Wikipedia as a Big Data source for Tourism

Global Forum on Tourism Statistics
Venice, November 23rd 2016

Serena Signorelli[1,2], Fernando Reis[2], Silvia Biffignandi[1]
[1]University of Bergamo
[2]EUROSTAT

Wikipedia as a Big Data source for Tourism

- Introduction on the study
- Big data source
- Official Statistics source
- Methodology and results
- Conclusions and future research

Introduction on the study

Source of data

Wikipedia page views, starting from Wikidata (linked data source of the Wikimedia Foundation)

Official Tourism data, composed by arrivals and overnight stays

Aim of the study

evaluate the use of these data 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.

Context

January 2012 to December 2015

three cities: Barcelona, Bruges, Vienna

Big data source

Description of data

Wikipedia page views: how many people have visited an article during a given time period

Languages considered: 31

- **24 official languages of the European Union**
- **7 languages in the top Wikipedia rankings in no. of page views**

Summary of data

City	Number of Wikidata items	Number of Wikipedia articles
Barcelona	1093	3996
Bruges	561	868
Vienna	2663	6315

Official Statistics source

Description of data

Official Tourism data:

- **arrivals (number of passengers)**
- **overnight stays (number of bookings)**

Collection of data

- **Available on the municipality of Barcelona website**
- **Available on the Flemish tourism website**
- **Available on Statistics Austria website**
  - **+ missing months provided by Statistics Austria**

Methodology and results

Type of results

Only using the big data source

- **Maps with points of interest in cities**
- **Charts with top six languages time series**
- **Maps with top six languages points of interest**
- **Classification of points of interest**

Using the two combined data sources

- **Modelling Official Tourism data using big data series as regressors**

Classification of Wikipedia articles

Method

Topic modelling using Latent Dirichlet Allocation (LDA) algorithm

For unclassified/uncertain categories, string match between some identified keywords and title of the article

Classified Wikipedia articles → classified Wikidata items

Quality of classification

95.5% of the total number of Barcelona Wikidata items

89.7% of the total number of Bruges Wikidata items

79.2% of the total number of Vienna Wikidata items

Classification of Wikipedia items

Barcelona – 14 categories

• Sport	64.3%
• Sagrada Familia	14.8%
• Buildings	9.0%
• Public transport	2.8%
• Streets and districts	2.4%

Bruges – 11 categories

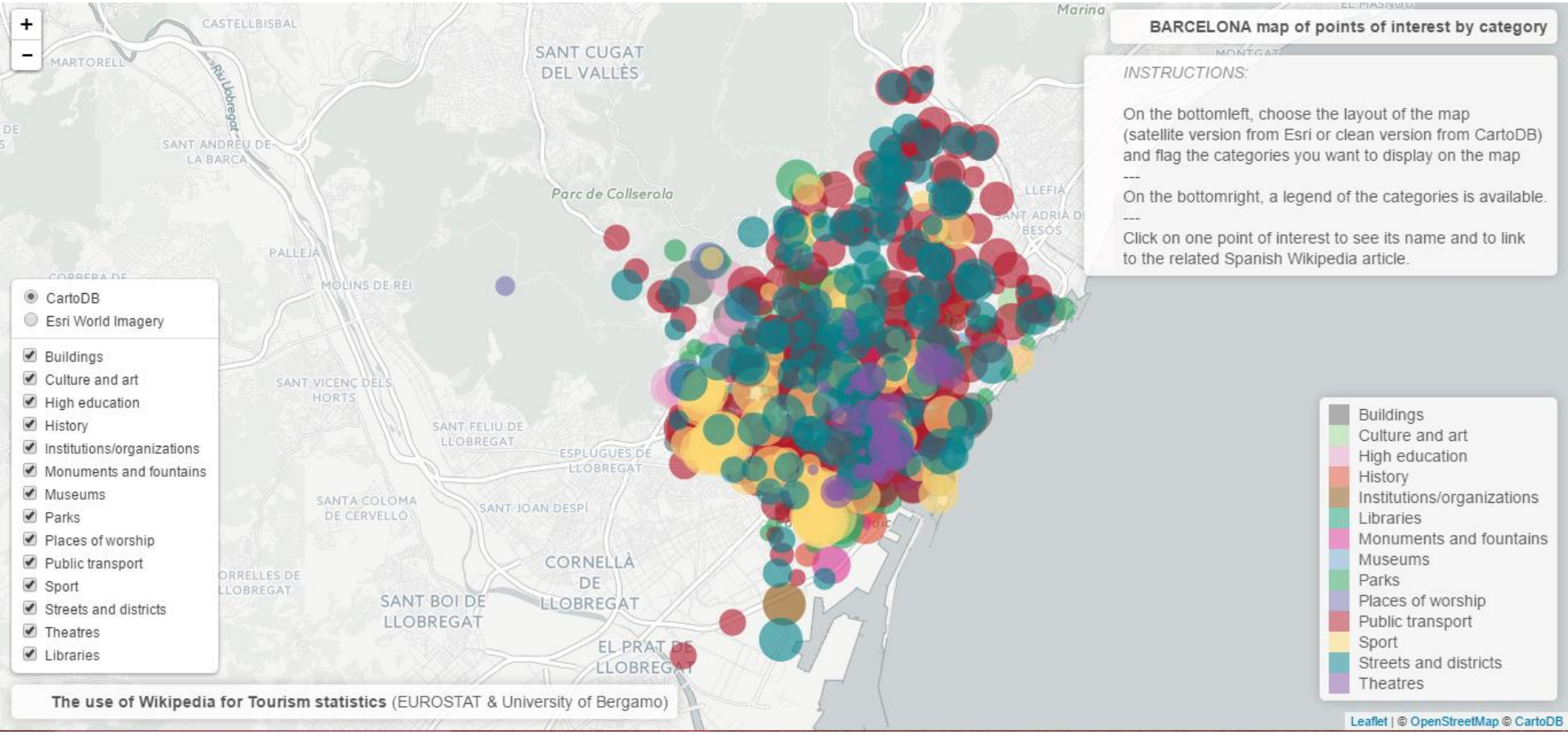
• Sport	40.1%
• Places of worship	23.3%
• Districts	13.7%
• Buildings	8.7%
• Streets and streams	6.8%

Vienna – 23 categories

• History	38.4%
• Institutions/organizations	22.8%
• Buildings	7.9%
• Museums	6.7%
• Sport	3.7%

European Commission

Barcelona classified points of interest



Combined data sources

Type of model

Autoregressive Integrated Moving Average with Explanatory Variables (ARIMAX)

Data

We try to model number of passengers (arrivals) and number of bookings (overnight stays) using the previously classified series as external regressors

Procedure

We run the models and identified the significant parameters:

- No doubts on interpretation of positive parameters
- Uncertainty about the negative ones

Further analysis into each significant category with negative parameter

- Identify of the Wikipedia article that collects the highest number of page views
- Remove that article from the category and put into the model separately

Combined data sources

Barcelona

$$\text{Number of passengers} = \text{Number of } passengers_{t-1} + 0.362687\ \text{parks} - 0.482018\ \text{sport} - 0.372502\ \text{Gran Teatro del Liceo} + \epsilon_t$$

$$\text{Number of bookings} = \text{Number of bookings}_{t-1} + 0.451294\ \text{parks} - 0.388938\ \text{sport} + \epsilon_t$$

Bruges

$$\text{Number of passengers} = -0.7456239\ \text{Number of passengers}_{t-1} + 0.1976131\ \text{buildings} + \epsilon_t$$

$$\text{Number of bookings} = -0.8040564\ \text{Number of bookings}_{t-1} + 0.3175176\ \text{buildings} + 0.5243503\ \text{companies} - 0.5548691\ \text{districts} - 0.2993752\ \text{Zeebrugge} - 0.2037880\ \text{Belfort Van Brugge} + \epsilon_t$$

Combined data sources

Vienna

Number of passengers
$= \text{Number of passengers}_{t-1} + 0.328203$ places of worship
$+ 0.289614$ mountains $+ 0.240543$ cemeteries
$- 0.362990$ institutions organizations $- 0.316615$ buildings
$- 0.277552$ Schloss Schonbrunn $- 0.301240$ Universität Wien
$- 0.245028$ Österreich $-$ Ungarn $+ \epsilon_t$

Number of bookings
$= \text{Number of bookings}_{t-1} + 0.344185$ places of worship
$+ 0.219496$ companies $+ 0.638848$ museums $+ 0.322709$ theatres
$- 0.529876$ embassies $- 0.936683$ rivers and parks $- 0.432915$ buildings
$- 0.335415$ high education $- 0.409401$ Schoss Schonbrunn
$- 0.271926$ Universität Wien $- 0.571975$ Östrerreich $-$ Ungarn $+ \epsilon_t$

Conclusions and future research

Future developments

Edit history of a Wikipedia article

Multicollinearity effects between page views series

Lags in the Wikipedia series

Split the tourism flow into residents' tourism and foreigners' tourism

Thank you for your attention

Serena Signorelli *serena.signorelli@unibg.it*

Fernando Reis *Fernando.REIS@ec.europa.eu*

Silvia Biffignandi *silvia.biffignandi@unibg.it*

Barcelona

Bruges

Vienna