Master Thesis

Grigor Dimitrov

19 October, 2016

Abstract

Write an abstract!

Contents

1 Introduction

In the last few years we created 90% of the worldâ€™s data, this claim has been made in 2013 and it still consistent today. More precisely it is estimated that every two years in the last three decades the amount of data increases by ten times. Chatfield (2016) . The advantages of such “big data” include possibilities to uncover heterogeneities and subtle population patterns” which are difficult to obtain using self-reported data Fan, Han, and Liu (2014). However, “big data” has been underutilized due to computational and statistical challenges Ammu and Irfanuddin (2013) Tole and others (2013) Labrinidis and Jagadish (2012). Moreover, Nunan and Di Domenico (2013) claim that there are high costs associated with analyzing and storing big data. One of the instances of “big data” is called behavioral data. In the context of this research “behavioral data” refers to recording human behavior, more precisely I am referring to “online behavioral data” which is clickstream data from subject’s browser interactions and the Internet.

The survey questionnaires are foremost used when conduction primary market research data collection. Since the emergence of Web 2.0, the online survey made its way as it is a cheap and convenient form or data collection Evans and Mathur (2005). With the increased penetration of the internet, there were more people online and this availability gave survey market research a huge reach. For example, in 2002 500millon dollars were spent on online surveys while this figure doubled in 2004 Evans and Mathur (2005).

However, 20 years ago when the survey plunged, our economic behavior wasnâ€™t focused that much online and the Internet was used mainly for information purpose. Surveys conducted back then explored topics related to our consumer offline behavior replacing standard paper and pencil interviewing popular in the 80â€™s. Now more than ever our economic activity is focused online, our decision-making process is reflected online and it happens that online we can observe and capture information with ease. Consequently, there is an increasing availability of data that reflects this behavior and our online activities are becoming more valuable to researchers. Due to the economic value that online behavior can drive, there are emerging passive data collection technologies. Such technologies allow us to observe and capture all the online consumer behavior.

Behavioral data is purely observational, such data is more accurate in terms of describing actual behavior i.e. information gathering, research of alternatives and transactions which are the ultimate reflection of preference. On the other hand, survey data is relevant when we have to assess unobservable information such as peopleâ€™s attitudes, motivations, and opinions. Furthermore, it is primary data collected for the specific research goals Hox and Boeije (2005) Glass (1976). These two types of data sources have their limitations and researchers are utilizing them separately, however, up to date, there is a limited amount of research that combines both sources of data. With the aid of passive data collection panels, this project provides a rare opportunity to match an actual online behavior with a subject behind it. That is, by combining passive online behavioral data with survey data, in this research I will be able to access information of both consumers' actual online behavior as well access unobservable information about consumersâ€™ personality traits.

The context of this research is the online travel industry amongst Dutch consumers. Tourism is highly competitive and fragmented market. It has been disrupted by the Internet at large, the disruption process has started with the emergence of the Internet and continues today. Before the penetration of the Internet the market was dominated by "high street" travel agencies, afterwards the market was revolutionized by online travel agencies and direct distribution of travel services as accommodation and transportation.

We are witnessing dynamic segments of consumers emerged because of the technological advancements with constantly changing needs (Xiang, Magnini, and Fesenmaier 2015). This domain has been chosen due to fact is one of the fastest growing online industry. For example, in the Netherlands for 2 out of 3 trips (69%) consumers book accommodations online according to Eurostat data (Eurostat 2016) Moreover, 40% of the Europeans used the internet for travel related purposes, in the Netherlands this number is higher than the average. Online presence of accommodation businesses in 2015 accounts for 95% of all enterprises compared to 75% for the whole economy. According to Euromonitor data, the sector is one of the fastest growing, rising from 899 millon trips in 2009 to 1.4 billon trips in 2019 (Euromonitor 2015). Furthermore, there is an increasing availability of options and all sorts of services bringing the consumers and offers closer together manifesting the ongoing process of disintermediation. Xiang, Magnini, and Fesenmaier (2015) notes that there is growing "bifurcation" or a split among the traditional online travelers to users of traditional travel products and people seeking deeper and authentic experiences. The authors point out that understanding how contemporary travelers use the internet is an important foundation for building successful communication strategies by the business stakeholders. The importance of the internet has been attributed to three main factors, the extensive amount of travel related information, the development of social networks, travel related social networks and peer-to-peer travel offerings where user can exchange travel services and experiences and the mobile computing, smartphones in particular (Euromonitor 2015).

The increasing availability of travel related options online and increasing disintermediation lowers the costs of the travel related services and activities. However, this also makes the decision-making process more demanding as consumers themselves are primarily responsible for choosing the best option rather than using an advice from a travel service provider. Thus, the decisions related to travel involve a lot of risk and uncertainties. Arguably one of the risk and uncertainty reducing instruments that consumers use in decision making related to travel is an information search. Urbany, Dickson, and Wilkie (1989) introduces body of research to support the link between uncertainty and information search. When it comes to perceived risks, in a meta-analysis Gemünden (1985) review one hundred empirical findings to conclude that information search is one of the risk reducing instruments consumers rely on when committing to purchases of complex products.

In an exploratory analysis, Roehl and Fesenmaier (1992) identify seven types of perceived risks related to leisure travel. Namely, equipment risks, financial risks, physical risks, psychological risks, satisfaction risks, social risks and time risks. Lepp and Gibson (2003) look at the general perceptions or risk and uncertainty from the perspective of travel motivations for novelty against familiarity. Their findings suggest that familiarity seekers are more risk averse comparing to less risk averse novelty seekers. With regards to risk attitude and tourism information search, Quintal, Lee, and Soutar (2010) examine the impact of risk and uncertainty avoidance on touristsâ€™ information search. The team distinguishes risk from uncertainty and investigate whether the two constructs have different impact on the information search in terms of number of sources. Their findings suggest that uncertainty avoidance has a positive significant relationship with the extent of information search while this is not valid for the risk averseness.

The following research further explores the relationship between risk and uncertainty attitudes and different instances of online information search behavior. More precisely, this research explores the relationship between risk and uncertainty attitudes, and information search behavior in tourism context while controlling for trip characteristics and demographics. The main objectives of this paper are to investigate the relationship between different risk and uncertainty attitudes, captured by means of self-reported data along with activities on travel related domains such as depth, breadth of the information search, travel micromoments and total time spend on travel-related websites captured by means of online passive behavioral data.

One of the main motivations of this paper lays in the methodological and managerial contributions. Namely, demonstrating a full cycle of research, where the research design utilizes observable online behavioral data collected via passive tracking technology combined with subjects' unobservable characteristics collected via online questionnaire, would provide guidelines to managers in how to best process and exploit their tracking data to extract valuable insights. Up to date, there is a limited amount of literature exploring these constructs in relationship to one another and in a relationship with non-self-reported metrics of information sought. Moreover, the research has theoretical contributions as it combines literature streams that have been researched only using self-reported data. Self-reported is prone to biases as opposed to passive data, which is further discussed in the next chapter.

To improve the research quality, I combine both self-reported, recalled information with objective, passively measured data. Using passive metering technology, I can capture the full footprint of consumers across their devices. This data can help understand the time spent planning the travel, the amount of information search with great accuracy and detail. Although prone to biases, survey data is still important as it can capture unobservable information which is still important and valuable. Merging the two datasets can provide an even deeper layer of insights into consumers' travel behavior.

The study findings are relevant to all travel stakeholders such as businesses, travelers and researchers. The increased availability of data from consumers and its right exploitation can have huge impact on the design of travel related products, their personalization and cross-platform usage, in-line with the trends reviewed in the industry.

2 Literature review

In this chapter I will review the main literature stream relevant for the hypnotized relationships in this paper. First, I will review online travel related activities and information search. Second, I will focus on risk and uncertainty attitudes and expand to risk and uncertainty attitudes related to travel and tourism. Third, I will briefly explore the big five inventory and its utilization in tourism research. In the final section of this chapter, I will derive the hypotheses of this research.

2.1 Travel related activities and information search

Tourism related information search and planning has been a widely researched topic. Internet search behavior in tourism context has been examined primarily from the perspective of demographic variables, motivation and prior knowledge about the destination (Jani, Jang, and Hwang 2014). In a paper exploring longitudinal data for 12 years, Xiang, Magnini, and Fesenmaier (2015) identified as a key trend number one that the internet penetration among consumers using the Internet for travel planning already reached the level of saturation. However, in a research exploring the online behavior from a generation perspective, H. Kim, Xiang, and Fesenmaier (2015) noted that there is no sufficient amount of research with regards to how the newly emerged segments of consumers behave online and use the internet for travel planning.

Xiang, Magnini, and Fesenmaier (2015) turn special attention on travel planning process as a specific type of information search that is an important component of the decision making in tourism. During the planning consumers obtain information based on which they choose their destinations and form their expectations. There is a substantial amount of research that looks at travel planning from different perspectives, some aim to identify the characteristics of the travelersâ€™ demographic characteristics (e.g. D.-Y. Kim, Lehto, and Morrison (2007)), other investigate the way travelers are conducting purchases and navigate in the information stream (e.g. Jun, Vogt, and MacKay (2007)), and recently social media and its influence on travel has been in the focus of researchers as well (e.g. Xiang and Gretzel (2010)).

From practitioner standpoint, the more consumers are being active online the more prerequisites this creates for the tourism stakeholders marketeers to reach them during their decision making process. Therefore, identifying the travel planning process i.e. the decision making journey is a critical step for the brands to intervene and influence the consumers in their direction (Chatfield 2014). Practitioners define that the decision making process related to the tourism as an array of micro moments which often are not even conscious to the consumers. A micromoment is essentially a user session with particular goal of obtaining information or committing to a purchase.

For the purpose of this research it is important to define the main components of the travel planning. In a meta-analysis, Jun, Vogt, and MacKay (2007) point out that there is a consensus among researchers that travel planning cannot be simplified to a single goal-oriented rational action but it is rather viewed as a complex task involving multiple goals and decisions around the different goals and characteristics of the trip. The authors define a conceptual model for travel planning which has three main sequential interrelated components, pre-trip, during trip and post-trip. This research focuses on pre-trip phase. Pre-trip phase itself consists of information search and planning (decision making), furthermore, travel related purchases also occurs at the end of this phase as well as during the trip itself.

Jun, Vogt, and MacKay (2007) define travel plan as a complex decision involving an assessment of multiple alternatives organized around the travel goals in mind. The planning process includes setting goals and considering alternatives in order to achieve that goals including an evaluation of different alternativesâ€™ outcomes. Planning is dependent on the all information search behavior, utilization of the obtained information, purchase behavior and activities including past experience. Pan and Fesenmaier (2006) define vacation planning over the internet as an interaction between the user and the â€œonline spaceâ€ related to destinations and tourism. The online space contains content provided by diverse sources and the technology that facilitates the communication. Userâ€™s â€œsituation, knowledge and skillsâ€ combined with the â€œonline spaceâ€ contribute to the effective search.

It is important to be noted that trip planning is an important and enjoyable part of the vacation experience itself (Stewart and Vogt 1999) and it is likely to be high costs and high involvement decision (Bonn, Furr, and Susskind 1998). Furthermore, travel information search behavior explains travel purchase behavior (Woodside and MacDonald 1994). Quintal, Lee, and Soutar (2010) review numerous aspect of from which information search has been researched including amount of search, number of sources, the search process, involvement, socio-demographic differences, culture etc.

Pan and Fesenmaier (2006) review the consumer vacation planning process from micro level perspective. Their research is motivated by the fact that previous research has been mostly focused on exploring planning and information search on macro level i.e. motivation, need, determinants and outcomes. Their research focuses on a â€œsnapshotâ€ of travel planning where subjects make choices regarding a hypothetical holiday trip to San Diego. Using such setting it was possible to observe different chapters containing many episodes on how consumers adapt in their online search and come up with final decision. These episodes or micromoments defined above represent one of the main variables of interest of this research.

Additionally, I will use the metrics defined by Ho and Liu (2005) in â€œAn exploratory investigation of web-based tourist information search behaviorâ€ . The researchers explored and characterized the online behavior of 96 subjects in laboratory research by assigning them one tourism related task and recording their desktop activity. Ho and Liu (2005) empirically analyzed the recorded videos and reported on different measures, representing the information search behavior.

Additionally to micromoments, in this paper, I will adapt measures utilized in Ho and Liu (2005). The full list of dependent variables representing the internet search behavior includes:

- Total number of micromoments on travel related websites
- Total number of unique travel related domains (breadth of search)
- Total number of pageviews on travel related domains (depth of search)
- Total time spend on travel related domains
- Total sum of all travel related micromoments (this measure has been introduced due to one particular limitations of the mobile measurement discussed in the limitations section)

Ho and Liu (2005) conducted the research in laboratory environment exploring a single goal-directed session of activity. However, the whole information search planning and purchasing usually happens on many occasions within a longer period. Planning is fragmented on many episodes each characterized with its own purpose reflecting a specific problem (Pan and Fesenmaier 2006). In this research, I will use real-life data representing the actual subjectsâ€™ behavior. In this sense, the data is highly accurate and gives detailed representation of the internet search behavior.

In conclusion, information search behavior is an important phase of the of the overall tourism behavior and more specifically the travel planning. From tourism stakeholdersâ€™ perspective, it is a crucial phase where the consumer can be influenced with effective communication strategies and communication systems.

2.2 Risk and uncertainty

Risk attitudes are a central part of the economic theory. Classical economic theory of decision under risk states that the risk is related to the probability of the occurrence of specific outcome. For example according to expected utility a prospect with probability P to win x amount of money opposed to 1-P to win nothing, will be evaluated as follows: p.u(x)+(1-p)u(0) where u is the utility function of money. Risk attitudes are defined as follows, risk aversion is an attitude which is manifested by the preference of the sure outcome over a prospect with higher expected value that involves risk. Whereas, risk seeking attitude will occur when the prospect is preferred over the sure amount.

Later economic theory evolves by distinguishing individual level probability weighing and utility weighting by taking into account different psychological variables. (Woodside and MacDonald 1994) propose â€œProspect theoryâ€ to explain choices among risky prospects that are inconsistent with the standard economic theory. In these recent developments the risk attitudes evolve. In order to explain decision making under risk scholars explore choices involving different amount of risk (high risk and low risk) and associated with outcome involving different monetary values (again high and low) as well gains or loses.

In a paper focused on risk measurement in consumer research Mandrik and Bao (2005) summarize that the measurement of risk attitudes typically has been assessed in three ways. The first method involved â€œchoice dilemmasâ€ where subjects are presented with several scenarios and asked for their preference between two courses of action, this results in computing an overall score which is used to determine respondentsâ€™ risk attitude. The second method involves gambles. Subjects are asked to choose an amount in order to participate in a gamble. Finally, researchers use self-reported measures. These include creation of different scales that are measure risk and uncertainty in specific decision situations. The authors validate a novel self-reported scale which measure general risk attitude as valid psychometrical measure. The construct proposed by Mandrik and Bao (2005) has been utilized in this research as it provides shorter and simple manner of assessing risk attitudes.

2.3 Risk and uncertainty in tourism

As noted above, one of the reasons researchers claim to cause the extensive information search is risk and uncertainty minimization. Stewart and Vogt (1999) attribute uncertainty as an implicit and universal characteristic of every planning process. Furthermore, the authors argue that in order to handle uncertainty the travelers prepare more than one plan for their trips. Sweeney, Soutar, and Johnson (1999) point out that consumers who are more sensitive to risk and uncertainty engage in more extensive search in order to avoid them. Sirakaya and Woodside (2005) claim that because of the intangible nature of tourism products the uncertainty in tourism is higher than comparing to other products or services.

An important remark related to risk and uncertainty is the difference between both constructs. The difference between them lays in the probabilities of their outcomes, while risk is associated outcomes with known probabilities, uncertainty is associated with outcomes with unknown probabilities. Quintal, Lee, and Soutar (2010) investigate the difference between risk and uncertainty on country level using Hofstede (1980) uncertainty avoidance index (UAI) and risk scale measurement on touristsâ€™ information search. The researchers claim that many other papers do not make the distinction and this is especially problematic when researchers are using country UAI scores to explain individual level behavior as individuals differ with regards to their attitudes of risk and uncertainty.

Quintal, Lee, and Soutar (2010) explain the relationship between uncertainty and risk in tourism and information search to be translated in the following way. In the early stages of their research, consumers search for information extensively and the outcomes are associated with uncertainty because the rate of occurrences of certain threatening events or undesirable outcomes is not known. In a later stage of the decision-making process, when travelers have already selected possible alternatives, the risk attitude is more likely to have an influence as consumers can assign relative probabilities to a few selected alternatives i.e. alternatives are being compared to one another providing a reference point.

Based on the relationship explained above, Quintal, Lee, and Soutar (2010) provide a hypothesis that the uncertainty attitude has an influence on the extend of the information search, while holding risk attitude constant. With regards to the risk attitude, Quintal, Lee, and Soutar (2010) provide a hypothesis that risk attitude doesnâ€™t influence the extend of the information search, holding uncertainty attitude constant. This hypothesis, however, has been defended under the notion that all the information search has been performed during the early stage of the trip planning and in the later stage consumers only compare alternatives based on the information they gathered in the early stage. However, in a conceptual model of trip planning presented by Jun, Vogt, and MacKay (2007), the information search is a component of all phases of the trip planning, including pre-trip, during trip and post-trip. Therefore, I find it credible to hypothesize that both risk and uncertainty attitudes influence the extend of the information search.

I further extend my hypotheses, regarding the direction of the effect. My hypotheses are based on Urbany, Dickson, and Wilkie (1989) and Gemünden (1985) claiming that information search is an instrument consumers use to minimize respectively the uncertainties and risks anticipated with regards to future purchases. For example, Gemünden (1985) finds that when the complexity of the decision making is increasing, so does the information search. They attribute this to higher risks involved in that decisions. As travel planning is regarded as complex task, I can expect the same relationship. Urbany, Dickson, and Wilkie (1989) outlines two types of uncertainty, choice and knowledge uncertainty, choice uncertainty is related to uncertainties choosing a particular alternative amongst many, while knowledge uncertainty is related to familiarity with the product features. Their finding suggest that choice uncertainty has a positive relationship with the extend of the information search.

Based on the literature review above about information search, travel planning and risk and uncertainty hereby I form the following hypotheses:

- H1. Risk seeking attitude decreases the extend of information search, keeping other factors constant.
- H2. Uncertainty seeking attitude decreases the extend of information search, keeping other factors constant.

2.4 Big Five Factors Inventory (BFI)

According to Leung and Law (2010) the usage of personality traits in travel related literature appears to be low even though the academics agreed upon their value in Marketing domain (Baumgartner 2002). Personality is a temperament or personâ€™s inherent qualities of mind and strategies according to which one behaves, dispositions and behavioral patterns that are stable across time and can be used to characterize oneâ€™s behavior. The trait perspective has been frequent utilized in consumer research because their ease of application as a self-reported measure and the measurement outputs can be easily applied in statistical analysis. (Jani, Jang, and Hwang 2014).

Big five inventory (BFI) is the most well-known and established factor structure to measure personality (Denissen et al. 2008). It has been validated on many occasions and widely utilized by researchers, creating prerequisites to compare findings across different studies. Another important factor for the popularity of BFI, is the fact they are freely available to use along with their validated translations in many languages. BFI has been proposed as a fundamental lexical hypothesis by Galton in 1884 (Goldberg 1993), which is a language taxonomy of human temperament based on adjectives describing different personality traits. The theory was put into practice by Goldberg (1993) and it has been greatly developed ever since, leading into the construction of five broad factors. BFI are based on factor analysis where a large group of traits is shown to be correlated and grouped into five universal traits.

BFI consists of Extraversion, Neuroticism, Conscientiousness, Agreeableness and Open to experience. Openness to experience is related to the degree of curiosity, inventiveness, adoption of novelty on the one hand and consistency and cautiousness on the other. That is, persons with high openness tend to be open-minded, adventurous while low openness can describe individuals that are more pragmatic. Conscientiousness reflects on the tendency for one to be organized, non-spontaneous, organized and efficient. Extroversion is related to traits such as outgoing personality, sociability, talkativeness. Personalities exhibiting low extroversion on the other hand, can be perceived as less open and reserved. Agreeableness is described as a person being more compassionate and cooperative. It measures whether a person can be trusted or not and if they are well-tempered. High agreeableness personalities are seen as more naÃ¯ve, while low are seen as more dominative and competitive. Finally, neuroticism explores the emotional stability of individuals. That is, a high need for stability results in individuals who are clam and stable, while low need for stability can describe emotionally unstable individuals.

These five factors are shown to be the main factors that drive human behavior, appear in different cultures, are relatively stable across the lifetime of subjects and have strong predictive validity (Jani, Jang, and Hwang 2014). Researchers using BFI often aim to measure big five personality dimensions using as less as possible items which lead to the development and validation of short scales of BFI [(Denissen et al. 2008).

2.5 Personality traits and tourism information search

In tourism domain, BFI has been examined with regards to general travel behavior, the undertaken activities in the travel destination, adventure tourism and pilgrim tourism (Jani, Jang, and Hwang 2014). There are two articles researching internet search behavior from the perspective of BFI. Jani (2011) proposes a model relating information needs and tourist information behavior from the perspective of BFI and travel personality. They define touristsâ€™ information needs as reasons for collecting information. Information search behavior is defined as breadth and depth of information sources that consumers use to obtain information. Furthermore, Jani, Jang, and Hwang (2014), research the relationship between BFI and internet search behavior. Using self-reported survey data, they observe that travel related information sought online varies with regards to the different personalities. The authors conclude that some of the factors from the BFI can improve the information search behavior predictability.

Jani, Jang, and Hwang (2014) address the research question whether BFI can be used as predictor of internet search behavior in terms of sources of information and the extent of the information sought. Using self-reported measures for internet information search and channels used, the authors confirm that personality traits can be used as a predictor of information search behavior.

As BFI in tourism have not been research thoroughly, in this research I will commit to hypotheses concerning the relationship between personality traits and information search behavior. However, the BFI will be used in the research design and analysis of this paper along with trip characteristics as control variables.

2.6 Hypotheses

Based on the literature review, here I define the full list of hypotheses relating each of them with the main variables of interest in regards to travel related activities as listed above. The following relationships between observed behavior and unobservable personality traits are researched:

	Hypotheses			
H1	Risk attitude	*Decreases* the amount of	a.	Micromoments
			b.	Domains
			c.	Pageviews
			d.	Time
			e.	Lenght
H2	Uncertainty attitude	*Decreases* the amount of	a.	Micromoments
			b.	Domains
			c.	Pageviews
			d.	Time
			e.	Length

3 Methodology

3.1 Introduction

The data utilized for this research has been collected from panelists who participate in a large consumer panel in the Netherlands. The data has two main components, behavioral data and survey data. The methodological section of this paper explains the sampling, data collection and analysis procedures that have been conducted, in the following way.

- Panel and sampling. Sampling based on behavioral data. A sample of respondents has been pulled out of the behavioral dataset based on their online behavior on popular tourism related websites in the Netherlands.
- Survey. An online survey has been administered among the sampled panelists to reveal the “unobservables” from the perspective of the available behavioral data. Namely, these include data points related to trip characteristics associated with respondents’ latest tourist related purchases as well as question regarding their personality.

- Data processing.The full behavioral data of the respondents who successfully completed the survey has been sampled out of the full behavioral panel dataset.
- Categorization. All of the unique websites of the sampled behavioral dataset have been classified into categories. In such a way it was possible to assess whether a certain website was travel related or not.
- Analysis. The aggregated information from the behavioral data has been regressed over the survey data to reveal the impact the unobservable personality attitudes and traits on the tourism related online behavior while controlling for the trip characteristics.

In the following sections I will first focus on explaining the behavioral data and the technology behind it. Afterwards, the 5 major steps of the methodology are explained in details. Then I proceed with the descriptive results of each of the datasets (survey and behavioral). Finally, I report the analysis results.

3.2 Behavioral data & Technology

The behavioral data, also referred to as observational data, reflects the online behavior of the consumers. It consists of records of the interaction made via consumers' digital devices and the Internet.

The behavioral data has been collected via a technology developed and provided for this research by Wakoopa. The company is a provider of a tracking technology. The technology is utilized primarily for market research purposes. Similarly to the market research consumer panels, where panelists enroll to participate in online surveys for incentives, Wakoopa provides its technology to market research consumer panel companies that are interested in tracking the online behavior of their panelists. After enrolling into the panel and giving their consent to be tracked, panelists install an application on their devices i.e. desktop, mobile and tablet. The tracking software collects every interaction of the panelists' devices on the Internet which consists of path or the address the panelists are reaching and the duration of the visit. The software works in different manner over the different operation systems platforms and devices, but the final result consists of recording raw data containing events. Each event has the address the participant accessed, the duration of the interaction and the client requesting the information i.e. browser, app etc.

3.3 Steps

3.3.1 Panel and sampling

Using Wakoopa's panel it was possible to 'pre-screen' relevant respondents for the purposes of the analysis, that could be invited to participate in the survey I have conducted. The respondents that I was looking for, should have been active on tourism related websites and should have conducted a purchase on such websites in the period of January 2015 to June 2016. Herein, I describe the process that was used to reach such subjects. Upon starting of the project the panel used for the research had 6682 active panelists and 7103 active devices. The majority of the panelists were being active only on desktop. See *Appendix #1. Panelists* and *Appendix #2. Devices* for reference.

First, I started looking into tourism related websites in the Netherlands. Initially I used 300 domains to account for the majority of the tourism related internet traffic in the Netherlands according to the internet analytics company SimilarWeb. Afterwards, based on this data I exported the activity of the whole panel over those websites ranging from 01/2015 to 06/2016. The data of this activity was manually analysed looking for the end pages of the payments, also referred to as "confirmation" pages or "thank you" pages. Confirmation pages are the pages where a customer has been redirected after conducting a purchase at a company website. In general, the analysis consists of looking up keywords within the travel domains URLs and marking down the common patterns with the aid of regular expressions (See Mitkov (2005) for discussion). Identified were the patterns of "confirmation" pages and data for the participants visited from the period of 01/03/2016 to 16/06/2016 was exported. Using this information, it was possible to identify a sample composed of 949 respondents which were to be invited into the online survey as described in the second step below.

Given the estimation of the incidence rate provided by the panel supplier, 20%, and the initially desired sample of 500 respondents more respondents were needed. Therefore, an additional random sample of 123 respondents was selected based on whether they were active on the 300 initial domains but without evidence for their purchases from the data.

3.3.2 Survey

The selected 1039 panelists were invited to participate in an online survey, that aimed to reveal their attitudinal and personal characteristics as well as the trip characteristics of their last travel. The fieldwork was conducted during the last week of June 2016. Out of the 1039 invitations sent 872 started the survey, which resulted in 495 completed interviews. (Note to self, move this to a separate cleaning section). The data was further cleaned by accounting for speeders and flatliners. Speeders were panelists that have finished the survey within less than half of the average length of the interview. Flatliners refer to respondents who have answered all of the grid questions in a straight line. Moreover, a few respondents attempted to participate in the survey multiple times and thus, they were excluded from further analysis. Thus, the final dataset consists of 426 observations.

The survey consisted of several parts. First, a screening criteria was used. It accounted for the number of travel related purchases since March 2016. Respondents with no travel related purchases were not allowed to further proceed. Then, subjects who have not purchased neither flight nor accommodation were also excluded from the questionnaire. Business only travellers were not relevant for the analysis, thus they were also screened out, leaving only panelists who have gone on a leisure trip or both on a leisure and business trip. The second part was a demographics section, which consisted of question regarding age, gender and income. The next part of the questionnaire, was in regards to the trip characteristics, which was adapted from Roehl and Fesenmaier (1992) and contained information about moment of booking, destination, planning horizon, information sources used i.e. Internet, advice from friend and relatives, tourist information office, travel agent etc., products purchased online, duration of the trip or number of nights spent away from home, whether the destination was visited before and how many times, number of travel companions, indication whether there was children on the trip and whether subjects visited friends or relatives during their trip.

The next two sections aimed to reveal more about the subjects personality and risk and uncertainty attitude. Items assessing risk and uncertainty originate from Quintal, Lee, and Soutar (2010). Respondents reported their risk attitude on three item scale which the authors adapt from Donthu and Gilliland (1996). Uncertainty attitude has been assessed on four items scale which authors adapt from Yoo and Donthu (2002) and the scale is based on Hofstede (1980) UA items. I have chosen these scales as they are the shortest reliable scales for self-assessing the risk and uncertainty attitudes. Using such scales I can make sure to not tire respondents and keep them engaged. Personality traits related to openness, consciousness, extraversion, agreeableness and neuroticism have been assessed using the short scale from the big five inventory proposed by S. D. Gosling, Rentfrow, and Swann (2003). In addition to the ten item scale an eleventh item was added as XXXX doesn't perform well in the ten item scale. The eleven items were derived from a validated Dutch translation of the BIG5 inventory from Denissen et al. (2008).

3.3.3 Data processing

After having collected and cleaned all the data, I further proceed with data processing and analysis. All of the data processing and the analyzation tasks of this paper were done in R Studio. All of the code, including libraries can be seen in the appendix.

The data processing includes merging down the two main behavioral streams of data that originate from desktop and mobile. Additional to that it is needed to classify and derive all the tourism related domains that will be further used for analysis. This is an important step as it will allow to analysis what the relationship between the online travel behavior and the traveller's attitudinal characteristics are. Categorization of the domains was done using machine learning classification algorithms provided by uClassify.com. The domains from the desktop and mobile behavioral data and all of the mobile apps are further classified using keywords.

Next step is, converting the raw data set onto micromoments. They are also further used in the analysis as I examine the relationship between one's micromoment frequency and subject's risk and uncertainty attitude. Micromoments are essentially user sessions where users were active within a given moment of time. For example, if a panelist access a certain domain and spend five minutes on it and then become inactive for more than 3 minutes, this will result in a factor variable grouping all of the observation within those five minutes of activity. If a travel related website has been visited during the micromoment, a dummy variable is assigned to this moment to indicate this. Furthermore, purchases of travel related products were included in the data based on the initial dataset used for pooling the sample out the panel for the survey. The final dataset includes aggregated information about panelists' activity over two main levels. 1. top-level i.e. total activity 2. low-level micromoment level.

This section includes the data processing tasks done on the behavioral data including descriptive statistics and variable derivation, classification and aggregation on the different levels intended for further analysis.

Starting point of the analysis of the behavioral data include procession the data and rendering it in a format suitable for running the analysis. In its initial form the data has been exported in a format containing the following variables: for the desktop data (sample data, see: *Appendix #3 Desktop data*) and for the mobile data (sample data, see: *Appendix #4 Mobile data*). The next step is adding *"host"* variable to both datasets. The host variable contains domains and the subdomains that are going to be used to for the categorization whether the domains were travel related or not. The next query on the data included identifying all of the micromoments in the data. First, the data has been subsetted on *"panelist_id"* level, sorted by the the timestamp *"used_at"* and assigned into list where each element is the full data for each individual respondent. The sum of *"used_at"* variables and *"active_seconds"*, the duration in seconds respondent spend on the page, were compared to *"used_at"* variable of the next observation. If differences larger than five minutes were found all of the variables prior to this difference were grouped together under a common factor variable.

Table 3.1: Desktop dataset

used_at	host	panelist_id	url	active_seconds	browser_name	mmid	Class_Travel	p

Table 3.2: Mobile dataset

panelist_id	device_id	scheme	url	domain	app_id	app_name	used_at	connection

3.3.4 Categorization

The categorization procedure includes using uClassify machine learning algorithms and also using keywords. The full activity coming from desktop and mobile devices resulted in 194,534 unique domains. The categorization algorithm has been responsible for classifying all domains that have more than ten visits or 47,818 domains in total works as follows. A web-scraper designed for these projects accesses a the collection of domains and collects all of the information on the page, then removes the HTML elements along with the punctuation and renders down the information only to a part that is visible to the website visitors. Then, it passes this information to an application programmable interface (API) that returns the probability of this text being into sixteen different categories including travel.

The full list of categories includes:

NULL

Websites with content with the highest probability to be travel were assigned value of a dummy variable 1 or 0 otherwise.

Due to technical and time contains the API service wasn't able to classify all of the domains. Therefore, the full dataset of unique domains names has also been scanned for keywords. Keywords include travel, tourism, accommodation, hotels, flight etc. (full list of the keywords can be found in the code in the appendix). If any of the keywords appear within the domain name, the respective domain has also been assigned to the list of travel domains. The same keyword approach has been used for classification of the application on mobile devices, where the apps have been classified using their names.

3.3.5 Analysis

The following section contains an explanation of the main techniques performed during the analysis along with their assumptions, followed by the results of the ordinary least squares diagnostic tests of the restricted model. Once the functional form of the restricted model has been selected I proceed with variable selection in order to come up with the final model. Finally, I ran the diagnostics of tests over the final model again.

3.3.5.1 Factor Analysis

Factor analysis is a widely used technique used for for explaining the variance in several variables by smaller set of latent variables. As in the current case it is often used to consolidate several survey variables onto their "underlying" factors in order to reduce the dimensionality of the data. Factor analysis groups variables together, that is, using a large amount of variables one can potentially reduce them to certain factors representing the latent underlying factors representing them by accounting the similar patterns in the variables. The intuition behind the anaysis is as follows. The analysis groups together observed, correlated variables into smaller groups of unobserved (latent) variables (Yong and Pearce 2013).

In this case I use factor analysis to to reduce the seven survey items rearding the risk and uncertainty attitude down to two constructs namely risk and unertainty. Also to reduce the eleven item scale of BIG5 to 5 factors representing each of on the of the five personlity traits.

3.3.5.2 Regression Analysis

For testing the hypotheses of this paper, regression analysis will be utilized. The regression model or ordinary least squares (hereafter OLS) is the "cornerstone of econometrics" (Verbeek 2008). It aims at explaining a variable, y, in terms of another variable, x. In other words, using OLS researchers are able to find how will y vary as x changes, the ultimate goal being to infer the causal effect x has on y. Using such models allows to find relationships between various variables, present the effect the independent variables, xi have on the dependent variable, y in order to be able to make predictions.

The general linear regression models is represented as follows:

$$y = \quad {}_0 + X_1 \; {}_1 + X_2 \; {}_2 + \qquad + X_k \; {}_k +$$

Where: y is the dependent variable

X_1 to X_k are the independent variables, which explain y

${}_0$ is the intercept, indicating the expected value of y when all the independent variables are equal to 0

${}_1$ to ${}_k$ are the coefficients which determine the effect x has on y

is the error term

3.3.5.2.1 Goodness of fit and model selection

The standard measures of fit include the R-squared and the adjusted R-squared, which measures the variance that is explained in the model for the independent variable by the dependent variables. The measure can be interpreted directly. For example if the R-squared is equal to 0.45, it means that the variables included in the model explain 45% of the variation of the independent variable, y. The higher the value, the higher its predictive power. However, it should be noted that adjusted R-squared penalizes for the additional number of parameters. Thus, applying additional variables to the model, I should test if they are jointly significant in order to assess whether they are relevant or not in the model.This is typically applied by using the F-test (Wald test). Using both the R-squared, the overall F-test and applying the F-test to certain variables I can compare best which model fits the data best.

3.3.5.2.2 Akaike information criterion (AIC)

Model selection has been done over Akaike information criterion (AIC) introduced by Akaike (1974). AIC is a metric traditionally used for model selection. It compares the goodness of fit for a number of explanatory variables and penalizes for each additional explanatory variable.

3.3.5.2.3 BLUE Assumptions

There are several assumptions that need to be met when applying OLS explained in the section below. Namely Gauss-Markov assumptions for full ideal conditions for OLS. The model needs to be best linear unbiased estimator" ("BLUE") (Verbeek 2008). It is crucial for the assumptions to be met as to compute unbiased and consistent estimates that explain the variation in the dependent variable. Now, I will go through each assumption: Linear in parameters This implies that the model should have linear parameters, b, however, there can be nonlinearities in the variables, x. This assumption is met as my specified model does not include non-linearities in the parameters.

3.3.5.2.4 Normality

The error term's should follow a normal distribution. In large datasets, however, even if the error term does not follow a normal distribution the regression estimators are 'asymptotically normally distributed', meaning that following non-normal distribution is not crucial as the estimates will still be consistent and unbiased. The Shapiro-Wilk test can is adopted here and results presented below. The test works under null hypothesis: “the sample comes from normally distributed population” Shapiro and Wilk (1965)

3.3.5.2.5 Random sample

The data collection should be done randomly, meaning that the each subject should have the same probability of being selected. In this research, both in the behavioral and survey data collection parts, I can say that subjects were randomly selected for further analysis.

3.3.5.2.6 Multicollinearity

Multicollinearity implies that there is no perfect linear relationship between the independent (explanatory) variables as this can lead to ‘unreliable regression estimated’ (Verbeek 2008). For example, adding both male and female in the analysis would lead to perfect collinearity (as male + female =1) and the estimations would not work. In this example, removing one of the variables would solve the problem, however there can be other variables that are highly correlated. Having multicollinearity would not lead to biased estimates, but to inaccurate estimates. In such a case, excluding variables from the model should be considered. There are no tests that specifically look for multicollinearity, however there are certain indications. For instance, having two variables that are jointly significant (have big F-statistics), but independently are not significant can be a sign of multicollinearity

3.3.5.2.7 Homoscedasticity

Homoscedasticity implies that the variance of the error term should be the same for all values of the independent variables. If this does not hold, there is problem with heteroscedasticity meaning that the estimates of the regression are inconsistent due to inaccuracy of their standard errors, meaning that the t-statistics and thus the significance level of the estimates is not valid anymore. To test for homoscedasticity, I perform the Breusch-Pagan test, which hypothesize that there is constant variance of the error terms.

3.3.5.2.8 Endogeneity

The last assumption is crucial to be met as otherwise the regression estimates are biased and inconsistent. Endogeneity implies that there is correlation between an independent variable and the error term. There are several reasons why this assumption does not fold:

1. The model is misspecified. That is, nonlinearities are missing from the model or interaction effects are not accounted for. To account for that I perform the Ramsey-Reset test. The tests adds fitted values on power and re-estimates the model. The intuition behind it is that if non linear combination of independent variables can explain the dependent variable there are evidence the model is misspecified. The Ramsey-Reset test work under null hypothesis that the model has no important omitted non-linearities (Ramsey 1974)
2. Endogeneity, meaning that we are either missing important variables that explain the variance in the independent variables or we have reverse causality, that means that there can be a loop of causality between the independent and dependent variable.

3.3.5.2.9 Stepwise regression

The idea of stepwise regression has been introduced by Hastie and Pregibon (1992) and further improved by Ripley (2002). It is an iterative function ran over a restricted model and a set of candidate models. Each candidate model consists of a different set of explanatory variables. The function computes iteratively Akaike information criterion (AIC) values for the models comparing them to the best performing models from the previous iteration and based on the performance chooses whether to continue the loop with the new model or remain with the old one. The final output is the best performing model.

3.3.6 Model selection

3.3.6.1 Restricted Model

There are two distinct sources of data resulting in three data sets to be investigated.

1. Dataset from desktop
2. Dataset from mobile

3. Combined dataset from desktop and mobile

Furthermore, there are five dependent variables that are intended to be examinated during this research. Namely,

1. Number of travel micro-moments
2. Number of unique travel domains visited
3. Number of travel pageviews
4. Total time in seconds spent on travel domains
5. Total length in seconds of travel micro-moments

The difference between the last two (*4* and *5*) is as follows. Total time in seconds (*4.*) has been measured once a panelist arrives at a certain website that has been classified as travel, whereas length (*5.*) is the total length (i.e. the difference between the start and end) of micro-moments amongst which a participant visited a travel related domain.

The restricted model of each one of the three datasets takes the following form:

Table 3.3: Restricted model(s)

Model	Dependent variables	Independent variables	Control *	Control2
# 1	Number of travel micro-moments	Risk + Uncertainty + Interaction	Total micro-moments	Total purchases + Days active in the panel
# 2	Number of unique travel domains visited		Total domains	
# 3	Number of travel pageviews		Total pageviews	
# 4	Total time in seconds spent on travel domains		Total time	
# 5	Total length in seconds of travel micro-moments		Total micro-moments length	

Model	Dependent variables	Independent variables	Control *	Control2

* As there was only one purchase detected on mobile devices this term wasn't present in mobile model

Running Ramsey-Reset test per each on of the basic models showed evidence that the functional form of the models is not well specified, thus the models were rejected. Consequently, all of the dependent variables along with their corresponding controls were transformed into logs accounting for additional five models per each dataset. The results from consequent running of Ramsey-Reset test still were still not satisfying, as the test still showed significant results indicating that the functional form of the model is not well specified. Consequently, the corresponding control variables accounting for total activity and purchases were transformed into binary variables with values 1 indicating a participant belongs to a highly active groups of participants (having values above the mean of the sample) and 0 vise versa. The latter transformation accounted for investigating ten more models per each dataset reaching the total number of thirty models to be reviewed. The functional form of the model has been selected based on the dataset containing desktop data due to the fact 1. this dataset is the most complete in terms of number of observation, 2. the participants were selected based on their desktop behavior.

Among all tested models, based on the results from Ramsey-Reset test, I select the following functional form of the model for Desktop, Mobile and Combined datasets. Shapiro-Wilk test for normality also performs best for this functional form, yet the sample is large enough so we can relax the normality assumption.

Table 3.4: Restricted model(s) functional form

Model	Dependent variables	Independent variables	Control *	Control2 **
# 1	log(#) of travel micro-moments	Risk + Uncertainty + Interaction	(D) Total micro-moments	(D) Total purchases + Days active in the panel

Model	Dependent variables	Independent variables	Control [*]	Control2 [**]
# 2	log(#) of unique travel domains visited		(D) Total domains	
# 3	log(#) of travel pageviews		(D) Total pageviews	
# 4	log(#) time in seconds spent on travel domains		(D) Total time	
# 5	log(#) length in seconds of travel micro-moments		(D) Total micro-moments length	

[*] Note: as there was only one purchase detected on mobile devices this term wasn't present in mobile model.

[**] Dummy variables take value 1 is the respondent is activity is above the average activity of the sample and 0 otherwise.

A table with the results from the restricted model along with the model performance metrics can be found in the appendix:

Table 4.1: Desktop data, restricted model Table 4.2: Mobile data, restricted model *Table 4.3: Combined data, restricted model* Table 4.4: Restricted models tests

Due to poor model performance shown in *Table 4.4: Restricted models tests* row **11 to 15**, the model shown in *Table 4.3: Combined data, restricted model* has been rejected. See Appendix *Table 4.4: Restricted models tests* for reference.

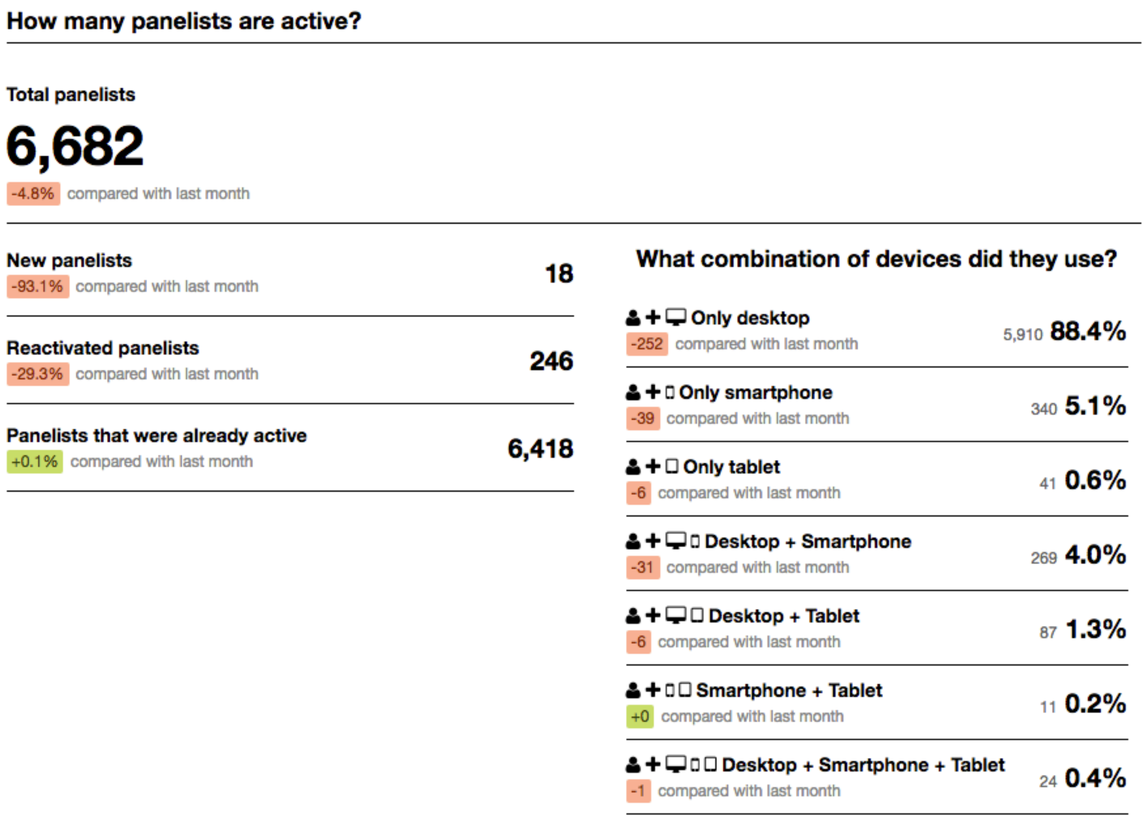
3.3.6.2 Final Model

Based on the selected functional form of the model(s) I proceed with stepwise selection in order to select the final model across desktop, mobile and combined dataset. First, the models are presented then results of the OLS performance tests are shown and discussed.

Results: https://www.dropbox.com/s/x6y4xl3sh4elr3u/Screenshot%202016-09-15%2018.58.17.png?dl=0 https://www.dropbox.com/s/fas9fw1y5jgo1pz/Screenshot%202016-09-15%2018.58.34.png?dl=0 https://www.dropbox.com/s/axaanyn7mfpctbt/Screenshot%202016-09-15%2018.58.45.png?dl=0

Tests: https://www.dropbox.com/s/dsl9e1zhton9w02/Screenshot%202016-09-15%2018.57.21.png?dl=0

4 Appendix



How many panelists are active?

Total panelists

6,682

-4.8% compared with last month

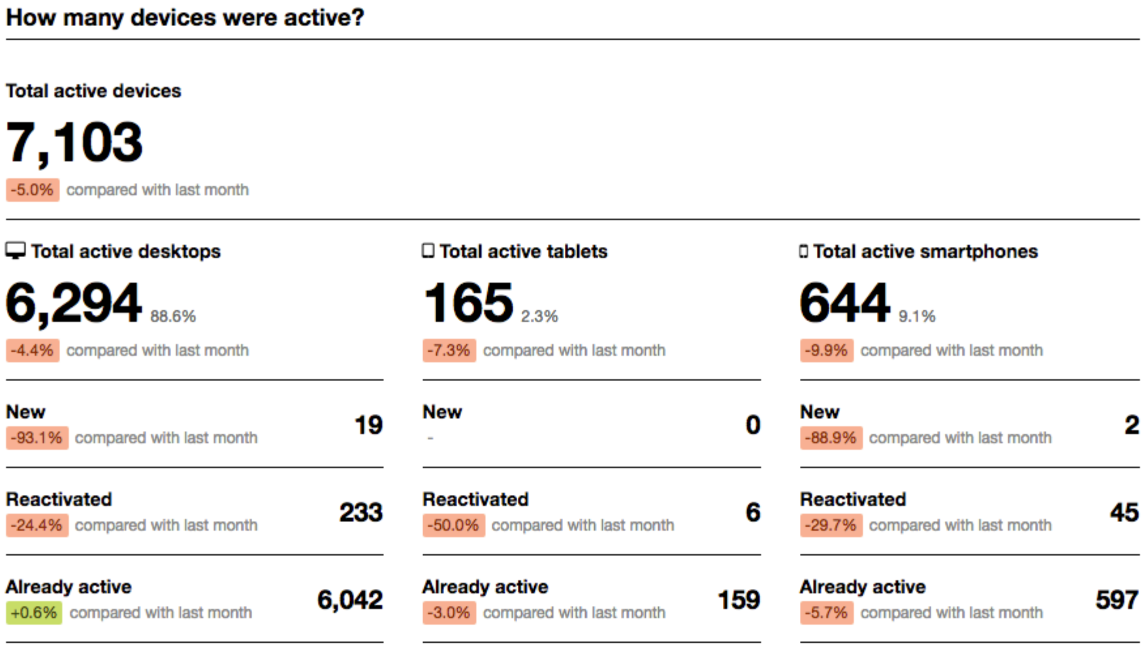
New panelists -93.1% compared with last month	**18**
Reactivated panelists -29.3% compared with last month	**246**
Panelists that were already active +0.1% compared with last month	**6,418**

What combination of devices did they use?

+ Only desktop -252 compared with last month	5,910	**88.4%**
+ Only smartphone -39 compared with last month	340	**5.1%**
+ Only tablet -6 compared with last month	41	**0.6%**
+ Desktop + Smartphone -31 compared with last month	269	**4.0%**
+ Desktop + Tablet -6 compared with last month	87	**1.3%**
+ Smartphone + Tablet +0 compared with last month	11	**0.2%**
+ Desktop + Smartphone + Tablet -1 compared with last month	24	**0.4%**

Figure 4.1: Panelists

```
##                  used_at          host panelist_id           url
## 1  2016-01-01 00:47:06 telegraaf.nl        2112 telegraaf.nl
## 2  2016-01-01 00:48:30 telegraaf.nl        2112 telegraaf.nl
```



How many devices were active?

Total active devices

7,103

-5.0% compared with last month

Total active desktops		Total active tablets		Total active smartphones	
6,294 88.6% -4.4% compared with last month		165 2.3% -7.3% compared with last month		644 9.1% -9.9% compared with last month	
New -93.1% compared with last month	19	New -	0	New -88.9% compared with last month	2
Reactivated -24.4% compared with last month	233	Reactivated -50.0% compared with last month	6	Reactivated -29.7% compared with last month	45
Already active +0.6% compared with last month	6,042	Already active -3.0% compared with last month	159	Already active -5.7% compared with last month	597

Figure 4.2: Devices

```
## 3  2016-01-01 00:51:28 telegraaf.nl        2112 telegraaf.nl
## 4  2016-01-01 00:53:29 telegraaf.nl        2112 telegraaf.nl
## 5  2016-01-01 01:28:52 telegraaf.nl        2112 telegraaf.nl
## 6  2016-01-01 02:01:00 telegraaf.nl        2112 telegraaf.nl
## 7  2016-01-01 02:01:12  twitter.com        2112  twitter.com
## 8  2016-01-01 02:04:07  twitter.com        2112  twitter.com
## 9  2016-01-01 02:05:07  twitter.com        2112  twitter.com
## 10 2016-01-01 02:27:31  twitter.com        2112  twitter.com
##    active_seconds browser_name mmid Class_Travel purchase
## 1              72       safari    1            0       NA
## 2             152       safari    1            0       NA
## 3               0       safari    1            0       NA
## 4               0       safari    1            0       NA
## 5              71       safari    2            0       NA
## 6              11       safari    3            0       NA
## 7             168       safari    3            0       NA
## 8              60       safari    3            0       NA
## 9               0       safari    3            0       NA
## 10              15       safari    4           0       NA
```

Figure 4.3: Desktop data

```
##                      app_name host panelist_id device_id scheme  url domain
## 1  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 2  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 3  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 4  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 5  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 6  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 7  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 8  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 9  ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
## 10 ABN AMRO Mobiel Bankieren <NA>     1008505     10093   <NA> <NA>   <NA>
##                                    app_id             used_at connection
## 1  cb46bcba-7258-4b47-8557-de3ff607b456 2016-02-22 13:43:15       wifi
## 2  cb46bcba-7258-4b47-8557-de3ff607b456 2016-02-03 18:03:34   cellular
## 3  cb46bcba-7258-4b47-8557-de3ff607b456 2016-01-23 15:38:49       wifi
## 4  cb46bcba-7258-4b47-8557-de3ff607b456 2016-01-16 15:11:21       wifi
## 5  cb46bcba-7258-4b47-8557-de3ff607b456 2016-02-03 18:43:44       wifi
## 6  cb46bcba-7258-4b47-8557-de3ff607b456 2016-01-08 14:15:29       wifi
## 7  cb46bcba-7258-4b47-8557-de3ff607b456 2016-01-24 10:20:45       wifi
## 8  cb46bcba-7258-4b47-8557-de3ff607b456 2016-02-03 17:00:56   cellular
## 9  cb46bcba-7258-4b47-8557-de3ff607b456 2016-01-22 21:34:28   cellular
## 10 cb46bcba-7258-4b47-8557-de3ff607b456 2016-01-06 07:24:50       wifi
##    duration mmid Class_Travel TravelApp
## 1        21  799            0         0
## 2        19  551            0         0
## 3        34  392            0         0
## 4        11  283            0         0
## 5        21  552            0         0
## 6        23  139            0         0
## 7          25  396             0          0
## 8           9  549             0          0
## 9          45  381             0          0
## 10         15   98             0          0
```

Figure 4.4: Mobile data

References

Akaike, Hirotugu. 1974. “A New Look at the Statistical Model Identification.” *IEEE Transactions on Automatic Control* 19 (6). Ieee: 716–23.

Ammu, Nrusimham, and Mohd Irfanuddin. 2013. “Big Data Challenges.” *International Journal of Advanced Trends in Computer Science and Engineering* 2 (1). Citeseer: 613–15.

Baumgartner, Hans. 2002. “Toward a Personology of the Consumer.” *Journal of Consumer Research* 29 (2). The Oxford University Press: 286–92.

Bonn, M. a., H. L. Furr, and a. M. Susskind. 1998. “Using the Internet as a Pleasure Travel Planning Tool: an Examination of the Sociodemographic and Behavioral Characteristics Among Internet Users and Nonusers.” doi:10.1177/109634809802200307.

Chatfield, T. 2014. “Google Travel Survey 2014.” *Google Travel Research*. Google/Ipsos MediaCT. https://storage.googleapis.com/think/docs/2014-travelers-road-to-decision_research_studies.pdf.

———. 2016. “The Trouble of Big Data Its Called Recency Bias.” *BBC*. BBC. http://www.bbc.com/future/story/20160605-the-trouble-with-big-data-its-called-the-recency-bias.

Denissen, Jaap JA, Rinie Geenen, Marcel AG Van Aken, Samuel D Gosling, and Jeff Potter. 2008. “Development and Validation of a Dutch Translation of the Big Five Inventory (BFI).” *Journal of Personality Assessment* 90 (2). Taylor & Francis: 152–57.

Donthu, Naveen, and David Gilliland. 1996. “The Infomercial Shopper.” *Journal of Advertising Research* 36 (2). World Advertising Research Center Ltd.: 69–77.

Euromonitor. 2015. “Travel Industry and Online Travel Global Overview.” *Euromon-*

Table 4.1: Desktop data, restricted model

	Dependent variable:				
	log(MM)	log(Domains)	log(PV)	log(Time)	log(Length)
	(1)	(2)	(3)	(4)	(5)
Risk:seek	0.16	0.25	0.42	0.47	0.28
	(0.12)	(0.12)	(0.13)	(0.13)	(0.12)
Uncertainty:seek	0.15	0.26	0.31	0.32	0.21
	(0.11)	(0.11)	(0.12)	(0.12)	(0.11)
Days	0.01	0.01	0.004	0.004	0.01
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
D MM	1.17				
	(0.10)				
D Domains		1.08			
		(0.10)			
D PV			0.88		
			(0.11)		
D Time				0.86	
				(0.11)	
D Length					1.34
					(0.10)
D Purchase	0.50	0.62	0.86	0.87	0.48
	(0.11)	(0.11)	(0.11)	(0.11)	(0.10)
Risk x Uncertainty	0.27	0.32	0.66	0.83	0.48
	(0.30)	(0.31)	(0.32)	(0.32)	(0.29)
Constant	3.12	3.59	5.66	8.94	10.99
	(0.27)	(0.27)	(0.29)	(0.29)	(0.26)
Observations	426	426	426	426	426
R^2	0.36	0.35	0.30	0.31	0.41
Adjusted R^2	0.35	0.34	0.29	0.30	0.41
Residual Std. Error (df = 419)	0.97	0.98	1.02	1.03	0.93
F Statistic (df = 6; 419)	39.89	37.21	29.69	31.07	49.37
Note:					$p<0.1$; $p<0.05$; $p<0.01$

Table 4.2: Mobile data, restricted model

	Dependent variable:				
	log(MM)	log(Domains)	log(PV)	log(Time)	log(Length)
	(1)	(2)	(3)	(4)	(5)
Risk:seek	0.36	0.29	0.42	0.96	1.26
	(0.39)	(0.31)	(0.42)	(0.91)	(0.92)
Uncertainty:seek	0.84	0.66	0.54	1.49	2.26
	(0.38)	(0.30)	(0.41)	(0.89)	(0.89)
Days	0.01	0.01	0.004	0.02	0.02
	(0.003)	(0.002)	(0.003)	(0.01)	(0.01)
D MM	1.50				
	(0.36)				
D Domains		2.28			
		(0.29)			
D PV			2.55		
			(0.40)		
D Time				3.09	
				(0.82)	
D Length					3.13
					(0.82)
Risk x Uncertainty	2.50	2.07	2.07	7.86	8.74
	(1.58)	(1.25)	(1.72)	(3.71)	(3.75)
Constant	0.10	0.03	0.37	1.60	1.58
	(0.46)	(0.36)	(0.50)	(1.08)	(1.08)
Observations	101	101	101	101	101
R^2	0.36	0.55	0.42	0.37	0.36
Adjusted R^2	0.33	0.52	0.38	0.33	0.33
Residual Std. Error (df = 95)	1.50	1.19	1.64	3.52	3.56
F Statistic (df = 5; 95)	10.84	22.94	13.49	10.99	10.72
Note:			$p<0.1$; $p<0.05$; $p<0.01$		

Table 4.3: Combined data, restricted model

	Dependent variable:				
	log(MM)	log(Domains)	log(PV)	log(Time)	log(Length)
	(1)	(2)	(3)	(4)	(5)
Risk:seek	0.20	0.20	0.35	0.38	0.19
	(0.13)	(0.13)	(0.13)	(0.14)	(0.15)
Uncertainty:seek	0.13	0.15	0.23	0.19	0.17
	(0.12)	(0.12)	(0.13)	(0.13)	(0.14)
Days	0.01	0.01	0.01	0.01	0.01
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
D MM	0.91				
	(0.11)				
D Domains		1.01			
		(0.10)			
D PV			0.82		
			(0.11)		
D Time				0.77	
				(0.12)	
D Length					0.98
					(0.12)
D Purchase	0.59	0.62	0.90	0.85	0.51
	(0.11)	(0.11)	(0.12)	(0.12)	(0.13)
Risk x Uncertainty	0.25	0.27	0.58	0.58	0.48
	(0.32)	(0.31)	(0.34)	(0.35)	(0.37)
Constant	2.67	3.16	5.08	8.10	10.01
	(0.29)	(0.28)	(0.30)	(0.31)	(0.32)
Observations	429	429	429	429	429
R^2	0.30	0.34	0.30	0.29	0.29
Adjusted R^2	0.29	0.33	0.29	0.28	0.28
Residual Std. Error (df = 422)	1.03	1.00	1.07	1.13	1.16
F Statistic (df = 6; 422)	30.57	36.53	29.52	29.22	29.27
Note:			p<0.1;	p<0.05;	p<0.01

Table 4.4: Restricted models tests

data_name	Shapiro_Wilk_Stat	Shapiro_Wilk_P.val	Reset_Stat	Reset_P.val	BP_Stat	BP_P.val	Wald.F.Stat	Wald.P.Stat
DESKTOP:REST:log(MM)	100	094	203	013	465	059	4032	0
DESKTOP:REST:log(TD)	100	024	120	030	1048	011	3876	0
DESKTOP:REST:log(TDPV)	099	016	079	045	682	034	2961	0
DESKTOP:REST:log(TT)	098	0001	131	027	1088	009	2844	0
DESKTOP:REST:log(TL)	099	001	197	014	777	026	4872	0
MOBILE:REST:log(MM)	096	0002	089	042	290	072	6112	0
MOBILE:REST:log(TD)	097	001	240	010	337	064	7880	0
MOBILE:REST:log(TDPV)	091	0000	204	014	184	087	4413	0
MOBILE:REST:log(TT)	098	006	003	097	1662	001	19569	0
MOBILE:REST:log(TL)	097	002	045	064	1480	001	18507	0
COMBINED:REST:log(MM)	099	002	1019	0000	1152	007	2853	0
COMBINED:REST:log(TD)	099	0001	868	0002	1167	007	3632	0
COMBINED:REST:log(TDPV)	098	0000	992	0001	1972	003	2882	0
COMBINED:REST:log(TT)	093	0	1293	0000	2666	0002	2368	0
COMBINED:REST:log(TL)	085	0	394	002	3010	0000	3424	0

Note:
Based on the results of Reset-test, all of the models containing COMBINED:REST have been rejected

itor. Euromonitor.

Eurostat. 2016. “Statistics on ICT Use in Tourism.” *Eurostat*. Eurostat. http://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_ICT_use_in_tourism.

Evans, Joel R, and Anil Mathur. 2005. “The Value of Online Surveys.” *Internet Research* 15 (2). Emerald Group Publishing Limited: 195–219.

Fan, Jianqing, Fang Han, and Han Liu. 2014. “Challenges of Big Data Analysis.” *National Science Review* 1 (2). Oxford University Press: 293–314.

Gemünden, Hans Georg. 1985. “Perceived Risk and Information Search. a Systematic Meta-Analysis of the Empirical Evidence.” *International Journal of Research in Marketing* 2 (2). Elsevier: 79–100.

Glass, Gene V. 1976. “Primary, Secondary, and Meta-Analysis of Research.” *Educational Researcher* 5 (10). JSTOR: 3–8.

Goldberg, Lewis R. 1993. “The Structure of Phenotypic Personality Traits.” *American Psychologist* 48 (1). American Psychological Association: 26.

Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann. 2003. “A very brief measure of the Big-Five personality domains.” *Journal of Research in Personality* 37 (6): 504–28. doi:10.1016/S0092-6566(03)00046-1.

Hastie, TJ, and D Pregibon. 1992. “Statistical Models in S, Chapter Generalized Linear Models.” *Wadsworth & Brooks/Cole* 51.

Ho, Chaang-iuan, and Yung-ping Liu. 2005. “An exploratory investigation of web-based tourist information search behavior.” *Asia Pacific Journal of Tourism Research* 10 (4): 351–60. doi:10.1080/10941660500363645.

Hofstede, Geert. 1980. “Motivation, Leadership, and Organization: Do American Theories Apply Abroad?” *Organizational Dynamics* 9 (1). Elsevier: 42–63.

Hox, Joop J, and Hennie R Boeije. 2005. “Data Collection, Primary Vs. Secondary.” *Encyclopedia of Social Measurement* 1: 593–99.

Jani, Dev. 2011. “The influence of personality on tourist information behaviour.” *E-Review of Tourism Research* 9 (3): 88–95.

Jani, Dev, Jun-Ho Jang, and Yeong-Hyeon Hwang. 2014. “Big Five Factors of Personality and Tourists’ Internet Search Behavior.” *Asia Pacific Journal of Tourism*

Research 19 (5): 600–615. doi:10.1080/10941665.2013.773922.

Jun, S H, C A Vogt, and K J MacKay. 2007. “Relationships between Travel Information Search and Travel Product Purchase in Pretrip Contexts.” *Journal of Travel Research* 45 (April): 266–74. doi:10.1177/0047287506295945.

Kim, Dae-Young, Xinran Y Lehto, and Alastair M Morrison. 2007. “Gender Differences in Online Travel Information Search: Implications for Marketing Communications on the Internet.” *Tourism Management* 28 (2). Elsevier: 423–33.

Kim, Heejun, Zheng Xiang, and Daniel R Fesenmaier. 2015. “Use of The Internet for Trip Planning: A Generational Analysis.” *Journal of Travel & Tourism Marketing* 32 (3): 276–89. doi:10.1080/10548408.2014.896765.

Labrinidis, Alexandros, and Hosagrahar V Jagadish. 2012. “Challenges and Opportunities with Big Data.” *Proceedings of the VLDB Endowment* 5 (12). VLDB Endowment: 2032–3.

Lepp, Andrew, and Heather Gibson. 2003. “Tourist roles, perceived risk and international tourism.” *Annals of Tourism Research* 30 (3): 606–24. doi:10.1016/S0160-7383(03)00024-0.

Leung, Rosanna, and Rob Law. 2010. “A Review of Personality Research in the Tourism and Hospitality Context.” *Journal of Travel & Tourism Marketing* 27 (5). Taylor & Francis: 439–59.

Mandrik, Carter A, and Yeqing Bao. 2005. “Exploring the Concept and Measurement of General Risk Aversion.” *Advances in Consumer Research* 32 (32): 531–39.

Mitkov, Ruslan. 2005. *The Oxford Handbook of Computational Linguistics.* Oxford University Press.

Nunan, Daniel, and MariaLaura Di Domenico. 2013. “Market Research and the Ethics of Big Data.” *International Journal of Market Research* 55 (4). The Market Research Society: 2–13.

Pan, Bing, and Daniel R. Fesenmaier. 2006. “Online Information Search. Vacation Planning Process.” *Annals of Tourism Research* 33 (3): 809–32. doi:10.1016/j.annals.2006.03.006.

Quintal, Vanessa Ann, Julie Anne Lee, and Geoffrey N. Soutar. 2010. “Tourists’ information search: The differential impact of risk and uncertainty avoidance." *International Journal of Tourism Research* 12 (4): 321–33. doi:10.1002/jtr.753.

Ramsey, James B. 1974. "Classical Model Selection Through Specification Error Tests." *Frontiers in Econometrics*. Academic Press New York, 13–47.

Ripley, BD. 2002. "Modern Applied Sta Tistics with S." Springer-Verlag, New York.

Roehl, W. S., and D. R. Fesenmaier. 1992. "Risk perceptions and pleasure travel: an exploratory analysis." *Journal of Travel Research* 30 (4): 17–26. doi:10.1177/004728759203000403.

Shapiro, Samuel Sanford, and Martin B Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52 (3/4). JSTOR: 591–611.

Sirakaya, Ercan, and Arch G Woodside. 2005. "Building and Testing Theories of Decision Making by Travellers." *Tourism Management* 26 (6). Elsevier: 815–32.

Stewart, Susan I, and Christine A Vogt. 1999. "A Case-Based Approach to Understanding Vacation Planning." doi:10.1080/014904099273165.

Sweeney, Julian C, Geoffrey N Soutar, and Lester W Johnson. 1999. "The Role of Perceived Risk in the Quality-Value Relationship: A Study in a Retail Environment." *Journal of Retailing* 75 (1). Elsevier: 77–105.

Tole, Alexandru Adrian, and others. 2013. "Big Data Challenges." *Database Syst J* 4 (3): 31–40.

Urbany, Joel E, Peter R Dickson, and William L Wilkie. 1989. "Buyer Uncertainty and Information Search." *Journal of Consumer Research* 16 (2). The Oxford University Press: 208–15.

Verbeek, Marno. 2008. *A Guide to Modern Econometrics*. John Wiley & Sons.

Woodside, Arch G, and Roberta MacDonald. 1994. "General System Framework of Customer Choice Processes of Tourism Services." *Spoilt for Choice*. Kulturverl: Thaur, Germany, 30–59.

Xiang, Zheng, and Ulrike Gretzel. 2010. "Role of Social Media in Online Travel Information Search." *Tourism Management* 31 (2). Elsevier: 179–88.

Xiang, Zheng, Vincent P. Magnini, and Daniel R. Fesenmaier. 2015. "Information technology and consumer behavior in travel and tourism: Insights from travel planning using the internet.” *Journal of Retailing and Consumer Services* 22. Elsevier: 244–49. doi:10.1016/j.jretconser.2014.08.005.

Yong, An Gie, and Sean Pearce. 2013. “A Beginner’s Guide to Factor Analysis: Focusing on Exploratory Factor Analysis.” *Tutorials in Quantitative Methods for Psychology* 9 (2): 79–94.

Yoo, Boonghee, and Naveen Donthu. 2002. “Testing Cross-Cultural Invariance of the Brand Equity Creation Process.” *Journal of Product & Brand Management* 11 (6). MCB UP Ltd: 380–98.