Possibilities and limitations of data-mining in the CzechTourism research

Jan Herget, CzechTourism Institute director, Czech Republic
Jan Fait, Stem/ Mark

1. Introduction

Big data is a rapidly expanding research area spanning the fields of computer science and information management. The term Big Data is used almost anywhere these days; from news articles to professional magazines, from tweets to YouTube videos and blog discussions. The term coined by Roger Magoulas from O'Reilly media in 2005, refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity. [1]

The digital universe is growing 40% a year into the next decade, expanding to include not only the increasing number of people and enterprises doing everything online, but also all the "things" – smart devices – connected to the Internet. By 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes. Like the physical universe, it is diverse – created by everyone using a digital camera, by the more than 2 billion people and millions of enterprises living their lives and doing their work online, and by the millions of sensors and communicating devices sending and receiving data over the Internet. But unlike the physical universe, the digital universe is created and defined by software, a man-made construct. It is defined by software that analyzes this ever-expanding universe of digital data, finding the hidden value and new opportunities to transform and enhance the physical world – keeping the Mars Rover roving, shipping money, or storing the pictures of our loved ones. And it is software that will both create new opportunities and new challenges for us as we try to extract value from the digital universe that we have created.[2]

Companies running social-networking websites conduct "data mining" studies on immense amount of personal data in order to identify hidden consumer preferences and invent efficient marketing strategies. "Geo-location" data allows companies analyze the people's lives. Big Data can offer great insight into many complicated issues, in many instances with remarkable accuracy and timeliness. The quality of business decision-making, government administration, scientific research and much else can potentially be improved by analyzing data in better ways. But critics worry that Big Data may be misused and abused, and that it may give certain players, especially large corporations, new abilities to manipulate consumers or compete unfairly in the marketplace. [3]

As the deluge of data grows, a key question is how to make sense of the raw information. How can researchers use statistical tools and computer technologies to identify meaningful patterns of information? Chris Anderson, the Editor-in-Chief of Wired magazine, ignited a small firestorm in 2008 when he proposed that "the data deluge makes the scientific method obsolete." Anderson argued the provocative case that, in an age of cloud computing and massive datasets, the real challenge is not to come up with new taxonomies or models, but to sift through the data in new ways to find meaningful correlations.[4]

According Fernanda B. Viégas, Research Scientist at the Visual Communications Lab at IBM is the computer-aided visualization of data perhaps one of the best tools for identifying meaningful correlations and developing new models and theories. However Jesper Andersen, a statistician, computer scientist and Co-Founder of Freerisk, warned about the special risks of reaching conclusions from a single body of data. He is saying that it is generally safer to use larger data sets from multiple sources. He thinks that the visualization techniques do not solve the problem.

Within this context, the big data open an interesting space for tourism boards and destination managements to dig in and to uncover the black box of consumer buying behavior. It enables also to implement the innovative research methodology to monitor the domestic and international tourist flow.

2. Big data in tourism

Within the tourism sector, which is arguably one of the most advanced user of Big Data, companies are inventing new services such that give driving directions (MapQuest), provide satellite images (Google Earth) and consumer recommendations (Trip Advisor).

Booking.com B.V., part of the Priceline Group (NASDAQ: PCLN), owns and operates Booking.com™, the world leader in booking accommodation online. Each day, over 650,000 room nights are reserved on Booking.com. More than 2.8 million room nights are booked on our website every week. The Booking.com website and apps attract visitors from both the leisure and business sectors worldwide. [5]

Expedia, Inc. is an online travel company, empowering business and leisure travelers through technology with the tools and information they need to efficiently research, plan, book and experience travel. Expedia, Inc. incorporates dynamic portfolio of travel brands, including our majority-owned subsidiaries that feature the world's broadest supply portfolio — including more than 260,000 hotels in 200 countries, 400 airlines, packages, rental cars, cruises, as well as destination services and activities. Travel suppliers distribute and market products via the traditional desktop offerings, as well as through new distribution channels including mobile and social media, private label business and call centers in order to reach our extensive, global audience, including the approximately 60 million unique visitors that visit the websites on a monthly basis. The portfolio of brands includes Expedia.com®, a full service online travel agency with sites in 31 countries; Hotels.com®, a hotel-only booking service with more than 85 sites worldwide; Hotwire.com®, a discount travel provider with sites in 12 countries; Expedia® Affiliate Network ("EAN"), which powers travel for some of the world's largest travel and non-travel brands, as well as more than 10,000 active affiliates worldwide; Classic Vacations®, a luxury travel specialist; Expedia Local Expert® ("LX"), a destination services and concierge services provider. [6]

Trip Advisor® is the world's largest travel site, enabling travelers to plan and have the perfect trip. Trip Advisor offers advice from real travelers and a wide variety of travel choices and planning features with seamless links to booking tools. Trip Advisor branded sites make up the largest travel community in the world, reaching nearly 260 million unique monthly visitors, and more than 150 million reviews and opinions covering more than 4 million accommodations, restaurants and attractions. The sites operate in 39 countries worldwide, including China under daodao.com. Trip Advisor also includes Trip Advisor for Business, a dedicated division that provides the tourism industry access to millions of monthly Trip Advisor visitors. [7]

According to Trip Advisor the life cycle of the traveler includes following stages: Inspiration > Planning > Travel > Review and repeat. However, very different traveler behavior and data can be observed and leveraged at each spoke of the wheel. The part of the wheel called consideration provides an interesting space for the destination managements to take diverse marketing action on it.

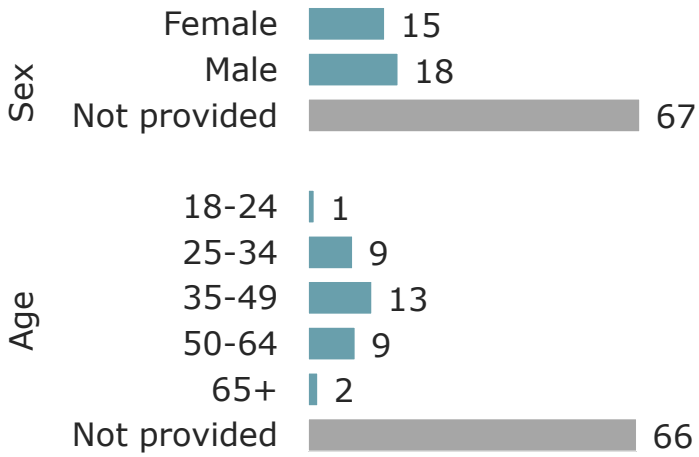
Figure I. The life cycle of the traveler according to Trip advisor

3. Project aim and methodology

The purpose of the project is to explore the possibilities of data-mining from the travel websites for tourism research. Doing this we attempt to investigate some hypotheses that would have been very troublesome and costly to research by traditional survey methods. This research project is solely limited to the users of one global reference website. Therefore without any doubt is clear that drawing conclusions about the state of tourism in a country based on such body of data has several grave methodological flaws that are easily spotted by even a novice sociologist.

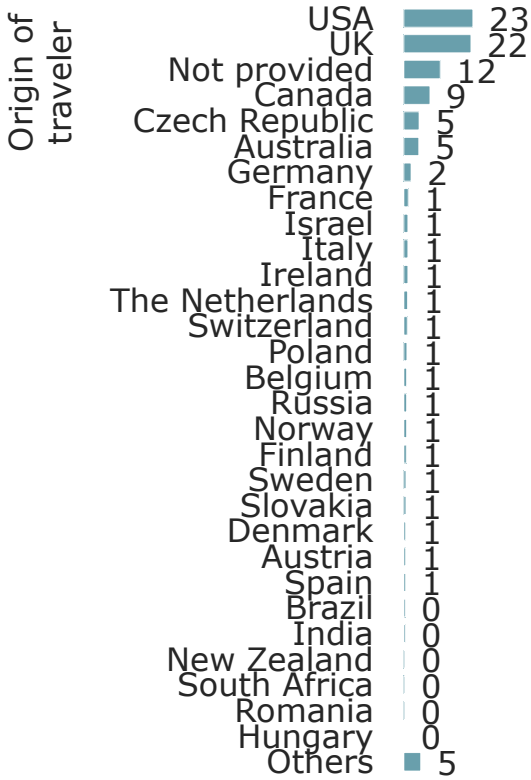
The data-collection has been running from early April – May 2014. Data was collected using a custom made crawler in R programming language. Cases that enter the analysis have been selected from reviews of hotel stays in the Czech Republic at global review website. – From each region we selected up to 20 hotels (not all regions have 20 hotels listed at global review website) and from each a minimum of 10 and a maximum of 100 most recent reviewers. In total, 1059 hotels were searched and 6837 user profiles extracted. The demographic distribution is described in the figures below.

Figure II. The demographic distribution of the sample



BASE: All travelers, n=6837 [chart data in %]

Figure III. The origin of the travelers from the sample



BASE: All travelers, n=6837 [chart data in %]

Each profile contains the user's demographics and all travels the user has recorded and all reviews she/he contributed. The resulting dataset contains 480 600 travels. No other sophisticated sampling methods or quotas were employed on - f.e. nationality of the users. More than a half of the cases have missing demographics

variables and thus constitute a nameless body of travelers, mostly likely less frequent users of the website.

The sample can be viewed as biased due to the following reasons. Not all visitors of the Czech Republic stay in a hotel – hereby we effectively exclude one-day trips and limit the analysis to overnight stays. Not all who stay in a hotel will leave a review – with most users, reviewing is a relatively rare activity. There is therefore a risk of reviewers representing a more active population of users and by the same token probably a population of frequent travelers. Oversampling big cities with lots of popular hotels is likely. We should not draw conclusions on the tourism patterns outside the Czech Republic and make comparisons with either tourists who have not been to the Czech Republic or body of users of the site. Creating a representative sample or allowing ourselves to do the above comparisons would require a selection of a substantially larger number of reviews -millions of cases - and their sampling based on a number of predefined factors. Such strategy was not viable due to budget/time constraints.

4. Research questions

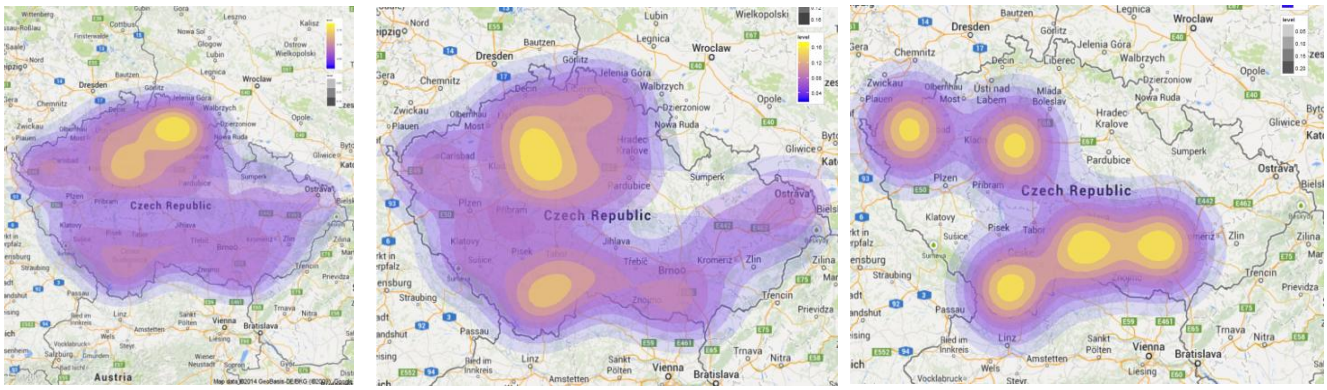
What can data that sits on big travel websites tell us about the tourism patterns in our country? Using the big data from the global review website we try to answer following research questions.

- What locations get visited and how the visits correlate with other locations?
- Do Europeans come to Czech Republic for a city break rather than an adventurous trip?
- Do travelers visit other countries together with the Czech Republic?
- Are people who travel outside of the capital qualitatively different or not?

What locations get visited and how the visits correlate with other locations?

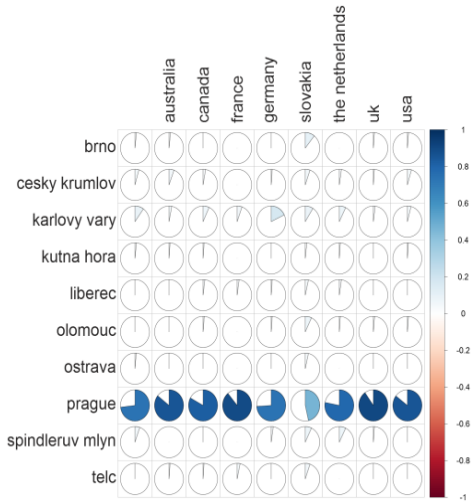
Figure IV. Visualization of the tourist flow in the Czech Republic based on the reviews

a. Base Total **b. Base North America** **c. Base South America**



Outside Prague, the most popular locations were the Krkonose mountains – mainly visited by the Dutch in winter, Karlovy Vary – mostly by Germans (very few Russian users in the sample) and the southern parts of Bohemia and Moravia with multiple locations. The least visited was the city of Pilsen and its immediate surroundings together with Vysocina region and the Jeseniky region.

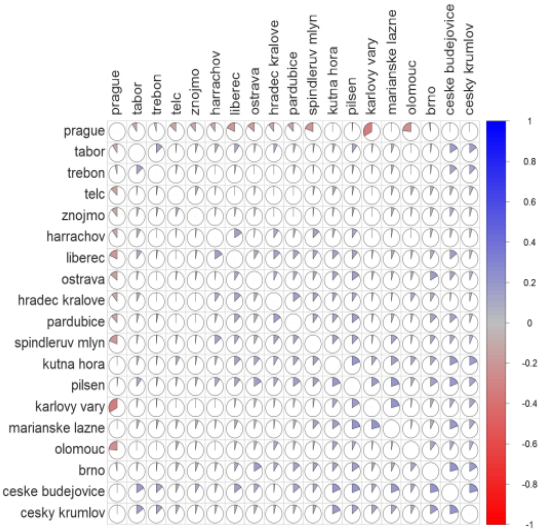
Figure V. Popular locations by the origin of traveler



BASE: All travelers except Czechs, n=6500

The Dutch are more likely to visit the mountains and other locations. They are the only non-neighboring nation with a relatively low share of visitors in Prague. Cesky Krumlov is popular with the visitors from the USA, Australia and Canada.

Figure VI. The correlation plot among the Czech destinations.

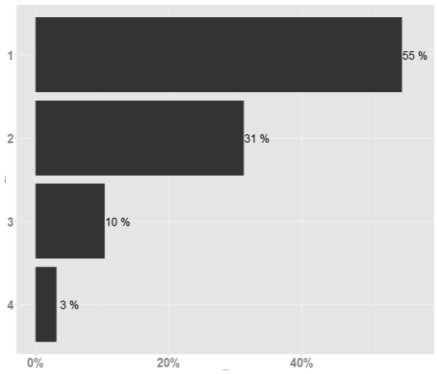


Blue = positive correlation.
BASE: All travelers except Czechs, n=6500

Prague as a destination is strongly isolated. This points to the absence of other major hubs of tourism or Prague‘s locking charm. Karlovy Vary as a „second best“ is still negatively correlated with Prague. It means that the visits of Prague and Karlovy Vary are often mutually exclusive. Even highlights such as Karlstejn are more often visited by tourists who are based in Pilsen (80km away vs. Prague 15km)

Do travelers visit other countries together with the Czech Republic?

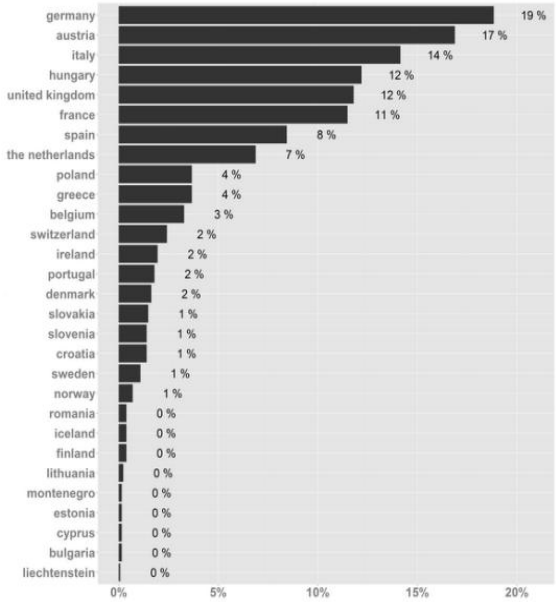
Figure VII. Visiting neighboring countries while in the Czech Republic



BASE: All travelers except Czechs and neighboring countries, n=2726

The plot shows a number of neighboring countries a visitor to Czech Republic has also visited. The visitors from neighboring countries and the domestic travelers were excluded. 55% of the travelers visited one neighboring country they do not originate from, 31% has been to two neighboring countries.

Figure VIII. Visiting European countries in the same year as the stay in Czech Republic

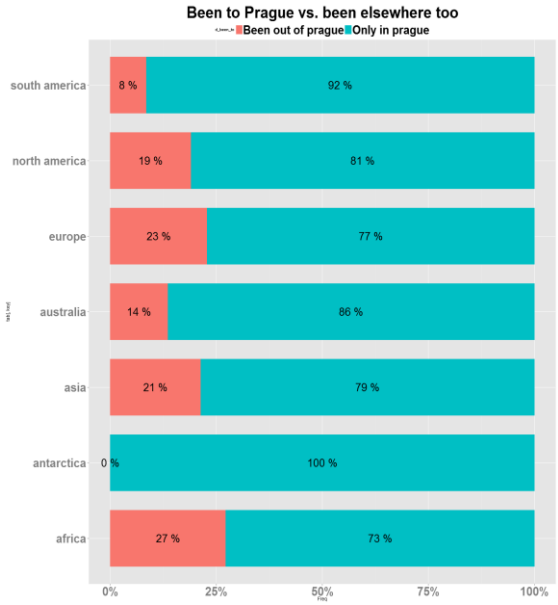


BASE: Travelers from North America, n=1279

There are 1279 travelers originating from North America in the sample that were linked to the countries they visited in 2013. 19% of them apart from visiting the Czech Republic also went to Germany, 17% to Austria and 14% to Italy. Only 4% went to Poland. Certain number of them may have visited these countries in two takes with a returning flight home. The assumption is that when the visits took place in a span of two months and less coming back home is unlikely.

Do Europeans come to Czech Republic for the city break and the rest of the world does the touring through the country?

Figure IX. Been only to Prague vs. been elsewhere too



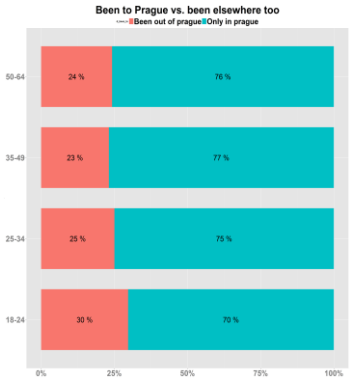
BASE: All travelers except Czechs and neighbors of the Czech Republic

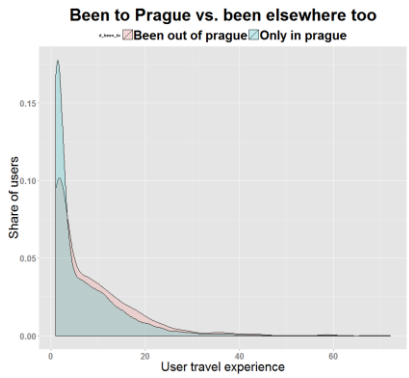
In total, 19% travelers also visit destinations outside of Prague. We cannot be certain that the others have not done so, but they did not record it in the travel diary online. Once neighboring countries are excluded, both travelers from Europe and North America are equally likely to visit other destinations than Prague. The hypothesis that only Europeans arrive for a city break cannot be confirmed from this sample.

Are people who travel outside of the capital qualitatively different or not?

Figure X. Been only to Prague vs. been elsewhere too

a. By the age **b. by the experience**





Those who made it out of Prague are younger and more experienced. As seen in the figure X.a The

age distribution is skewed towards the younger folks. The chart X. b shows the experience distribution. Experience is measure by the percentage of countries in the world that the user has visited. It is obvious that those who traveled outside of Prague are more experienced travelers – their distribution is less skewed towards 0.

5. Results and Discussion

Mining data from travel servers has a potential for answering questions we tend to research by traditional surveys at a fraction of the cost provided that a sound methodological setup has taken place. Validation of such methodology would be a difficult task unless we can compare the results with representative surveys.

Once the data-mining tools have been created, we can use them to collect more „interviews“ at a minimal cost. This method has a potential of bringing more accurate data on very specific groups of tourists which are hard to reach or which are underrepresented in the traditional surveys – f.e. tourism patterns among people from Dallas are nearly impossible to research – not in this case. For the sake of credibility, we should not concentrate so much on the „big“ in the data, but rather on the question of accuracy. Optimally, the sample we extract from the sites should mimic the tourist population inferred from traditional surveys. The sample size is secondary.

The important condition we have not discussed yet is the legal aspect. The data shall be extracted and presented only with a written permission of their proprietors.

6. References & Figures

Figures

- Figure I. The life cycle of the traveler according to Trip advisor
- Figure II. The demographic distribution of the sample
- Figure III. The origin of the travelers from the sample
- Figure IV. Visualization of the tourist flow in the Czech Republic based on the reviews
    - a. base total
    - b. base North America
    - c. base South America
- Figure V. Popular locations by the origin of traveler
- Figure VI. The correlation plot among the Czech destinations.
- Figure VII. Visiting neighboring countries while in the Czech Republic
- Figure VIII. Visiting European countries in the same year as the stay in Czech Republic
- Figure IX. Been only to Prague vs. been elsewhere too
- Figure X. Been only to Prague vs. been elsewhere too
    - a. by age
    - b. by the experience

References

[1] Gali Halevi, MLS, PhD, Dr. Henk Moed. The Evolution of Big Data as a Research and Scientific Topic, Research Trends Issue 30 September 2012, 3-6

[2] Vernon Turner, IDC, Data Growth, Business Opportunities, and the IT Imperatives, http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm, 2014

[3] David Bollier, The Promise and Peril of Big Data, The Aspen Institute, 2010,

[4] Chris Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, Wired, at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory, June 23, 2008

[5] Booking.com, About Booking, http://www.booking.com/content/about.en-gb.html?dcid=1&lang=en-gb&sid=ad5ae5a0627537f2d04befdd7cec8858 **,** Retrieved: May 2014.

[6] Expedia.com, About Expedia, http://ir.expediainc.com/events.cfm Retrieved: May 2014.

[7] Tripadvisor.com, About Tripadvisor, http://www.tripadvisor.com/PressCenter-c6-About_Us.html **,** Retrieved: May 2014.http://www.microsoft.com/en-us/news/press/2011/oct11/10-13skypepr.aspx, 10.13.2011