland. Qian and Soopramanien (2015) compared three types of diffusion models: Bass, Logistic, and Gompertz, to forecast

sales in the Chinese vehicle market. Lee and Huh (2017) employed a logistic diffusion model to evaluate the effects of policy and oil price on the growth of renewable energy. Diffusion models have been widely used for forecasting sales over product life cycles (Wu & Chu, 2010).

Despite the science research, management practitioners also pay a lot of attention to diffusion models in their search of the most accurate forecasts tool as well as the chance to estimate market reactions to new products and to improve their ability to effectively promote innovation. Sales volume forecasts have a special position because they are an important premise for making many decisions. In particular the launch of a new product on the market is usually accompanied by a lack of data on current sales and it is necessary to use qualitative or selected quantitative methods in forecasting (Dittmann, 2003).

The best-known Bass model, described the diffusion of innovation as a viral process initiated and driven by mass communication and word of mouth (Bass, 1969) – for this reason, sometimes it is referred to in the literature as the Bass epidemic model. The theoretical basis of the Bass model is the innovation diffusion theory, which explains how new products are launched to the market and received in different communities (Rogers, 1962, 1976; Baran & Davis, 2007). If an innovation is a product newly introduced to the market, diffusion of innovation should be understood as the process of popularization of a new product among consumers (Kosińska, 2008), however, the theory assumes that there is a group of consumers ("innovators") ready to buy new products and then recommend them to other consumers "who buy new the product first, and then informs other groups about its usefulness" (Prymon, 1999, p. 81). This means that, according to the innovation diffusion model, most potential customers are willing to buy a new product only after others have tried it (see: Mazurek-Łopacińska, 2002, p. 71).

Diffusion models have a particularly high potential, as they bridge the micro and macro levels because they promote a better understanding of the diffusion process. This approach is presented, among others, by Peres et al. (2010) as well as Chatterjee and Eliashberg (1990). However, the Bass diffusion model deserves special attention as its universality has been confirmed by many applications. The diffusion of innovation is the process of spreading product information among potential customers, regardless of the type of segment to which

the product belongs. In practice, this means that with fragmentary data on product sales, and only the relevant assumptions, you can forecast the volume of sales for each new product.

The main assumptions of the model (nowadays referred to as the classical one) were presented in 1969 (Bass, 1969, p. 215–227). Ultimately, the model had many versions and was repeatedly used in marketing research as it has a strong prediction power for long-term new product performance (Turk & Trkman, 2012). Bass et al. (1994) extended the original model to the Generalized Bass Model (GBM) by including exogenous variables that represent marketing effort and clarify external influences on consumers' decisions.

The latter model extends the simple Bass diffusion model by including additional exogenous variables, thus gaining greater prediction power. While considering different exogenous variables (representing marketing mix), researchers compared GMB's prediction accuracy with that of the classical model (Bass et al., 1994; Danaher et al., 2001; Boehner & Gold, 2012). For instance, in forecasting the adoption of solar photovoltaic systems, government subsidies and policies were included into the modes as exogenous variables (Guidolin & Mortarino, 2010; Yamaguchi et al., 2013; Wang et al., 2017).

Although diffusion modeling has been researched extensively, this field of study has much more to offer in terms of describing and incorporating current market trends, which include the opening up of markets in emerging economies, web-based services, online social networks, and complex product-service structures. On the basis of their review, Peres et al. (2010) suggest that the diffusion framework, if it is to remain a state-of-the-art paradigm for market evolution, must broaden in scope from focusing on interpersonal communications to encompass the following definition: "Innovation diffusion is the process of the market penetration of new products and services that is driven by social influences, which include all interdependencies among consumers that affect various market players with or without their explicit knowledge" (p. 91).

Following the belief, some researchers propose modified diffusion models that integrate big data, such as search traffic and blog posts (i.e., online reviews) as exogenous variables. Numerous studies indicated that search traffic might be useful in market behavior research (Kulkarni et al., 2012; Onishi & Manchanda, 2012; Hassani & Silva, 2016; Choi & Varian, 2012; Hand & Judge, 2012; Du et al., 2015). Onishi and Manchanda (2012) as well as Gopinath et al. (2013) proved

that blog post volume is a better prediction tool than advertising expenditures. The former also showed that TV advertisements increased blogging activity and sales during the pre-launch period, while blog post volume had more significant influence on the sales during the post-launch period (Onishi & Manchanda, 2012). Liang (2006) showed the potential impact of news volume on the stock price.

The potential use of web search and online reviews for improving market prediction accuracy as a part of Bass model has been also proved by Fan et al. (2017), who included a sentiment analysis of online reviews into the GBM and proved the better prediction accuracy of such a forecasting tool, as well as by Goel et al. (2010) in forecasting films and video games sales. Ettredge et al. (2005) indicated that the online search data were useful for predicting the unemployment rate in the USA. Kim et al. (2019) proved that their new model, including search data volume, is superior to the other ones in terms of prediction power. Lee et al. (2018) explored the diffusion of industrial robots by including the volume of search traffic and patent citations as exogenous variables. Yang et al. (2020) were probably the first to make effort to evaluate which types of big data are better for improving the prediction accuracy of GBMs (namely for Korean reverse mortgage sales forecast).

Following the preceding literature, the combination of big data and the Generalized Bass Model (GBM) is to be adapted for tourism packages sales prediction.

METHOD AND DATA COLLECTION

The purpose of this research is to estimate the sales volume of tour operators' offers as well as to evaluate the effects of utilizing big data on improving the prediction accuracy, with the use of the combination of the Generalized Bass Model (GBM) and big data (BD). To be precise, the Bass model will be extended by big data elements, including marketer-generated content (MGC) and user-generated content (UGC) within the exogenous variables.

The Bass model has considered indirect measurements (e.g., price) as exogenous variables to clarify consumers' purchase decisions. It is worth considering how to effectively integrate big data into the GBM to improve their prediction performance. According to the first hypothesis:

H1: *Including big data (blog posts, web search, and web news) in the Bass model as exogenous variables provides better prediction accuracy in predicting the sales volume of tourist packages.*

The most common examples of big data used in market prediction research include: web news (Cooper et al., 2005; Liang, 2006), search traffic (Goel et al., 2010; Kulkarni et al., 2012; Ginsberg et al., 2009; Hassani & Silva, 2016), and blog posts (Gopinath et al., 2013; Yang et al., 2020). The same set of data is to be considered in the following research to test the first hypothesis however the data have been divided into two subsets based on who generates the content. Companies' marketing activities (online news, web sites) are considered as a proxy for marketer-generated content (MGC), while consumers' online activities (blog posts and search traffic representing consumers' interest in new products) are considered user-generated content (UGC). The latter is treated as a tool to promote product information, as consumers' demand is highly sensitive to the data volumes of UGC shared on-line by potential consumers with informational utility (Paletz et al., 2019; Kim & Hanssens, 2017; Riz, 2013; Di Pietro & Pantano, 2012; Taylor et al., 2012). Paletz et al. (2019) and Kim and Hanssens (2017) demonstrated that the UGC had more influence on demand than did firms' marketing activities, however, firms' marketing activities promote consumers' interests by increasing the chance of exposure to products. In order to check which type of on-line marketing activities is more beneficial for improving prediction accuracy, the second hypothesis was formulated:

H2: *The volume of blog posts and web search traffic (representing user-generated content) provide better prediction accuracy in predicting the sales of tourist packages than marketer-generated content (company web news volume).*

This hypothesis has later been extended to Hypothesis 3:

H3: *Various types of UGC (search traffic and blog posts volume) have different prediction power within the field of tourist packages sales volume.*

In the Bass model, current users are referred to as innovators, in the case of potential users the name of followers is used. According to the theory of diffusion of innovation, a group of followers will be interested in the product after having tested it by innovators. This

means the sales process following the Bass curve that can be profiled using only three parameters. This is due to the following mathematical form of the conventional Bass model (Formula 1), which is to be read as: "The portion of the potential market that adopts at time *t* given that they have not yet adopted is equal to a linear function of previous adopters" (http://www.bassbasement.org/BassModel).
(Formula 1):

$$L(t) = \frac{f(t)}{1 - F(t)} = p + \frac{q}{M}[N(t)]$$

Where:

L(t) – the likelihood of purchase at time *t*, given that no purchase has yet been made.

f(t) – the likelihood of purchase at time *t*; the density function of F(t).

F(t) – the cumulative function describing the proportion of those adopting at time *t*.

p – denotes an innovation coefficient referring to external influences such as advertising.

q – denotes an imitation coefficient that refers to internal influences such as WOM.

N(t) – the cumulative function describing the proportion of those adopting at time *t*.

M – the potential market (the ultimate number of adopters).

The potential market M is the number of members of the social system within which word-of-mouth from past adopters is the driver of new adoptions. The Bass model assumes that M is constant, but in practice, M is often slowly changing. Because in the Bass model each adopter is assumed to make one and only one adoption, the terms mathematical term N(t) and n(t) can be thought of as either adoptions or adopters.

The coefficient of innovation *p* is so called because its contribution to new adoptions does not depend on the number of prior adoptions. Since these adoptions were due to some influence from outside the social system, the parameter is also called the "parameter of external influence", while the coefficient of imitation *q* received its moniker because its effect is proportional to cumulative adoptions N(t) implying that the number of adoptions at time *t* is proportional to the number of prior adopters. In other words, the more people are talking about

a product, the more other people in the social system will adopt it. This parameter is also referred to as the "parameter of internal influence".

The Bass model assumes some market potential in the form of the final number of possible buyers of a given product (M), while for each buyer the time of purchase is a random variable of the distribution function F(t) and the function of density f(t) determined by the Formula (1).

As one can easily see, next to the M parameter in the form of the final number of possible buyers of the product, the basic parameters of the model are the parameters *p* and *q* associated with innovators and followers, respectively. This means that forecasting the volume of sales after previous assumption of the size of the market potential is limited to controlling only the two parameters. This allows tailoring of the Bass curve to existing (often fragmentary) data illustrating the volume of past sales. In the absence of any data, sales forecasting is based on the principle of analog forecasting, which involves the use of a previously used combination of parameters for an identical or similar product whose sales has already taken place, thus the values of *p* and *q* parameters are known.

Then, the big data have been included into the conventional Bass model as additional exogenous variables. In order to do this, the basic equation has been converted:

(Formula 2):

$$\frac{f(t)}{1-F(t)} = \left[p + \frac{q}{M}F(t)\right]x\left(t\right)$$

According to Bass et al. (1994) *x(t)*, as shown in Formula (3), represents how sensitive the diffusion process is to price and advertising: *Pr(t)* and *Ad(t)* are the current price and level of advertising expenditure at time *t*. If *x(t)* is greater than 1, the timing of adoption will be accelerated. When it is lower than 1 – the adoption timing will be delayed.

(Formula 3):

$$x\left(t\right) = 1 + \beta_1 \frac{\Delta Pr(t)}{Pr(t-1)} + \beta_2 \frac{\Delta Ad(t)}{Ad(t-1)} = 1 + \beta_1 \frac{Pr(t) - Pr(t-1)}{Pr(t-1)} + \beta_2 \frac{Ad\left(t\right) - Ad(t-1)}{Ad(t-1)},\ t \geq 0$$

Equation 4 was derived to distinguish the cumulative market effort X (*t*) from the periodic market effort *x* (*t*):

$$x(t) = 1 + \beta_1 \ln\left(\frac{\Pr(t)}{Pr(0)}\right) + \beta_2 \ln\left(\frac{Ad(t)}{Ad(0)}\right), \geq 0$$

(Formula 4):

Reorganizing Equations (2) and (4) yields the sales volume at time t as Formula 5:

$$F(t) = m\frac{1 - e^{-(p+q)X(t)}}{1 + \frac{q}{p}e^{-(p+q)X(t)}}, \; t \geq 0$$

(Formula 5):

Where:

M − stands for market potential (saturation level), which is constant and not influenced directly by the marketing efforts

X(t) − represents exogenous variables.

Finally, Formula (3) has been modified to Formula (6) by replacing variables from the original Bass model (e.g., price and advertisement expenditure) with big data and tourism index (TPI). Referring to the literature (Bass et al., 1994), the percent change of each type of big data has been included the changes in *x(t)*. For example, [*Search(t)* – *Search(t*-1)]/*Search(t*-1*)* means the percent change of search traffic volume at time t, and ß4 is its corresponding parameter. The Nonlinear Least Squares (NLS) method was used to estimate the parameters in the model.

(Formula 6):

$$x(t) = 1 + \beta_1 \frac{TPI(t) - TPI(t-1)}{TPI(t-1)} + \beta_2 \frac{WN(t) - WN(t-1)}{WN(t-1)} + \beta_3 \frac{BP(t) - BP(t-1)}{BP(t-1)} + \beta_4 \frac{WS(t) - WS(t-1)}{WS(t-1)}$$

Where:

TPI – tourism price index

WN – web news volume

BP – blog posts volume

WS – web search traffic (relative change rate in the volume over time)

Finally, six models were developed to address the research hypotheses:

Model 1: *x(t)* includes only TPI ($\beta 1 \geq 0$, $\beta 2 = 0$, $\beta 3 = 0$, $\beta 4 = 0$)

Model 2: *x(t)* includes only web news volume, representing the MGC ($\beta 1 = 0$, $\beta 2 \geq 0$, $\beta 3 = 0$, $\beta 4 = 0$)

Model 3: *x(t)* includes only blog post volume, as a part of the UGC ($\beta 1 = 0$, $\beta 2 = 0$, $\beta 3 \geq 0$, $\beta 4 = 0$)

Model 4: *x(t)* includes only web search traffic volume, as a part of the UGC ($\beta 1 = 0$, $\beta 2 = 0$, $\beta 3 = 0$, $\beta 4 \geq 0$)

Model 5: *x(t)* comprises blog post volume and search traffic volume, representing the UGC ($\beta 1 = 0$, $\beta 2 = 0$, $\beta 3 \geq 0$, $\beta 4 \geq 0$)

Model 6: *x(t)* comprises all the types of big data (web news, blog posts and web search) for predicting the sales volume ($\beta 1 = 0$, $\beta 2 \geq 0$, $\beta 3 \geq 0$, $\beta 4 \geq 0$)

RESULTS

The prediction accuracy of each of the six models was evaluated in order to clarify the contribution of variables to improving the model's sales volume prediction power. Model 1 includes only TPI while the others include different types of marketer- and user-generated content (big data). The comparison of Model 1 versus the other ones was used to verify Hypothesis 1. To validate Hypothesis 2, Model 2 was compared with each of the remaining models (3, 4, and 5). Hypothesis 3 was validated by comparing Model 3 and the Model 4 is considered for the last hypothesis.

Prices (in this case: the prices of tourist packages) remain an important component of the analysis based on the Bass model. The TPI as a proxy measure for travel packages was obtained from the TravelDATA Institute for Tourism Market Research (traveldata.pl). Substantially, package prices are average real prices offered by tour operators in booking systems at a given moment. The average price includes 2/3 of the most attractive prices in a given category of hotels at a given date. This assumption is caused by the fact that the remaining ones are relatively rarely the subject of customer choice, while in some directions they significantly affect the price level, thus

disturbing the relative comparability of average prices between individual destinations.

The adjusted package price is created by including in its price elements that are additional payable by the customer, but are not included in the prices provided by the reservation systems. Such price-creating elements include: any fees for tourist visas and any other fees related to tourist stay (e.g., fees for safes, sunbeds, etc.).

Following the annual analysis of the TravelDATA Institute for Tourism Market Research (traveldata.pl), it was assumed that the comparisons will include the packages in the most popular version of the all-inclusive packages (reflecting the choice of 70% of tourists) and covering the most popular period of stay, i.e., one week (again 70%). The compared packages have been grouped into three categories of hotels, i.e., 3-star and 3.5-star, 4-star and 4.5-star and 5-star and 5.5-star (reflecting the choice of almost 95% of tourists). For every category twelve dates of stays have been adopted covering the second week of each month (despite April, May, and December, when the holiday period has been covered).

Regarding the issue of comparisons of tourist attractiveness and the VFM, 22 destinations (12 annual and 10 typically summer) were adopted for comparison, reflecting the most important travel destination (the choice of 88% of Polish tourists). The monthly travel packages sales volume and average packages' prices were obtained from four Polish tour operators.

Web news volumes were collected by crawling entire web pages related to the outbound travel packages offer from WirtualnaPolska.pl and Onet.pl, the dominant domestic portal sites in Poland, as well as the tour operators' web sites. The volume of blog posts covering the discussion of Polish tourists' travel opinions was obtained from Brand 24 (https://brand24.pl/). Meanwhile, Google Trends (http://www.google.com/trends) were used to obtain search traffic volume.

This research relies on the assumption that web news and blog post volumes (referred to as big data) are highly associated with the sales volume and improve the prediction power. The first step was to collect text-based documents by crawling web pages that contain the keywords "foreign holidays". The initial crawling process resulted in 3,292 documents. In the second step, the documents were preprocessed by removing stop words and repeated documents, and 2,656 documents were removed from the analysis.

Then, a set of 8 keywords were identified as strongly associated with the idea of foreign holidays: "holidays abroad", "where to go on holidays", "vacation trips", "seaside", and "beach", but also: "travel agency", "all-inclusive", and "organized trips". Finally, the web sites were re-crawled and finally 9,245 records containing "foreign holidays" and one out of the 8 keywords. The documents were used to measure monthly volumes of web news and blog posts related to tourist packages offer.

The research covered period from January 2019 to February 2020. The data were collected, processed, and analyzed with the R software version 3.6.1.

The search traffic volume was obtained by using Google Trends (http://www.google.com/trends/) and it was used to determine the popularity of query terms over time on Google. Search traffic volumes are adjusted by normalizing the query data on a scale of 0–100 to allow for comparisons over time. Although the absolute search traffic volume is not available at Google Trends, the Bass model requires only the relative change rate in the volume over time.

Figure 1 illustrates changes in tourist packages sales volumes and big data (web news, web search, and blog posts) in a given period and shows the relationship between them. In particular, the number of web searches over time behaved similarly to the volume of sales.

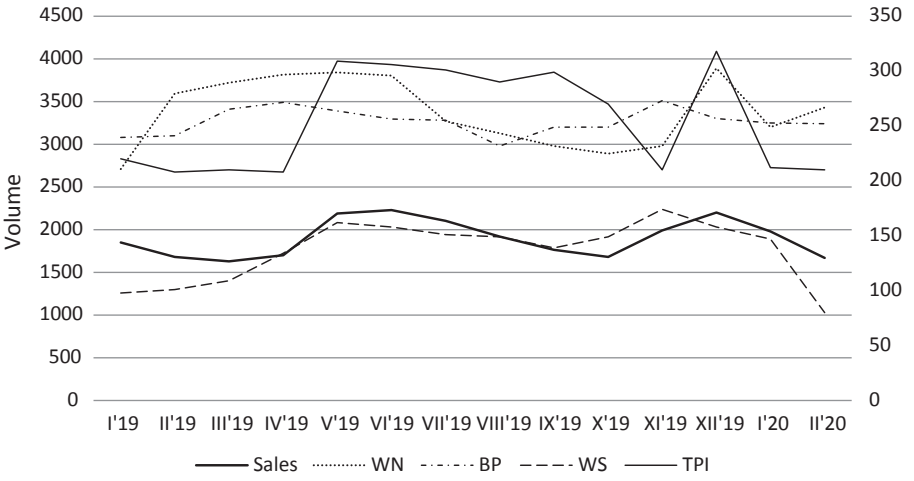

Figure 1. Trends in sales volume, web news, web search, and blog posts

Source: own description based on research results.

To statistically verify the relationships, a correlation analysis was performed, as summarized in Table 1. It was statistically proven that the tourism packages sales volume is significantly and positively correlated with other big data variables (web news, blog posts, and web search). It is, therefore, justified to conclude that the inclusion of big data as part of the Bass model increases the accuracy of forecasting sales volume. Moreover, Table 1 shows the high correlation between individual variables underlying the multi-collinearity in further analysis.

The next step was to measure the prediction accuracy of the six models proposed before. Using the results of MAPE (Mean Absolute Percentage Error, %) and MAE (Mean Absolute Error) the best prediction accuracy was observed for web search traffic variable (i.e., Model 4), being followed by cumulated user-generated content data (blog posts and web search traffic). Both MAPE and MAE results show that Model 1 (only TPI, no big data provided) represents the lowest prediction accuracy, while using big data (Model 6) improves the prediction significantly. Therefore, Hypothesis 1 was confirmed: Using big data as a variable in the Bass model improves the accuracy of tourist packages sales volume prediction. Moreover, the results revealed that Model 3 and Model 4 (i.e., the models utilizing blog post and/or search traffic volumes) represent much better prediction accuracy than Model 2 (based on company web news). However, Model 3 (including only blog posts volume) gives way to Model 5 (total UGC).

Table 1. Correlation analysis

	Sales	**TPI**	**WN**	**BP**	**WS**
Sales	1	0.7718	0.8105	0.8186	0.8834
TPI		1	0.9538	0.9091	0.9763
WN			1	0.9897	0.9894
BP				1	0.9801
WS					1

Source: own description based on research results.

When comparing the performance of Model 5 (including UGC) and Model 6 (big data, including both: UGC and MGC), it turns out that marketer-generated content lowered the model's prediction accuracy significantly. Therefore, it can be concluded that data on UGC provided better information on the diffusion of tourist packages in Poland

than big did data on company web news (MGC) which proves the truth of Hypothesis 2. It is worth pointing out that adding more types of big data does not always guarantee better performance.

Table 2. Prediction accuracy of research models

Models	Variables used in *x(t)*	MAPE (%)	MAE
Model 1	TPI	15.07	496.08
Model 2	web news volume (MGC)	12.03	481.14
Model 3	blog post volume	8.34	379.42
Model 4	web search traffic volume	7.86	362.08
Model 5	blog post and web search traffic volume (UGC)	8.18	367.02
Model 6	web news, blog post, and web search traffic (big data)	9.87	467.25

Source: own description based on research results.

According to the last hypothesis (H3), the two types of UGC (i.e., search traffic and blog posts volume) have a different prediction power within the field of tourist packages sales volume. In fact, Model 4 had lower both MAPE and MAE than those of Model 3. The prediction accuracy of UGC data differs significantly (Table 2).

DISCUSSION

The research showed that the use of big data in the Bass model improves the accuracy of forecasts in relation to the model based solely on the level of prices and expenditure on advertising. Empirical analysis has also shown that forecast accuracy is closely related to the type of data used: UGC provides higher forecast accuracy than MGC, and under UGC, less MAPE and IEA involves using web search than blog posts. On this basis, it can also be concluded that increasing the volume of big data can accelerate sales growth. To confirm such a far-reaching conclusion, it is necessary to continue research and determine which type of big data has the greatest impact on sales growth. To this end, it would be appropriate to conduct a series of sensitivity analyzes, taking into account individual types of exogenous variables (search traffic, blog posts, and web news volume), and observe their impact on sales volume. However, this requires an analysis of longer time series and an increase in the number of entities included in the study.

An indication for future research is also to carry out a similar analysis for other entities of the tourist market (in particular hotels). In addition, future research should be split into different consumer groups. It should be expected that the distinction of leisure tourism and business tourism can have a significant impact on the final result of the study.

An important challenge is also the so-called sentiment analysis of Internet users' comments posted in the form of blogs and microblogs (including comments and opinions regarding previous purchases). While maintaining the impact on sales volume, they can change the vector (positive or negative correlation) depending on the sentiment (positive or negative comment). This aspect was not included in the study.

Last but not least, the research covered the period from January 2019 to February 2020. In March 2020, the economy closed due to the COVID-19 pandemic. In the case of the hospitality industry, this meant a complete exclusion of both domestic and international tourist traffic. Discussion on social forums and Internet searches have been dominated by the issue of withdrawing advances for tourist travel and the prospects of ending the pandemic. As a result, further research according to the assumed researcher method lost its meaning until the economy was re-opened. Research replication will only be possible if the tourism economy is fully open and international connections are restored.

CONCLUSIONS

Marketing is a decision area where support from big data analysis is assessed as particularly useful. Numerous large market entities with access to big data and adequate capital, technological and human capabilities provide evidence of the business utility of big data analysis. Progress in this field is fast, although it requires overcoming many methodological, technical and organizational problems. In the case of tourism sector companies, the lack of ICT competence is an additional barrier.

Future research fields related to big data include, among others an identification of the relationship between the use of the Internet, social media, and mobile devices with customer journey, recognition

of the ways of purchasing and using products, examination of buyers' behavior in terms of shaping the offer, selection of distribution and marketing channels, ways of personalizing offers created on the basis of big data, determining the impact of marketing algorithmization on financial results, identification of factors affecting the processes of creation and diffusion of innovation, etc.

This article first verified whether big data can be effectively used to predict sales of tourist packages based on the well-known tool, the Bass model. It has been shown that the use of big data in the Bass model improves the accuracy of forecasts in terms of forecasting market penetration by tourist packages. To evaluate the effects of using big data, the accuracy of TPI-based forecasts and big data-based forecasts was compared. There are two types of big data reflecting the level of consumer involvement: MGC and UGC. Comparison of forecast accuracy has shown that big data is more effective in predicting sales of tourist packages than TPI. Both MAPE and MAE levels were significantly lower for all big data variables than for TPI. The analysis also showed differences in the suitability of different types of big data for predicting the level of sales of tourist packages: UGC has a greater impact than MGC, and the use of web search is characterized by higher accuracy than the use of blog posts. Therefore, it can be concluded that the forecasting of sales of tourist packages will improve if both variables are included in the forecasting model. It should be emphasized, however, that web search has a particular impact on the forecasting accuracy. Tourist market entities – apart from generating information independently – should attach much greater importance to network traffic and stimulate the generation of information by Internet users.

Future research should consider the sensitivity analysis for individual big data types (based on different scenarios), which would determine whether increasing the volume of UGC will have a greater impact on the volume of sales of tourist packages than MGC, and how big the synergy effect is between blog entries and traffic associated with search. An important aspect of the future research will also be the analysis of online news sentiment and blog posts. Research into the sale of tourist packages should be extended to other segments of the tourist market (including hotel services), taking into account various groups of buyers (leisure tourism and business tourism).

REFERENCES

Aggarwal, C.C. & Wang, H. (2010). Graph Data Management and Mining: A Survey of Algorithms and Applications, *Managing and Mining Graph Data*, Series: Advances in Database Systems, *40*, 13−68.

Baran, S.J. & Davis, D.K. (2007). *Teorie komunikowania masowego*, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Bass, F.M. (1969). A new product growth model for consumer durables. *Management Science*, *15*, 215–227.

Bass, F.M., Krishnan, T.V. & Jain, D.C. (1994). Why the bass model fits without decision variables. *Marketing Science*, *13*, 203–223.

Bean, R. (2017). How companies say they're using big data. *Harvard Business Review*. https://hbr.org/2017/04 (Accessed: 12th April 2020).

Boehner, R. & Gold, S. (2012). Modeling the impact of marketing mix on the diffusion of innovation in the generalized bass model of firm demand. *Developments in Business Simulation and Experiential Learning*, *39*, 75–91.

Bosch, V. (2016). Big data in market research: why more data does not automatically mean better information. *GfK Marketing Intelligence Review*, *8*(2), 56–63.

Cisco (2016). Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011−2016. http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html (Accessed: 23rd February 2020).

Chang, C., Kayed, M., Girgis, M.R. & Shaalan, K.F. (2006). A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1411−1428.

Chatterjee, R. & Eliashberg, J. (1990). The Innovation Diffusion Process in a Heterogeneous Population: A Micromodeling Approach. *Management Science*, *36*(9), 1057–1079.

Choi, H. & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, *88*, 2–9.

Cooper, C., Mallon, K., Leadbetter, S., Pollack, L. & Peipins, L. (2005). Cancer internet search activity on a major search engine, United States 2001−2003. *Journal of Medical Internet Research*, *7*.

Cox, M. & Ellsworth, D. (1997). *Managing Big Data for Scientific Visualization*. ACM SIGGRAPH '97, Los Angeles.

Danaher, P.J., Hardie, B.G. & Putsis Jr, W.P. (2001). Marketing-mix variables and the diffusion of successive generations of a technological innovation. *Journal of Marketing Research*, *38*, 501–514.

Di Pietro, L. & Pantano, E. (2012). An Empirical Investigation of Social Network Influence on Consumer Purchasing Decision: The Case of Facebook. *Journal of Direct Data and Digital Marketing Practice*, *14*, 18–29.

Dittmann, P. (2003). *Prognozowanie w przedsiębiorstwie. Metody i ich zastosowanie*. Kraków: Oficyna Ekonomiczna.

Doug, L. (2001). Data Management: Controlling Data Volume, Velocity, and Variety. *Application of Delivery Strategies*, META Group (currently with Gartner).

Du, R.Y., Hu, Y. & Damangir, S. (2015). Leveraging trends in online searches for product features in market response modeling. *Journal of Marketing*, *79*, 29–43.

Erevelles, S., Fukawa, N. & Swayne, L. (2015). Big data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904.

Ettredge, M., Gerdes, J. & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*, 87–92.

Fan, W. & Bifet, A. (2012). Mining big data: current status, and forecast to the future, *ACM SIGKDD Explorations Newsletter*, *14*(2), 1–5.

Fan, Z.P., Che, Y.J. & Chen, Z.Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of Business Research*, *74*, 90–100.

Fourt, L.A. & Woodlock, J.W. (1960). Early prediction of market success for new grocery products. *Journal of Marketing*, *25*, 31–38.

Frank, L.D. (2004). An analysis of the effect of the economic situation on modeling and forecasting the diffusion of wireless communications in Finland. *Technological Forecasting and Social Change*, 2004, *71*, 391–403.

Frizzo-Barker, J., Chow-White, P.A. & Mozafari, M. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management*, *36*, 403–413.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. & Brilliant, L. (2009). Detecting Influenza Epidemics Using Search Engine Query Data. *Nature,* 457, 1012–1014. doi:10.1038/nature07634.

Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M., & Watts, D.J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences of the USA*, *107*, 17486–17490.

Gopinath, S., Chintagunta, P.K. & Venkataraman, S. (2013). Blogs, advertising, and local-market movie box office performance. *Management Science*, *59*, 2635–2654.

Guidolin, M. & Mortarino, C. (2010). Cross-country diffusion of photovoltaic systems: modelling choices and forecasts for national adoption patterns. *Technological Forecasting and Social Change*, *77*, 279–296.

Hand, C. & Judge, G. (2012). Searching for the picture: forecasting UK cinema admissions using Google Trends data. *Applied Economics Letter*, *19*, 1051–1055.

Hassani, H. & Silva, E.S. (2016). Forecasting energy data with a time lag into the future and Google trends. *International Journal of Energy Statistics*, *4*, 1650020.

He, B., Patel, M., Zhang, Z. & Chang, K.C. (2007). Accessing the deep web, *Communications of the ACM*, ACM, New York, USA, *50*(5), 94−101.

Hofacker, Ch.F., Malthouse, E.C. & Sultan, F. (2016). Big data and consumer behavior: imminent opportunities. *Journal of Consumer Marketing*, *33*(2), 89–97.

Hurvitz, J., Nugent, A., Halper, F. & Kaufman, M. (2013). *Big Data for Dummies*. Hoboken, NJ: John Wiley & Sons, Inc.

Kalkowski, J. (2016). *Handling a flood of data*. https://www.brandpackaging.com/articles/85628 (Accessed: 11th March 2020).

Katal, A., Wazid, M. & Goudar, R.H. (2013). *Big Data: Issues, Challenges, Tools and Good Practices*, Sixth International Conference on Contemporary Computing (IC3), IEEE, Noida, 404−409.

Kiesling, E., Günther, M., Stummer, Ch. & Wakolbinger, L.M. (2012). Agent-based simulation of innovation diffusion: a review. *Central European Journal of Operations Research*, *20*(2), 183−230.

Kim, D., Woo, J., Shin, J., Lee, J. & Kim, Y. (2019). Can search engine data improve accuracy of demand forecasting for new products? Evidence from automotive market. *Industrial Management & Data Systems*, *119*, 1089–1103.

Kim, H. & Hanssens, D.M. (2017). Advertising and word-of-mouth effects on prelaunch consumer interest and initial sales of experience products. *Journal of Interactive Marketing*, *37*, 57–74.

Kosińska, E. (2008). *Marketing międzynarodowy. Zarys problematyki*. Warszawa: PWE.

Kulkarni, G., Kannan, P. & Moe, W. (2012). Using online search data to forecast new product sales. *Decision Support System*, *52*, 604–611.

Lee, C.Y. & Huh, S.Y. (2017). Forecasting the diffusion of renewable electricity considering the impact of policy and oil prices: The case of South Korea. *Applied Energy*, *197*, 29–39.

Lee, H. & Sohn, I. (2016). *Fundamentals of Big Data Network Analysis for Research and Industry*, Hoboken, NJ: John Wiley & Sons.

Lee, W.S., Choi, H.S. & Sohn, S.Y. (2018). Forecasting new product diffusion using both patent citation and web search traffic. *Public Library of Science ONE*, *13*, e0194723.

Liang, X. (2006). Mining associations between web stock news volumes and stock prices. *International Journal of Systems Science*, *37*, 919–930.

Mazurek-Łopacińska, K. (2002). *Orientacja na klienta w przedsiębiorstwie*, Warszawa: PWE.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, *29*(4), 741–766.

Marcus, J. (2004). *Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut*. New York: The New Press.

Marr, B. (2016). *Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results*. New York: Wiley.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work and think*. New York: Houghton-Mifflin Harcourt.

McKinsey Global Institute (2011). *Big data: The next frontier for innovation, competition, and productivity.* http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (Accessed: 3rd March 2020).

Minelli, M., Chambers, M. & Dhiraj, A. (2013). *Big Data, Big Analytics. Emerging Business Intelligence and Analytic Trends for Today's Businesses*. Hoboken, NJ: John Wiley & Sons Inc.

Onishi, H. & Manchanda, P. (2012). Marketing activity, blogging and sales. International *Journal of Research in Marketing*, *29*, 221–234.

OSC − Open Sap Courses. (2015). Driving business results with Big Data, https://open.sap.com/courses (Accessed: 15th March 2020).

Paharia, R. (2013). *Loyalty 3.0: How to Revolutionize Customer and Employee Engagement with Big Data and Gamification*. Uitgever: McGraw-Hill Education.

Paletz, S.B.F., Auxier, B.E. & Golonka, E.M. (2019). *A Multidisciplinary Framework of Information Propagation Online*. Cham: Springer International Publishing.

Peres, R., Muller, E. & Mahajan, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions, *International Journal of Research in Marketing, 27*, 91–106.

Prymon, M. (1999). *Menedżerskie i społeczne aspekty współczesnego marketingu*, Wrocław: Wydawnictwo Ekspert.

Przanowski, K. (2014). *Credit Scoring w erze Big Data*. Warszawa: Oficyna Wydawnicza SGH.

Qian, L. & Soopramanien, D. (2015). Incorporating heterogeneity to forecast the demand of new products in emerging markets: Green cars in China. *Technological Forecasting and Social Change*, *91*, 33–46.

Riz, L.V. (2013). Online marketing through consumers: a study of effectiveness of various tools and techniques across industries. *British Journal of Marketing Studies*, *1*(4), 11−19.

Rogers, E.M. (1962). *Diffusion of Innovations* (1st edition). London: The Free Press.

Rogers, E.M. (1976). New product adoption and diffusion. *Journal of Consumer Research,* 1976, *2*, 290−301.

Ryan, B. & Gross, N. (1950). The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology*, *27*(372), 15–24.

Saboo, A.R., Kumar, V. & Park, I. (2016). Using big data to model time-varying effects for marketing resource (re)allocation. *MIS Quarterly*, *40*(4), 911–939.

Taylor, D.G., Strutton, D. & Thompson, K. (2012). Self-Enhancement as a Motivation for Sharing Online Advertising. *Journal of Interactive Advertising*, *12*, 13−28.

Turk, T. & Trkman, P. (2012). Bass model estimates for broadband diffusion in European countries. *Technological Forecasting and Social Change*, *79*, 85–96.

Vriens, M. & Kidd, P. (2014, November/December). The big data shift: what every marketer needs to know about advanced analytics. *Marketing Insights*, 23–29.

Wang, W., Yu, N. & Johnson, R. (2017). A model for commercial adoption of photovoltaic systems in California. *Journal of Renewable and Sustainable Energy*, *9*, 025904.

Wedel, M. & Kannan, P.K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121.

Wei, L., Xiaofeng, M. & Weiyi, M. (2010). A Vision-Based Approach for Deep Web Data Extraction, *IEEE Transactions on Knowledge and Data Engineering*, IEEE, 447−460.

Wu, F.S. & Chu, W.L. (2010). Diffusion models of mobile telephony. *Journal of Business Research*, 2010, *63*, 497–501.

Xu, Z., Frankwick, G.L. & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: a knowledge fusion perspective. *Journal of Business Research*, *69*(5), 1562–1566.

Yamaguchi, Y., Akai, K., Shen, J., Fujimura, N., Shimoda, Y. & Saijo, T. (2013) Prediction of photovoltaic and solar water heater diffusion and evaluation of promotion policies on the basis of consumers' choices. *Applied Energy*, *102*, 1148–1159.

Yang, J., Min, D. & Kim, J. (2020). The Use of Big Data and Its Effects in a Diffusion Forecasting Model for Korean Reverse Mortgage Subscribers. *Sustainability*, *12*(3):979. doi:10.3390/su12030979.

Zhan, Y., Tan, K., Ji, G., Chung, L. & Tseng, M.L. (2017). A big data framework for facilitating product innovation processes. *Business Process Management Journal*, *23*(3), 518–536.

Zikopoulos, P., deRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D. & Giles, J. (2013). *Harness the Power of Big Data: The IBM Big Data Platform*, New York: McGraw-Hill.

ANALIZY *BIG DATA* JAKO NARZĘDZIE PROGNOZOWANIA SPRZEDAŻY PAKIETÓW TURYSTYCZNYCH

Abstrakt

Wprowadzenie. Eksplozja nieustrukturyzowanych wielkich zbiorów danych (*big data* – BD), automatyzacji i uczenia maszynowego pozwala współczesnym przedsiębiorcom lepiej przewidywać zachowania poszczególnych podmiotów. W badaniach naukowych duże zbiory danych są szeroko wykorzystywane do badania zachowań i opinii konsumentów. Jednym z narzędzi umożliwiających prognozowanie wielkości sprzedaży jest model dyfuzji Bassa, którego uniwersalny charakter udowodniono w licznych zastosowaniach w prognozowaniu sprzedaży produktów w różnych segmentach rynku. W niniejszym artykule zaproponowano zastosowanie *big data* jako zmiennych egzogenicznych w modelu Bassa w celu zwiększenia dokładności prognozowania wolumenu sprzedaży pakietów turystycznych.

Cel badań. Celem badań jest ocena wpływu wykorzystania wielkich zbiorów danych na poprawę dokładności prognoz sprzedaży pakietów turystycznych. Uogólniony model Bassa został w tym celu rozszerzony o zbiory *big data*. Zmienne egzogeniczne obejmują: (1) treści generowane przez marketerów (en. *marketer-generated content*, MGC) i (2) treści generowane przez użytkowników (en. *user-generated content*, UGC), w tym wyszukiwania w sieci i posty zamieszczane na blogach i mikroblogach.

Metodologia. W artykule analizowane są dane obejmujące wiadomości online, posty na blogach i wolumen wyszukiwań w Internecie (ruch sieciowy) związane z informacjami dotyczącymi pakietów turystycznych polskich touroperatorów. Informacje te zostały zintegrowane z modelem Bassa jako część zmiennych egzogenicznych reprezentujących działania marketingowe touroperatorów. Założono, że wolumen informacji zamieszczanych online przez touroperatorów stanowi odzwierciedlenie treści generowanych przez marketerów (MGC), podczas gdy posty na blogach i ruch związany z wyszukiwaniem w sieci stanowią treści generowane przez użytkowników (UGC).

Wyniki. Analiza empiryczna wykazała, że włączenie dużych zbiorów danych do modelu Bassa zapewnia dokładniejsze prognozowanie wielkości sprzedaży pakietów turystycznych. Ponadto UGC (jako zmienna egzogeniczna) lepiej służy prognozowaniu wielkości sprzedaży niż MGC. UGC jest dość dobrym narzędziem wyjaśniającym poziom zainteresowania i zaangażowania potencjalnych turystów. Wykazano jednak, że skuteczność prognozowania jest różna w przypadku wpisów na blogach i liczby wyszukiwań w sieci.

Słowa kluczowe: model Bassa, *big data*, prognozowanie sprzedaży, rynek turystyczny, treści generowane przez użytkowników.