

Virtual vs. real visits: an analysis of three cities through Wikipedia page views and tourism data

Serena Signorelli (serena.signorelli@unibg.it)¹, Fernando Reis (Fernando.REIS@ec.europa.eu)² and Silvia Biffignandi (silvia.biffignandi@unibg.it)¹

Keywords: big data, Wikipedia page views, tourism, official statistics.

1. INTRODUCTION

Just as big data are becoming one of the main topics in the scientific world, official statistics bodies are trying to assess the potential of the use of these new sources of data. Eurostat, as the statistical office of the European Union, set up a task force which is performing some pilot studies on different big data sources.

One of these sources is represented by Wikipedia. The aim of the paper is the analysis of virtual visits (represented by Wikipedia page views) compared to real visits (represented by tourism visits). This will allow to evaluate the use of Wikipedia page views as a source of information for the identification of factors that drive tourism to an area and whether it is possible to predict tourism flows using these data. Assessing the potential of building some lead indicators is another issue to be explored.

2. METHODS

The analysis is performed on three cities (Barcelona Bruges and Vienna), considering all the points of interest of the area (culture, heritage, athletic, nature, leisure, etc.) on a five-year period (January 2012 – December 2016). Starting from Wikidata, the linked data source of the Wikimedia Foundation, we identify all the points of interest of the cities, the related Wikipedia articles and get their monthly page views.

The attention is then devoted to the study of these pages' number of visualizations, with the construction of heat maps that graphically highlight the main points of interest of the area and show which types of points of interest are more popular.

The research continues with the analysis of time series of the page views combined with official tourism data, to identify possible factors that drive tourism to that specific area. We will evaluate the possibility of predicting tourists' flows with them.

In this study, we use two different sources of data: Wikipedia page views of all the articles with geo-coordinates that relate to the three cities (the big data source) and official tourism data, i.e. arrivals (number of passengers) and overnight stays (number of bookings).

The languages considered in our study are 31: the 24 official languages of the European Union plus other 7 languages that were in the top Wikipedia rankings in terms of number of page views.

2.1. Big data source

Big data sources that are potentially relevant for official statistics are those which cover large portions of populations of interest and which can potentially provide answers to

¹ University of Bergamo

² EUROSTAT

questions raised by policy makers and the civil society. Some work on assessing the quality of this big data source has been done in the paper by Reis et al. (2016), where they tried to take the principles of three different statistical quality frameworks (from UNECE, Eurostat and AAPOR) and apply them to this specific data source.

Wikipedia page views represent the source of data we use in our analysis, but they are not immediately available and they require a bit of work to getting them. Page view statistics is a tool available for Wikipedia pages, which allows to know how many people visited an article during a given period (usually hourly counts).

We first must select the articles we want to include in the study. To do so, we decided not to start directly on the selection of articles on Wikipedia, but to use the Wikimedia Foundation linked data source, Wikidata.

In our study, we want to get all Wikidata items with geo-coordinates that fall into the area of the cities. This is made possible through the Wikidata Query Service, an interface that allows to query its database using the SPARQL language. After having obtained the list of Wikidata items around the cities geo-coordinates, we must filter them to consider only those items that fall into the Urban Audit shapefiles provided by Eurostat. Then, we need the list of articles in each Wikidata item in the chosen 31 languages. Some ad hoc built R functions and some filtering methods allow this procedure and the result is available in Table 1.

Table 1. Number of Wikipedia articles in each city at different Urban Audit levels

<i>City</i>	<i>Urban Audit level</i>	<i>Number of points of interest</i>	<i>Number of Wikipedia articles</i>
Barcelona	C	1093	3996
	K	1450	5256
Bruges	C	561	868
	F	649	1127
Vienna	C	2663	6315

After having defined the final list of articles (considering also redirect articles), we extract the Wikipedia monthly page views through the Page views statistics, a tool available for Wikipedia pages. We consider the monthly page views from January 2012 to December 2016.

2.2. Official statistics source

Official tourism data are available on the Web for the three cities:

- Barcelona: data are available from the Website of the municipality of Barcelona (<http://ajuntament.barcelona.cat/en/>). They concern monthly arrivals and overnight stays by country of residence.
- Bruges: data are provided by the Flemish tourism Website (<http://www.toerismevlaanderen.be/>). They concern monthly arrivals and overnight stays by country of origin.
- Vienna: data provided by Statistics Austria (http://www.statistik.at/web_en/statistics/index.html), they concern monthly arrivals and overnight stays by country of origin.

As we will extract Wikipedia page views from January 2012 to December 2016, the same period has been used as a reference for official tourism data.

3. RESULTS

Once we extracted the Wikipedia page views and we got the data from the tourism offices, it is time to start analysing them. We decided first to focus on the big data source as it is, because we think it is important to identify what people are interested in when searching information on a city. After that, we try to classify the time series of the Wikipedia page views, to identify the factors that drive tourism to an area, and we try to combine the two datasets to identify common patterns.

3.1. Points of interest in cities

First, we decided to build some interactive maps to visualize how the points of interest are distributed across the three cities and their importance according to their number of page views. We only show the map for Barcelona (Urban Audit Level C) and leave the other maps to the curiosity of the reader³. In Figure 1 a screenshot of the map is available. Each circle represents a Wikidata item. The size of the circle and the intensity of its colour are according to the sum of page views of the related Wikipedia articles that refer to that specific Wikidata item, considering the 31 languages of our analysis.

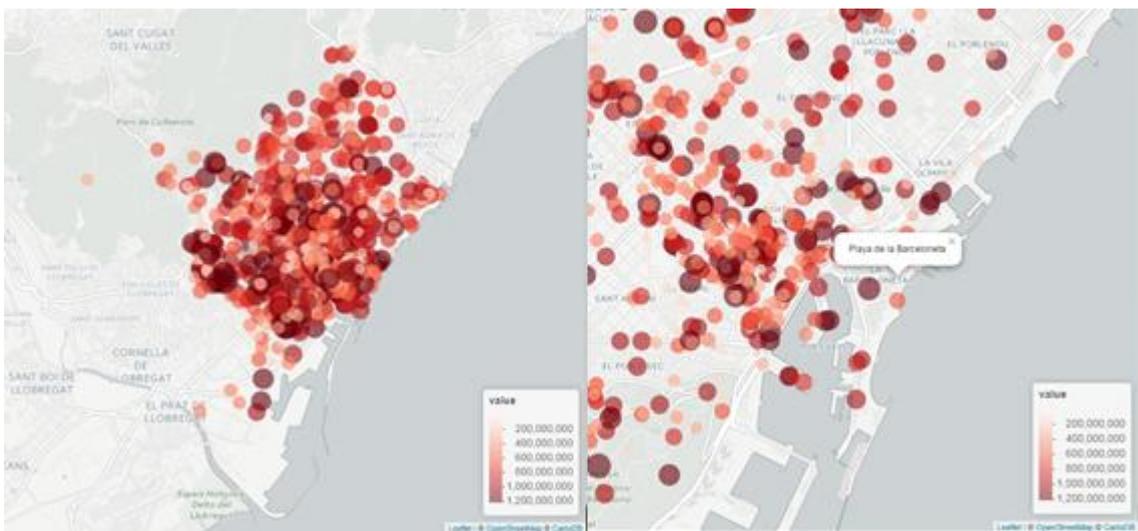


Figure 1. Map of Barcelona (Urban Audit Level C) points of interest according to Wikipedia page views

Other kind of maps and plots are available; they map and show time series of the points of interest per top languages and per category (see next section).

3.2. Categories (classification)

After this first visual analysis on the page views, we need to classify the time series according to the topic of each Wikipedia article. We decided to build a classification that

³ http://serenasignorelli.altervista.org/Barcelona_C/Barcelona_C.html
http://serenasignorelli.altervista.org/Barcelona_K/Barcelona_K.html
http://serenasignorelli.altervista.org/Bruges_C/Bruges_C.html
http://serenasignorelli.altervista.org/Bruges_F/Bruges_F.html
http://serenasignorelli.altervista.org/Vienna_C/Vienna_C.html

could reflect the real content of the article. Topic modelling represents the area we are dealing with, and we use an algorithm called Latent Dirichlet Allocation (LDA).

After the classification of the Wikipedia articles, we group them by Wikidata item (or point of interest in the city). This approach allows us in the end to classify 95.5% of the total number of Barcelona Wikidata items, 89.7% of Bruges items and 79.2% of Vienna items. We identified 14 categories for Barcelona, 11 for Bruges and 23 for Vienna, plus an 'unclassified' residual category for each of the three cities. This classification is fundamental in the next step of the study, the combination of the big data source with official tourism data, which is described in the next paragraph.

3.3. Combined data sources

In this final phase, the goal is trying to predict tourism flows using a big data source.

Before using the time series of the Wikipedia categories as regressors, we pre-process them to consider some phenomenon that could create strange issues. These are represented by edits in Wikipedia articles (as contributors make some changes and keep opening the article to check if it is ok) or even the phenomenon of edit warring occurring when editors who disagree about the content of a page repeatedly override each other's contributions. Plotting the time series, we noticed some strange peaks, so it is necessary to detect the edits in the time series and adjust the page views number accordingly.

The model that we thought was suitable, at least for a first analysis, is an Autoregressive Integrated Moving Average with Explanatory Variables (ARIMAX). We first run the ARIMA model on number of passengers and number of bookings, considering the categories time series as external regressors. This allows to identify which are the significant parameters (categories) and to look at their sign. Following a general-to-specific approach, these will be the only categories that will be considered from now on. Furthermore, as the negative sign of some parameters seems a little bit ambiguous to interpret, we decided to split the significant categories in two, taking aside the articles that collected the highest number of page views and considering it as a unique category. We want to verify if this significant effect is due just to one single article or not. We run again the model with this limited number of parameters and we can identify which are the categories of virtual visits that drive real visits to the cities.

CONCLUSIONS

The analysis seems promising, but a lot of work still must be done and some issues must be considered. We did not consider any multicollinearity effects between page views series. An interesting issue would also be to consider some lags in the Wikipedia series, to see whether it is possible to identify how much time before the trip the tourists look for tourism information from Wikipedia. It would also be interesting to split the tourism flow into residents' tourism and foreigners' tourism, to identify if something changes in the patterns.

REFERENCES

- [1] F. Reis, L. di Consiglio, B. Kovachev, A. Wirthmann, M. Skaliotis, Comparative assessment of three quality frameworks for statistics derived from big data: the cases of Wikipedia page views and Automatic Identification Systems, European Conference on Quality in Official Statistics (Q2016), Madrid, 31 May-3 June 2016