



UNIVERSITY OF BERGAMO

School of Doctoral Studies

Doctoral Degree in Analytics for Economics and Business

XXIX Cycle

SSD: SECS-S/03 – Economic Statistics

**THE USE OF BIG DATA  
IN OFFICIAL STATISTICS**

Advisor:

Prof. Silvia Biffignandi

Doctoral Thesis

Serena **SIGNORELLI**

Student ID 52978

Academic year 2015/16



# Summary

Introduction.....	7
1 Applications with Big Data: a literature review .....	11
1.1 Introduction on Big Data.....	11
1.2 The use of big data for operational purposes .....	13
1.2.1 Tourism sector.....	14
1.2.2 Marketing .....	15
1.2.3 Social science .....	16
1.2.3.1 Analysis of human personality .....	17
1.2.3.2 Study of traffic conditions.....	18
1.2.3.3 Medical field .....	19
1.3 The use of big data for statistical purposes .....	19
1.3.1 Mobile communication .....	22
1.3.2 World Wide Web .....	23
1.3.3 Sensors .....	25
1.3.4 Transactions of process generated data .....	26
1.3.5 Crowdsourcing .....	27
1.4 Conclusions .....	27
2 From Big Data to Information: Statistical Issues Through Case Studies.....	29
2.1 Introduction on Big Data.....	29
2.2 Statistical and quality issues.....	30
2.3 Some empirical studies.....	34
2.4 Case study .....	38
2.4.1 Outflow maps .....	40
2.4.1.1 ISTAT dataset .....	40
2.4.1.2 Telecom Italia dataset.....	43
2.4.2.1 ISTAT dataset .....	44

2.4.2.2	Telecom Italia dataset.....	48
2.4.2	Method of comparison of the two datasets.....	49
2.4.3	Results.....	52
2.5	Conclusions.....	55
2.6	Acknowledgements.....	55
3	Case study: Phone calls and mobility in the Lombardy region.....	57
3.1	Introduction.....	57
3.2	Description of the datasets.....	58
3.3	Preliminary analysis.....	59
3.3.1	Correlations.....	60
3.3.2	Linear regressions.....	64
3.3.3	Non-linear regressions.....	65
3.3.4	Ranks.....	68
3.4	Conclusions.....	71
4	What attracts tourists while planning for a journey? An analysis of three cities through Wikipedia page views.....	73
4.1	Introduction.....	74
4.2	Data collection and preprocessing.....	76
4.2.1	Big data source.....	77
4.2.1.1	Wikidata items and Wikipedia articles.....	77
4.2.2	Official Statistics source.....	83
4.3	Methodology and results.....	84
4.3.1	Big data source.....	84
4.3.1.1	Points of interest in cities.....	84
4.3.1.2	First attempt of classification using Wikidata properties.....	93
4.3.1.3	Second attempt of classification using clustering.....	97
4.3.1.3.1	Hierarchical clustering with Dynamic Time Warping on categories.....	97
4.3.1.3.2	Hierarchical clustering with Dynamic Time Warping on languages.....	104

4.3.1.3.3	Cluster analysis on items.....	107
4.3.1.4	Third attempt of categorization using LDA and string match.....	120
4.3.2	Combined data sources.....	126
4.3.2.1	Initial models.....	127
4.3.2.2	Inspection on significant series .....	130
4.3.2.3	Final models .....	134
4.4	Conclusions and future research .....	139
4.5	Appendix.....	140
4.5.1	Quality of Wikidata classification.....	140
4.5.2	Quality of official points from municipality of Vienna.....	141
4.5.3	Change in the number of items.....	148
4.5.4	Check for cities boundaries .....	153
4.6	Disclaimer .....	155
5	Conclusions .....	157
	Bibliography.....	159



# INTRODUCTION

Big data represents a kind of data that is becoming more and more popular these days. Official Statistics institutions are trying to deal with this new source to identify possible uses, as a support to existing sources or to produce new statistics.

This new data presents some advantages, as well as disadvantages. As a benefit, big data is already available for other purposes, so it should (at least theoretically) come at no cost. Moreover, it represents a timely updated source, that allows to have information nearly in real time. On the other side, the coverage of the data is not complete, as a big data source may represent just one part of the population. Furthermore, as data is not designed specifically for a statistical purpose, it is necessary to perform some pre-processing and manipulation before being able to use it.

To better understand how an Official Statistics institution deals with big data, I had the opportunity to spend six months as a trainee in the Eurostat Big Data Task Force.

This thesis was developed with the aim to bring some new experimental studies into the big data literature. We wanted to state if this particular kind of data could be used in Official Statistics, considering two different meanings:

1. combining a traditional data source with a big data source to verify the potential of the latter to replicate official results;
2. analysing a big data source per se and then trying to combine with an Official Statistics source to identify common patterns.

The outline of the thesis is the following:

- a literature review of various definitions of big data and experiments, in particular concerning the use of the new source combined with traditional data sources;
- three chapters about our studies on big data:
  - o the first two concern mobility in Lombardy region using mobile phone data. They both refer to the same issue, but they differ in the traditional data source used: the Origin/Destination matrix in the first case, an integrated version of the O/D matrix in the other;
  - o the third shows the pilot that was carried out during the traineeship at Eurostat. It concerns the use of Wikipedia, free online encyclopedia, for Tourism Statistics;
- a final chapter, with conclusions and future remarks on the use of big data in Official Statistics.

A first version of the literature review merged into a poster that was presented at the Webdatanet Conference 2015 in Salamanca, Spain.

An extraction of the first mobility pilot has become a paper that has first been presented at the 10<sup>th</sup> Scientific Meeting of the Classification and Data Analysis Group (CLADAG 2015) in Cagliari, Italy, and has now been accepted for publication on the book “Advances in Classification and Data Analysis”, in the Series “Studies in Classification, Data Analysis, and Knowledge Organization” edited by Springer.

The Wikipedia pilot has been presented several times. In Eurostat (during an R users’ group meeting (June 14<sup>th</sup>, 2016), during the Working Group on Tourism Statistics (October 10<sup>th</sup>, 2016) and during a lunchtime presentation (November 28<sup>th</sup>, 2016)); at a conference, the 14<sup>th</sup> Global Forum on Tourism Statistics in Venice, Italy (November 23<sup>rd</sup>, 2016), at the University of Bergamo, Italy (December 5<sup>th</sup>, 2016) and at the DIGITEC conference in Brussels (November 29<sup>th</sup>, 2016); an updated version of the pilot, considering also an extension to 2016 data has been accepted to the 2017 New Techniques and Technologies for Statistics (NTTS), an international biennial scientific conference series organised by Eurostat on new techniques and methods for Official Statistics, and the impact of new technologies on statistical collection, production and dissemination systems. It will take place in March in Brussels, Belgium.

A little note about reproducibility of studies: the R scripts of the study carried out in Eurostat are available online at <https://github.com/avirgillito/unece-sandbox2015-wikistat>. The experience gained through the Query Wikidata Service, allowed also the construction of an R package that is downloadable from my GitHub page (<https://github.com/serenasignorelli>, instructions available on the main page).

The big data analysis of the third pilot has been carried out through the access that Eurostat has to the UNECE Sandbox. It is an environment that has been created with support from the Central Statistics office (CSO) of Ireland and the Irish Centre for High-End Computing (ICHEC) that provides a technical platform to load big data sets and tools. More information is available at <http://www1.unece.org/stat/platform/display/bigdata/Sandbox>.



# **1 APPLICATIONS WITH BIG DATA: A LITERATURE REVIEW**

## **1.1 Introduction on Big Data**

In the last few years, a huge amount of digital data has become available and the term big data has been used in various fields, especially in statistics. The best "guesstimates" of the total amount of data in the world suggest that from 1987 to 2007 the size of analogue and digital data in the world grew from 3 billion gigabytes to 300 billion gigabytes, an increase by a hundred in two decades (Haire & Mayer-Schönberger, 2014). The data originate from various and heterogeneous different sources like people, machines or sensors. As emphasized by the UN Global Pulse, big data is both the information that is passively generated as by-products of people's everyday use of technologies and the information people willingly communicate about themselves on the web<sup>1</sup>.

Unfortunately, a precise and unique definition of big data does not exist, as it is a general concept related to many disciplines and to a wide amount of different data. Anyway, multiple definitions have been proposed in literature; De Mauro et al. (2015) collected and summarized them into four groups:

1. Many definitions describe big data through its characteristics. In particular, three main features can be identified (Laney, 2001), known as the three Vs: Volume, Velocity, and Variety. Others have added other features (Japiec et al, 2015): Variability, Veracity, and Complexity. As this research will focus in particular

on the use of big data in Official Statistics, it is important to highlight that the main dimensions for this purpose are Volume, Velocity, Variety and Veracity. Volume is related to the large amount of data to be processed to obtain statistical indicators; Velocity states the ability to provide timely results; Variety attains the need to process different big data sources and to possibly integrate them; Veracity means the need for evaluation of different aspects of the quality of the data source. The most critical dimensions to evaluate the use of big data in Official Statistics are Variety and Veracity.

2. A second group of definitions emphasizes the technological needs behind the processing of large amount of data. As an example, Microsoft (2013) defines big data as “the application of serious computing power to massive sets of information”, and the National Institute of Standards and Technology (NIST) highlights the need for a scalable architecture for efficient storage, manipulation and analysis when dealing with them.
3. A third group links big data to the crossing of some sort of threshold regarding the processing capacity of conventional database systems (Dumbill, 2013).
4. The fourth group of definitions highlights the impact of big data advancement on society.

Even if the aim of this thesis is not the description and definition of the technology used to process big data, it is useful to introduce an IT concept that is deeply linked with big data: Hadoop. According to Google Trends, it represents the most related query to “big data” in users searches (De Mauro et al., 2015).

Hadoop is an open source framework that enables the distributed processing of big quantities of data by using a group of dispersed machines and specific computer programming models. The main components of Hadoop are:

1. its file system HDFS, that allows access to data scattered over multiple machines without having to cope with the complexity inherent to their dispersed nature;
2. MapReduce, a programming model designed to implement distributed and parallel algorithms in an efficient way.

---

<sup>1</sup> <http://www.unglobalpulse.org/about/faqs>

When we talk about big data, we refer to different kind of data. According to Japec et al. (2015), the main classifications are:

- Social media data
- Personal data
- Sensor data
- Transactional data
- Administrative data (there is a debate whether this category can be considered as big data).

In some cases, survey data quickly collected using technical tool and contacting a large number of units could also be considered in the frame of the big data concept.

Having described the main features and classifications of this new kind of data, it is now the time to consider which are the aims that encourage us to analyze this information. Big data can be useful for two main different purposes:

- 1) operational: used by businesses in order to analyze their management performance and to improve it. They consist in databases for managerial purposes; there is no special task to obtain statistical indicators. The data user might just count the numbers in the database and compute some measures, with no need to extend results to a collectivity;
- 2) statistical: should provide statistical information, i.e. data which are representative of the whole target population and are of good quality.

The next two paragraphs will show some applications in literature concerning each of the two abovementioned purposes. The experiments are classified by branch for the operational aim and by category of data for the statistical purpose.

## **1.2 The use of big data for operational purposes**

Big data has already been used for managerial purposes in many sectors; this happened because it is easier to use this kind of data without any aim of generalization to the whole population. We decided to classify the experiments available in literature by

branch of application; in particular, we will consider three sectors: tourism, marketing and social sciences.

### ***1.2.1 Tourism sector***

Its recognized socio-economic relevance, as one of the largest industry in the world, makes tourism an interesting field of research. In fact, the recent interest of scholars and researchers from different areas demonstrate the existence of several unexplored areas as well as its multidisciplinary nature (Del Vecchio et al., 2014).

There is a huge amount of data available online, coming from different sources: social networks, databases of booking transactions, reviews on specialized Websites as Tripadvisor, Holiday check, Gogobot, etc.

A study by Pan and Yang (2014) proposes a conceptual framework that connects different types of big data with stages of travel: pre-trip, on-route, on destination and post-trip. This experiment lists each kind of data that can be collected: search queries, Web analytics data, GPS logs and mobile positioning, Bluetooth and infrared tracking, customer reviews, other user-generated content, transaction data, appl logs and smart cards. It finally proposes five future directions for the use of big data in tourism sector:

1. Understanding tourist behaviour.
2. Forecasting tourist activities and the future performance of tourism businesses.
3. Personalizing of service and improving customer experience.
4. Optimizing business operations.
5. Allocating resources and facilities at destinations.

Some applications on the use of big data have been done, especially by new start-ups. Two Italian examples are Travel Appeal and Localler. The former has developed a semantic algorithm through which it is possible to collect all data regarding a specific place, to filter and to analyse them in order to construct some indices of Web reputation. The latter has built a tool that allows hospitality managers to supervise in the best way all the selling channels.

There is also a Spanish example, Amadeus, which is similar to Localler, that built a tool for hospitality managers in order to measure their booking performance and to find

new business opportunities. This start-up has also written an independent study (Davenport, 2013) on the impact of big data on the global travel industry.

Another research provided by RocaSalvatella and Telefónica (2014) develops a new methodology for improved analysis and knowledge of the Spanish tourism industry. The aim of their report is to make use of opportunities for the sector of incorporating macrodata collected from electronic activity of anonymous foreign tourists into market research. The kind of data that they used consists in foreigner's smartphone data and electronic payments by foreign cards.

An additional study that uses smartphone data has been carried out by Frias-Martinez et al. (2010). The application tried to work on the characterization and automatic identification of the gender of a cell phone user based on behavioural, social and mobility variables.

### ***1.2.2 Marketing***

In marketing, in particular in the advertising field, two different applications (Duong T., Millman S., 2015; Porter S., Lazaro C.G., 2014) have added big data to traditional survey data in order to check the effectiveness of mobile advertisements and brands.

The former authors have implemented a method to record human-mobile interaction with the ad campaign as well as providing the opportunity for users to complete the surveys on mobile devices. In their experiment, they sought to learn what advertisement style is the most effective among the available ad units, how it compares to non-advertisement environment, and whether interaction with the ad will improve customer's perception of the brand. In order to achieve these goals, they coupled mobile survey research and interaction measurement. Authors found that the methodology to collect interaction points to different ad types on mobile devices using additional pixel calls provides additional information to the traditional online survey. Furthermore, by adding interaction data, they were able to have a clearer picture and to come closer to the truth in measuring the effect of ad types on mobile devices.

The second paper aims at illustrate, through a series of cases, the variety of possibilities for combining survey and non-survey data. Authors present four experiments:

1. making comparisons by ad: direct response advertising data combined with copy test survey data;
2. making comparisons in trends over time: competitive intelligence and sales performance data combined with brand tracking survey data;
3. making comparisons by respondent: consumer behaviour data from website activity and transactions combined with survey data capturing perceptions, attitudes, life events, and offsite behaviour trends;
4. analysis with multiple levels of data: using survey data as a wide (but thin) overview of the market, to contextualize the deep (but narrow) pockets of non-survey data.

They found some commonalities among the four cases:

- in each case, the question was better answered by the combination of both the survey and non-survey data than could have been accomplished by either alone;
- the non-survey data is generally used to get to a more detailed answer to what than would have been possible in a survey, and the survey data is generally used to explore evidence about why (although this is a fuzzy distinction, and some pieces of evidence fall in the alternate category);
- predictive modeling is used to enable the blending of data in a way that makes it strategically actionable.

### ***1.2.3 Social science***

Social science is a major category of academic disciplines, concerned with society and the relationships among individuals within a society. It constitutes a very heterogeneous field; various applications with big data have been carried out, but most of them are oriented to two directions: analysis of human personality and study of traffic conditions. It is worth citing also an interesting experiment in medical field.

### *1.2.3.1 Analysis of human personality*

A research by Statistics Netherlands (Daas et al., 2013) analysed social networks data from two points of view: content and sentiment. Studies of the content of Dutch Twitter messages revealed that nearly 50% of the messages are composed of personal insights, while in the remainder spare time, activities, work, media (TV & radio) and politics are predominantly discussed. This suggested them that these messages could be used to extract opinions, attitudes, and sentiments towards these topics. The other potential use of social media messages is sentiment analysis (defined as the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials). Authors sourced messages from the largest social media platforms including Twitter, Facebook, Hyves, Google+ and LinkedIn, but also from numerous public weblogs and forums. The sentiment of each message was automatically determined by counting the number of positive and negative words, so that messages were classified as positive, negative or neutral depending on their overall score. After that, they used this classification of the data in order to discover correlations among words.

Social media data have been used also by Tostes et al. (2014) in order to study traffic conditions. In their study, they analysed two social sensing sources: Foursquare and Instagram, two social networks in which every data shared is referred to as a 'check-in'. Their aim is to investigate whether it is possible to use data from social sensors to better understand the traffic condition, represented by the traffic flow retrieved from Bing Maps. The objective is to verify if check-ins can be used as a hint of traffic conditions changes or current situation.

A similar application by Noulas et al. (2011) analysed Foursquare data (in particular user check in dynamics) in order to demonstrate how they could reveal meaningful spatio-temporal patterns and offer the opportunity to study both user mobility and urban spaces.

Another application in the field of social sciences has been carried out by De Montjoye et al. (2013); the goal of the researchers is to show that users' personality can

be reliably inferred from standard mobile phone logs<sup>2</sup>. In particular, they introduced five sets of psychology - informed metrics (basic phone use, active user behaviours, mobility, regularity, diversity) that can be easily extracted from standard phone logs to predict how extroverted, agreeable, conscientious, open to experience, and emotionally stable a user is.

### *1.2.3.2 Study of traffic conditions*

Cici et al (2014) did an experiment in order to assess the potentiality of ride-sharing for reducing traffic in a city, using mobility data extracted from Call Detail Records (CDRs)<sup>3</sup> of the cities of Madrid and Barcelona, and social media data from Twitter and Foursquare of New York and Los Angeles. They first analysed the dataset to understand mobility patterns, home and work locations and social ties between users; they then developed an algorithm to match users with similar mobility patterns, in order to predict an upper bound to the potential decrease in the number of cars in a city that can be achieved by ride-sharing.

CDRs have been used also by a Brazilian start up called ‘Cignifi’, which developed a technology to recognize patterns in the usages of mobile devices. In particular, the system recognizes phone-calls, text messages and data usage and through this information it reconstructs someone’s lifestyle and his/her corresponding credit risk profile.

The study of mobile phone data (in the form of CDRs) and GPS data have helped Pappalardo et al. (2015) discovering the existence of two distinct classes of individuals: returners and explorers, and the existence of a correlation between their mobility patterns and social interactions.

---

<sup>2</sup> A phone log is metadata collected from telephone or mobile phones that may include various information: length of calls, phone numbers of both parties, phone-specific identification information, gps location, call proximity, and/or computer converted voice-to-text transcripts of the phone call conversation

<sup>3</sup> A Call Detail Record is a data record produced by a telephone exchange or other telecommunications equipment that documents the details of a telephone call or other communications transaction (e.g., text message) that passes through that facility or device. The record contains various attributes of the call, such as time, duration, completion status, source number, and destination number

### 1.2.3.3 *Medical field*

Centers for Disease Control and Prevention have used big data in order to study Diabetes (Day H.R., Parker J.D., 2013). In particular, the research compares self-reported diabetes in the National Health Interview Survey (NHIS) with diabetes identified using the Medicare Chronic Condition (CC) Summary file. This experiment was carried out because people who self-report diabetes may not always be identified by Medicare claims, and not all people who have Medicare claims for diabetes will self-report the disease. The records of 2005 NHIS participants aged 65 and over were linked to 2005 Medicare data. Of the Medicare beneficiaries in the 2005 NHIS, 20.0% self-reported diabetes and 27.8% had an indicator for diabetes in the CC Summary file. Of those who self-reported diabetes in NHIS, the percentage with a CC Summary indicator for diabetes was high (93.1%). Of those with a CC Summary indicator for diabetes, the percentage self-reporting diabetes was comparatively lower (67.0%). Some differences exist in subgroups, as the self-reported diabetes and the CC Summary indicator for diabetes were compared and described by demographics, socioeconomic status, health status indicators, and geographic characteristics.

## **1.3 The use of big data for statistical purposes**

Big data could represent an opportunity for Official Statistics (Kitchin, 2015); in fact, they open to the possibility for *nowcasting* (the prediction of the present), they represent a rich source of granular data to complement and extend micro-level and small area analysis, and they potentially ensure comparability of phenomena across countries.

Many Global, European and National official entities have already moved into the direction of big data. Some of the initiatives and projects of the main organizations are shown below (Japac and Kreuter, 2015).

Eurostat, the statistical office of the European Union, launched the project “Big Data Action Plan and Roadmap”, which includes some pilot studies exploring the potential of selected big data sources.

The United Nations Economic Commission for Europe (UNECE) developed the ‘Sandbox’, which provides a computing environment to load big datasets and tools; the experiments currently underway regards:

- Consumer price indices;
- Mobile phone data;
- Smart meters;
- Traffic loops;
- Social media;
- Job portals;
- Web scarping.

Statistics Netherlands, the Dutch Central Bureau of Statistics, has developed the so-called “Roadmap Big Data” which rests on two focus projects:

- The use of traffic loop data for transportation statistics;
- The use of mobile phone data for daytime population and tourism statistics.

Other six projects are included into the “Roadmap”; they concern the analysis of internet data for price statistics, bank and credit card transactions, social media data for detecting trends in social cohesion, internet data for encoding enterprise purchases and sales, smartcards of public transport for statistics and internet data for statistics about job vacancies.

At Statistics Sweden, the Swedish government agency responsible for producing Official Statistics, researchers have also carried out some applications. They used scanner data in order to improve the Household Budget Survey, they used Web scraping techniques to build job vacancies statistics, and they collaborated with the agency for Transport Analysis in order to evaluate the use of Automatic Identification System (AIS) data.

The European Statistical System Committee (ESSC), an agency chaired by the European Commission (Eurostat) and composed of the representatives of EU Member States' National Statistical Institutes, has identified several Official Statistics domains

that could be profitably augmented by the use of different kinds of big data (ESSC, 2014); its findings are exposed in Table 1.

*Table 1. Potential use of big data in Official Statistics*

<i>No</i>	<i>Data source</i>	<i>Data type</i>	<i>Statistical domains</i>
1	Mobile communication	Mobile phone data	Tourism statistics Population statistics
2	WWW	Web searches	Labour statistics Migration statistics
		e-commerce websites	Price statistics
		Businesses' websites	Information society statistics Business registers
		Job advertisements	Employment statistics
		Real-estate websites	Price statistics (real-estate)
		Social media	Consumer confidence GDP and beyond Information society statistics
3	Sensors	Traffic loops	Traffic/transport statistics
		Smart meters	Energy statistics
		Satellite images	Land use statistics Agricultural statistics Environment statistics
		Automatic vessel identification	Transport and emission statistics
4	Transactions of process generated data	Flight movements	Transport and emission statistics
		Supermarket scanner and sales data	Price statistics Household consumption statistics
5	Crowdsourcing	Volunteered geographic information (VGI) websites (OpenStreetMap, Wikimapia, Geowiki)	Land use
		Community pictures collections (Flickr, Instagram, Panoramio)	-

Source: ESSC, 2014.

After having introduced the initiatives of some institutions, we will now briefly highlight many more or less explorative applications that are under study by researchers and Official Statistics agencies. We decided to classify the experiments using the

categories highlighted by the ESSC, as they will give a clearer view of the whole situation.

### ***1.3.1 Mobile communication***

Mobile phone data, in particular the abovementioned Call Detail Records (CDRs), have been used in several different applications published between 2011 and 2015.

Soto et al. (2011) tried to evaluate the use of aggregated cell phone data to model and predict different socioeconomic levels of a population. The experiment concerns a main city in a Latin-American country and lead to the result of correct prediction rates of over 80%.

Frias-Martinez et al. (2013) studied whether the time series of socioeconomic indicators computed by National Statistical Institutes (NSIs) using surveys can be forecasted using behavioural information extracted from calling records. This experiment concerns again Latin America and the results indicate that some of the multivariate time-series models are not able to predict approximate real values, but can forecast changes in the trends of the NSI series.

Heerschap et al. (2014) conducted a pilot study in the Netherlands to see whether aggregated mobile phone metadata based on CDRs could be used for tourism statistics. Even if the dataset had some limitations (it was referred to only one telecommunication provider and each mobile phone was followed only for one day), this study represents an attempt to reconstruct Statistics Netherlands's survey based tourism accommodation statistics starting from mobile phone data.

Liang and Frias-Martinez (2015) tried to create a technique to automatically compute official traffic counts using mobility features extracted from CDRs. They were able to rebuild with good accuracy Senegal's official traffic counts computed in 2002. This approach seems promising for emerging and poor economies, as Official computations are costly and not frequently updated, whereas CDRs could come theoretically at no cost.

Bogomolov et al. (2014) presented a novel approach to predict crime in a geographic space from multiple data sources, in particular mobile phone and demographic data. In

their experimental results with real crime data from London, they obtained an accuracy of almost 70% when predicting whether a specific area in the city will be a crime hotspot or not. This application could be useful for crime prevention and monitoring. As well, the collected data could be the basis for statistical indicators computed on a detailed geo-grid level.

Deville et al. (2014) used CDRs from Portugal and France in order to show how spatially and temporarily explicit estimations of population densities can be produced at national scales, and how these estimates compare with output produced using alternative human population mapping methods.

An attempt of use of CDRs has been carried out also at the Italian Statistical Institute (ISTAT) (Barcaroli, 2015). Mobile phone data have been integrated into the project “Persons and places” that aims at the construction of the Origin-Destination matrix at municipality level. The reason of this inclusion lies in the fact that administrative data don’t allow either to distinguish between dynamic residents and commuters nor to estimate occasional visitors, as they do not contain information on the frequency of the mobility.

### ***1.3.2 World Wide Web***

This second category is highly extended, as it includes very different kind of data. We will show applications that make use of social media data, businesses’ Websites, Internet queries and Wikipedia page views.

Regarding social media, the University of Michigan (Antenucci et al., 2014) built the Social Media Job Loss Index to monitor the unemployment trend using Twitter data. In particular, they used data from Twitter to create indexes of job loss, job search and job posting. Signals are derived by counting job-related phrases in Tweets such as “lost my job”. The social media indexes are constructed from the principal components of these signals. The Social Media Job Loss Index tracks initial claims for unemployment insurance at medium and high frequencies and predicts 15 to 20% of the variance of the prediction error of the consensus forecast for initial claims.

A different technique is represented by Web scraping. Nathan et al. (2013) collected all the data they could find on the Internet about English digital enterprises in order to rewrite the official classification of English enterprises (in which digital companies are underestimated). In particular, the digital economy comprises almost 270,000 active companies in the UK (14.4% of all companies as of August 2012) that compares to 167,000 companies (10.0%) when the Government's conventional SIC-based<sup>4</sup> definitions are used. SIC-based definitions of the digital economy miss out a large number of companies in business and domestic software, architectural activities, engineering, and engineering-related scientific and technical consulting, among other sectors.

Other experiments using the Web scraping technique have been carried out at the Italian National Statistical Institute (ISTAT).

The first one (Barcaroli et al., 2014) concerns the sampling survey on "Information and Communication Technology in enterprises", which aims at producing information on the use of ICT and, in particular, on the use of Internet by Italian enterprises. Web scraping techniques were applied concerning the questionnaire section containing a query about the existence of the enterprise's website (located on servers belonging to the enterprise or third party website). Authors had a twofold aim: to verify the capability to access the websites indicated by enterprises participating to the sampling survey, and to collect all the relevant information, and to use the information collected from the Internet in order to predict the characteristics of the websites, not only for surveyed enterprises, but for the whole population of reference. The approach presents opportunities and disadvantages, but it revealed to be promising as a whole.

The second experiment (Polidoro et al., 2015) conducted at ISTAT was applied on consumer price survey, in particular on two groups of products: consumer electronics (goods) and airfares (services). Web scraping procedures have been developed and tested for both these groups of products, with two aims: to replace the download of the lists of prices with the automatic download (for goods), and to record prices for air transport services. For consumer electronics, the adoption of Web scraping techniques

---

<sup>4</sup> The Standard Industrial Classification is a system for classifying industries by a four-digit code.

led to a reduction of the workload necessary to manage the survey from about 23 working days to 16 working days, while for airfares there were a few advantages in terms of time saving without expanding the data collection.

Using Web scraping techniques, the improvements prevail on the drawbacks, even though both improvements and drawbacks should be better investigated.

ISTAT also tried to estimate the monthly unemployment rate using Internet queries, in particular Google Trends (Barcaroli, 2015). In order to do so, the time series related to the monthly unemployment rate calculated from the Labour Force survey have been compared with those from Google Trend data.

Another kind of Web data available is represented by Wikipedia article traffic statistics. Page view statistics is a tool available for Wikipedia pages, which allows to know how many people have visited an article during a given time period (usually hourly counts). Some applications have been carried out in the last two years.

McIver and Brwnstein (2014) developed a novel method of estimating, in near real-time, the level of influenza-like illness in the United States by monitoring the rate of particular Wikipedia article views on a daily basis.

Yasseri and Bright (2015) developed a theoretical model which highlights why people might seek information online at election time, and how this activity might relate to overall electoral outcomes. The experiment is based upon the individuation of a relationship between Wikipedia traffic patterns around election time and the overall electoral turnout.

Finally, Munzert (2015) carried out an application in order to evaluate public issue attention, which is a topic that usually gets measured through Most Important Problem (MIP) surveys. He proposes the use of Web data, in particular Google Trends (that present some limitations) and Wikipedia page view statistics.

### ***1.3.3 Sensors***

Statistics Netherlands, already mentioned before, used traffic sensors to predict traffic intensity (Daas et al., 2013). In particular, they used traffic loop detection data, consisting of measurements of traffic intensity. Each loop counts the number of vehicles

per minute that pass at that location, and measures speed and length. They created maps indicating the number of vehicles for each measurement location (by using different colours) and, by combining these maps, a movie that displays the changes in vehicle counts for all locations during the day. They also studied the number of vehicles in various length categories.

Another application tested in the region of Eindhoven (the Netherlands) tries to figure out whether traffic intensity contains relevant information on regional economic activity (Braaksma and Zeelenberg, 2015). This study is currently under research, and it is carried out using traffic loop data and the manufacturing sentiment survey as a benchmark. Up to now, the evolution of the traffic intensity indicator tracks that of expected production development very well.

A different kind of sensors is represented by satellite data, in particular those about lights during the night. Henderson et al. (2009) developed a statistical framework that uses lights growth to augment existing income growth measures, under the assumption that measurement error in using observed light as an indicator of income is uncorrelated with measurement error in national income accounts. They found that for countries with good national income accounts data, the information on growth of lights is of marginal value in estimating the true growth rate of income, while it becomes an important tool in poor-data countries in order to measure income growth. This information is also important to evaluate sub- and supranatural regions.

#### ***1.3.4 Transactions of process generated data***

Concerning this category, we will consider data that comes from flight movements.

Each year a plane can create 2.5 billion terabytes of sensor data<sup>5</sup>. Southwest Airlines have been working with NASA and have created algorithms in order to pinpoint small anomalies that could represent a safety issue on planes. This means both improved safety on their planes and decreased equipment malfunctions, so fewer delays.

---

<sup>5</sup> Big Data in air travel, <https://channels.theinnovationenterprise.com/articles/27-big-data-in-air-travel>

This kind of data are also considered at the Frankfurt airport<sup>6</sup>, where all the processes that occur in the hub generate data that can be classified according to four different views:

- a. General data like weather or punctuality information from ATM
- b. Data related to aircraft movements (e.g. time stamps, load and performance data)
- c. Data related to passengers – ‘average’ figures only and not related to individuals
- d. Data related to individual but ‘anonymous’ passengers.

Considering these views, four different implementations have been carried out; they relate to aircraft noise emissions, enhancing airside operations, reducing waiting times and understanding shopping trends.

### ***1.3.5 Crowdsourcing***

Regarding this last category, we cite an application by Barchiesi et al. (2015); the research team used geotagged photographs uploaded to the photo-sharing website Flickr to quantify international travel flows, by extracting the location of users and inferring trajectories to track their movements across time. They found that Flickr based estimates of the number of visitors to the United Kingdom significantly correlate with the official estimates released by the UK Office for National Statistics, for 28 countries for which official estimates are calculated.

## **1.4 Conclusions**

Big data potentialities for Official Statistics need a huge amount of experimentation and of economic statistics studies in order to set up a suitable metadata framework and to evaluate Veracity, Validity and Value of the considered big data. Surely deeper insight in quality of this big data and in the variety of aspects and sources which could be integrated to setup the potential use as a statistical information is needed.

---

<sup>6</sup> Airport Big Data: Reality or hype? <http://www.internationalairportreview.com/advent-calendar/6-december-2014/>



## 2 FROM BIG DATA TO INFORMATION: STATISTICAL ISSUES THROUGH CASE STUDIES

Silvia Biffignandi<sup>1</sup>, Serena Signorelli<sup>1</sup>

<sup>1</sup> University of Bergamo

(e-mail: [silvia.biffignandi@unibg.it](mailto:silvia.biffignandi@unibg.it), [serena.signorelli@unibg.it](mailto:serena.signorelli@unibg.it))

**ABSTRACT:** This paper gives a short overview on the use of big data for statistical purposes. The introduction on the characteristics of big data and their classifications highlights the problems that arise when trying to use them in a statistical way. They are mainly related to veracity of big data. We focus on quality aspects and representativeness. After that, we show some applications in literature that use big data in combination with traditional survey data. Fields considered are health, marketing, social media, Web. Finally, a small-scale case study is presented by critically highlighting problems and solutions arising in the transition from big data to information; the experiment combines Official Statistics data from Italian NSI with a telecommunication provider dataset. The aim is trying to put in a unique interpretative framework one traditional statistical source and one typical kind of big data in order to evaluate some informative potentialities of this approach.

**KEYWORDS:** ‘big data’, ‘quality’, ‘representativeness’, ‘communication’, ‘mobility’

### 2.1 Introduction on Big Data

In the last few years, the term big data has been used in various fields, especially in statistics. Unfortunately, a precise definition of big data does not exist, as it is a general concept related to many disciplines and to a wide amount of different data. However, three main characteristics can be identified (Laney, 2001), known as the three Vs: Volume, Velocity, and Variety. Others have added other features (Japiec et al., 2015): Variability, Veracity, and Complexity. Volume, Velocity, Variety and Veracity are the main dimensions for big data in Official Statistics. Volume is related to the large amount of data to be processed to obtain statistical indicators; Velocity states the ability to provide timely results; Variety attains the need to process different big data sources and to possibly integrate them; Veracity means the need for evaluation of different

aspects of the quality of the data source. The most critical dimensions to evaluate the use of big data in Official Statistics are Variety and Veracity. Having the abovementioned characteristics as a background information, research has to be undertaken in order to define more specific concepts and methodologies. One interesting operational starting point consists of considering various types of data relying on the general concept and of evaluating advantages and problems regarding different disciplines that should be engaged in the collection, treatment and use of this data.

In this paper, we highlight some key issues disentangled from the statistical point of view, especially from business and social statistics. Big data can be classified into (Japéc et al., 2015):

- Social media data
- Personal data
- Sensor data
- Transactional data
- Administrative data (there is a debate whether this category can be considered as big data).

In some cases, survey data quickly collected using technical tool and contacting a large number of units could be considered in the frame of the big data concept.

The paper discusses at first some statistical and quality issues (par. 2), introduces some recent empirical experiences (par. 3) and focuses on a case study (par. 4) by pointing out an original overview and analysis of existing databases for the use of big data for statistical purposes. Thus, it contributes to increase literature on research devoted to match experimental studies with the statistical features of the provided information.

## **2.2 Statistical and quality issues**

From a statistical point of view, a huge amount of data could be considered as a positive aspect for the information provided through the data collection. Big data, as the term

suggests, carry a great quantity of data but quality is a characteristic to look at before using them for statistical purposes.

Big data can be useful for two main different purposes:

- 1) operational: used by businesses in order to analyze their management performance and to improve it. They consist in databases for managerial purposes; there is no special task to obtain statistical indicators. The data user might just count the numbers in the database and compute some measures, with no need to extend results to a collectivity;
- 2) statistical: should provide statistical information, i.e. data which are representative of the whole target population and are of good quality.

We refer to this second purpose. The statistical context has some particular issues, as prior characteristics: quality and representativeness.

Big data, differently from traditional probability based survey data, are not collected and designed to a specific statistical purpose, but are 'harvested' as they are. Therefore, traditional statistical approaches (like inference or modeling techniques) used in survey data collection are not immediately applicable. There are many risks in using big data. As regards Veracity, they could contain errors of different nature and they need appropriate error categorizations and statistical methods, which are still under study. Most errors are related to representativeness. In literature, many experiences focus on the potentiality of big data but most of them are not focused on their statistical properties.

Another issue concerns their volatility and instability; big data coming from social networks could become incomparable from one day to the other, due to some recurring changes that providers introduce in order to improve the structure and the user experience. Moreover, transactional or administrative data could change their structure and the way they are collected for operational and efficiency reasons.

Other problems concern big dimensionality. Big data have a big dimensionality but they represent only a specific part of the general population, and big effort is necessary to make them representative to the whole population. Different definitions of

representativeness exist in literature; we consider it as the attempt to generalize the results to the target population. This task is not easy, as big data collected through a variety of formats (mostly through IT technology) could catch units or phenomena that differ from the units or phenomena that are not collected (i.e. considering mobile data, people who use smartphones could behave differently from people who do not use them). An attempt of generalization has been made by Elliott (2009) who built pseudo-weights in order to combine probability and non-probability samples. Representativeness is only one aspect of the statistical issues relevant in order to obtain statistical socio-economic indicators taking big data as a source.

The other main statistical issue is represented by quality. If big data quality is in doubt and cannot be specified a priori, one approach consists in just let the “data to speak for itself” in order to discover its semantics or to detect correlations among variables, but not causal relationship. The risks of poor big data quality arise at three steps (Bellmore, 2014) affecting:

- i. initial data loading: in addition to the six classic data quality dimensions (completeness, conformity, consistency, accuracy, duplication and integrity), relevancy of the specific big data as a data source has to be considered;
- ii. application integration: various sources of data are available and their integration has to be done carefully. This is a rather critical point to be implemented since each source has its own quality characteristics and different sources generally have heterogeneous characteristics. A great effort in finalizing metadata for possible data sources integration requires a great knowledge and skills in the variables generation process;
- iii. data maintenance: agents like private businesses or public bodies, outside the Official Statistics organizations, are providing big data; there is a need to check the persistence of the characteristics and quality of the data.

The data gathered using big data technology is much more vulnerable to statistical errors (non-sampling and sampling errors) than using traditional data sources. User entry errors, redundancy, corruption, noise accumulation and the uncorrelation of model covariates with the residual error) are problems that affect the value of the data (Saha et al., 2014; Fan et al., 2014).

Two solutions are recommended in literature in order to deal with quality issues:

1. Japiec et al. (2015) suggest the introduction of a Total Error Framework specific for big data, based on the Total Survey Error framework that already exists (Biemer, 2010). Biemer (2014) has created the “Big Data process map” that contains three phases:
  - a. generation, in which data are generated from some source either incidentally or purposively;
  - b. extraction / transformation / loading, a phase that brings all the data together in a homogeneous computing environment;
  - c. analysis, when data are converted to information.

For each phase, he individuates which errors arise.

2. Dufty et al. (2015) propose a framework that aims at assessing the quality of these data at three stages:
  - a. input, when the data are acquired;
  - b. throughput, when the data get transformed, analyzed or manipulated;
  - c. output, the phase of reporting.

In order to obtain socio-economic statistical information from these data it is necessary to specify in detail the characteristics of the above-mentioned framework and to apply it. This framework focuses on the specific quality requirements and challenges for the use of big data in Official Statistics.

Big data could represent an opportunity for Official Statistics (Kitchin, 2015); in fact, they open to the possibility for “nowcasting” (the prediction of the present), they represent a rich source of granular data to complement and extend micro-level and small area analysis and they potentially ensure comparability of phenomena across countries.

On the other hand, big data carry some challenges, like representativeness, both of phenomena and populations, as these data are not generally representative of an entire population, but they only relate to whomever uses a service. Furthermore, data quality dimensions (OECD, 2011) are largely unknown with respect to various forms of big data and generators of the data are often reluctant to share methodological transparency

in how they were produced and processed. Finally, frames within big data are generated are mutable, changing over times.

## **2.3 Some empirical studies**

Big data tentative experiences rely on different fields. In some cases, they were used in combination with traditional survey data.

In medical field, Centers for Disease Control and Prevention have used big data in order to study Diabetes (Day H.R., Parker J.D., 2013). In particular, the research compares self-reported diabetes in the National Health Interview Survey (NHIS) with diabetes identified using the Medicare Chronic Condition (CC) Summary file. This experiment was carried out because people who self-report diabetes may not always be identified by Medicare claims, and not all people who have Medicare claims for diabetes will self-report the disease. The records of 2005 NHIS participants aged 65 and over were linked to 2005 Medicare data. Of the Medicare beneficiaries in the 2005 NHIS, 20.0% self-reported diabetes and 27.8% had an indicator for diabetes in the CC Summary file. Of those who self-reported diabetes in NHIS, the percentage with a CC Summary indicator for diabetes was high (93.1%). Of those with a CC Summary indicator for diabetes, the percentage self-reporting diabetes was comparatively lower (67.0%). Some differences exist in subgroups, as the self-reported diabetes and the CC Summary indicator for diabetes were compared and described by demographics, socioeconomic status, health status indicators, and geographic characteristics. This is a advantageous use of big data for statistical purposes and could satisfy high level of Veracity. In this case, data appear more like an administrative source; moreover, due to very specific item to which the indicator refers to (diabetes), classification and definition criteria are less important or clearly defined.

In marketing, in particular in the advertising field, two different applications (Duong T., Millman S., 2015; Porter S., Lazaro C.G., 2014) have added big data to traditional survey data in order to check the effectiveness of mobile advertisements and brands. The former authors have implemented a method to record human-mobile interaction with the ad campaign as well as providing the opportunity for users to complete the

surveys on mobile devices. In their experiment, they sought to learn what advertisement style is the most effective among the available ad units, how it compares to non-advertisement environment, and whether interaction with the ad will improve customer's perception of the brand. In order to achieve these goals, they coupled mobile survey research and interaction measurement. Authors found that the methodology to collect interaction points to different ad types on mobile devices using additional pixel calls provides additional information to the traditional online survey. Furthermore, by adding interaction data, they were able to have a clearer picture and come closer to the truth in measuring the effect of ad types on mobile devices. The second paper aims at illustrate, through a series of cases, the variety of possibilities for combining survey and non-survey data. Authors present four experiments:

1. making comparisons by ad: direct response advertising data combined with copy test survey data;
2. making comparisons in trends over time: competitive intelligence and sales performance data combined with brand tracking survey data;
3. making comparisons by respondent: consumer behaviour data from website activity and transactions combined with survey data capturing perceptions, attitudes, life events, and offsite behaviour trends;
4. analysis with multiple levels of data: using survey data as a wide (but thin) overview of the market, to contextualize the deep (but narrow) pockets of non-survey data.

They found some commonalities among the four cases:

- in each case, the question was better answered by the combination of both the survey and non-survey data than could have been accomplished by either alone;
- the non-survey data is generally used to get to a more detailed answer to *what* than would have been possible in a survey and the survey data is generally used to explore evidence about *why* (although this is a fuzzy distinction, and some pieces of evidence fall in the alternate category);
- predictive modelling is used to enable the blending of data in a way that makes it strategically actionable.

This study is an example of operational use of big data and it is not aiming at Official Statistics construction; nevertheless, evidence obtained in better answering questions by the combination of both the survey and non-survey data is an interesting point to be considered and further investigated in the Official Statistics, too.

Regarding social media, the University of Michigan (Antenucci et al., 2014) built the Social Media Job Loss Index to monitor the unemployment trend using Twitter data. In particular, they used data from Twitter to create indexes of job loss, job search, and job posting. Signals are derived by counting job-related phrases in Tweets such as “lost my job.” The social media indexes are constructed from the principal components of these signals. The Social Media Job Loss Index tracks initial claims for unemployment insurance at medium and high frequencies and predicts 15 to 20% of the variance of the prediction error of the consensus forecast for initial claims. An application by Statistics Netherlands (Daas et al., 2013) analyses social networks data from two points of view: content and sentiment. Studies of the content of Dutch Twitter messages revealed that nearly 50% of the messages are composed of personal insights, while in the remainder spare time, activities, work, media (TV & radio) and politics are predominantly discussed. This suggested them that these messages could be used to extract opinions, attitudes, and sentiments towards these topics. The other potential use of social media messages is sentiment analysis (defined as the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials). Authors sourced messages from the largest social media sites including Twitter, Facebook, Hyves, Google+, and LinkedIn, but also from numerous public weblogs and forums. The sentiment of each message was automatically determined by counting the number of positive and negative words, so that messages were classified as positive, negative or neutral depending on their overall score. After that, they used this classification of the data in order to discover correlations among words. They also used traffic sensors to predict traffic intensity. In particular, they used traffic loop detection data, consisting of measurements of traffic intensity. Each loop counts the number of vehicles per minute that pass at that location, and measures speed and length. They created maps indicating the number of vehicles for each measurement location (by using different colours) and, by combining these maps, a movie that

displays the changes in vehicle counts for all locations during the day. They also studied the number of vehicles in various length categories. Bogomolov et al. (2014) presented a novel approach to predict crime in a geographic space from multiple data sources, in particular mobile phone and demographic data. In their experimental results with real crime data from London they obtained an accuracy of almost 70% when predicting whether a specific area in the city will be a crime hotspot or not. This application could be useful for crime prevention and monitoring. As well, the collected data could be the basis for statistical indicators computed on a detailed geo-grid level. Metadata definition could be rather simply decided.

A different technique is Web scraping. Nathan et al. (2013) collected all the data they could find on English digital enterprises in order to rewrite the official classification of English enterprises (in which digital companies are underestimated). In particular, the digital economy comprises almost 270,000 active companies in the UK (14.4% of all companies as of August 2012) that compares to 167,000 companies (10.0%) when the Government's conventional SIC-based definitions are used. SIC-based definitions of the digital economy miss out a large number of companies in business and domestic software, architectural activities, engineering, and engineering-related scientific and technical consulting, among other sectors.

Other experiments using Web scraping have been carried out at the Italian National Statistical Institute (ISTAT). The first one (Barcaroli et al., 2014) concerns the sampling survey on "Information and Communication Technology in enterprises", which aims at producing information on the use of ICT and, in particular, on the use of Internet by Italian enterprise. Web scraping techniques were applied on the questionnaire section containing a query about the existence of the enterprise's website (located on servers belonging to the enterprise or third party website). Authors had a twofold aim: to verify the capability to access the websites indicated by enterprises participating to the sampling survey and to collect all the relevant information, and to use the information collected from the Internet in order to predict the characteristics of the websites, not only for surveyed enterprises, but for the whole population of reference. The approach presents opportunities and disadvantages, but it revealed to be promising as a whole. The second experiment (Polidoro et al., 2015) conducted at ISTAT was applied on

consumer price survey, in particular on two groups of products: consumer electronics (goods) and airfares (services). Web scraping procedures have been developed and tested for both these groups of products, with two aims: to replace the download of the lists of prices with the automatic download (for goods), and to record prices for air transport services. For consumer electronics, the adoption of Web scraping techniques led to a reduction of the workload necessary to manage the survey from about 23 working days to 16 working days, while for airfares there were a few advantages in terms of time saving without expanding the data collection. In conclusion, for Web scraping the improvements prevail on the drawbacks, even though both improvements and drawbacks should be better investigated.

## **2.4 Case study**

The aim of our case study is trying to put in a unique interpretative framework one traditional statistical source (2011 ISTAT Origin/destination matrix from the 15th Population and housing census) and one typical kind of big data (2014 1st Telecom Italia Big Data Challenge datasets) in order to evaluate some informative potentialities of this approach. Origin/destination matrix is released every ten years, whereas Telecom Italia data are collected continuously. It is evident that identifying the informative potentialities of big data is a great challenge for understanding social trends over short run.

As just mentioned, the case study is based on the use of two datasets:

- the 2011 ISTAT Origin/destination matrix from the 15th Population and housing census, that contains data on the number of persons that commute between municipalities – or inside the same municipality – classified by gender, mean of transportation, departure timeslot and journey duration. The spatial aggregation is represented by Italian municipalities. The 15<sup>th</sup> Census was carried out on October 9th 2011, the questions regarding commuting pattern referred to ‘last Wednesday’ or a typical working/studying day.
- One of the 2014 1st Telecom Italia Big Data Challenge datasets. In particular, we use the one named “Telecommunications - MI to Provinces”, which provides

information regarding the level of interaction between the areas of the city of Milan and the Italian provinces. The level of interaction between an area of Milan and a province is given as a number; it represents the proportion of calls issued from the area of Milan to the province (and vice versa). For each area, the dataset contains two proportions: one representing the proportion of inflow telephone traffic and the other one representing the proportion of the outflow. The spatial aggregation of the dataset is the Milano GRID squares<sup>7</sup> and the Italian provinces. The values are aggregated in timeslots of ten minutes.

Our analysis is limited to Lombardy region, divided into twelve administrative provinces. We build commuting patterns concerning the city of Milan and all provinces (excluded flows from the city of Milan to the city of Milan itself). According to the 2014 Report by the Italian Media Safeguards Authority (AGCOM), Telecom Italia is the market leader in mobile telecommunications with an amount of 33,2% in 2013 (the year our data refer to), of which 29,8% are residential and 47,9% business. A good users' coverage is guaranteed. Obviously, the findings might hold for all residential users under the hypothesis that the consuming behaviour does not differ with respect to the telephone provider. It seems reasonable to suppose that the consuming behaviour does not differ with respect to the telephone provider.

At first, this work consists in the analysis of each dataset separately. We show the results regarding the outflow from the municipality of Milan to each of the Lombardy provinces, both of people (ISTAT dataset) and phone calls (Telecom Italia dataset). The results are presented as geographic maps created using CartoDB<sup>8</sup>. The twelve provinces are ranked considering the amount of the outflow and are then split into seven buckets and coloured from dark (first position) to pale green (last position).

We then build the same visualizations considering the inflow from the Lombardy provinces to the city of Milan. In this case the provinces are coloured from dark (first position) to pale blue (last position).

---

<sup>7</sup> Some of the datasets of the 1<sup>st</sup> Telecom Big Data Challenge referring to the Milano urban area are spatially aggregated using a grid composed by 10.000 cells over the municipality of Milan.

<sup>8</sup> <https://cartodb.com/>

## 2.4.1 Outflow maps

### 2.4.1.1 ISTAT dataset

Fig. 1 shows how the provinces are ranked considering the general outflow of ISTAT dataset.

You can notice as the major outflow goes to the nearest (Milan, Monza e Brianza, Varese) and biggest (Brescia, Bergamo and Pavia) Lombardy provinces.

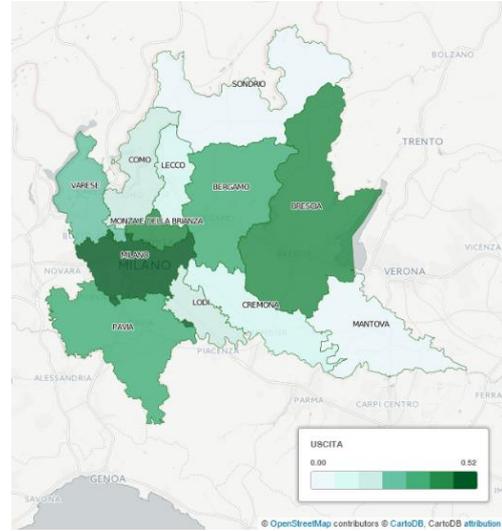


Fig. 1 General outflow from Milan to provinces – ISTAT dataset

It is possible to filter the results by commuting purpose: work or study (Fig. 2 and Fig. 3). In this case, again, the nearest and biggest provinces reveal major flows, but with

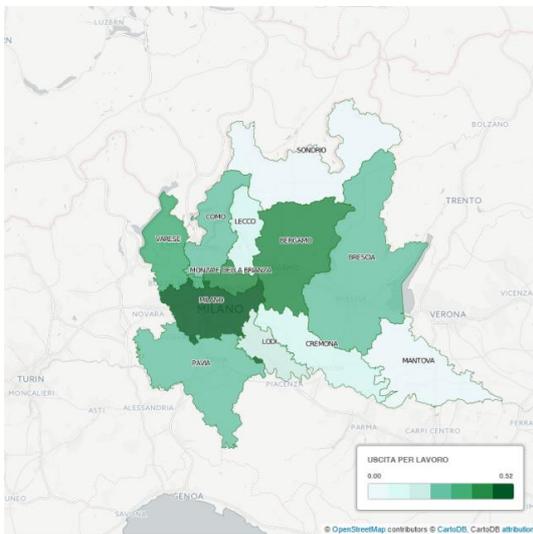


Fig. 2 Work outflow from Milan to provinces – ISTAT dataset

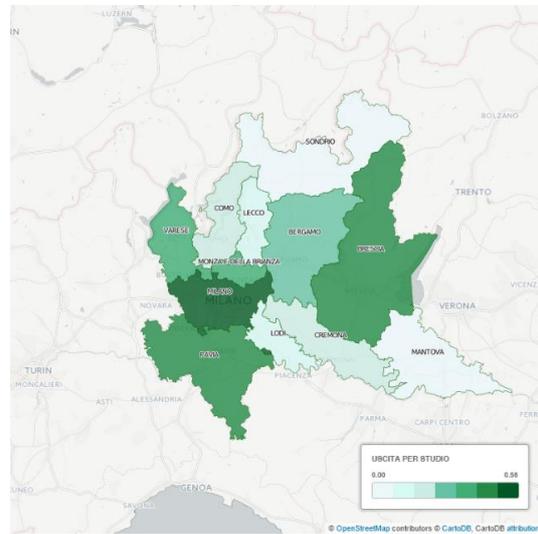


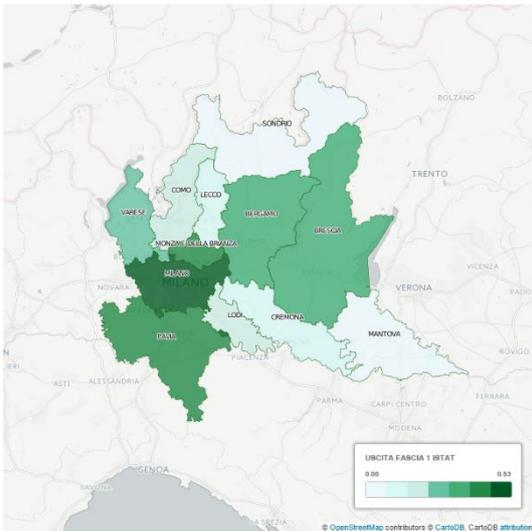
Fig. 3 Study outflow from Milan to provinces – ISTAT dataset

some changes in the rankings. For work purposes, the nearest provinces are the most linked to the city of Milan (Bergamo, Monza e Brianza and Varese), while for study you can notice how the universities in Brescia and Pavia make a big attraction for students coming from the city of Milan.

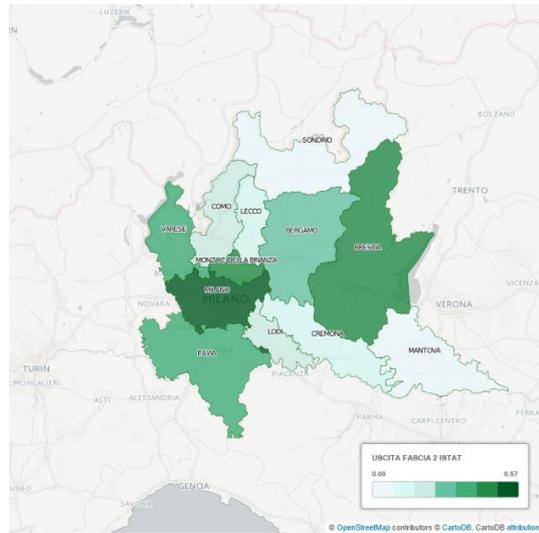
It is also possible to split results into four departure timeslots, as asked into the Census question:

- timeslot 1: before 7,15
- timeslot 2: from 7,15 to 8,14
- timeslot 3: from 8,15 to 9,14
- timeslot 4: after 9,14

Results regarding timeslot are shown in Figures from 4 to 7.

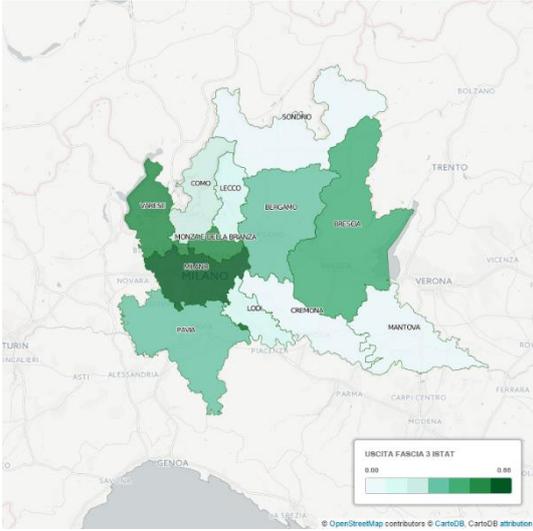


*Fig. 4 Outflow in timeslot 1 from Milan to provinces – ISTAT dataset*

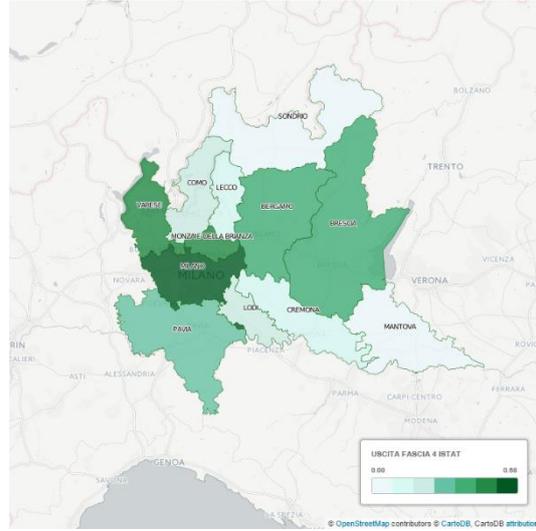


*Fig. 5 Outflow in timeslot 2 from Milan to provinces – ISTAT dataset*

In this case, it seems that the distance manages the rankings. In fact, the biggest provinces (Pavia, Brescia and Bergamo) appear darker early in the morning, probably due to the heavy traffic situation, while they are lighter in timeslots 3 and 4. For the provinces of Varese and Monza e Brianza, the situation seems not to change during the morning; this could be caused by the small distance that links them to the city of Milan and to different habits of commuters (traditional workers vs. people that work in offices).

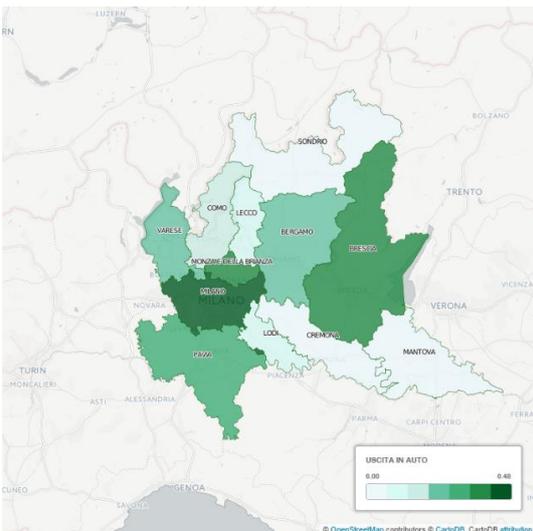


*Fig. 6 Outflow in timeslot 3 from Milan to provinces – ISTAT dataset*

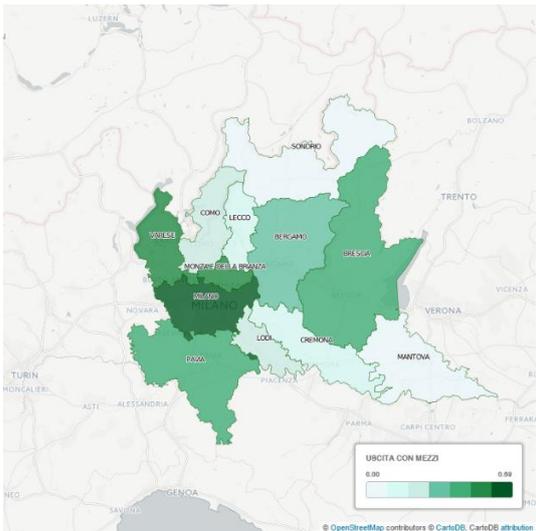


*Fig. 7 Outflow in timeslot 4 from Milan to provinces – ISTAT dataset*

The last filter that can be applied is the means of transportation used: car or other transport (Fig. 8 and Fig. 9).



*Fig. 8 Outflow by car from Milan to provinces – ISTAT dataset*



*Fig. 9 Outflow by other transport from Milan to provinces – ISTAT dataset*

The situation appears different only in some provinces: Brescia, Varese and Monza e Brianza. In particular, it seems that people are mostly driving to Brescia, while they rely on public transport to the other two provinces. This fact could be influenced by two causes: the distance that divide Brescia from the city of Milan, that could justify the choice of taking the car, and the not so efficient situation and links through public transport between the two cities.

### 2.4.1.2 Telecom Italia dataset

The Telecom Italia dataset represents the big data source of the analysis; if we follow the classification from the European Statistical System Committee (2014), we are dealing with the data source identified as “Mobile communication” in the form of mobile phone data, in particular Call Detail Records (in an aggregated way). This dataset is provided with data of November and December 2013. We use only weekdays of the four complete weeks of November and compute the average over 20 days in order to have a mean value comparable to the one from ISTAT.

In Fig. 10 we present the result of the general outflow. Compared to the ISTAT dataset, the situation here appears slightly different, as nearest provinces are the ones at the top of the ranking (Brescia is in pale green).

It is not possible here to split calls by commuting purpose. It is possible to split into timeslots as similar as possible to ISTAT ones. Data are originally grouped into slots of ten minutes, so we build the following timeslots:

- timeslot 1: 6,20 – 7,10
- timeslot 2: 7,20 – 8,10
- timeslot 3: 8,20 – 9,10
- timeslot 4: 9,20 – 10,10

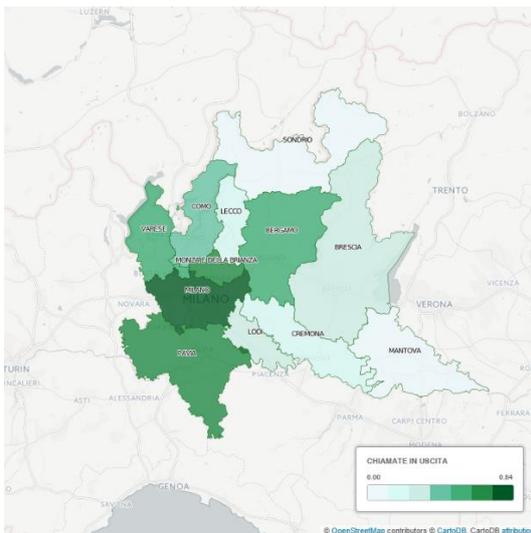
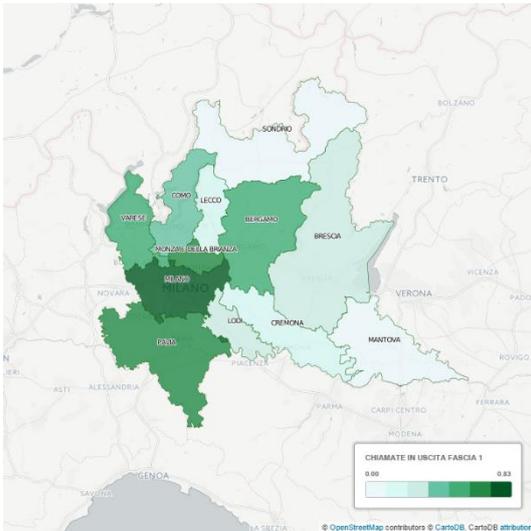
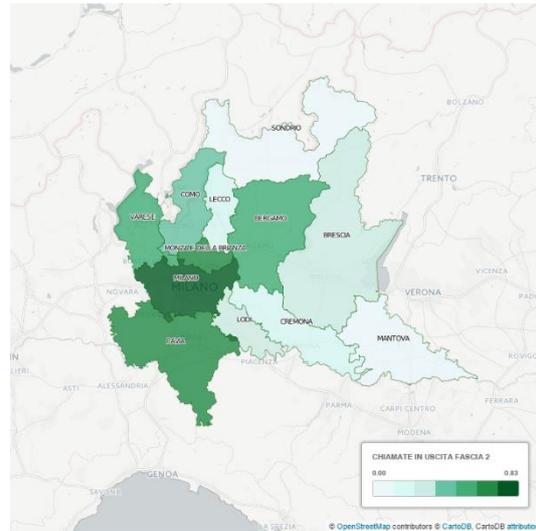


Fig. 10 General outflow from Milan to provinces – Telecom Italia dataset

Results regarding timeslots are shown in Figures from 11 to 14.

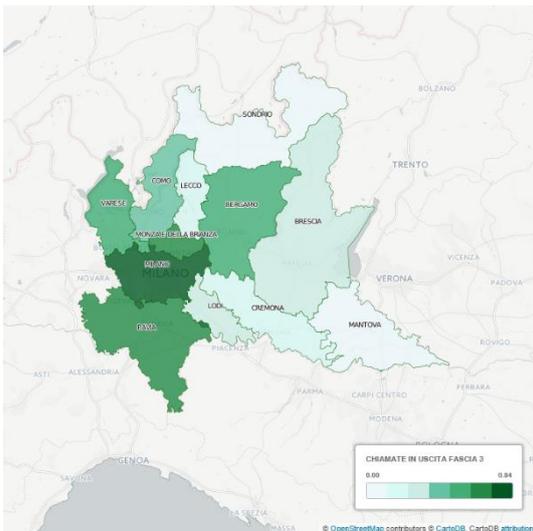


*Fig. 11 Outflow in timeslot 1 from Milan to provinces – Telecom Italia dataset*

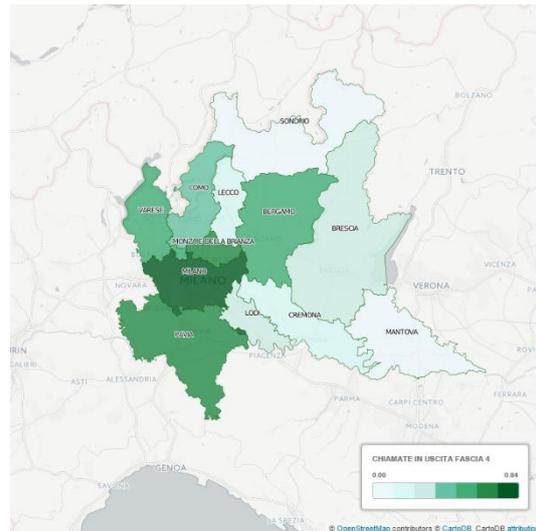


*Fig. 12 Outflow in timeslot 2 from Milan to provinces – Telecom Italia dataset*

As you can notice, there are no differences in the calling behaviour in the four timeslots.



*Fig. 13 Outflow in timeslot 3 from Milan to provinces – Telecom Italia dataset*



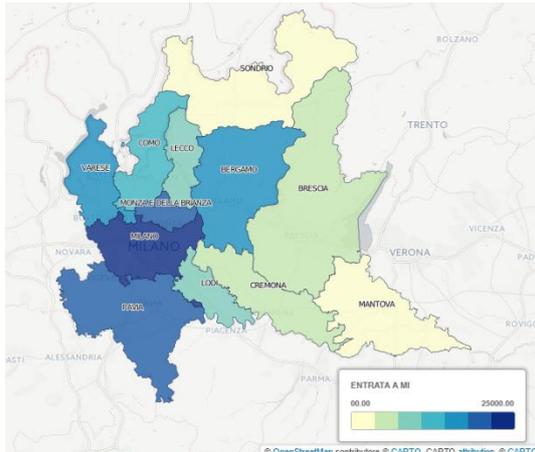
*Fig. 14 Outflow in timeslot 4 from Milan to provinces – Telecom Italia dataset*

## 2.4.2 Inflow maps

We created the same kind of maps for the inflow from the Lombardy provinces to the city of Milan, you can find them in the following figures.

### 2.4.2.1 ISTAT dataset

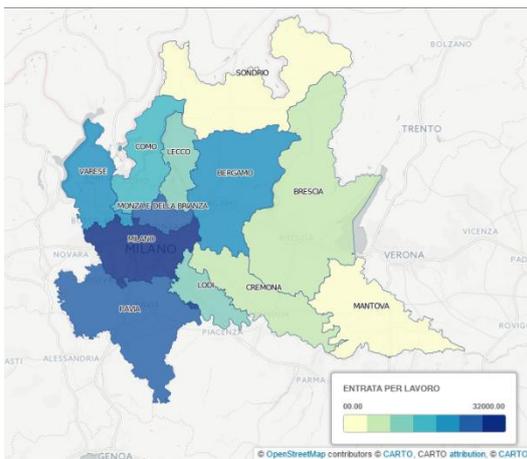
Let's start with maps created using the origin/destination matrix. Figure 15 shows the general inflow from Lombardy provinces to the city of Milan in ISTAT dataset.



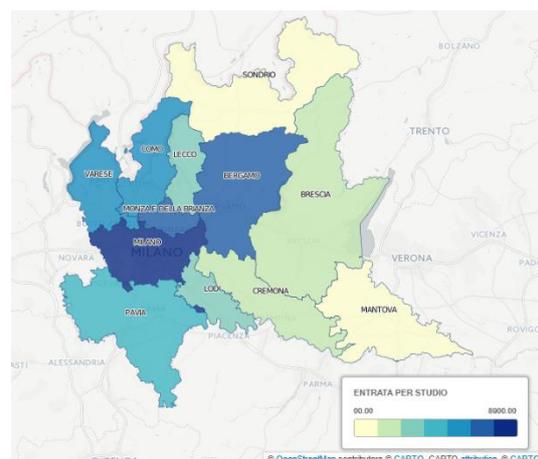
*Fig. 15 General inflow from provinces to Milan - ISTAT dataset*

Differently from the outflow situation, again we can notice how the nearest provinces attract most of the commuters, especially from Pavia, Monza e Brianza and Bergamo.

Then, we divided the inflow considering the purpose: work or study (Figures 16 and 17). The effect of the university, as for the outflow, is again visible: for study purpose, Bergamo and Monza e Brianza attract a lot of students, while Pavia does not (as there is a very famous university).



*Fig. 16 Work inflow from provinces to Milan – ISTAT dataset*



*Fig. 17 Study inflow from provinces to Milan – ISTAT dataset*

The following four maps show the rankings of provinces considering the abovementioned timeslots (Figures 18 to 21).

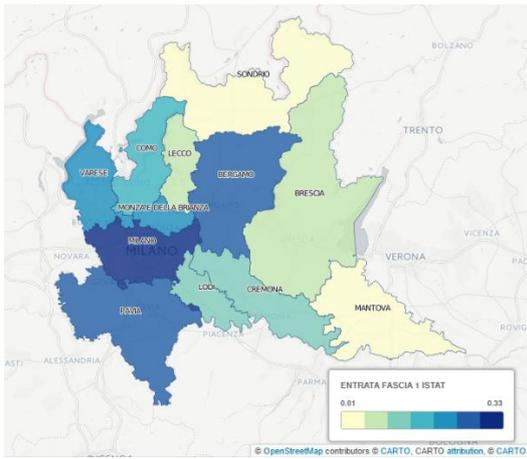


Fig. 18 Inflow in timeslot 1 from provinces to Milan – ISTAT dataset

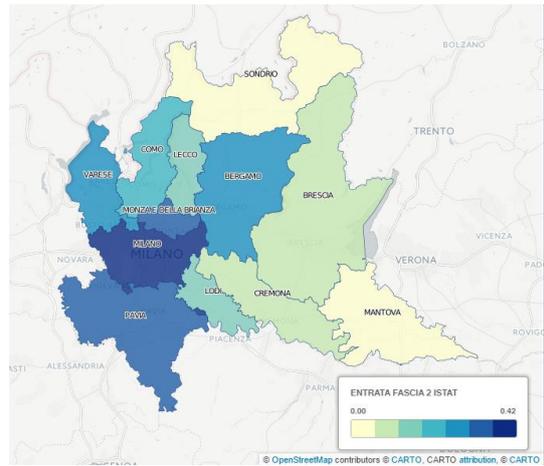


Fig. 19 Inflow in timeslot 2 from provinces to Milan – ISTAT dataset

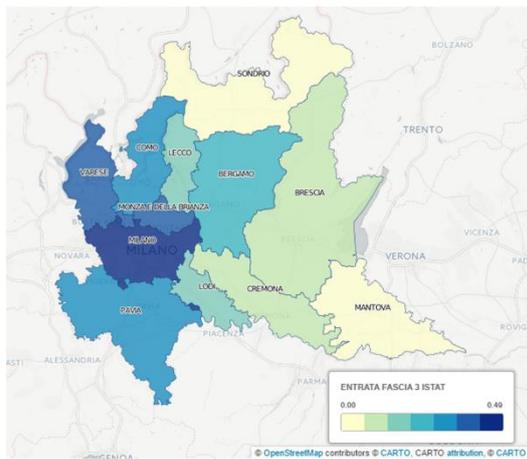


Fig. 20 Inflow in timeslot 3 from provinces to Milan – ISTAT dataset

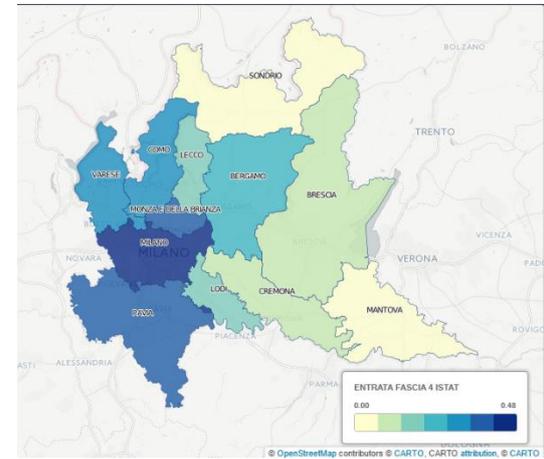


Fig. 21 Inflow in timeslot 4 from provinces to Milan – ISTAT dataset

As for the outflow, the timeslots reveal different behaviours from provinces; for example, Bergamo and Cremona appear darker early in the morning, probably due to heavy traffic, while Varese and Monza e Brianza show darker colours late in the morning.

Regarding the inflow, this time we decided to inspect a little bit about the purpose of the commute. In Figures 22 to 25, you can see the situation in the four timeslots for work inflow.

You can notice that the situation remains more or less stable in all the timeslots, except for Bergamo (darker early in the morning) and Monza e Brianza (lighter early in the morning). We can suspect that different kind of jobs create this steady situation, with differences in the time people start working and, consequently, leave home in the morning.



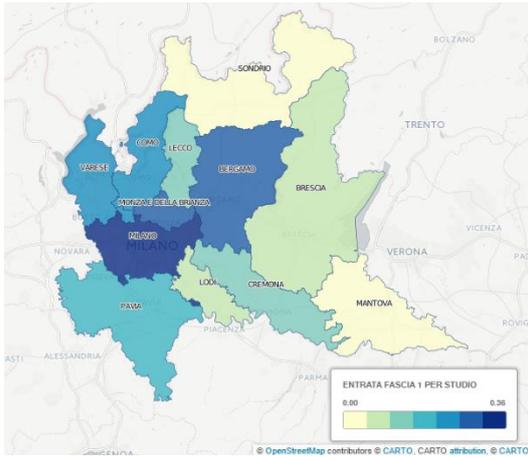


Fig. 26 Study inflow in timeslot 1 from provinces to Milan – ISTAT dataset

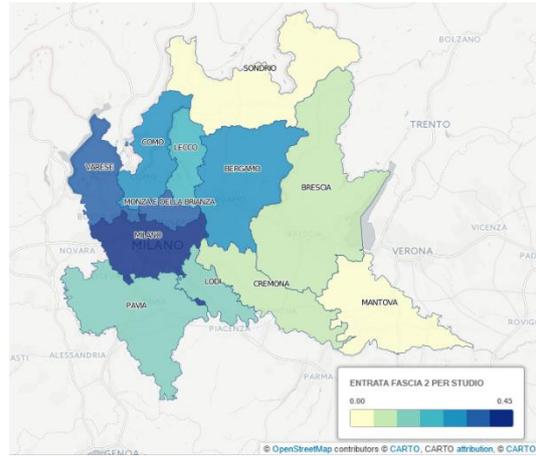


Fig. 27 Study inflow in timeslot 2 from provinces to Milan – ISTAT dataset

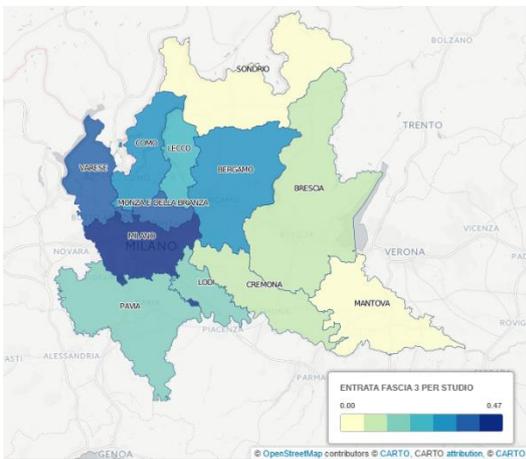


Fig. 28 Study inflow in timeslot 3 from provinces to Milan – ISTAT dataset

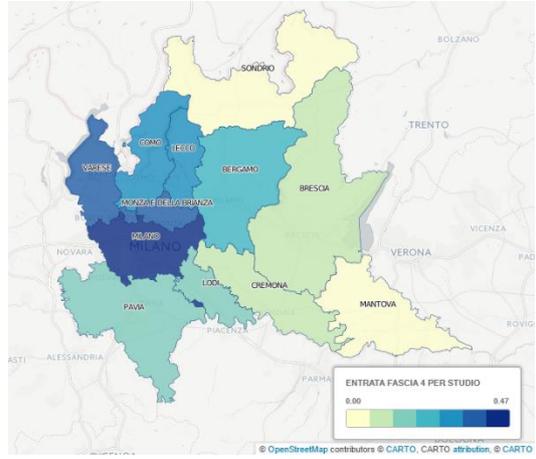


Fig. 29 Study inflow in timeslot 4 from provinces to Milan – ISTAT dataset

#### 2.4.2.2 Telecom Italia dataset

As for the outflow, we built maps for the general outflow of phone calls (Figure 30) as well as for the outflow split into timeslots (Figures 31 to 34). Data are originally grouped into slots of ten minutes, so we build the following timeslots:

- timeslot 1: 6,20 – 7,10
- timeslot 2: 7,20 – 8,10

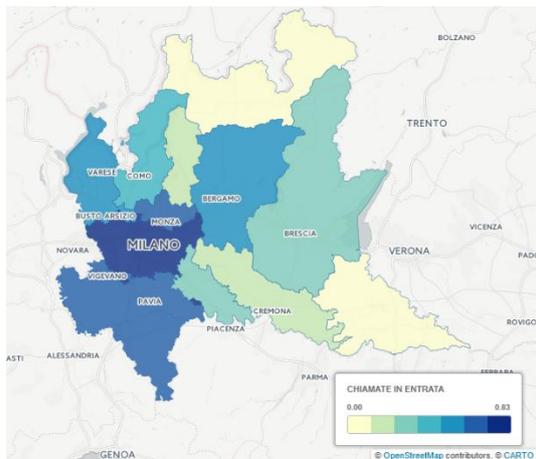
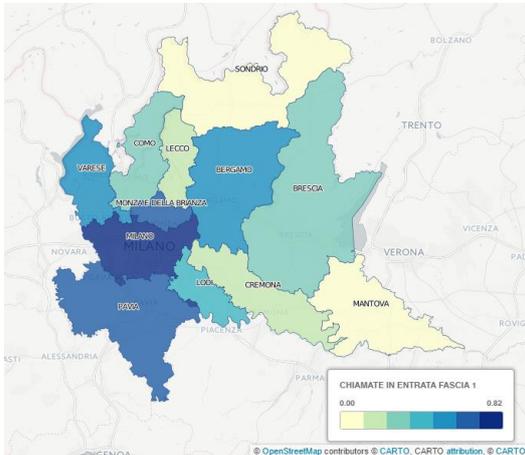


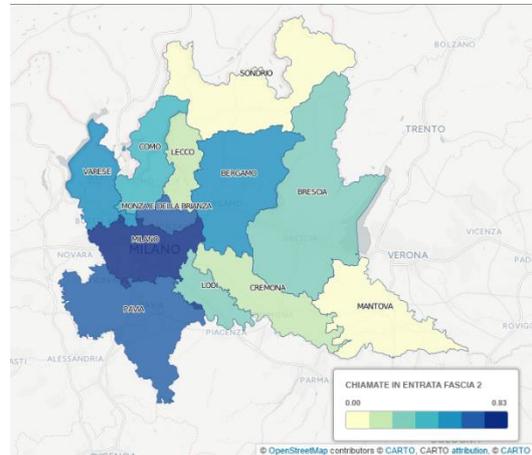
Fig. 30 General inflow from provinces to Milan – Telecom Italia dataset

- timeslot 3: 8,20 – 9,10
- timeslot 4: 9,20 – 10,10

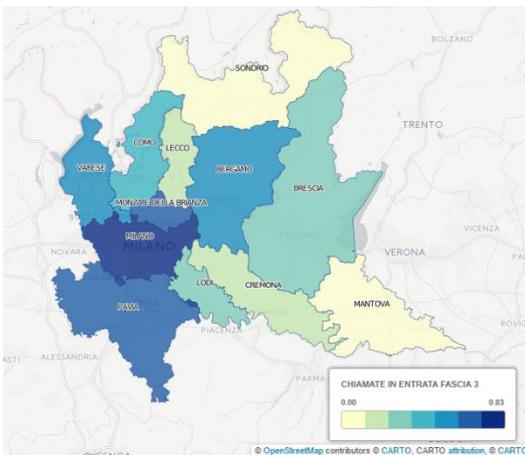
Considering the amount of incoming phone calls into the city, the situation is again driven by the distance: the nearest provinces are at the top of the ranking.



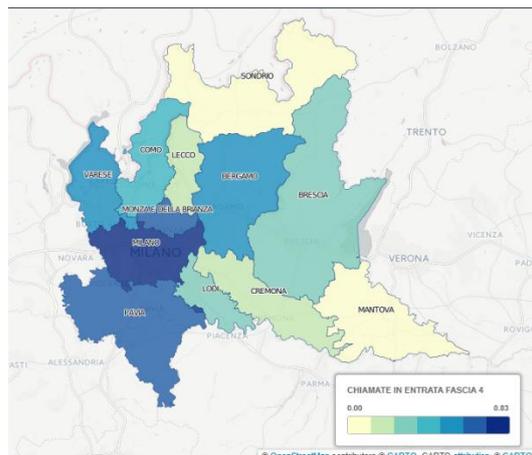
*Fig. 31 Inflow in timeslot 1 from Milan to provinces – Telecom Italia dataset*



*Fig. 32 Inflow in timeslot 2 from Milan to provinces – Telecom Italia dataset*



*Fig. 33 Inflow in timeslot 3 from Milan to provinces – Telecom Italia dataset*



*Fig. 34 Inflow in timeslot 4 from Milan to provinces – Telecom Italia dataset*

You can notice as no significant differences appear considering the inflow of phone calls during the four timeslots.

#### **2.4.2 Method of comparison of the two datasets**

After the individual analysis of each dataset, we proceed to the comparison of the two. Table 1 presents the codification of the Lombardy provinces provided by ISTAT. The codes are useful for results in tables and figures.

*Table 1 ISTAT Codification of Lombardy provinces*

<i>Province</i>	<i>ISTAT code</i>	<i>Province</i>	<i>ISTAT code</i>
Bergamo	16	Mantova	20
Brescia	17	Milano	15
Como	13	Monza e della Brianza	108
Cremona	19	Pavia	18
Lecco	97	Sondrio	14
Lodi	98	Varese	12

The first step in our analysis is represented by the construction of rankings of provinces, both in the Telecom Italia and in the ISTAT dataset. To explain the methodology, we will show in this paragraph only examples about the outflow.

Table 2 presents the comparison of rankings for each timeslot for the general outflow (each number presents a province, according to the codification of Table 1).

*Table 2 Provinces' ranking for general outflow for each dataset by timeslot*

<i>Timeslot 1</i>		<i>Timeslot 2</i>		<i>Timeslot 3</i>		<i>Timeslot 4</i>	
Telecom	Istat	Telecom	Istat	Telecom	Istat	Telecom	Istat
14	14	14	14	20	20	14	14
20	20	20	20	14	14	20	20
97	97	19	19	19	19	97	97
19	19	97	97	97	97	19	19
98	98	98	98	98	98	98	98
13	13	13	13	13	13	13	13
12	12	16	16	16	16	18	18
17	17	18	18	18	18	16	16
16	16	12	12	17	17	17	17
18	18	17	17	12	12	12	12
108	108	108	108	108	108	108	108
15	15	15	15	15	15	15	15

Comparing each position in the ranking in every timeslot (e.g. first position in Telecom Italia dataset in timeslot 1 vs. first position in ISTAT dataset in timeslot 1), we build another table (Table 3) in which an equal appears if the position in ISTAT



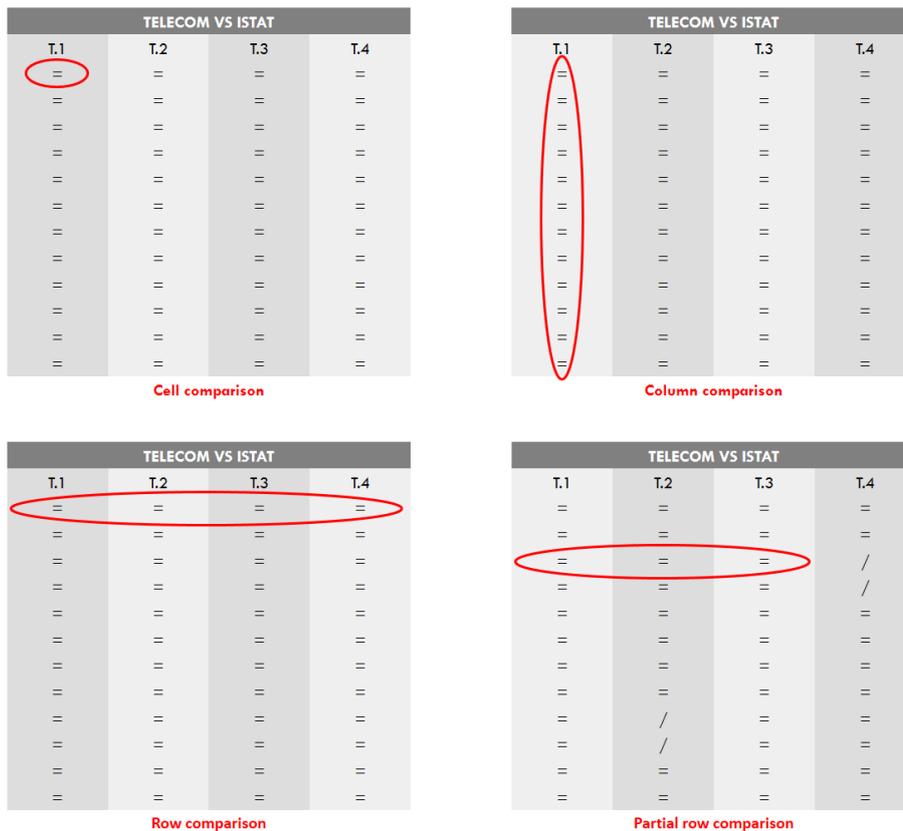


Fig. 35 Possible comparisons in general outflow tables

Similar table have been built also for each of the filters that can be applied on ISTAT dataset (work/study, means of transportation, work by means of transportation), comparing Telecom Italia dataset by the general outflow (the only filter that it is possible to apply). The match is not as perfect as it appears in the showed Table 3.

### 2.4.3 Results

The comparison is between Telecom Italia dataset and ISTAT dataset broken down by timeslots. For the former, we can only consider the general outflow, while for the latter different filters can be applied: the general outflow, the outflow divided based on the purpose (work and study), the general outflow by means of transportation (car and other transport), the work outflow by means of transportation (car and other transport). The abovementioned four types of comparison (cell, column, row, partial row) are performed in the following.

Outflow results (in terms of number of matches) are shown in Table 4. It is worth noting that a perfect match is found only in the general outflow. Rather high matches exist in the work outflow and in the outflow by car.

*Table 4 Outflow results (number of matches)*

<b>TELECOM VS ISTAT</b>	<b>CELL (48)*</b>	<b>COLUMN (4)*</b>	<b>ROW (12)*</b>	<b>PARTIAL ROW (12)*</b>
<b>OUTFLOW</b>	48	4	12	12
<b>WORK OUTFLOW</b>	44	2	8	12
<b>STUDY OUTFLOW</b>	16	0	1	2
<b>OUTFLOW BY CAR</b>	40	0	7	6
<b>OUTFLOW BY OTHER TRANSPORT</b>	28	0	2	3
<b>WORK OUTFLOW BY CAR</b>	31	0	4	10
<b>WORK OUTFLOW BY OTHER TRANSPORT</b>	22	0	2	4

\*maximum number of matches

A similar analysis with the same filters is performed for inflow. The results are presented in Table 5.

*Table 5 Inflow results (number of matches)*

<b>TELECOM VS ISTAT</b>	<b>CELL (48)*</b>	<b>COLUMN (4)*</b>	<b>ROW (12)*</b>	<b>PARTIAL ROW (12)*</b>
<b>INFLOW</b>	18	0	3	5
<b>WORK INFLOW</b>	22	0	1	6
<b>STUDY INFLOW</b>	13	0	1	2
<b>INFLOW BY CAR</b>	23	0	1	4
<b>INFLOW BY OTHER TRANSPORT</b>	21	0	3	4
<b>WORK INFLOW BY CAR</b>	23	1**	1	4
<b>WORK INFLOW BY OTHER TRANSPORT</b>	21	0	1	5

\*maximum number of matches

\*\*two missing matches

The situation is very different from the previous one; in general, very low matches appear. Anyway, some matches arise in the cell comparison, especially in inflow by car,

work inflow by car, work inflow. The lack of matches could be influenced by the very heavy traffic situation of people commuting every day to the city of Milan.

The findings of this case study show some opportunities of the use of this source of big data, like providing increased information on the social exchanges between provinces. In particular, a vision of the exchanges in a physical (people mobility) perspective and in a communication (measured through mobile phone data) perspective is highlighted. If the matching between the two datasets is sufficiently satisfactory, big data could represent a useful source in Official Statistics, cheap and up-to-date.

Furthermore, the absolute value that we adopted in our analysis is an interesting measure of the size of the flows and of the communications. Standardized values, i.e. with respect to the whole population (in the case of work commuting, with respect to the work age range) could highlight other interesting aspects regarding communication and mobility behaviours. Moreover, the possibility to carry out an analysis of the phone calls that takes into account also the user profile, could give a remarkable knowledge contribution even though in full compliance with privacy issues. Unfortunately, it is not possible to handle this kind of data due to privacy constraints in releasing.

In the future, it could be evaluated integrating the considered data in Official Statistics. For instance, these data could be useful to create indicators of social exchange and interaction. They could be used as a base for a proxy of the demographic mobility in the time interval of the Census data collection. Our study is a preliminary feasibility analysis; further analyses will be planned. Our analysis has some limits which further studies could try to overcome:

- the two flows (people and phone calls) have different aims: Telecom Italia dataset does not specify a specific purpose (so it contains every possible aim of the communication), while ISTAT dataset only contains flows for work and study purposes;
- for a more extended users' coverage, other providers should be considered. This could help in evaluating the stability of the behaviour across users of different providers, i.e. in understanding if our findings could apply to target population; Telecom Italia dataset only contains traditional phone calls, other type of calls (i.e. Skype, WhatsApp), which are now very spread, are not considered;

- Telecom Italia dataset only contains province reference, it would be useful to have municipalities' references in order to do a better match with ISTAT data and to map commuting patterns;
- The ISTAT dataset also contains other variables, such as gender and age, so further inspections on these series would be needed, in order to find some common patterns with mobile phone data.

## **2.5 Conclusions**

Big data potentialities for Official Statistics need a huge amount of experimentation and of economic statistics studies in order to set up a suitable metadata framework and to evaluate Veracity, Validity and Value of the considered big data. Surely, deeper insight in quality of big data and in the variety of aspects and sources which could be integrated to setup their potential use as a statistical information is needed.

The overview of potentiality and problems presented in our paper highlights most critical research points and the present case study shows some innovative ideas on how to go through the tentative use of big data in Official Statistics.

Some potentialities seem to be expected. In particular, our case study shows how mobile phone data could be investigated with respect to mobility in Official Statistics and highlights how these data could catch jointly social communication aspects and physical mobility aspects. Obviously further research is needed; for instance: more detailed analyses on similarities and differences between the two datasets.

Moeover, the search for more possible data is to be considered in the comparison and the identification of alternative case studies on big data analysis.

## **2.6 Acknowledgements**

The authors acknowledge financial support by the ex 60% University of Bergamo, Biffignandi grant.



## **3 CASE STUDY: PHONE CALLS AND MOBILITY IN THE LOMBARDY REGION**

### **3.1 Introduction**

The aim of our case study is trying to put in a unique interpretative framework one traditional statistical source (2014 Lombardy Origin/destination) and one typical kind of big data (2014 1st Telecom Italia Big Data Challenge datasets) in order to evaluate some informative potentialities of this approach. In particular, we try to verify if it is possible to model the daily commuters' mobility in the Lombardy region using a mobile phone dataset. Origin/destination matrix is released every ten years, whereas Telecom Italia data (in the form of Call Detail Recors, CDRs) are collected continuously. It is evident that identifying the informative potentialities of big data is a great challenge for understanding social trends over short run.

This application represents an update of the experiment carried out in the paper "From Big Data to Information: Statistical Issues Through a Case Study" by Biffignandi and Signorelli, accepted for publication in the Springer Series "Studies in Classification, Data Analysis, and Knowledge Organization". In that paper, we used the Origin/destination matrix produced by Istat as an Official Statistics source, while this time we are using its integrated version created by Lombardy region. Moreover, we used the two sources of data to build some maps of the rankings of provinces and to compare those rankings in the two datasets. In this case, we decided to do a step further in the analysis, not building maps, but studying correlations among series, identifying possible linear and non-linear regressions and building, again, rankings of provinces.

## 3.2 Description of the datasets

As just mentioned, the case study is based on the use of two datasets, a big data source and a more traditional source (in this case, corrected with some other data):

- the 2014 Lombardy Origin/destination matrix, which has been built using a transportation model that integrated the results of a survey (February – May 2014) with data from the 15th Population and housing census and contributions from local entities and stakeholders from mobility sector. This dataset contains data on the number of commutes between municipalities – or inside the same municipality – classified by:
  - means of transportation: car (as driver), car (as passenger), motorcycle, public transport on iron (trains, metro, tram), public transport on wheels (bus and autobuses), bicycle, on foot (only if the journey takes more than 10 minutes), and a residual category called ‘others’;
  - reason of the commute: work, occasional, business, return, study.

The data are given in hourly slots. The spatial aggregation is the one given by the original dataset, the Origin/Destination matrix by the 2011 Census, and it is represented by Italian municipalities. This experiment is focused on the Lombardy region. This dataset aims at represent an average working day of people aged 14 years and older.

- One of the 2014 1st Telecom Italia Big Data Challenge datasets. In particular, we use the one named “Telecommunications - MI to Provinces”, which provides information regarding the level of interaction between the areas of the city of Milan and the Italian provinces. The level of interaction between an area of Milan and a province is given by a number; this represents the proportion of calls issued from the area of Milan to the province (and viceversa). For each area, the dataset contains two numbers: one representing the proportion of inflow telephone traffic and the other one representing the proportion of the

outflow. The spatial aggregation of the dataset is the Milano GRID squares<sup>9</sup> and the Italian provinces. The values are aggregated in timeslots of ten minutes.

Our analysis is limited to Lombardy region, divided into twelve administrative provinces. We build commuting patterns concerning the city of Milan and all provinces (excluded flows from the city of Milan to the city of Milan itself) considering the 24 hours of the day.

According to the 2014 Report by the Italian Media Safeguards Authority (AGCOM), Telecom Italia is the market leader in mobile telecommunications with an amount of 33,2% in 2013 (the year our data refer to), of which 29,8% are residential and 47,9% business. A good users' coverage is guaranteed. Obviously, the findings might hold for all residential users under the hypothesis that the consuming behaviour does not differ with respect to the telephone provider. It seems reasonable to suppose that the consuming behaviour does not differ with respect to the telephone provider.

### **3.3 Preliminary analysis**

In order to make the two datasets comparable, a preliminary manipulation of the two is necessary. This first analysis is performed using RStudio.

First of all, the experiment has been divided into outflow and inflow; the former considers all the commutes from the city of Milan to all the twelve Lombardy provinces, while the latter considers the opposite flow, from the provinces to the city of Milan. This filter can be applied on both data sources.

After this, the analysis on the datasets is different, as on the Telecom Italia we can just aggregate the observations by province and by time slot. As the time slots contain 10 minutes' data, it is necessary to merge them six by six in order to perform the comparison with the O/D matrix and have hourly counts.

On the Lombardy matrix, indeed, other filters may be applied. In fact, it is possible to create a dataset that highlights the means of transport used by province and another one

---

<sup>9</sup> Some of the datasets of the 1<sup>st</sup> Telecom Big Data Challenge referring to the Milano urban area are spatially aggregated using a grid composed by 10.000 cells over the municipality of Milan.

that considers the aim of the commute by province. We cross these two to have a more detailed dataset that shows aim and means of transport by province.

After this preliminary analysis, we were able to build the final datasets, that contain the filtered variables from the Lombardy matrix and the Telecom Italia data. In summary, our analysis contains the following datasets:

- inflow by means of transport;
- inflow by reason of commute;
- inflow by means of transport and reason of commute;
- outflow by means of transport;
- outflow by reason of commute;
- outflow by means of transport and reason of commute.

On these datasets, we performed the following analysis:

- analysis of correlations between variables from Lombardy matrix and Telecom Italia data by province;
- identification of possible linear regression between variables from Lombardy matrix and Telecom Italia data by province;
- identification of possible non-linear regression between variables from Lombardy matrix and Telecom Italia data by province;
- creation of ranks of provinces on the Lombardy matrix by means of transport, reason of commute and the combination of the two, then comparison with Telecom Italia dataset and identification of equal positions in rankings.

These analyses have been carried out using the SAS Enterprise Guide software.

### ***3.3.1 Correlations***

We performed the analysis of correlations between variables from Lombardy matrix and Telecom Italia data by province for each of the six datasets.

Regarding the outflow by means of transport, the only two variables that highlight correlations are Sondrio and Mantova using car as driver. It is worth noting that,

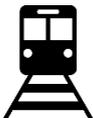
together with Brescia, these are the only Lombardy provinces which do not border with Milano.

Considering the outflow by reason of commute, business and occasional reasons correlate with Telecom Italia series for each province (with the only exception of Sondrio for business).

Going deeper into the analysis of the outflow, crossing two filters (means of transport and aim), we find that the variables that show correlations are again referred to business and occasional aims, together with the use of different means of transport. In each province, different means of transport show correlations, probably depending on the distance of the province itself to the city of Milano. In fact, as shown in Table 1, the nearest provinces show correlations with means of transport like bicycle and on foot, while the most distant ones highlight only car as a driver and public transport on iron.

*Table 1 – Correlations for outflow by reason of commute and means of transport*

		<b>Occasional</b> 	<b>Business</b> 
<b>On foot</b> 		Milano	Milano
<b>Bicycle</b> 		Milano Monza e Brianza Pavia	Milano Monza e Brianza
<b>Motorcycle</b> 		Bergamo Como Cremona Lecco Lodi Milano Monza e Brianza Pavia Varese	Cremona Lodi Milano Monza e Brianza Pavia Varese

<p><b>Car as a driver</b></p> 	<p>Bergamo Brescia Como Cremona Lecco Lodi Mantova Milano Monza e Brianza Pavia Sondrio Varese</p>	<p>Bergamo Brescia Como Cremona Lecco Lodi Mantova Milano Monza e Brianza Pavia Varese</p>
<p><b>Car as a passenger</b></p> 	<p>Bergamo Brescia Como Cremona Lecco Lodi Milano Monza e Brianza Pavia Varese</p>	<p>Brescia Lodi Milano Monza e Brianza Pavia Varese</p>
<p><b>Public transport on wheels</b></p> 	<p>Bergamo Como Cremona Lecco Lodi Milano Monza e Brianza Pavia Varese</p>	
<p><b>Public transport on iron</b></p> 	<p>Bergamo Brescia Como Cremona Lecco Lodi Mantova Milano Monza e Brianza Pavia Sondrio Varese</p>	<p>Bergamo Brescia Como Cremona Lodi Mantova Milano Monza e Brianza Pavia Varese</p>

Considering the inflow by means of transport, only motorcycle shows correlations with mobile phone data, but not in all provinces, only in Bergamo, Brescia, Como, Milano, Mantova, Sondrio and Varese.

The inflow by reason of commute, as for the outflow, highlights correlations for occasional (in all provinces) and business reasons (except in Monza e Brianza and Sondrio).

As highlighted before for the outflow, the analysis performed by reason of commute and means of transport shows correlations only for business and occasional reasons, and for different means of transport, depending, we suppose, on the distance of the province from the city of Milano. Results are shown in Table 2.

It is worth noting that Monza e Brianza and Sondrio are the only provinces that do not show correlations for business. These are emblematic cases, as Monza e Brianza is very close to Milano, while Sondrio is one of the most distant provinces.

Table 2 – Correlations for inflow by reason of commute and means of transport

		<b>Occasional</b>	<b>Business</b>
			
<b>On foot</b>		Milano	Milano
<b>Bicycle</b>		Milano Monza e Brianza Pavia	Milano Monza e Brianza Pavia
<b>Motorcycle</b>		Bergamo Brescia Como Cremona Lecco Lodi Milano Monza e Brianza Pavia Varese	Como Cremona Lecco Lodi Milano Monza e Brianza Pavia Varese

<p style="text-align: center;"><b>Car as a driver</b></p> 	<p>Bergamo Brescia Como Cremona Lecco Lodi Mantova Milano Monza e Brianza Pavia Sondrio Varese</p>	<p>Bergamo Brescia Como Cremona Lecco Lodi Mantova Milano Pavia Varese</p>
<p style="text-align: center;"><b>Car as a passenger</b></p> 	<p>Bergamo Brescia Como Cremona Lecco Lodi Milano Monza e Brianza Pavia Varese</p>	<p>Brescia Lodi Milano Pavia Varese</p>
<p style="text-align: center;"><b>Public transport on wheels</b></p> 	<p>Bergamo Brescia Como Cremona Lecco Lodi Milano Monza e Brianza Pavia Varese</p>	<p>Brescia Pavia Varese</p>
<p style="text-align: center;"><b>Public transport on iron</b></p> 	<p>Bergamo Brescia Como Cremona Lecco Lodi Mantova Milano Monza e Brianza Pavia Sondrio Varese</p>	<p>Bergamo Brescia Como Cremona Lecco Lodi Mantova Milano Pavia Varese</p>

### 3.3.2 *Linear regressions*

Our analysis continues considering possible linear regressions between each variable of the Lombardy matrix and the Telecom Italia dataset variables by province. We will not include the variables that already showed correlations into this step.

The aim of this analysis is to predict the variable of the Origin/destination matrix starting from the Telecom Italia dataset; in particular:

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

where

- $y_i$  is the dependent variable, extracted from the OD dataset;
- $x_{i1}, x_{i2}, \dots, x_{ip}$  are the independent variables (in our case only one variable regarding the phone calls from the Telecom Italia dataset);
- $\boldsymbol{\beta}$  is a p-dimensional parameter vector;
- $\varepsilon_i$  is called the error term, disturbance term, or noise.

The variables of the outflow by means of transport dataset does not appear to be significant parameters in the linear regression model, and the same happens considering the dataset by reason of commute.

Considering the inflow datasets, the situation does not change, as no linear regressions appear.

### 3.3.3 *Non-linear regressions*

The experiment continues with the study of non-linear regressions. As for the linear regression, the model will be composed by a variable of the Origin/destination matrix as dependent variable and the Telecom Italia variable as independent variable.

If we consider the dataset of the outflow by means of transport, the Telecom variable represents a significant parameter of the model in several cases summarized in Table 3.

If we consider the dataset by reason of commute, the Telecom Italia variables represent a significant parameter in the model and show a high  $R^2$  in all provinces considering the aim 'return' (Table 4).

Table 3 – Means of transport for which Telecom Italia variable represent a significant parameter for outflow

Province	Means of transport							
	Car as driver	Car as passenger	Motorcycle	Public transport on iron	Public transport on wheels	Bicycle	On foot	Other
Bergamo	X	X	X	X	X	X		X
Brescia	X	X	X	X				X
Como	X	X	X	X	X	X		X
Cremona	X	X	X	X	X	X		X
Lecco	X	X	X	X	X	X		X
Lodi	X	X	X	X	X	X		X
Mantova	X	X			X			X
Milano	X	X		X	X	X	X	X
Monza e Brianza	X	X	X	X	X	X	X	X
Pavia	X	X	X	X	X	X		X
Sondrio	X	X	X					X
Varese	X	X	X	X	X	X	X	X

Table 4 – Aims of commute for which Telecom Italia variable represent a significant parameter for outflow

Province	Reason of commute				
	Work	Occasional	Business	Return	Study
Bergamo				X	
Brescia				X	
Como				X	
Cremona				X	
Lecco				X	
Lodi				X	
Mantova				X	
Milano				X	
Monza e Brianza				X	
Pavia				X	
Sondrio				X	
Varese				X	

Regarding the inflow, the Telecom Italia variable appears as a significant parameter only for some means of transport in the provinces (Table 5), in particular:

- car as a driver for Brescia and Sondrio;
- car as a passenger for Mantova;

- motorcycle for Cremona, Lecco, Monza e Brianza and Pavia.

*Table 5 – Means of transport for which Telecom Italia variable represent a significant parameter for inflow*

Province	Means of transport							
	Car as driver	Car as passenger	Motorcycle	Public transport on iron	Public transport on wheels	Bicycle	On foot	Other
Bergamo								
Brescia	X							
Como								
Cremona			X					
Lecco			X					
Lodi								
Mantova		X						
Milano								
Monza e Brianza			X					
Pavia			X					
Sondrio	X							
Varese								

The situation, indeed, is the same as the outflow if we look at the aim of the commute: in all provinces, the Telecom Italia variables represent a significant parameter in the model considering the aim ‘return’ (Table 6).

Table 6 – Aims of commute for which Telecom Italia variable represent a significant parameter for inflow

Province	Reason of commute				
	Work	Occasional	Business	Return	Study
Bergamo				X	
Brescia				X	
Como				X	
Cremona				X	
Lecco				X	
Lodi				X	
Mantova				X	
Milano				X	
Monza e Brianza				X	
Pavia				X	
Sondrio				X	
Varese				X	

### 3.3.4 Ranks

The way to compare the two datasets (Telecom Italia and O/D matrix) is the same we performed in the previous paper using rankings. We will now briefly resume the method.

For each dataset, the provinces' ranking for each timeslot have been computed. We then compare the rankings, in order to find out whether the placement in one dataset in a specific timeslot is the same as in the other dataset. Putting the results in a table, it is possible to perform three kinds of comparison (due to the lower number of matches, this time we are not considering the *partial row comparison*):

- *cell comparison*: how many equals appear over the cells of the table;
- *column comparison*: if a column contains only equals. This means that all provinces behave in the same way in that specific timeslot;
- *row comparison*: if all equals appear in a row. This means that the specific province behaves in the same way in the two datasets in all the twenty-four timeslots.

The same analysis has been performed for each of the filters that can be applied on ISTAT dataset (work/study, means of transportation, work by means of transportation),

comparing Telecom Italia dataset by the general outflow (the only filter that it is possible to apply).

The comparison is between Telecom Italia dataset and ISTAT dataset broken down by timeslots. For the former, we can only consider the general outflow, while for the latter different filters can be applied: the general outflow, the outflow divided by means of transport (car as driver, car as passenger, motorcycle, public transport on iron (trains, metro, tram), public transport on wheels (bus and autobuses), bicycle, on foot (only if the journey takes more than 10 minutes), and a residual category called ‘others’), by aim of the commute (work, occasional, business, return, study) and by a combination of these two. The abovementioned three types of comparison (*cell*, *column*, *row*) are performed in the following. Outflow results (in terms of number of matches) are shown in Table 7, while for the inflow are in Table 8.

*Table 7 Outflow results (number of matches)*

<b>TELECOM VS ISTAT</b>	<b>CELL (288)*</b>	<b>COLUMN (24)*</b>	<b>ROW (12)*</b>
<b>OUTFLOW</b>	133	0	2
<b>OUTFLOW BY CAR AS A DRIVER</b>	156	0	2
<b>OUTFLOW BY CAR AS A PASSENGER</b>	160	0	2
<b>OUTFLOW BY MOTORCYCLE</b>	121	0	2
<b>OUTFLOW BY PUBLIC MEANS ON IRON</b>	112	0	1
<b>OUTFLOW BY PUBLIC MEANS ON WHEELS</b>	109	0	2
<b>OUTFLOW BY BICYCLE</b>	124	0	2
<b>OUTFLOW ON FOOT</b>	71	0	1
<b>OUTFLOW BY OTHER MEANS</b>	92	0	0
<b>WORK OUTFLOW</b>	137	0	2
<b>OCCASIONAL OUTFLOW</b>	97	0	2
<b>BUSINESS OUTFLOW</b>	67	0	0
<b>RETURN OUTFLOW</b>	110	0	0
<b>STUDY OUTFLOW</b>	72	0	0

\*maximum number of matches

Table 8 Inflow results (number of matches)

<b>TELECOM VS ISTAT</b>	<b>CELL (288)*</b>	<b>COLUMN (24)*</b>	<b>ROW (12)*</b>
<b>INFLOW</b>	157	0	1
<b>INFLOW BY CAR AS A DRIVER</b>	157	0	1
<b>INFLOW BY CAR AS A PASSENGER</b>	126	0	1
<b>INFLOW BY MOTORCYCLE</b>	109	0	1
<b>INFLOW BY PUBLIC MEANS ON IRON</b>	118	0	1
<b>INFLOW BY PUBLIC MEANS ON WHEELS</b>	111	0	1
<b>INFLOW BY BICYCLE</b>	115	0	0
<b>INFLOW ON FOOT</b>	72	0	1
<b>INFLOW BY OTHER MEANS</b>	106	0	1
<b>WORK INFLOW</b>	121	0	1
<b>OCCASIONAL INFLOW</b>	147	0	1
<b>BUSINESS INFLOW</b>	73	0	0
<b>RETURN INFLOW</b>	83	0	0
<b>STUDY INFLOW</b>	114	0	0

\*maximum number of matches

It is worth notice that for the outflow nor for the inflow, no *column* match appears. The comparison performed over *rows* is very low, as it shows a maximum of 2 for the outflow and 1 for the inflow, over a maximum of 12.

The comparison by *cell* appears slightly better, even if only in a few cases it exceeds 50% of possible matches. This happens in the case of outflow by car (in both possibilities, as driver and as passenger) and in three cases of inflow: general, by car as a driver and occasional.

In the introduction of the case study we stated the comparison of the combination of purposes, but due to the very low number of matches, we decided not to go over in the analysis. Going into more detailed flows seems not giving more to the results.

### **3.4 Conclusions**

The results that can be extracted from this case study appear limited, first considering one of the data sources. This modified version of the Origin/destination matrix was produced by the Lombardy region, but the exact methodology was not released. In particular, we do not have details about the survey that was carried out and about the weights that were given to the different kind of data (Census and questionnaire results) that compose the matrix.

On the other hand, we decided to perform different analysis with respect to the previous case study, without creating maps but considering correlations and regressions. These new steps could open to new possibilities for the study using the Census O/D matrix. The nice characteristic of the modified O/D matrix was the fact of having a 24-hours coverage, something not available from Census data. A possibility for the future could be an improvement of the questions about commuting pattern in the Census, not limiting the possible response to just a few hours in the morning. Moreover, it would be also useful to have an idea of evening flows, regarding people that go home after work/study.

As already mentioned for the previous study, the analysis is limited also because it only considers CDRs, which just record events such as calls (as in this case) or text messages. It would also be important to catch other new communication channels, such as apps that allow to call/video call. The evolution of this analysis should be done leaving CDRs and using pings, which allow to locate a mobile phone without having to wait for an event (such as a call) to happen. Through this system, it is way too easy to map and observe real moving patterns. The only problem, again, remains the access to the data, especially for privacy concerns.



## 4 WHAT ATTRACTS TOURISTS WHILE PLANNING FOR A JOURNEY? AN ANALYSIS OF THREE CITIES THROUGH WIKIPEDIA PAGE VIEWS

Serena Signorelli<sup>1,2</sup>, Fernando Reis<sup>2</sup>, Silvia Biffignandi<sup>1</sup>

<sup>1</sup> University of Bergamo  
(email: serena.signorelli@unibg.it, silvia.biffignandi@unibg.it)

<sup>2</sup> EUROSTAT  
(email: Serena.SIGNORELLI@ec.europa.eu, Fernando.REIS@ec.europa.eu)

**ABSTRACT:** 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 is Wikipedia data. As a free and open source of information, more and more people use it to get informed about different topics, from medicine to history, passing through geography and politics and many more. One common use of this source is to gain information about places to visit during day trips, journeys and/or holidays. It is worth keeping into consideration this behaviour, as it could represent a hint for the different cultural interests that people have in visiting a certain place. Just as Wikipedia is openly available, also page views data are made available by the Wikimedia Foundation, and this is the source of information that is used in this experiment. The aim is 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 potentiality of building some lead indicators is another issue to be explored. The analysis will be performed at a city level, considering all the points of interest of the area (culture, heritage, athletic, nature, leisure, etc.). We decided to focus our attention on three European cities: Barcelona (Spain), Vienna (Austria) and Bruges (Belgium). The choice fell on these cities as they have different sizes and characteristics. We decided to perform the experiment starting from Wikidata, the collaboratively edited knowledge base operated by the Wikimedia Foundation. We will perform a query about all Wikidata items related to the area through geo-coordinates, and from each item we will extract all the Wikipedia articles in the languages that we are considering in the analysis. The attention will then be moved 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 looking which types of points of interest are more popular. The research will continue with the analysis of time series of the page views, in order to study correlations among them and to identify possible factors that drive tourism to that specific area. We will evaluate the possibility of predicting tourists' flows through them. The benefit of this approach is new statistical developments framed by existing tourism Official Statistics directed to better guide policy making in the tourism sector with higher detail and timeliness than it is currently possible. It will also assess the possibility of building some lead indicators, and it represents a tool to better understand the effective usage of Wikipedia by individuals as a cultural source while planning journeys. Given the worldwide dimension of Wikipedia, the methodology and the outcomes of this experiment could be promising for many users and researchers across the globe.

**KEYWORDS:** 'big data', 'Wikidata', 'Wikipedia page views', 'tourism', 'Official Statistics'

## 4.1 Introduction

Before going into the details of the experiment, it is worth giving a short introduction on literature on the use of Wikipedia article traffic statistics. Page view statistics is a tool available for Wikipedia pages, which allows to know how many people have visited an article during a given time period (usually hourly counts). Some applications have been carried out in the last few years.

Munzert (2015) performed an application in order to evaluate public issue attention, which is a topic that usually gets measured through Most Important Problem (MIP) surveys. He proposes the use of Web data, in particular Google Trends (that present some limitations) and Wikipedia page view statistics. The use of the latter approach is justified by the fact that increased attention to a specific topic on Wikipedia indicates risen public interest. Moreover, Web search data are particularly useful to identify short-term public attention that is induced by focus events. Web search activity represents a measure less confounded than previously used media-based measures and more flexible than MIP surveys.

He highlights which are the main advantages of Wikipedia page view statistics:

- data are made available for free;

- page view statistics come at a daily basis;
- this kind of data can be used to track awareness towards a wide range of topics simultaneously.

Compared to Google trends, Wikipedia has some other advantages:

- the access volume at Wikipedia reflects interest for a certain issue, while some search keywords used may threaten validity if they are used with another intention;
- it is possible to identify articles with unambiguous meaning that are closer to the issue of interest;
- the Wikipedia Analytics Team provides raw data.

Munzert also identifies some drawbacks, like the fact that Wikipedia page views statistics does not provide further information on individuals. Moreover, time series are not available before December 2007 for English Wikipedia, and even later for other languages. Finally, the count statistics are language, not country-specific.

Yasseri and Bright (2015) developed a theoretical model which highlights why people might seek for information online at election time, and how this activity might relate to overall electoral outcomes. The experiment is based upon the individuation of a relationship between Wikipedia traffic patterns around election time and the overall electoral turnout.

McIver and Brownstein (2014) developed a novel method of estimating, in near real-time, the level of influenza-like illness in the United States by monitoring the rate of particular Wikipedia article views on a daily basis.

Yucesoy and Barabási (2015) developed an application to quantify the relationship between performance and success by focusing on tennis. They built a predictive model, relying only on a tennis player's performance in tournaments, that can accurately predict an athlete's popularity, both during a player's active years and after retirement. Their model establishes also a direct link between performance and momentary popularity. The visibility of a player is measured through the number of hourly visits to its Wikipedia page.

This big data source can be analysed not only by the number of page views, but also by the editing activity. This is what Göbel and Munzert did (2016), assembling data covering editing activity for the articles on all 1.100 members of the German parliament for the three last legislatures. They investigated if and how edit histories were linked to offline political events, such as elections campaigns. Then, they tested whether strategic incentives rooted in the electoral system, as well as sociodemographic characteristics, affected the editing behaviour.

Other different experiments try to understand the behaviour that people have when surfing on Wikipedia. Examples of these experiments are Reinoso et al. (2009), Gyllstrom and Moens (2012), Reinoso et al. (2012), Tian and Agrawal (2015), Kämpf et al. (2015), Moat et al. (2013).

## 4.2 Data collection and preprocessing

We decided to perform this experiment on three different European cities: Barcelona (Spain), Vienna (Austria) and Bruges (Belgium). The choice was made considering the different sizes and characteristics of those cities and the different reasons that can attract tourists there.

Before starting with the analysis of the data, we have to define precisely what a city is. In order to answer to this question, we decided to consider the Urban Audit dataset provided by Eurostat<sup>10</sup>. According to it, cities are identified at three different levels:

1. a City is a local administrative unit (LAU) where the majority of the population lives in a urban centre of at least 50.000 inhabitants ('C' in Urban Audit dataset);
2. the Greater City is an approximation of the urban centre when it stretches far beyond the administrative city boundaries ('K' in dataset);
3. the Functional Urban Area consists of a city and its commuting zone (formerly known as larger urban zone (LUZ)) ('F' in dataset).

As a starting point of this experiment, we decided the following:

---

<sup>10</sup> <http://ec.europa.eu/eurostat/web/cities/data/database>

- for Barcelona, we'll consider two levels: City (C) and Greater City (K), as some tourists' attractions could be situated over the city boundaries;
- for Vienna, we'll consider only the City (C), as this area is highly extended (four times the surface of Barcelona City);
- for Bruges, we'll consider the City (C) and the Functional Urban Area (F).

In this analysis, we will consider two kind of data: a big data source (Wikipedia) and an Official Statistics source (tourism data).

#### ***4.2.1 Big data source***

##### *4.2.1.1 Wikidata items and Wikipedia articles*

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 questions raised by policy makers and the civil society (Reis et al., 2016).

As a new data source, big data offers great advantages and challenges for Official Statistics. On the one hand, it offers the possibility of higher timeliness, increased efficiency of statistical production systems, much higher detail and significant development in regional and spatial statistics. Big data sources consist of direct measurements of phenomena and not on indirect reporting by a survey respondent, which may result in improved accuracy.

On the other hand, the complexity of production systems will increase, as single big data sources cannot be expected to answer all statistical needs and will need to be combined with survey data, administrative sources and other big data sources.

One important challenge of big data sources is the assessment of the quality of the statistics with it. They very often suffer from selectivity, which may lead to biased results if not measured and accounted for. However, they stress in particular the existing quality frameworks in statistics by bringing new factors which were not present in traditional sources.

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 source.

Wikipedia page views represent the source of data that we use in our analysis, but they are not immediately available and they require a bit of work in getting them. We first must select the articles that we want to include in the study. In order 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. It is worth giving a short introduction to it.

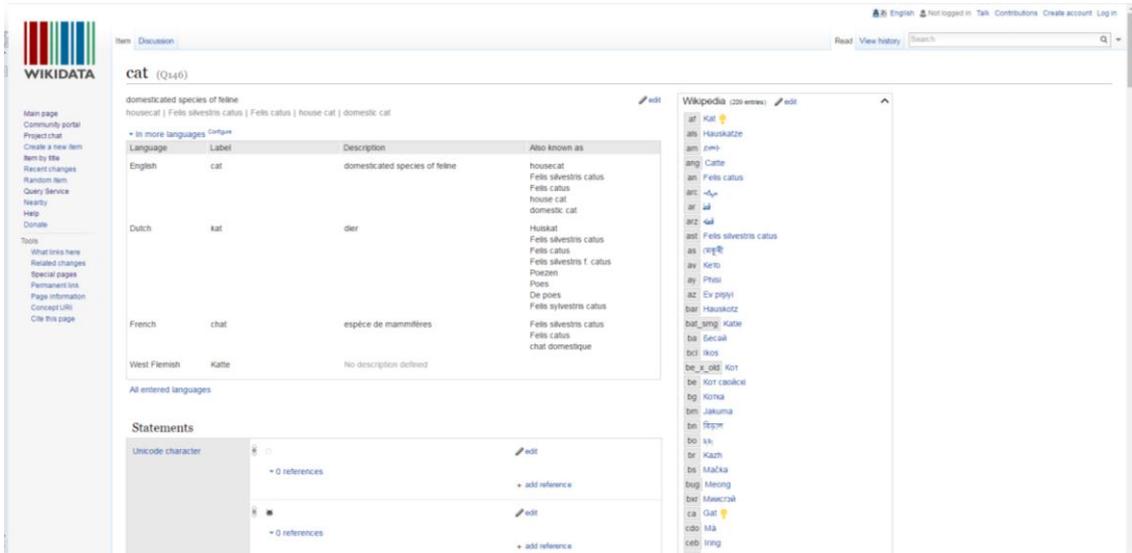
First, we are in the area of Semantic Web, which "provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries", according to the World Wide Web Consortium (W3C) (2011). This source represents one of the best examples of linked data, a method of publishing structured data so that it can be interlinked and become more useful through semantic queries.

Wikidata is a document-oriented database, focused on items. Each item represents a topic (or an administrative page used to maintain Wikipedia) and is identified by a unique number, prefixed with the letter Q. This enables the basic information required to identify the topic the item covers to be translated without favouring any language.

It is useful to add a small example that could explain better how an item is composed. Let's search for the item 'cat' (in the sense of the animal) on Wikidata and see how it appears. In Figure 1 there is a screenshot of how the item looks like.

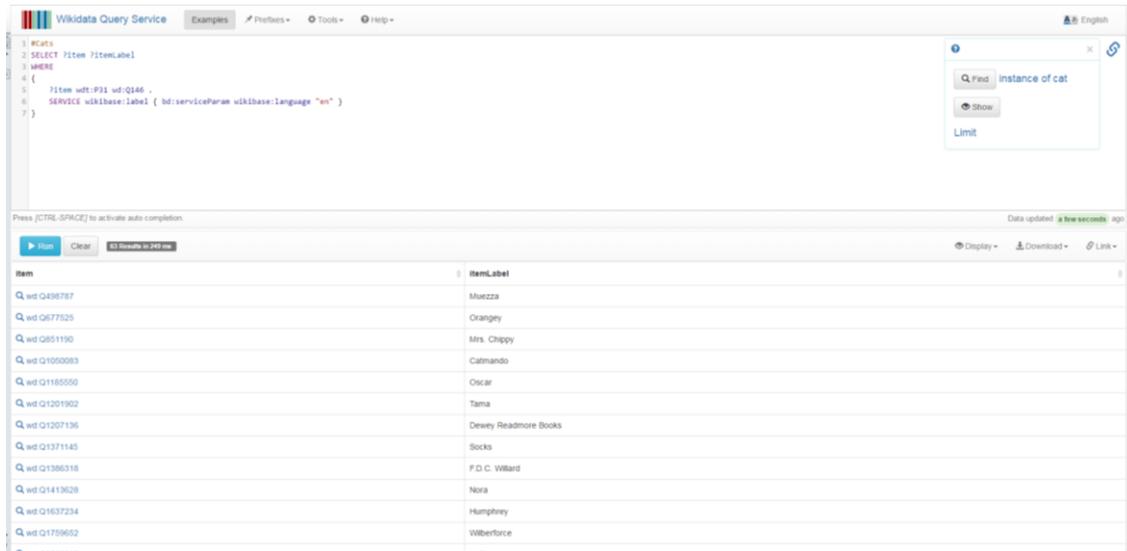
First of all, it has a title ('cat'), the Wikidata identifier (Q146) and a little description ('domesticated species of feline'). Then we find a lot of boxes that contain different informations; on the left side, the item's translation in other languages, the statements about the item and the identifiers. These last two elements allow querying from the database. On the right side, we have indeed the links to all the Wikimedia Foundation projects articles that concern this item, in this case, the cat.

Figure 1 – Structure of a Wikidata item



In order to get data from this source, the Wikimedia Foundation provides the interface "Wikidata Query Service"<sup>11</sup> that allows to query its database using Structured Query Language (SQL). This infrastructure presents some examples, so that also people who are not familiar with this language can immediately perform queries. In Figure 2 an example of how the Query Service looks like in the first example provided, 'cats'.

Figure 2 – Example of a query on Wikidata



<sup>11</sup> <https://query.wikidata.org/>

On the upper part of the page there is the query that we want to perform. In this case, you can recognize the identifier of the item 'cat' that we saw before, Q146. The infrastructure also provides a help service and allows to understand the meaning of each identifier just passing the mouse on it. After having run the query, the results appear in the lower part of the page, showing the elements that were asked (in this case the item identifier and the item label that appear on the second line of the query) and it is possible to download in various formats the output.

In our experiment, we want to get all items that fall into the area of the cities, and to do so, there are three different kinds of queries:

1. using the administrative entity of a location; in this case, it is required to specify the Wikidata identifier of the city. The Query Service will return all items that have the chosen city as 'location'. This method returns you fewer Wikidata items than the following two methods;
2. using a radius around the location; in this case the Wikidata identifier of the location and a radius in kilometers are required. Wikidata will return all items that fall into a circle with the specified radius and with the center located into the city's geo-coordinates;
3. using a box around the location. In this case, it is required to choose two items that are placed at two opposite corners with respect to your target location. Then, to identify at which corner they are placed, it requires to specify it in the format: SouthWest, SouthEast, NorthWest, NorthEast.

For our purpose, the best query to perform is the second one. The SQL query that we run was the following (the examples is about the city of Barcelona, Q1492):

```
SELECT ?item ?name ?coord
WHERE {
  wd:Q1492 wdt:P625 ?mainLoc .
  SERVICE wikibase:around {
    ?item wdt:P625 ?coord .
    bd:serviceParam wikibase:center ?mainLoc .
    bd:serviceParam wikibase:radius "10" .
  }
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en", "cs", "da", "de", "el", "bg", "es", "et", "fi", "fr", "ga", "hr", "hu", "is", "it", "lt", "lv",
```

```
"mk", "mt", "nl", "no", "pl", "pt", "ro", "ru", "sk", "sl", "sq", "sr",  
, "sv", "tr".  
  ?item rdfs:label ?name  
}  
}
```

where:

- at line 3, 'Q1492' is the Wikidata identifier of the location (in the example, Barcelona);
- 'P625' is the requirement on the item to have geo-coordinates;
- then, two different services are used by the query:
  - 'wikibase:around', which queries for items that fall into a circle around the location;
  - 'wikibase:label', which gets the labels (names) of the Wikidata items in some chosen languages. We decided to use 31 languages in the experiment, which represent the 24 official languages of the European Union plus other 7 languages that were in the top rankings considering the amount of page views.

First of all, it is necessary to choose a radius that will contain all the borders of the cities. To compute this, we used the software ArcGIS. We loaded the cities' shapefiles and drew a circle around each. The results are the following (they refer to the bigger level considered):

- Barcelona: 39 kilometers (level K)
- Vienna: 47 kilometers (level C)
- Bruges: 55 kilometers (level F).

Then, for the sake of simplicity, a radius of 30 kilometers would fit for all the cities in the radius approach.

The query can be run on the Website provided by the Wikimedia Foundation (<https://query.wikidata.org/>), but we decided to automate the whole process by creating some R functions that perform the extraction. The R scripts of the newly created functions and of the whole analysis are available at <https://github.com/avirgillito/unece-sandbox2015-wikistat>. The whole analysis was performed on the "Sandbox", an

environment created at UNECE with support from the Central Statistics office (CSO) of Ireland and the Irish Centre for High-End Computing (ICHEC).

After having obtained the list of Wikidata items around the cities' geo-coordinates, we had to filter them in order to consider only those points that fall into the Urban Audit shapefiles. This procedure can be performed using ArcGIS or in RStudio using some overlay functions. Once this filtering procedure is performed, the number of points in each city is shown in Table 1.

*Table 1 – Number of Wikidata items per city level*

<b>City</b>	<b>Level</b>	<b>No of points</b>
Barcelona	C	1103
	K	1463
Bruges	C	563
	F	652
Vienna	C	2728

These represents just Wikidata items with geo-coordinates that fall into the shapefile of the city. We are interested into the Wikipedia articles related to these items in the chosen 31 languages. Some ad hoc built R functions and some filtering allowed this procedure and the result is in Table 2.

*Table 2 – Number of Wikidata items and Wikipedia articles per city level*

<b>City</b>	<b>Level</b>	<b>No of points</b>	<b>No of Wikipedia articles</b>
Barcelona	C	1093	3996
	K	1450	5256
Bruges	C	561	868
	F	649	1127
Vienna	C	2663	6315

You can notice that fewer items appear, as some from the original list do not have Wikipedia articles, or they have but not in the considered languages.

The total number of Wikipedia articles is 12.698. It is necessary to add to this list also the so-called "redirect articles", which are pages that have no content themselves but they send the reader to another page, usually another article or section of an article. The list of Wikipedia articles plus the redirects is composed by 27.850 elements.

After having defined the final list of articles, we extracted the Wikipedia monthly page views. This is made possible through the Page views statistics, which is a tool available for Wikipedia pages. It allows to see how many people have visited an article during a given time period. We have extracted the monthly page views from January 2012 to December 2015.

#### ***4.2.2 Official Statistics source***

Some tourism data are available for the three cities, in particular:

- Barcelona: they are made available on the Website of the municipality of Barcelona (<http://ajuntament.barcelona.cat/en/>). They concern monthly arrivals and overnight stays by country of residence from January 2012 to May 2016.
- Bruges: they are provided by the Flemish tourism Website (<http://www.toerismevlaanderen.be/>). They concern monthly arrivals and overnight stays by country of origin from January 2012 to December 2014.
- Vienna: provided by Statistics Austria, they concern monthly arrivals and overnight stays by country of origin from January 2015 to April 2016.

As we extracted Wikipedia page views from January 2012 to December 2015, the same period has been used as a reference for official tourism data. For Vienna, we contacted the Directorate for Spatial Development at Statistics Austria that provided us with the missing data (January 2012-December 2014).

Official tourism data are represented by arrivals (number of passengers) and overnight stays (number of bookings). These data are collected through the hotel occupancy survey, which is carried out by National Statistical Institutes.

## 4.3 Methodology and 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, in order to identify the factors that drive tourism to an area, and we try to combine the two datasets to identify common patterns.

### 4.3.1 Big data source

#### 4.3.1.1 Points of interest in cities

Before trying to combine the two sources, we decided to look at the big data source separately. First, we built some maps to visualize the page views by Wikidata item. As each item has its own geo-coordinates, we can define it as a point of interest inside the city. The attention each point receives from Web readers is represented by the number of visualizations the articles inside that point of interest receive.

The maps are interactive (they are html files) and were built in RStudio using the package *leaflet*, which is an open source JavaScript library that can also be used in R.

In the following maps, the first visualization we build, each circle represents a Wikidata item. The size of the circle and the intensity of its colour is according to the pageviews of the related Wikipedia articles that refer to that specific Wikidata item, considering the 31 languages of our analysis. You can see screenshots of the maps in Figures 3 to 7<sup>12</sup>.

The screenshots do not say much about the maps, but exploring them online gives you an idea of the places that attracts most of the virtual visits (we can call in this way the interest of online readers). The bigger and the darker a point of interest appears, the more page views it got in this four years' period. We can also notice how the points of interest are not limited over the city center, but they are spread all around.

---

<sup>12</sup> [http://serenasignorelli.altervista.org/Barcelona\\_C/Barcelona\\_C.html](http://serenasignorelli.altervista.org/Barcelona_C/Barcelona_C.html)  
[http://serenasignorelli.altervista.org/Barcelona\\_K/Barcelona\\_K.html](http://serenasignorelli.altervista.org/Barcelona_K/Barcelona_K.html)  
[http://serenasignorelli.altervista.org/Bruges\\_C/Bruges\\_C.html](http://serenasignorelli.altervista.org/Bruges_C/Bruges_C.html)

Figure 3 – Points of interest in Barcelona (Urban Audit Level C)

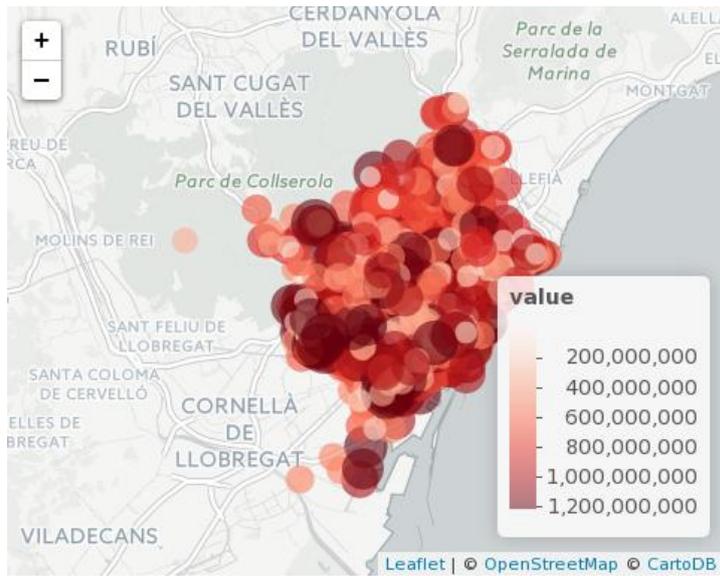
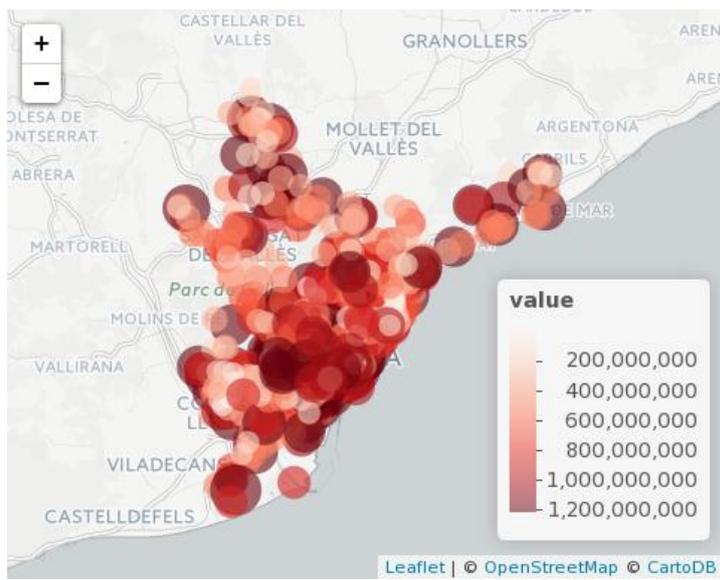


Figure 4 – Points of interest in Barcelona (Urban Audit Level K)



---

[http://serenasignorelli.altervista.org/Bruges\\_F/Bruges\\_F.html](http://serenasignorelli.altervista.org/Bruges_F/Bruges_F.html)  
[http://serenasignorelli.altervista.org/Vienna\\_C/Vienna\\_C.html](http://serenasignorelli.altervista.org/Vienna_C/Vienna_C.html)

Figure 5 – Points of interest in Bruges (Urban Audit Level C)

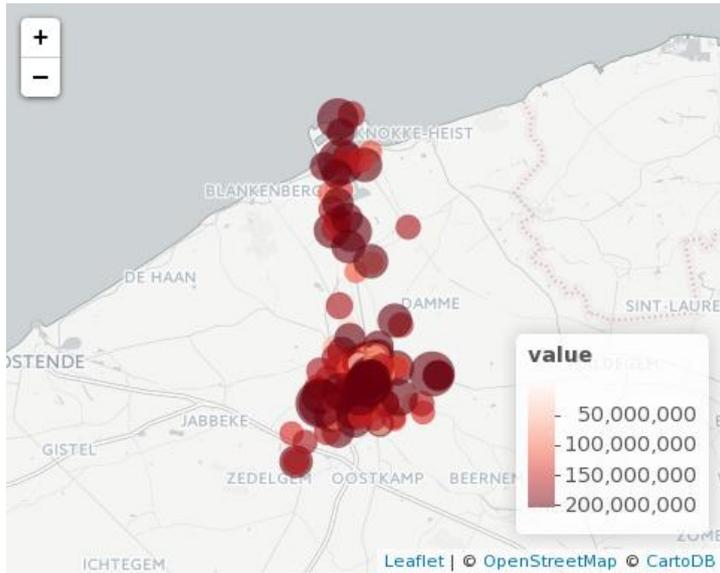


Figure 6 – Points of interest in Bruges (Urban Audit Level F)

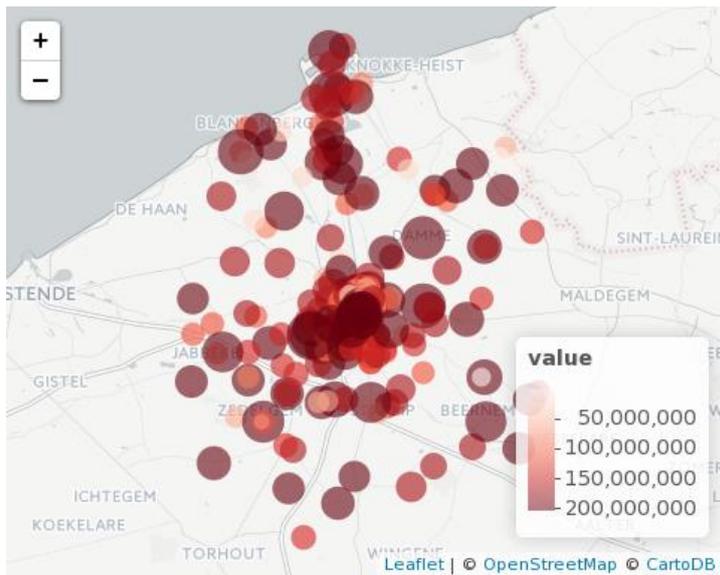
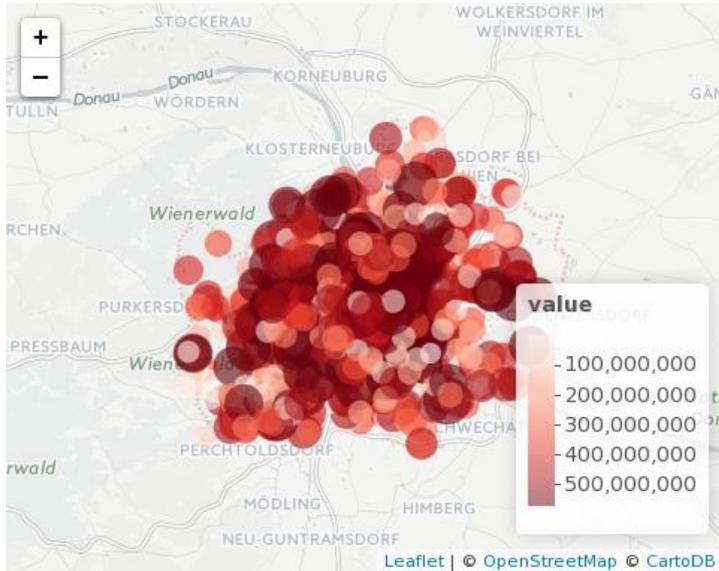


Figure 7 – Points of interest in Vienna (Urban Audit Level C)



After the first visualization, we grouped the page views by language. We performed a computation of rankings based on the page views, the selection of the top six and the plotting into a chart (Figures 8 to 12). The set of interactive maps is available online<sup>13</sup>.

Again, we decided to build interactive charts using another R package, *dygraphs*, which is an R interface to the *dygraphs* JavaScript charting library. The series here are non-standardized (in fact you can notice the big difference between English and other languages) and the chart is interactive in the sense that you can choose to display four years of data or smaller time periods, just moving the slider below the chart. The charts give their best online, where it is possible to identify peaks in one or more languages, that can be linked with some events in the news.

---

<sup>13</sup> [http://serenasignorelli.altervista.org/Barcelona\\_C\\_top\\_6\\_lang\\_ts/Barcelona\\_C\\_top\\_6\\_lang\\_ts.html](http://serenasignorelli.altervista.org/Barcelona_C_top_6_lang_ts/Barcelona_C_top_6_lang_ts.html)  
[http://serenasignorelli.altervista.org/Barcelona\\_K\\_top\\_6\\_lang\\_ts/Barcelona\\_K\\_top\\_6\\_lang\\_ts.html](http://serenasignorelli.altervista.org/Barcelona_K_top_6_lang_ts/Barcelona_K_top_6_lang_ts.html)  
[http://serenasignorelli.altervista.org/Bruges\\_C\\_top\\_6\\_lang\\_ts/Bruges\\_C\\_top\\_6\\_lang\\_ts.html](http://serenasignorelli.altervista.org/Bruges_C_top_6_lang_ts/Bruges_C_top_6_lang_ts.html)  
[http://serenasignorelli.altervista.org/Bruges\\_F\\_top\\_6\\_lang\\_ts/Bruges\\_F\\_top\\_6\\_lang\\_ts.html](http://serenasignorelli.altervista.org/Bruges_F_top_6_lang_ts/Bruges_F_top_6_lang_ts.html)  
[http://serenasignorelli.altervista.org/Vienna\\_C\\_top\\_6\\_lang\\_ts/Vienna\\_C\\_top\\_6\\_lang\\_ts.html](http://serenasignorelli.altervista.org/Vienna_C_top_6_lang_ts/Vienna_C_top_6_lang_ts.html)

Figure 8 – Time series plot for top 6 languages in Barcelona (Urban Audit Level C)

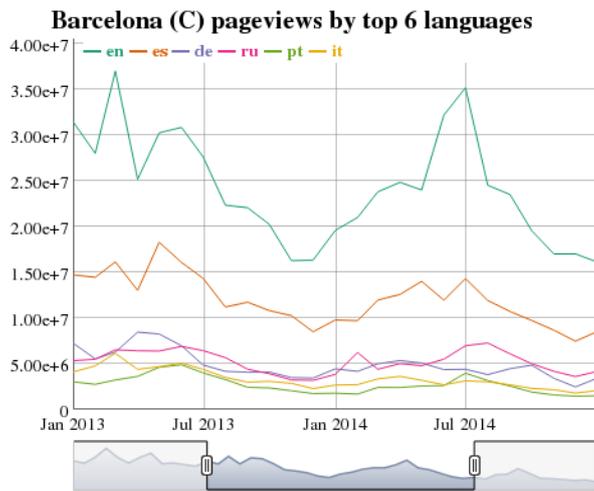


Figure 9 – Time series plot for top 6 languages in Barcelona (Urban Audit Level K)

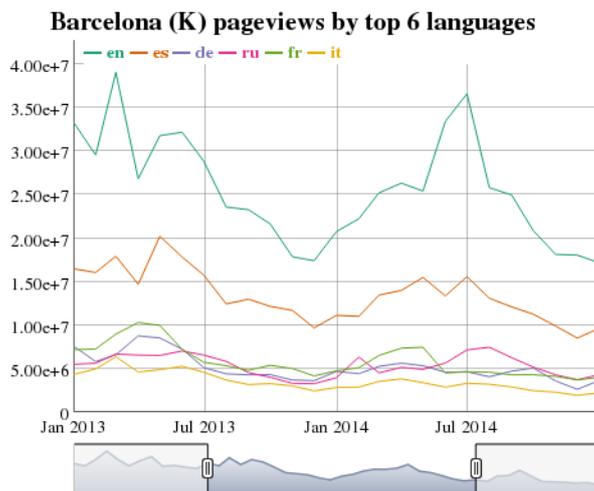


Figure 10 – Time series plot for top 6 languages in Bruges (Urban Audit Level C)

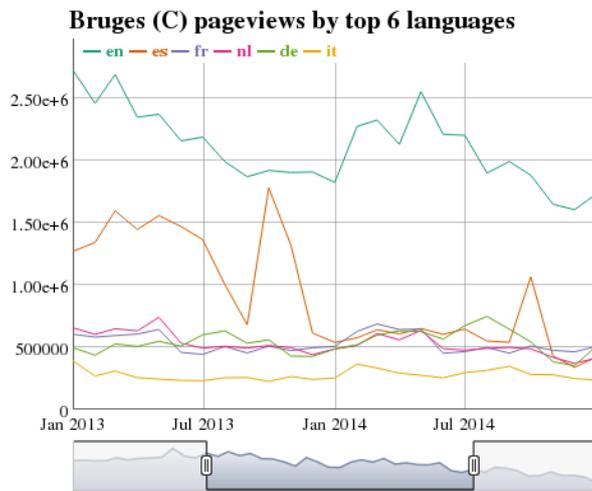


Figure 11 – Time series plot for top 6 languages in Bruges (Urban Audit Level F)

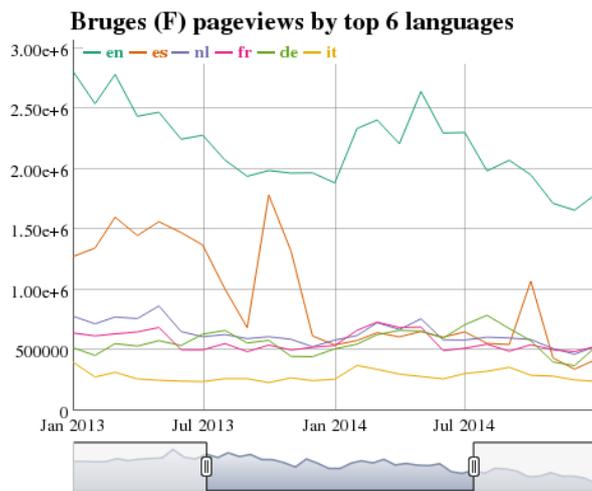
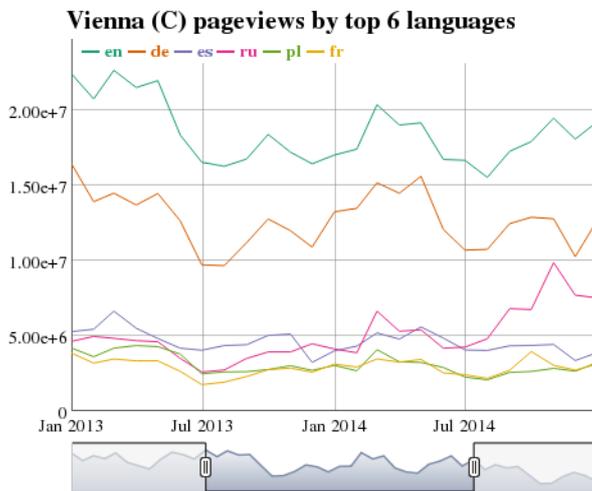


Figure 12 – Time series plot for top 6 languages in Vienna (Urban Audit Level C)



The above-mentioned charts have also been built with standardized-series, so that it is possible to explore better the behaviour of the series in different languages. We do not provide any screenshot of the charts, but leave the links to investigate them<sup>14</sup>.

The same time series have been put in some maps, where you can choose which of the six languages to display. We used again *leaflet* to create interactive maps<sup>15</sup> (Figures 13 to 17). In this case, the different languages are displayable on the choice of the user, just by flagging the box on the upper right side of the map. Each language has a different colour and each circle represents a Wikidata item. As seen before, the size of

<sup>14</sup> [http://serenasignorelli.altervista.org/Barcelona\\_C\\_top\\_6\\_lang\\_ts\\_scaled/Barcelona\\_C\\_top\\_6\\_lang\\_ts\\_scaled.html](http://serenasignorelli.altervista.org/Barcelona_C_top_6_lang_ts_scaled/Barcelona_C_top_6_lang_ts_scaled.html)  
[http://serenasignorelli.altervista.org/Barcelona\\_K\\_top\\_6\\_lang\\_ts\\_scaled/Barcelona\\_K\\_top\\_6\\_lang\\_ts\\_scaled.html](http://serenasignorelli.altervista.org/Barcelona_K_top_6_lang_ts_scaled/Barcelona_K_top_6_lang_ts_scaled.html)  
[http://serenasignorelli.altervista.org/Bruges\\_C\\_top\\_6\\_lang\\_ts\\_scaled/Bruges\\_C\\_top\\_6\\_lang\\_ts\\_scaled.html](http://serenasignorelli.altervista.org/Bruges_C_top_6_lang_ts_scaled/Bruges_C_top_6_lang_ts_scaled.html)  
[http://serenasignorelli.altervista.org/Bruges\\_F\\_top\\_6\\_lang\\_ts\\_scaled/Bruges\\_F\\_top\\_6\\_lang\\_ts\\_scaled.html](http://serenasignorelli.altervista.org/Bruges_F_top_6_lang_ts_scaled/Bruges_F_top_6_lang_ts_scaled.html)  
[http://serenasignorelli.altervista.org/Vienna\\_C\\_top\\_6\\_lang\\_ts\\_scaled/Vienna\\_C\\_top\\_6\\_lang\\_ts\\_scaled.html](http://serenasignorelli.altervista.org/Vienna_C_top_6_lang_ts_scaled/Vienna_C_top_6_lang_ts_scaled.html)

<sup>15</sup> [http://serenasignorelli.altervista.org/Barcelona\\_C\\_top6/Barcelona\\_C\\_top6.html](http://serenasignorelli.altervista.org/Barcelona_C_top6/Barcelona_C_top6.html)  
[http://serenasignorelli.altervista.org/Barcelona\\_K\\_top6/Barcelona\\_K\\_top6.html](http://serenasignorelli.altervista.org/Barcelona_K_top6/Barcelona_K_top6.html)  
[http://serenasignorelli.altervista.org/Bruges\\_C\\_top6/Bruges\\_C\\_top6.html](http://serenasignorelli.altervista.org/Bruges_C_top6/Bruges_C_top6.html)  
[http://serenasignorelli.altervista.org/Bruges\\_F\\_top6/Bruges\\_F\\_top6.html](http://serenasignorelli.altervista.org/Bruges_F_top6/Bruges_F_top6.html)  
[http://serenasignorelli.altervista.org/Vienna\\_C\\_top6/Vienna\\_C\\_top6.html](http://serenasignorelli.altervista.org/Vienna_C_top6/Vienna_C_top6.html)

the circle and the intensity of its colour is proportional to the pageviews of the Wikipedia articles related to that specific Wikidata item in the 31 languages.

Figure 13 – Points of interest in top 6 languages in Barcelona (Urban Audit Level C)

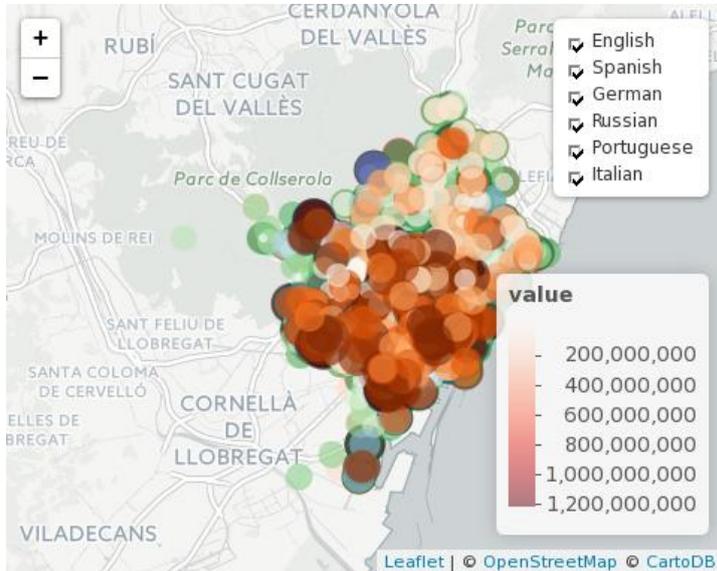


Figure 14 – Points of interest in top 6 languages in Barcelona (Urban Audit Level K)

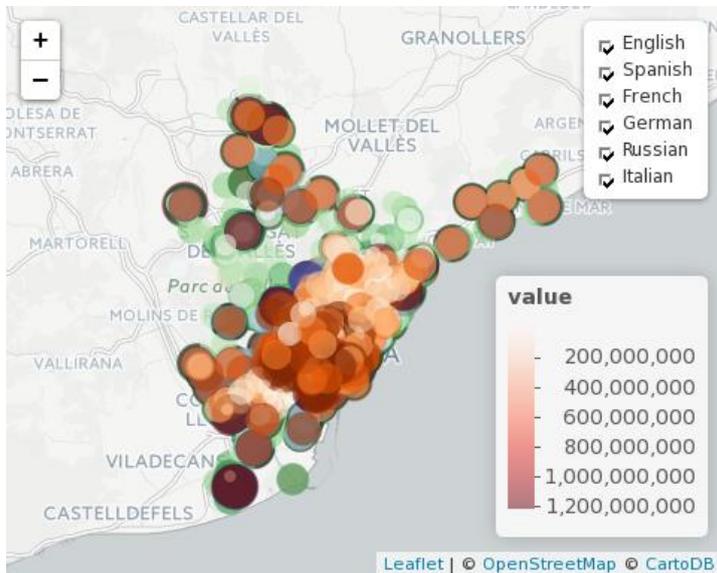


Figure 15 – Points of interest in top 6 languages in Bruges (Urban Audit Level C)

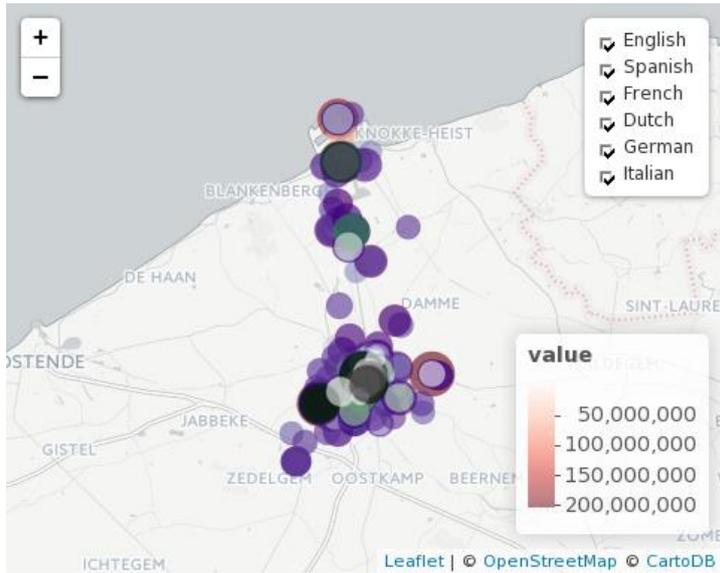


Figure 16 – Points of interest in top 6 languages in Bruges (Urban Audit Level F)

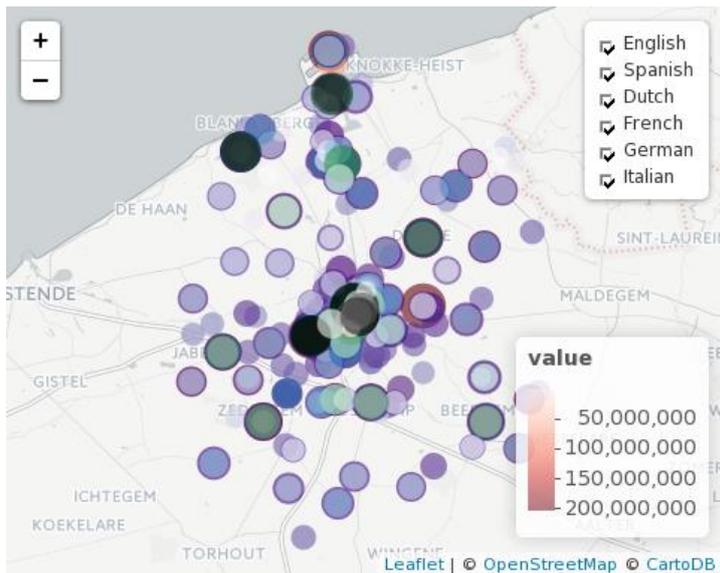
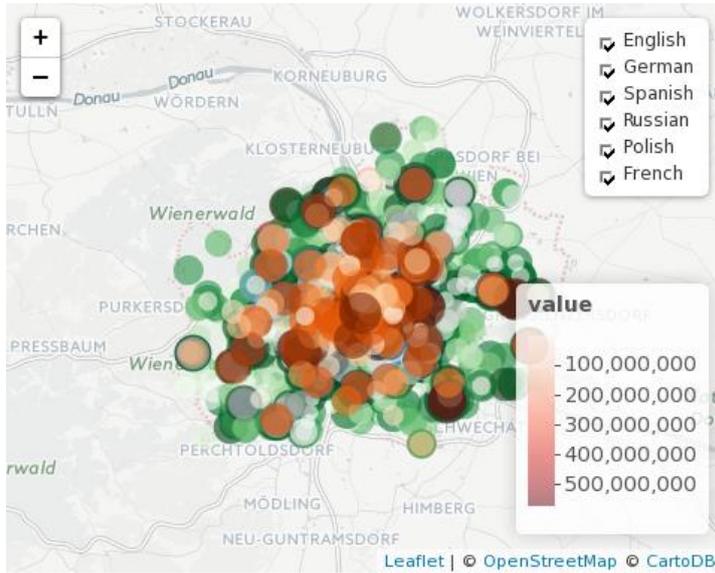


Figure 17 – Points of interest in top 6 languages in Vienna (Urban Audit Level C)



#### 4.3.1.2 First attempt of classification using Wikidata properties

As the aim of this experiment is trying to identify the factors that drive tourism to a city, it would be useful to have a sort of "category" for each of the Wikidata items. The categories are already available, as each item has inside some statements that show many characteristics of the item itself.

In our case, the category we are interested in is the property P31 "instance of". We modified the query in order to get the same items as before plus the property P31. In Table 3, we highlighted the number of items with property P31 and the total number of different properties available in the city area.

Table 3 – Number of Wikidata items, items with property P31 and number of properties P31 per city level

<i>City</i>	<i>Level</i>	<i>No of points</i>	<i>No of points with property</i>	<i>No of properties</i>
Barcelona	C	1.093	850	196
	K	1.450	1.170	214
Bruges	C	561	264	65
	F	649	336	80
Vienna	C	2.663	1.882	242

Not all items have property P31, and some items have more than one property P31. So, we decided to consider only one property per item (the first one that appear in the data frame) and not considering for the moment the items with no categorization. We then grouped items by property and sum up the number of page views. Displaying the page views in decreasing order, we obtain the top tens shown in Tables 4 to 8.

*Table 4 – Top ten properties P31 per number of page views in Barcelona (Urban Audit Level C)*

<i>Property</i>	<i>Pageview</i>
association football club	1.215.300.908
city	776.415.274
church	372.488.865
stadium	204.675.118
building	122.287.340
Summer Olympic Games	106.862.448
park	75.219.412
mountain	26.768.030
neighborhood	24.605.866
architectural structure	23.088.181

*Table 5 – Top ten properties P31 per number of page views in Barcelona (Urban Audit Level K)*

<i>Property</i>	<i>Pageview</i>
association football club	1.228.442.513
city	823.673.343
church	372.517.140
stadium	204.689.584
building	133.761.151
Summer Olympic Games	106.862.448
park	75.219.412
municipality of Spain	58.832.648
international airport	32.162.953
mountain	28.104.588

*Table 6 – Top ten properties P31 per number of page views in Bruges (Urban Audit Level C)*

<b><i>Property</i></b>	<b><i>Pageview</i></b>
city	202.311.900
association football club	14.142.370
village	6.911.664
sculpture	6.817.030
university	6.233.836
church	5.148.857
neighborhood	4.578.096
football stadium	4.225.740
museum	3.340.650
art museum	1.560.800

*Table 7 – Top ten properties P31 per number of page views in Bruges (Urban Audit Level F)*

<b><i>Property</i></b>	<b><i>Pageview</i></b>
city	202.311.900
association football club	14.142.370
municipality of Belgium	9.714.282
village	7.141.008
sculpture	6.817.030
university	6.233.836
church	5.154.512
neighborhood	4.587.368
football stadium	4.225.740
museum	3.340.650

Table 8 – Top ten properties P31 per number of page views in Vienna (Urban Audit Level C)

<i>Property</i>	<i>Pageview</i>
capital	570.564.420
state	338.381.352
international organization	254.440.650
human	180.933.871
palace	153.036.503
intergovernmental organization	122.724.462
former country	105.799.882
district of Vienna	87.524.529
organization	73.490.121
museum	64.932.026

As you can notice, the categorization is not much satisfactory; moreover, the property 'city' is at the top of the rankings more or less every time, and we suspect that these are the visits to the article about the city itself. The number of properties P31 is very high in almost all cities, so we decided to extract from the Wikidata items another property, the P279, which correspond to "Subclass of". We act in this way in order to get a sort of 'upper property', a class to which the property P31 belongs to. Same as before, there's the possibility for a property to belong to multiple classes. After having joined the list of items and properties with the list of classes, we considered only one class per item, in order to avoid having the same item multiple times (Table 9).

Table 9 – Number of Wikidata items, items with property P31, number of properties P31 and number of property P279 (class) per city level

<i>City</i>	<i>Level</i>	<i>No of points</i>	<i>No of points with property</i>	<i>No of properties</i>	<i>No of classes</i>
Barcelona	C	1093	850	196	111
	K	1450	1170	214	120
Bruges	C	561	264	65	44
	F	649	336	80	52
Vienna	C	2663	1882	242	121

It is worth noting that a lot of properties and classes are defined in the same way (for example, the property 'museum' could represent the class 'museum' for another item). So, we decided to first check these correspondences, using the function *semi\_join* from the dplyr R package on properties and classes, and *left\_join* the result to the classes data frame, in order to reduce their number. From now on we will use a new variable called 'category' (cat), that represents the joined version of properties and classes (Table 10).

*Table 10 – Number of Wikidata items, items with property P31, number of properties P31, number of property P279 (class) and number of newly defined categories per city level*

<i>City</i>	<i>Level</i>	<i>No of points</i>	<i>No of points with property</i>	<i>No of properties</i>	<i>No of classes</i>	<i>No of categories</i>
Barcelona	C	1093	850	196	111	103
	K	1450	1170	214	120	113
Bruges	C	561	264	65	44	46
	F	649	336	80	52	53
Vienna	C	2663	1882	242	121	120

Even after this third step, we decided not to use directly this classification, but trying to refine it with clustering methods, as shown in the following paragraphs.

#### *4.3.1.3 Second attempt of classification using clustering*

As the categorization already available from Wikidata is not satisfactory, we have to build our own, starting from the categories we just identified. The simplest way to achieve this is applying a clustering technique. In particular, we tried two different clustering methods: hierarchical clustering (on categories identified from Wikidata and on languages) and k-means clustering (directly on Wikidata items' series).

##### *4.3.1.3.1 Hierarchical clustering with Dynamic Time Warping on categories*

First of all, we standardized the time series of the categories (the ones coming out from the three steps on Wikidata properties), in order to be able to compare them. After that, we performed a hierarchical agglomerative clustering to see which categories can be grouped together and which topic they cover (in order to answer our research question on factors that drive tourism).

To decide which clusters should be combined, a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

**Metric.** In this case, we chose to compute the distance matrix with Dynamic Time Warping, an algorithm for measuring similarity between two temporal sequences which may vary in speed. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity, independently from certain non-linear variations in the time dimension. This method is often used in time-series classification.

**Linkage criteria.** We decided to use average linkage clustering, also called UPGMA (Unweighted Pair Group Method with Arithmetic Mean). It constructs a rooted tree (dendrogram) that reflects the structure present in a pairwise similarity matrix (or a dissimilarity matrix). At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters  $A$  and  $B$ , each of size (i.e., cardinality)  $|A|$  and  $|B|$ , is taken to be the average of all distances  $d(x, y)$  between pairs of objects  $x$  in  $A$  and  $y$  in  $B$ , that is, the mean distance between elements of each cluster:

$$\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y)$$

The results of hierarchical clustering are usually presented in a dendrogram.

You can see the results in the following Figures (18 to 22).

Figure 18 – Result of hierarchical clustering on categories in Barcelona (Urban Audit Level C)

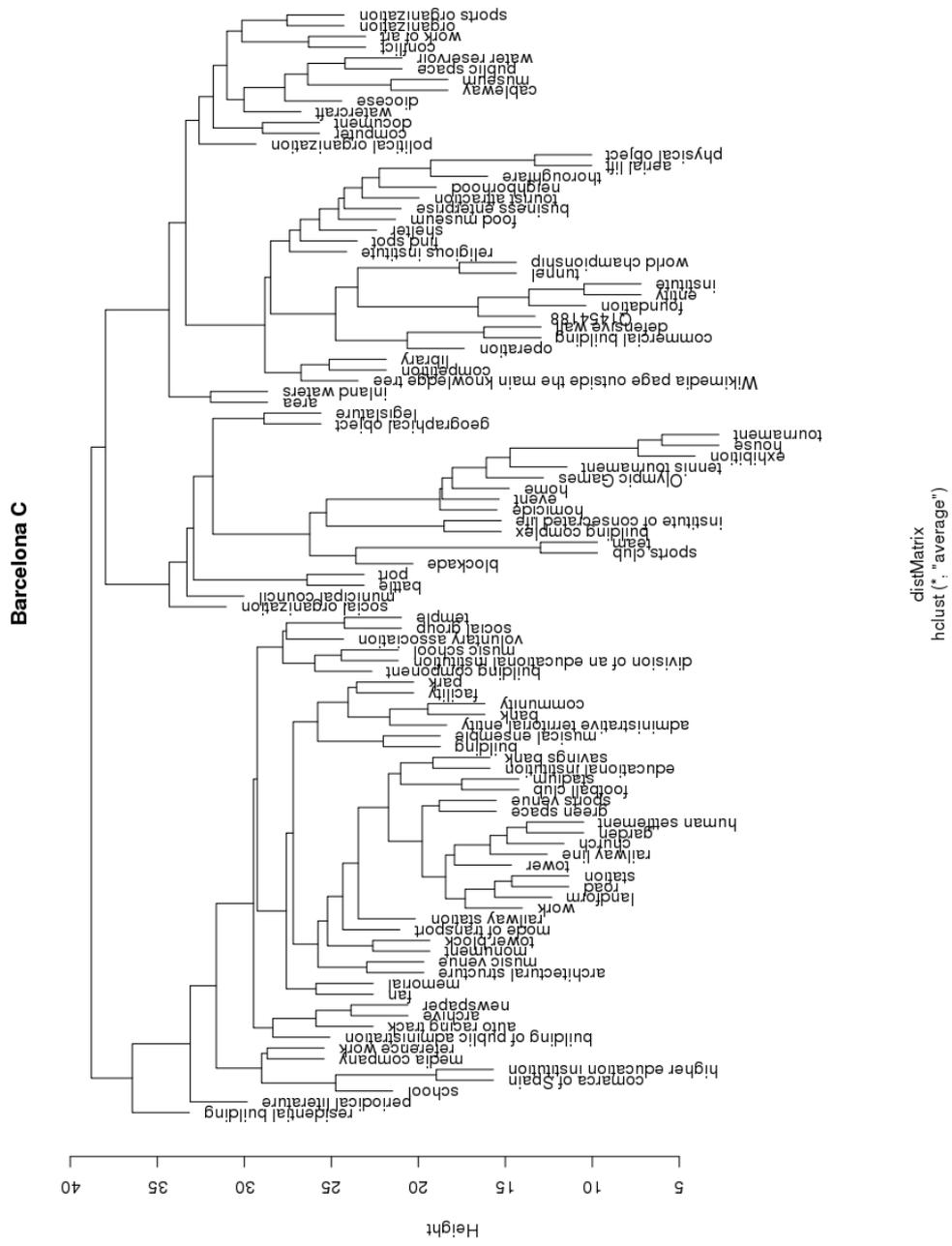


Figure 19 – Result of hierarchical clustering on categories in Barcelona (Urban Audit Level K)

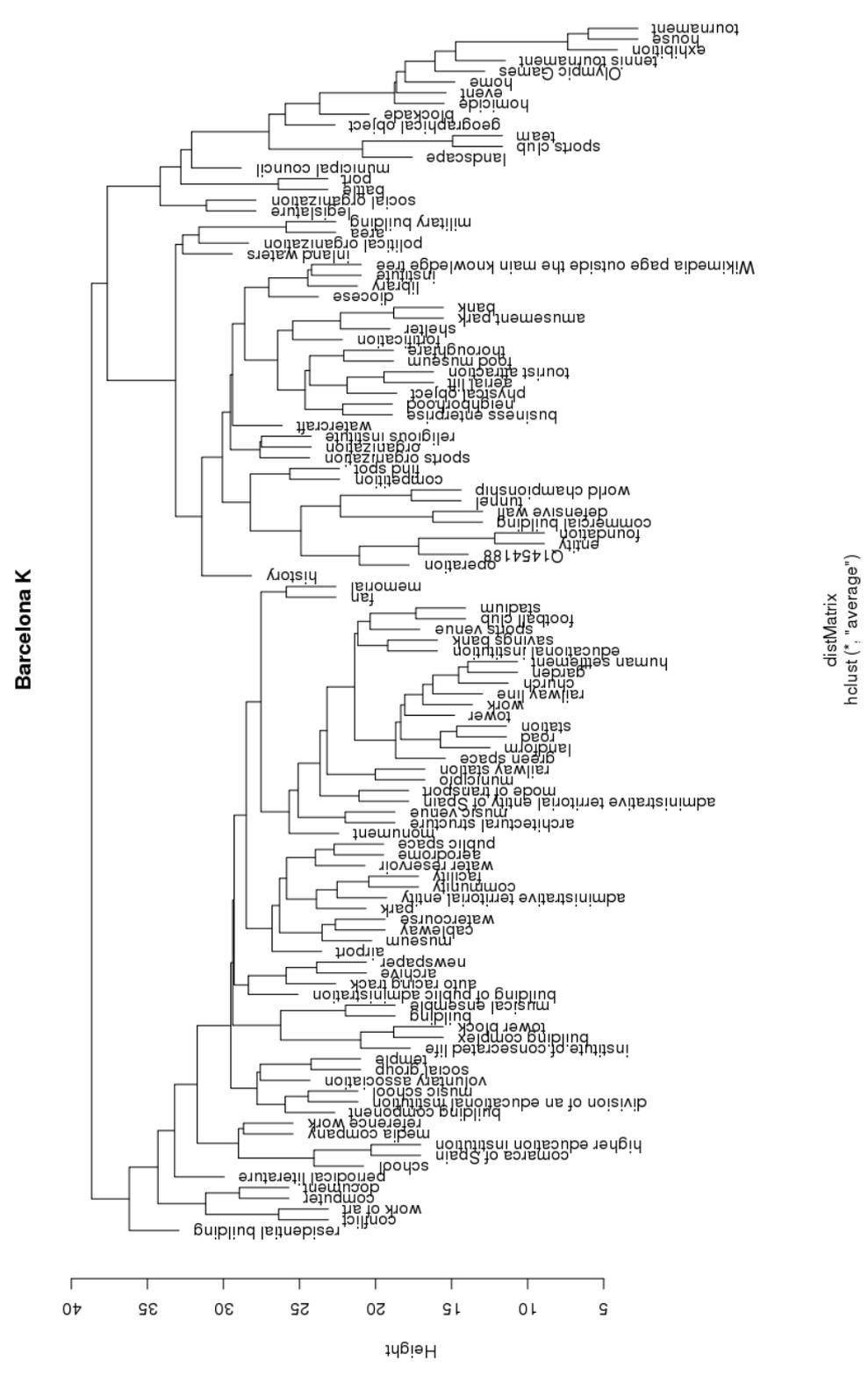


Figure 20 – Result of hierarchical clustering on categories in Bruges (Urban Audit Level C)

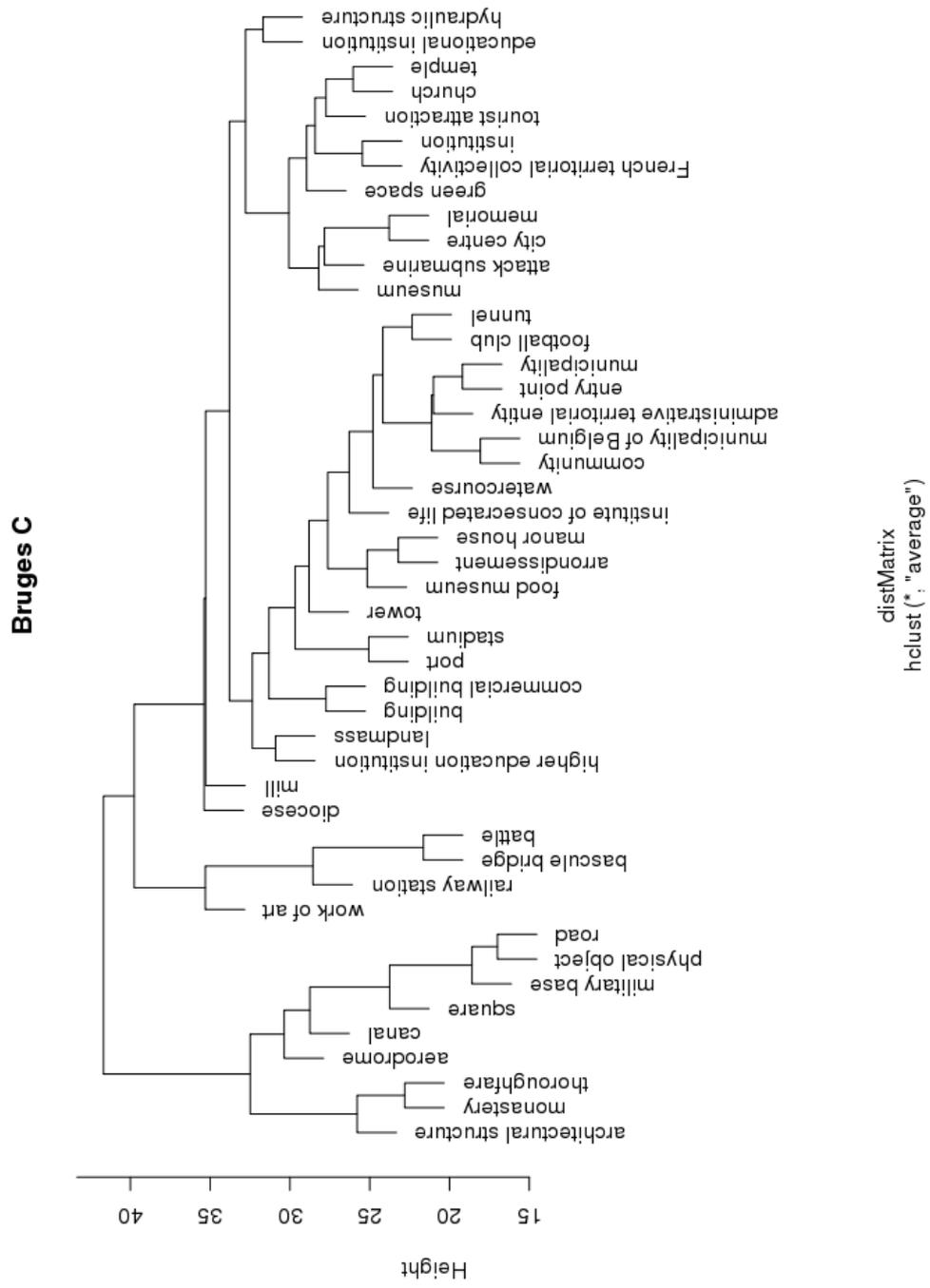


Figure 21 – Result of hierarchical clustering on categories in Bruges (Urban Audit Level F)

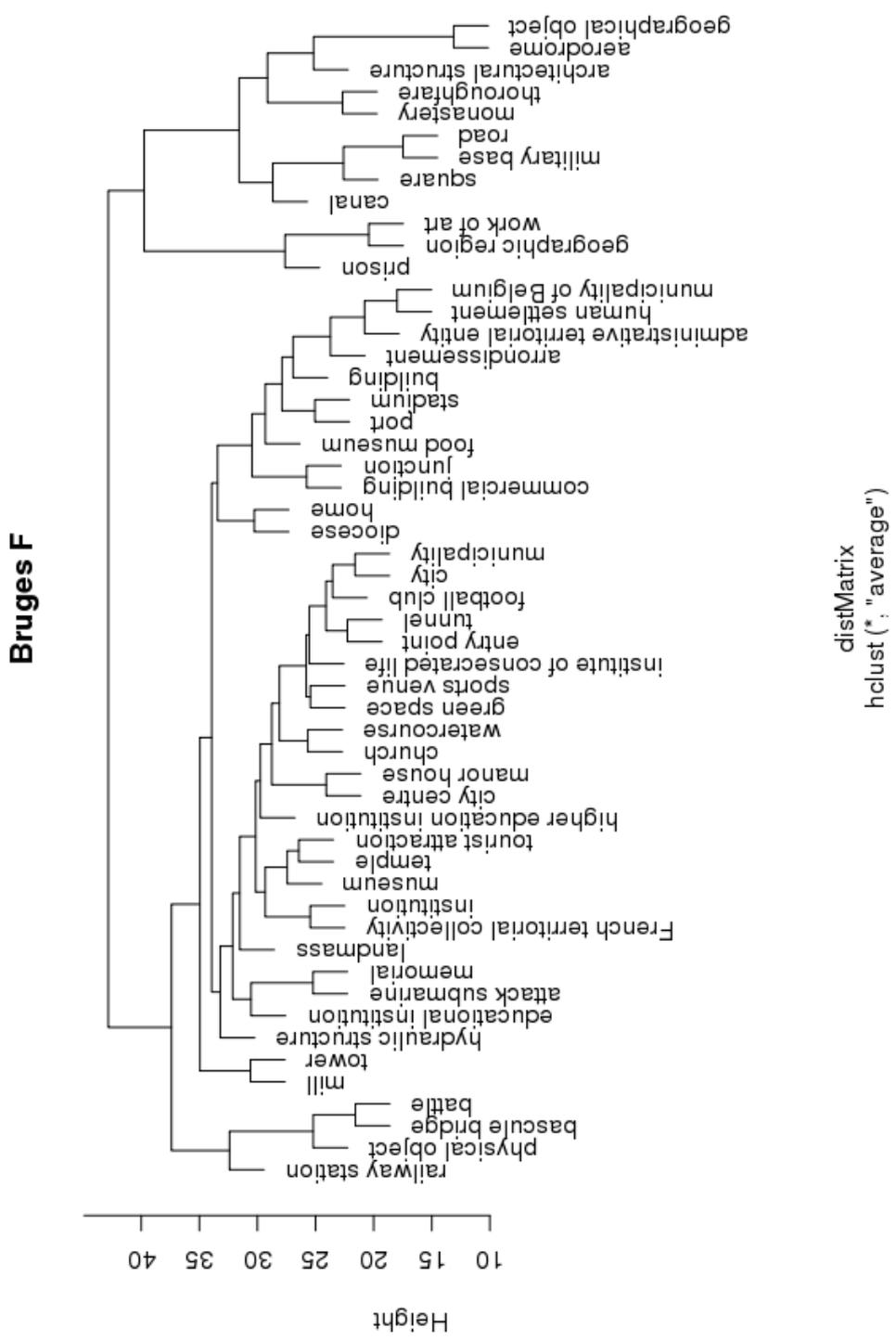
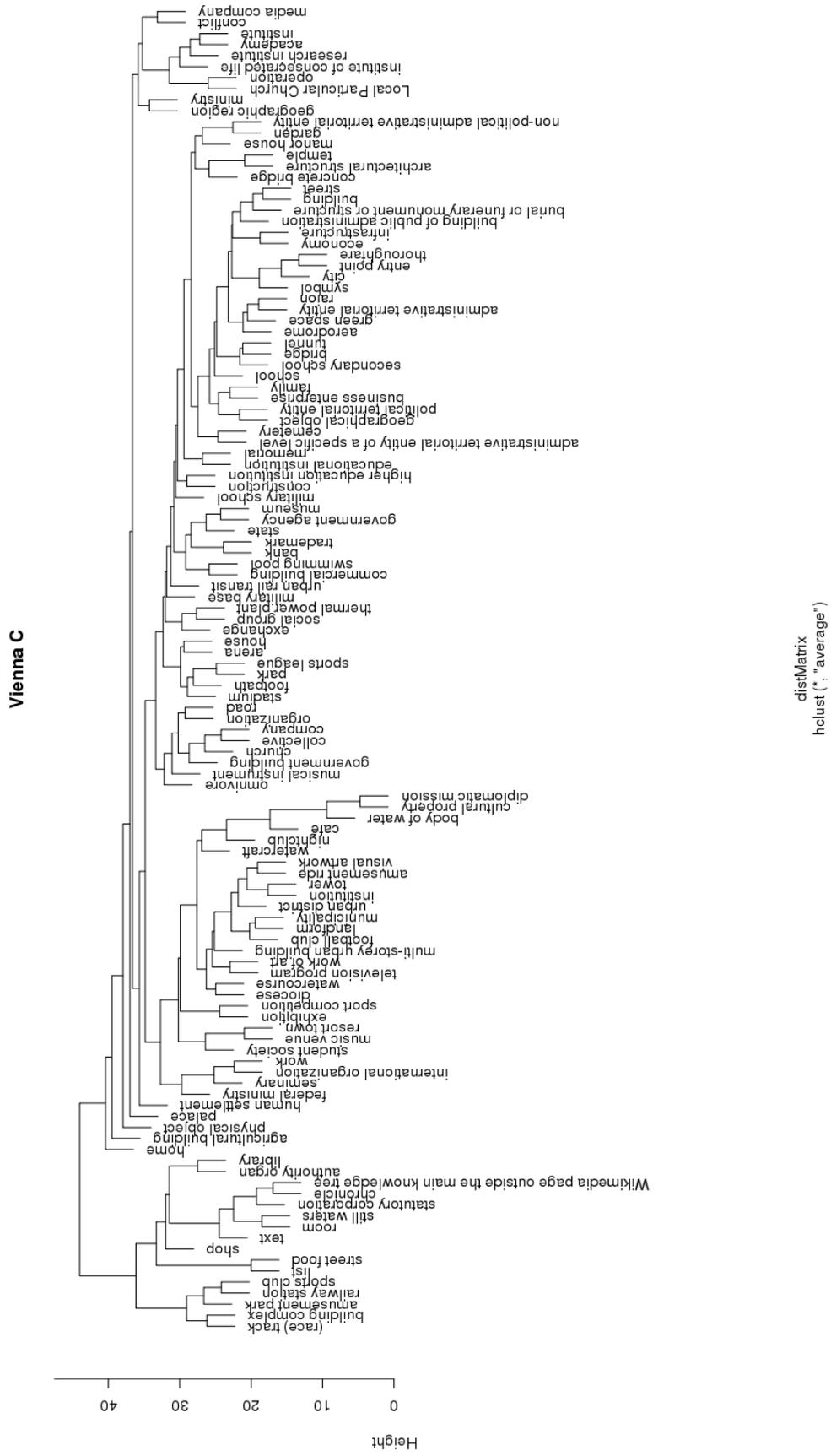


Figure 22 – Result of hierarchical clustering on categories in Vienna (Urban Audit Level C)



These dendrograms show the proximity of Wikidata items based on their category, but this kind of results seems not to be appropriate for our purpose. In fact, we want to identify factors that drive tourism to an area, but if, for example, we identify that a museum and a soccer team series are close to each other, how could we call that factor?

So, the categorization from Wikidata seems not to be useful; we decided then to study proximity between languages.

#### 4.3.1.3.2 Hierarchical clustering with Dynamic Time Warping on languages

We performed the same cluster procedure on time series by language, to see which languages behave in the same way. The results are plotted in Figures 23 to 27.

Figure 23 – Result of hierarchical clustering on languages in Barcelona (Urban Audit Level C)

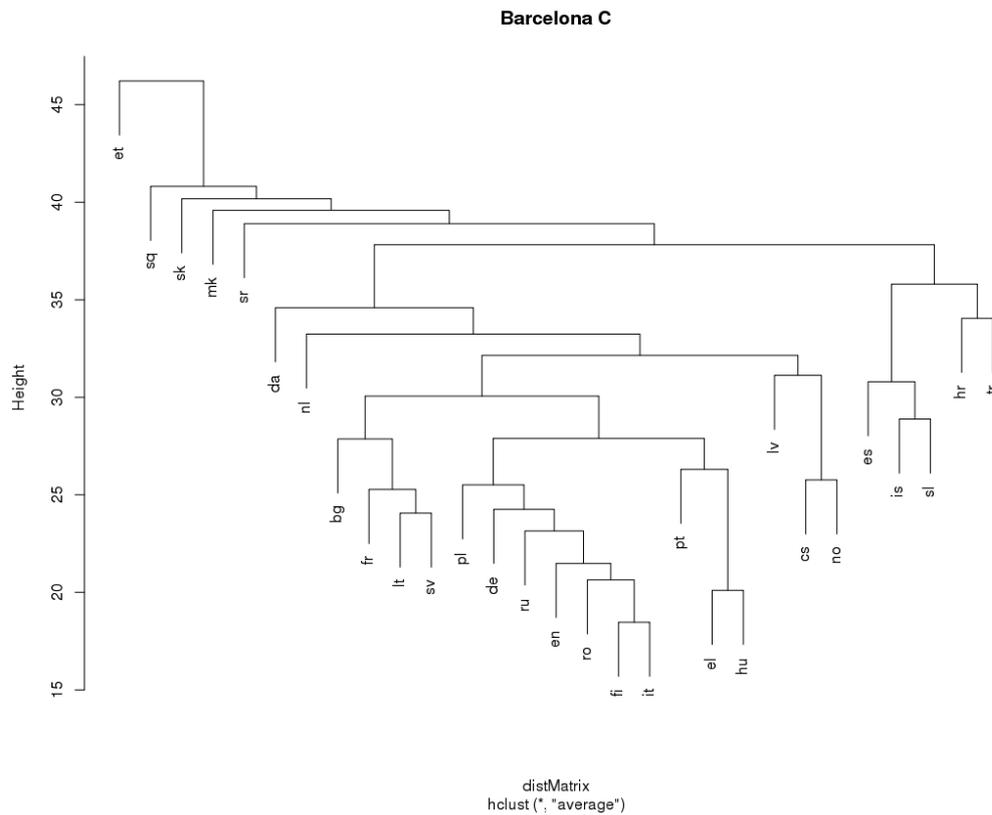


Figure 24 – Result of hierarchical clustering on languages in Barcelona (Urban Audit Level K)

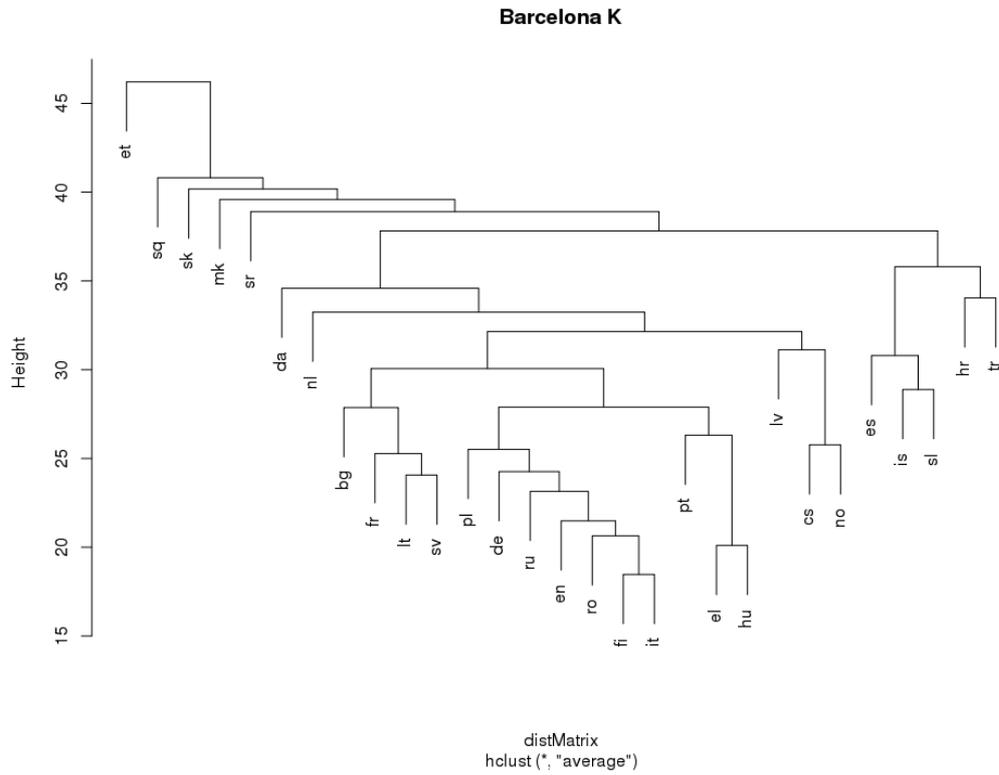


Figure 25 – Result of hierarchical clustering on languages in Bruges (Urban Audit Level C)

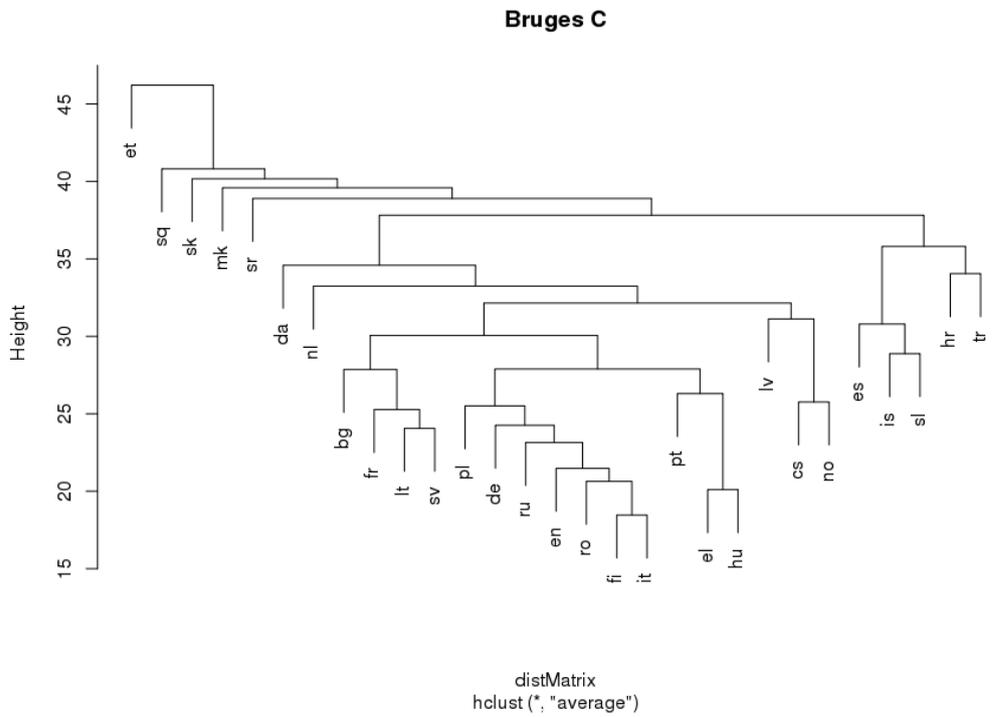


Figure 26 – Result of hierarchical clustering on languages in Bruges (Urban Audit Level F)

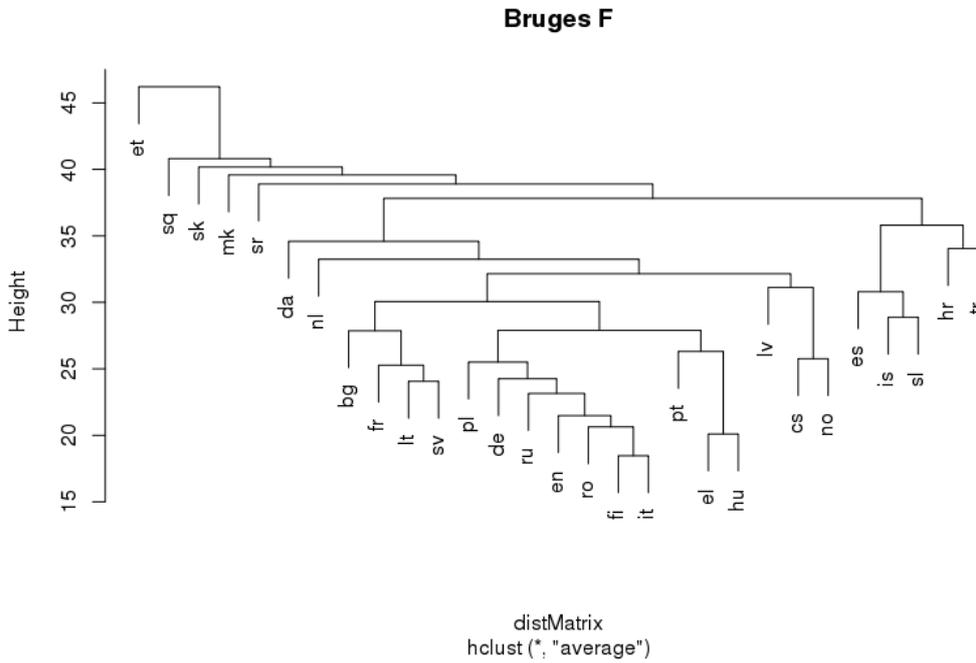
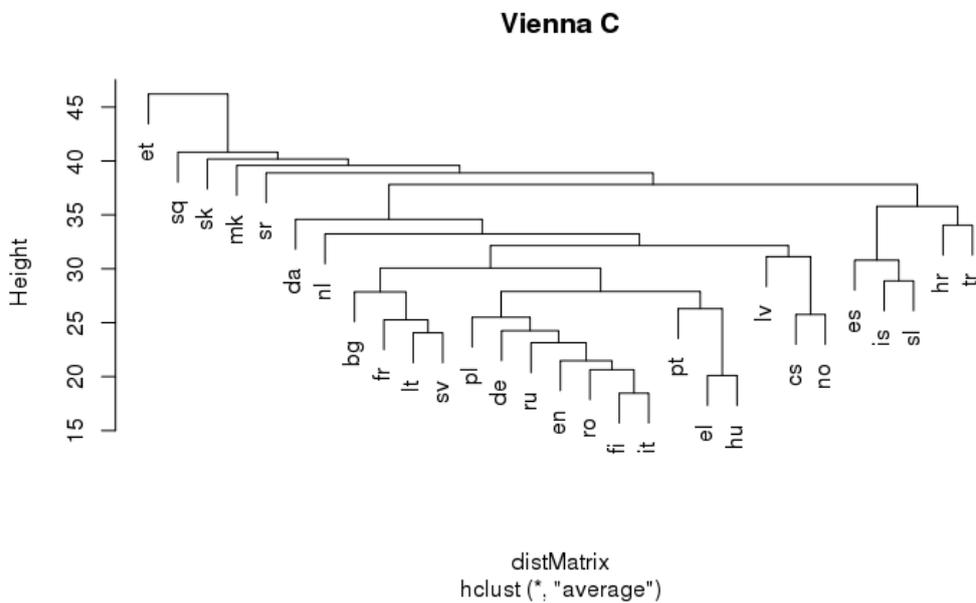


Figure 27 – Result of hierarchical clustering on languages in Vienna (Urban Audit Level C)



Also after this clustering analysis we are not satisfied, as proximity in languages could not help us in the purpose of identifying factors.

#### 4.3.1.3.3 Cluster analysis on items

We decided to perform a cluster analysis directly on the time-series of the Wikidata items themselves. We made this to be able to group the points of interest that present a similar pattern. In this case, we used the k-means clustering, as we do not need to visualize the results in a dendrogram, but we need to define groups.

K-means clustering aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

Formally, given a set of observations  $(x_1, x_2, \dots, x_n)$  where each observation is a  $d$ -dimensional real vector, k-means clustering aims to partition the  $n$  observations into  $k$  ( $\leq n$ ) sets  $S = \{S_1, S_2, \dots, S_k\}$  so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

$$\operatorname{argmin}_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2$$

where  $\mu_i$  is the mean of points in  $S_i$ .

We found that the ideal number of clusters for each city in each level is as shown in Table 11.

*Table 11 – Number of Wikidata items, clusters and cumulative reduction WSS per city level*

<b>City</b>	<b>Level</b>	<b>No of items</b>	<b>No of clusters</b>	<b>Cumulative reduction WSS</b>
Barcelona	C	1093	14	82,3
	K	1450	18	84,2
Bruges	C	561	9	80,9
	F	649	12	85,2
Vienna	C	2663	28	81,2

In the following pages, you can see the charts that show the division of each city in clusters (Figures 28 to 32).

Figure 28 – Result of cluster analysis in Barcelona (Urban Audit Level C)

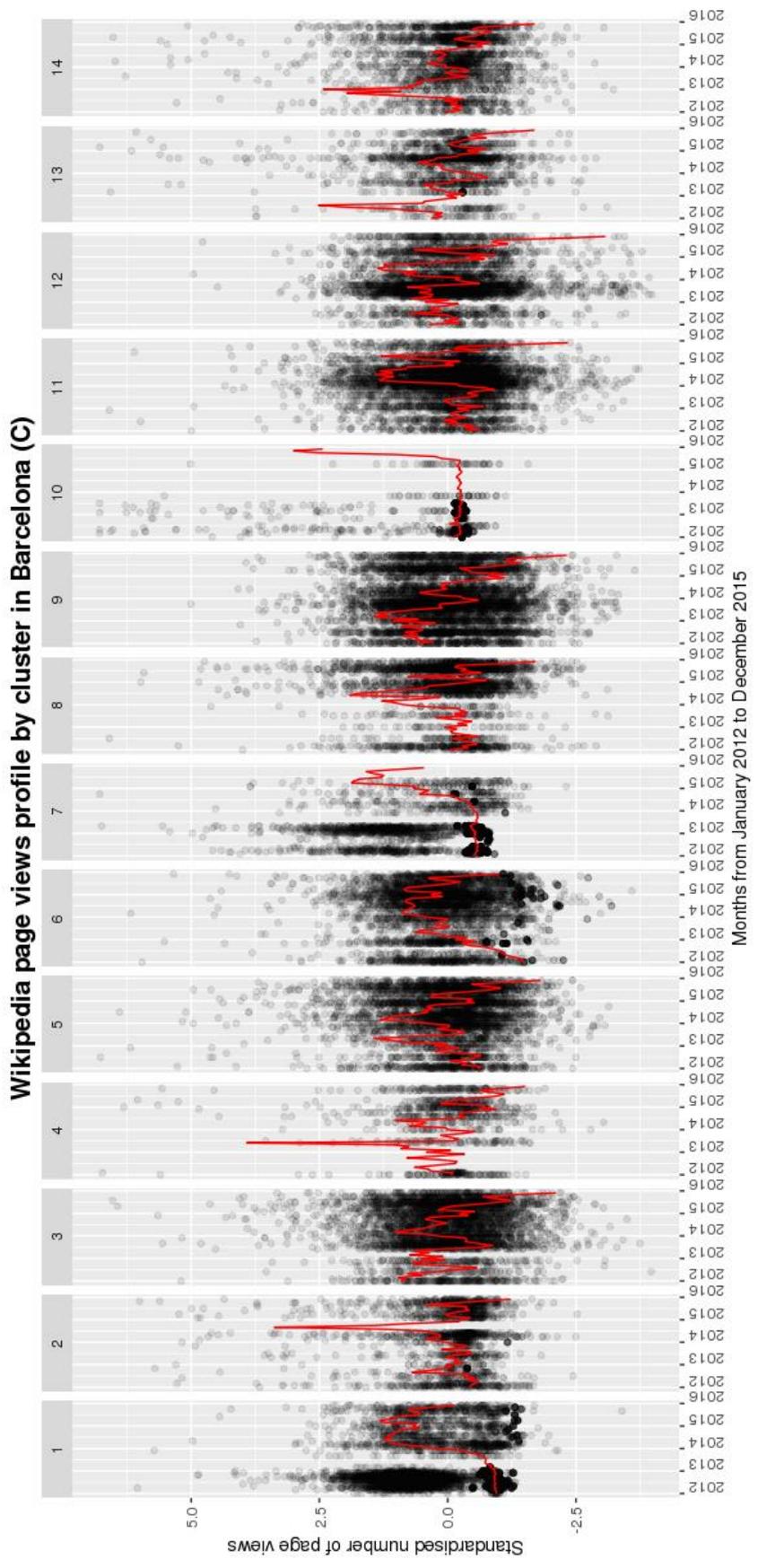


Figure 29 – Result of cluster analysis in Barcelona (Urban Audit Level K)

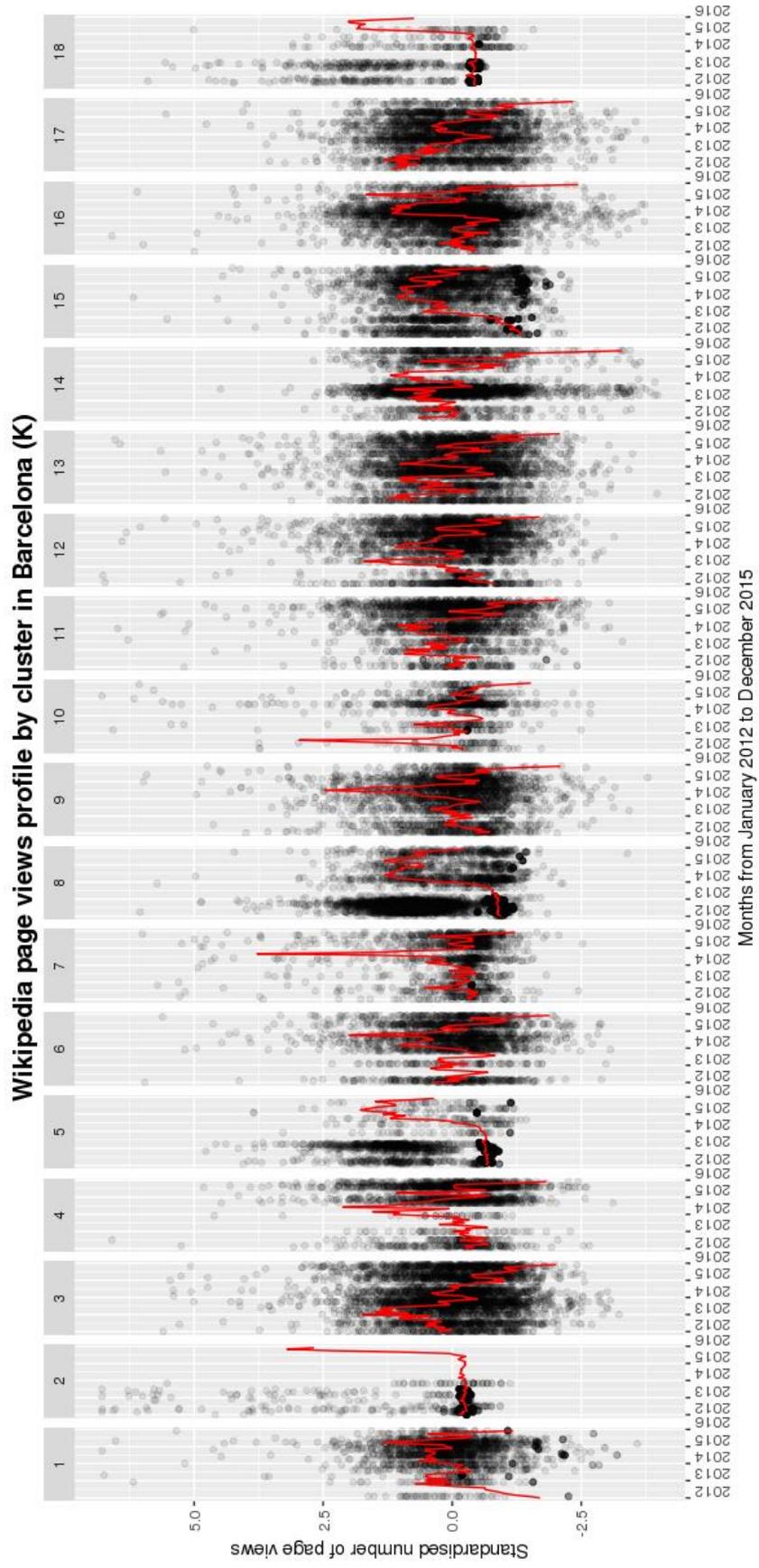


Figure 30 – Result of cluster analysis in Bruges (Urban Audit Level C)

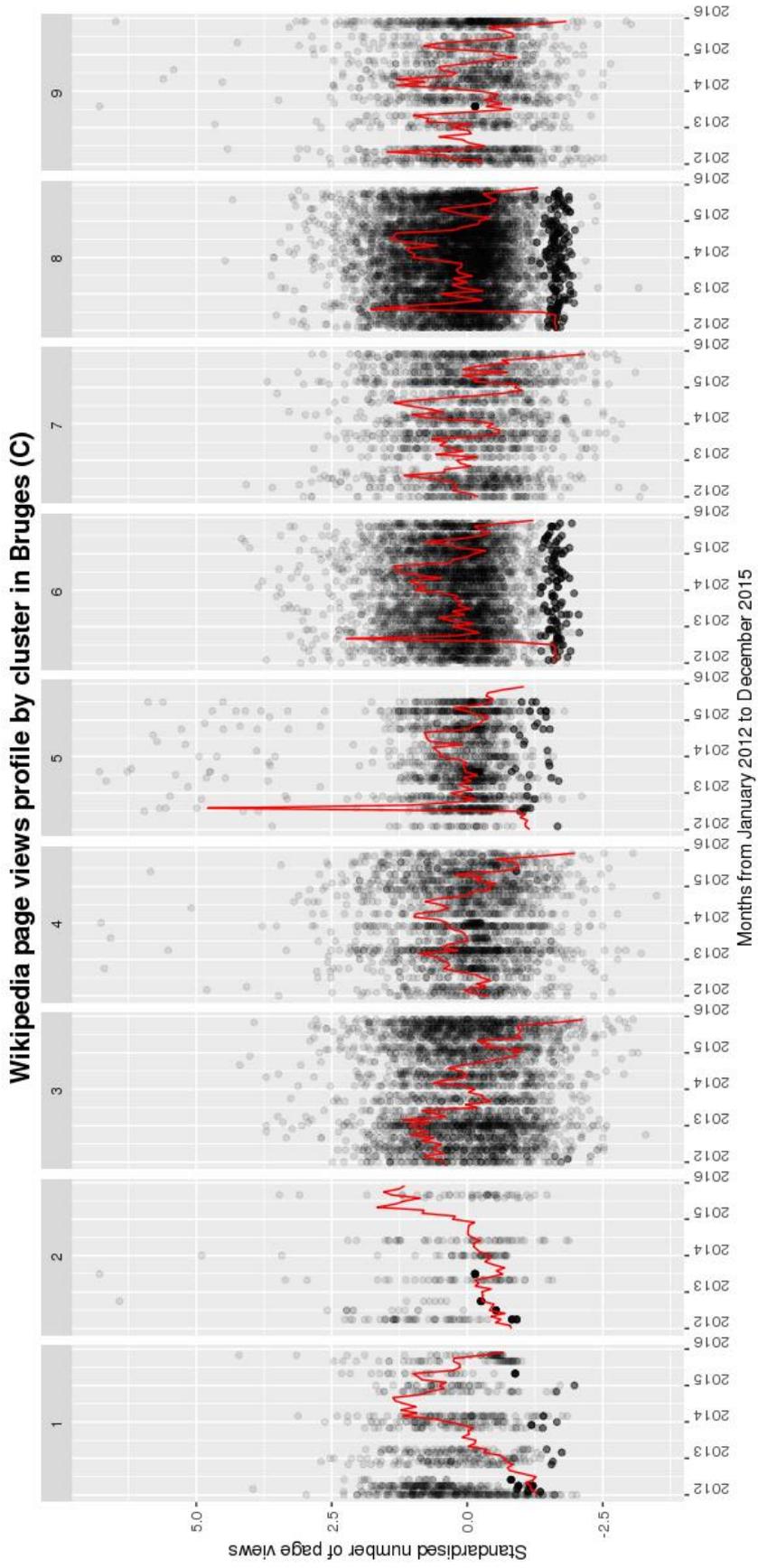


Figure 31 – Result of cluster analysis in Bruges (Urban Audit Level F)

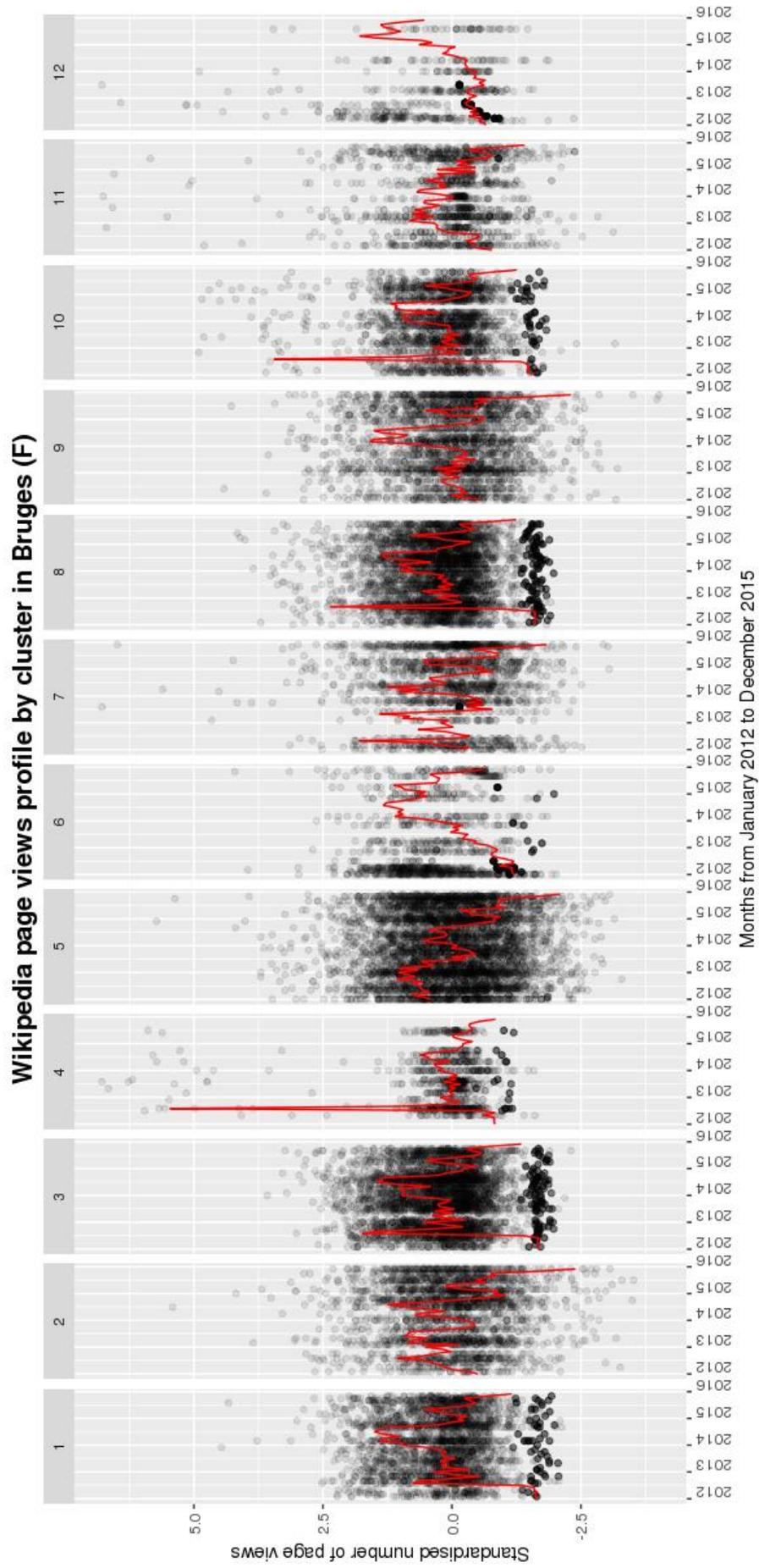
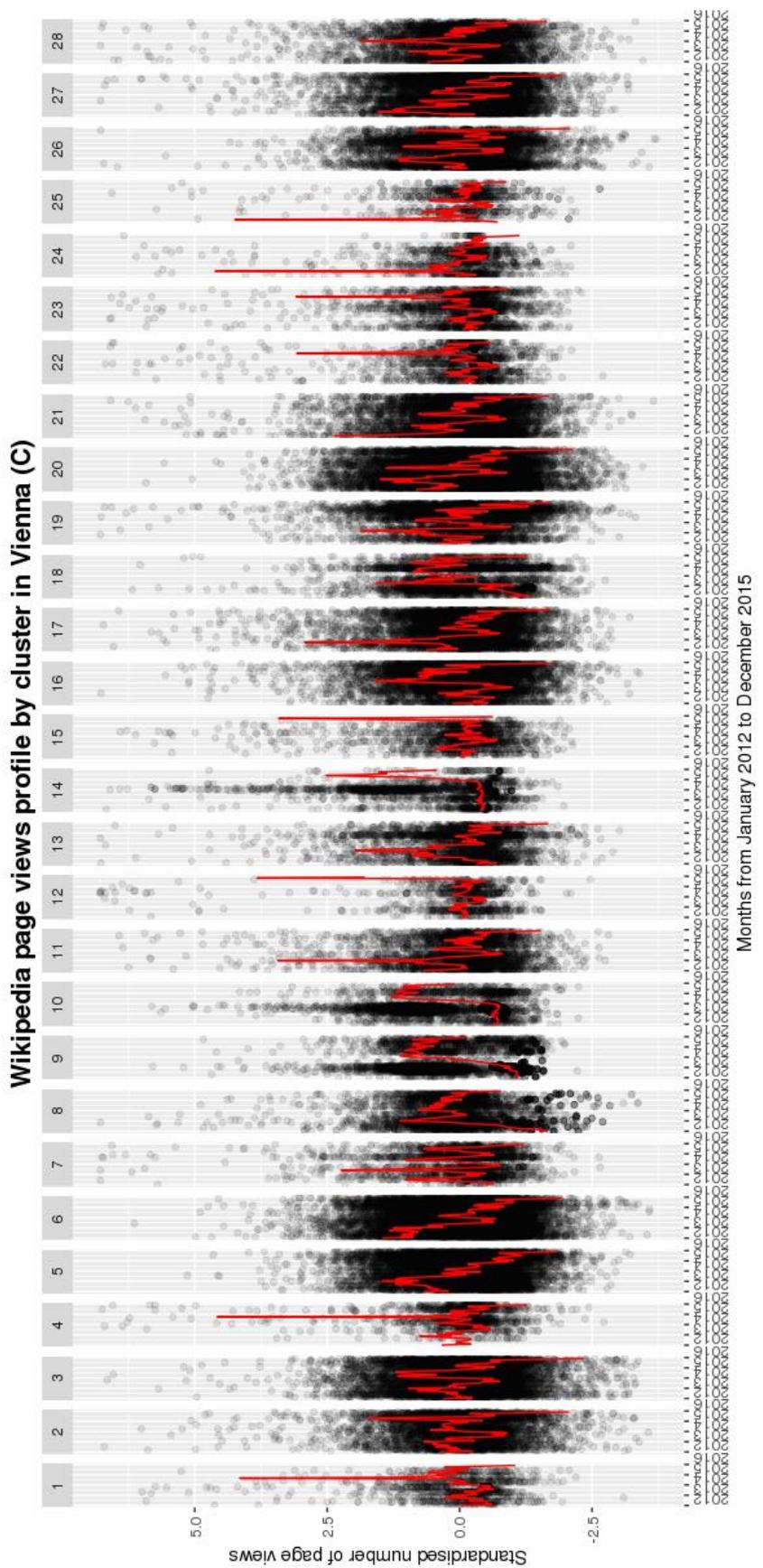


Figure 32 – Result of cluster analysis in Vienna (Urban Audit Level C)



As you can notice, some clusters present a similar pattern in page views, and this allow to group them together. In particular, we have:

- Barcelona C:
  1. Descending pattern with high peak in the first half of the period (clusters 4, 13, 14);
  2. Growing pattern with peaks in the second half of the period (clusters 1 and 7);
  3. Descending pattern with distributed peaks (clusters 3, 9, 12);
  4. Constant pattern with peaks in the second half of the period (clusters 2, 8, 11);
  5. “U” pattern (clusters 5 and 6);
  6. Constant pattern with only one final peak (cluster 10).
  
- Barcelona K:
  1. Descending pattern with high peak in the first half of the period (clusters 3, 10);
  2. Growing pattern with peaks in the second half of the period (clusters 5, 8, 15);
  3. Descending pattern with distributed peaks (clusters 11, 13, 14, 17);
  4. Constant pattern with peaks in the second half of the period (clusters 7 and 16);
  5. “U” pattern (clusters 1 and 12);
  6. Constant pattern with only one final peak (clusters 2 and 18);
  7. “V” pattern (clusters 4, 6, 9).
  
- Bruges C:
  1. Growing pattern (cluster 2);
  2. Descending pattern (clusters 3, 4, 7, 9);
  3. Constant pattern with high peak in the beginning (cluster 5);
  4. “M” pattern (clusters 6 and 8);

5. “V” pattern (cluster 1).
- Bruges F:
    1. Growing pattern (cluster 12);
    2. Descending pattern (clusters 2, 5, 7);
    3. Constant pattern with high peak in the beginning (cluster 4);
    4. “M” pattern (clusters 1, 3, 8, 10);
    5. “V” pattern (cluster 6);
    6. Constant pattern with high peak in the end (cluster 9);
    7. “U” pattern (cluster 11).
  - Vienna C:
    1. Descending pattern (clusters 2, 3, 5, 6, 19, 20, 21, 26, 27, 28);
    2. Constant pattern with peaks in the second half of the period (clusters 14, 15, 22, 23);
    3. Descending pattern with high peak in the first half of the period (clusters 11, 13, 17, 24, 25);
    4. Constant pattern with only one final peak (cluster 12);
    5. “V” pattern (cluster 1);
    6. Descending pattern with high peak in the second half of the period (cluster 4);
    7. Growing pattern with peaks in the second half of the period (clusters 8, 9, 10, 18);
    8. Constant pattern with a lot of peaks (cluster 7);
    9. “N” pattern (cluster 16).

So now we can perform the cluster analysis again, specifying a fewer number of groups (Figures 33 to 37).

Figure 33 – Result of cluster analysis in Barcelona (Urban Audit Level C)

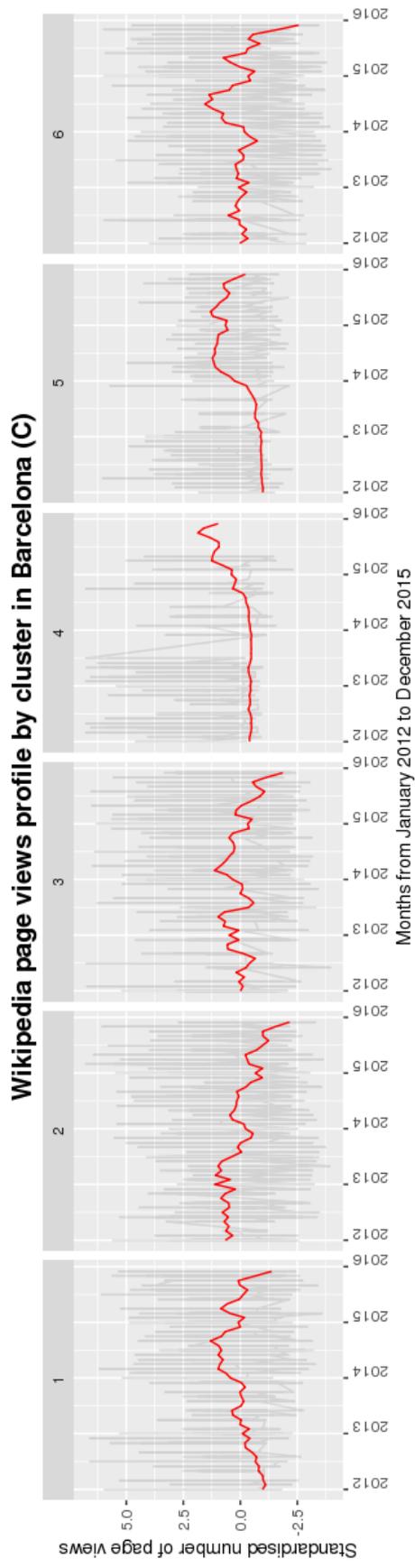


Figure 34 – Result of cluster analysis in Barcelona (Urban Audit Level K)

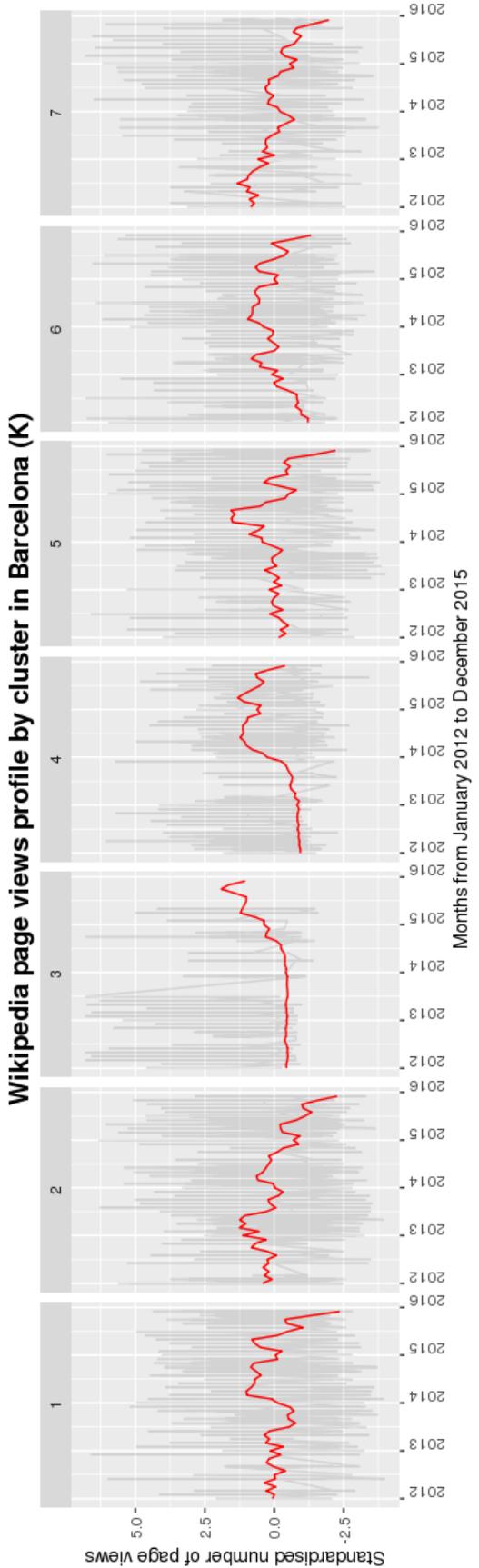


Figure 35 – Result of cluster analysis in Bruges (Urban Audit Level C)

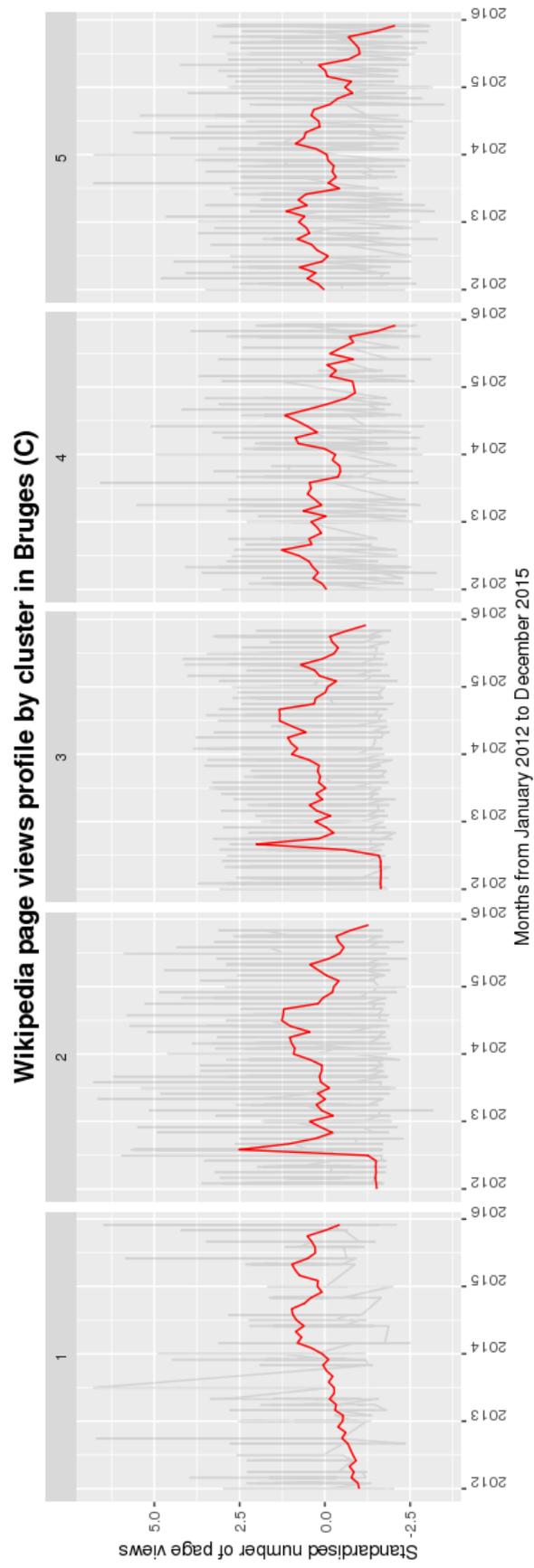


Figure 36 – Result of cluster analysis in Bruges (Urban Audit Level F)

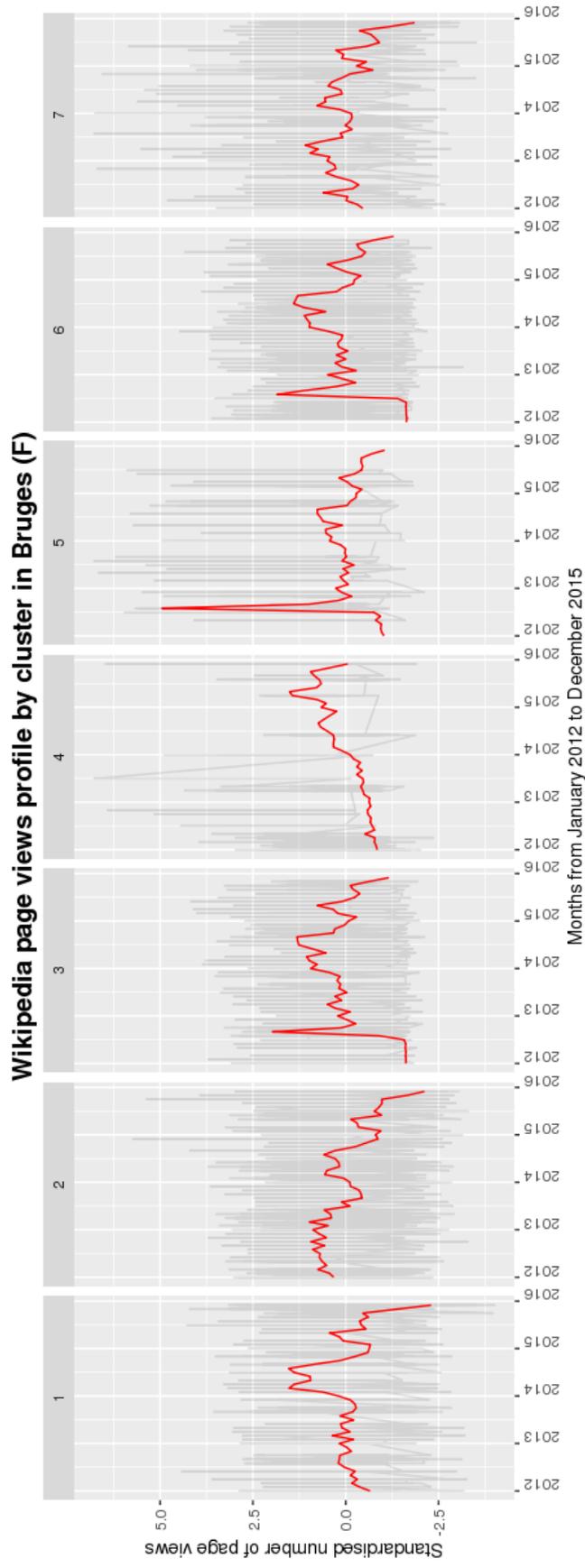
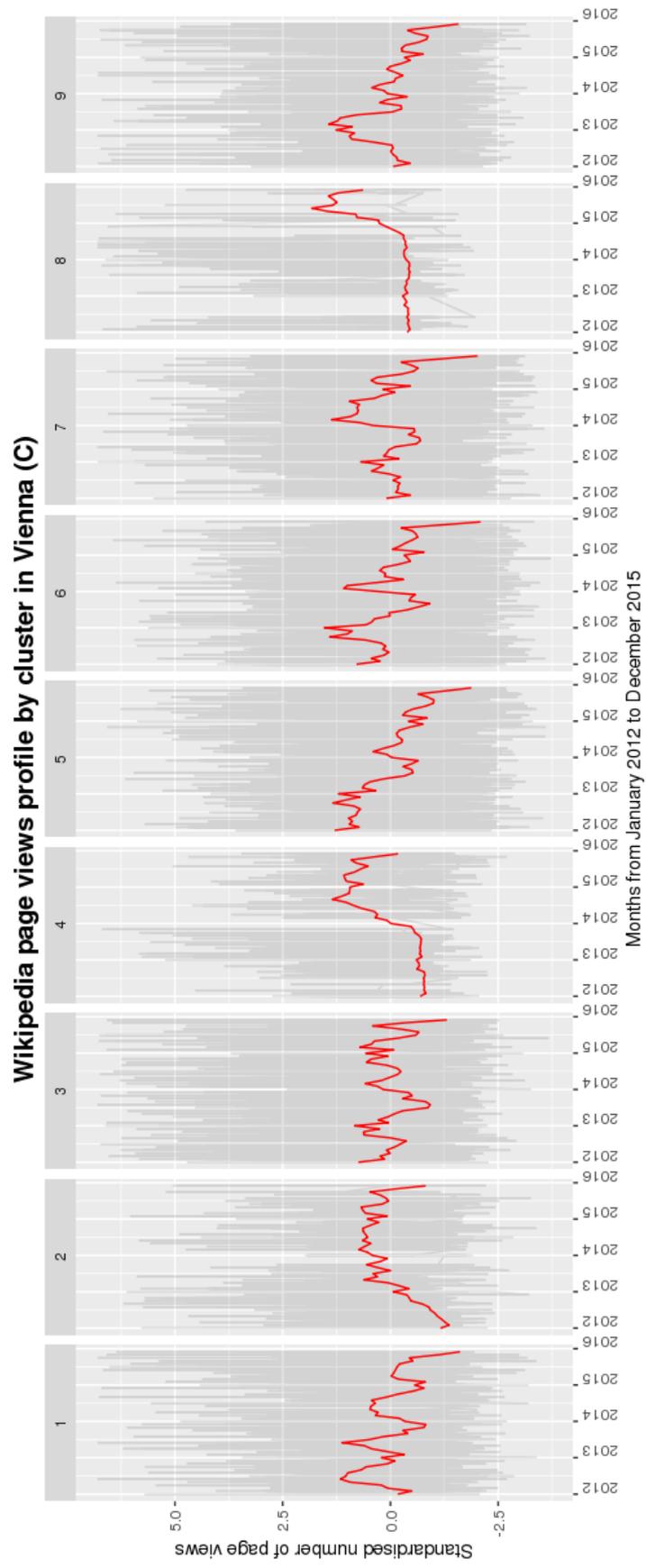


Figure 37 – Result of cluster analysis in Vienna (Urban Audit Level C)



From the second series of charts, you can notice that a few common patterns appear, but, as in the previous case, the similarity between points of interest does not give us an idea of the factors that attract tourists. We decided to completely change the procedure, performing some topic modelling.

#### *4.3.1.4 Third attempt of categorization using LDA and string match*

We decided, then, to build a classification that could reflect the real content of the Wikipedia articles. Topic modelling represents the area we are dealing with, and in particular, we used an algorithm called Latent Dirichlet Allocation (LDA). It is defined as a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic Latent Semantic Analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior.

In particular, we assume that there are  $j$  underlying latent topics according to which documents are generated, and then each topic is represented as a multinomial distribution over the  $|V|$  words in the vocabulary. A document is generated by sampling a mixture of these topics and then sampling words from that mixture.

The number of topics  $k$  has to be fixed a-priori. The LDA model assumes the following generative process for a document  $w = (w_1, \dots, w_N)$  of a corpus  $D$  containing  $N$  words from a vocabulary consisting of  $V$  different terms,  $w_i \in \{1, \dots, V\}$  for all  $i = 1, \dots, N$ .

The generative model consists of the following three steps.

**Step 1:** The term distribution  $\beta$  is determined for each topic by

$$\beta \sim \text{Dirichlet}(\delta)$$

**Step 2:** The proportions  $\theta$  of the topic distribution for document  $w$  are determined by

$$\theta \sim \text{Dirichlet}(\alpha)$$

**Step 3:** For each of the  $N$  words  $w_i$

(a) Choose a topic  $z_i \sim \text{Multinomial}(\theta)$

(b) Choose a word  $w_i$  from a multinomial probability distribution conditioned on the topic  $z_i$ :  $p(w_i|z_i, \beta)$

$\beta$  is the term distribution of topics and contains the probability of a word occurring in a given topic

In R, a dedicated package is available on CRAN, which is called *topicmodels*<sup>16</sup> and allows using LDA models and Correlated Topics Models (CTM).

The data on which we perform the analysis is the content of the articles, called *Wikimarkup*. This is made available again by the Wikimedia Foundation. The function that was built to download the Wikimarkup of each article is available in the GitHub repository.

The classification has to be done in one single language, but this does not represent a problem in our study. As we are working on Wikidata items, which they group in a single place the same article in different languages, once we are able to classify one Wikipedia article, the classification is easily extendable to the same article in other languages, and to the Wikidata item. We computed the number of articles we have in each language and we chose the language with the highest number of articles. In our case, the languages chosen were Spanish for Barcelona, Dutch for Bruges and German for Vienna.

Before applying the model, it is necessary to perform some text mining on the content of the articles. This is done through the *tm* R package<sup>17</sup>; in particular:

- transformation of the Wikimarkup into corpus;
- transformation of characters into lower letters;
- removal of some symbols, as €, ", etc., punctuation, numbers, whitespaces;
- removal of stopwords for the specific language (the default ones provided into the package);
- stemming<sup>18</sup> of all words in the corpus;

---

<sup>16</sup> <https://cran.r-project.org/web/packages/topicmodels/index.html>

- removal of some particular stopwords that often appeared and biased the results (i.e. "wien", "austria", "osterreich", "vienna" for Vienna).

After these cleaning operations, we build a document term matrix and apply the LDA algorithm. The only thing we have to specify in advance is the number of topics we want to identify from the articles. We tried different numbers, and looking every time at the results we found out which was the best choice.

After a check on the official tourism data, we decided to perform the analysis only considering the Urban Audit Level C, as it represents the borders the tourism offices refer to in their data.

This algorithm is able to group the articles and identify the main categories in the city. Unfortunately, the method is not 100% precise, as some categories are a sort of a mixture and we always have one unclear category for each city. We decided then to identify some keywords (computed by the LDA algorithm - words with the highest frequency – and identified by us) that defined the categories and use them to classify unmatched cases. This is possible performing string match between each of the keywords and the title of the unclassified articles.

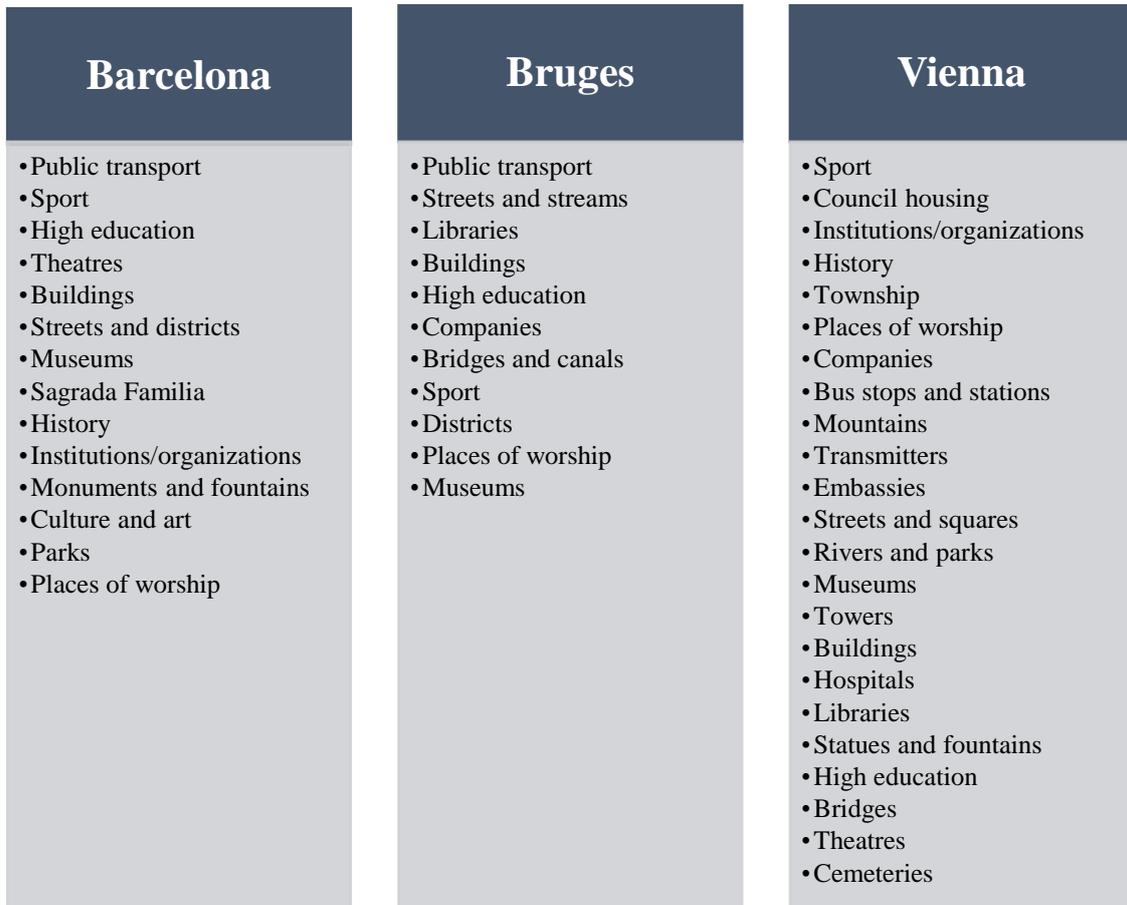
After the classification on Wikipedia articles, we group them by Wikidata item (or point of interest in the city). This approach allows us in the end to classify 95.5% of the total number of Barcelona Wikidata items, 89.7% of Bruges items and 79.2% of Vienna items. We identified 14 categories for Barcelona, 11 for Bruges and 23 for Vienna, plus an 'unclassified' residual category for each of the three cities. The results of the classification are shown in Figure 38.

---

<sup>17</sup> <https://cran.r-project.org/web/packages/tm/index.html>

<sup>18</sup> The process of reducing inflected (or sometimes derived) words to their word stem, base or root form.

Figure 38 – Results of the classification on the Wikipedia articles for the three cities



After the classification is completed, we are able to join the points of interest to the Wikipedia page views, so that we can build rankings among the categories. Considering just the top five for each city, these are the results:

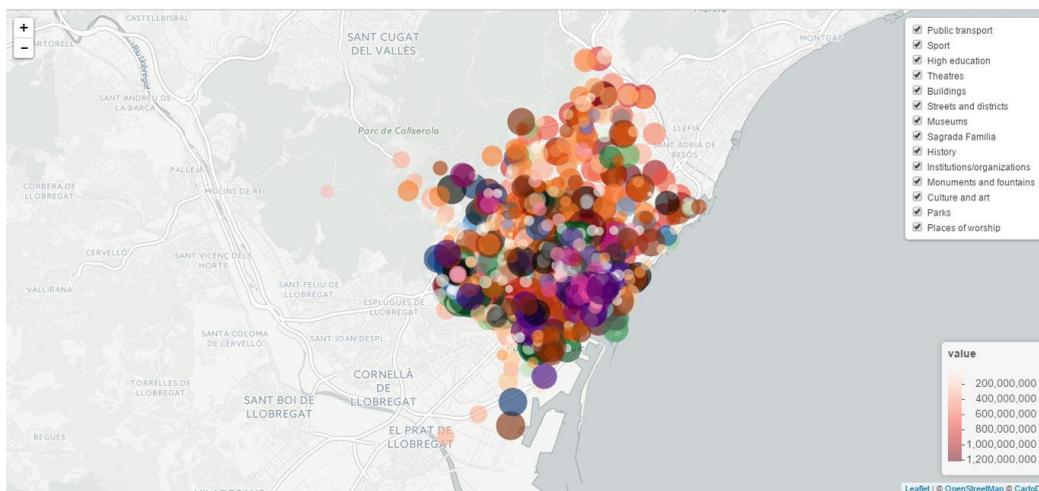
- |                               |                             |
|-------------------------------|-----------------------------|
| <b>Barcelona:</b>             | 3. Districts 13,7%          |
| 1. Sport 64,3%                | 4. Buildings 8,7%           |
| 2. Sagrada Familia 14,8%      | 5. Streets and streams 6,8% |
| 3. Buildings 9%               |                             |
| 4. Public transport 2,8%      | <b>Vienna:</b>              |
| 5. Streets and districts 2,4% | 1. History 38,4%            |
|                               | 2. Institutions 22,8%       |
| <b>Bruges:</b>                | 3. Buildings 7,9%           |
| 1. Sport 40,1%                | 4. Museums 6,7%             |
| 2. Places of worship 23,3%    | 5. Sport 3,7%               |

These rankings were built considering the 31 languages together. Looking at the rankings for each language, no big differences appear, except for those languages that do not have many articles (where only some categories appear in the rankings, due to the lack of Wikipedia articles). We decided to visualize these series in two ways:

- Through interactive maps, like the ones with the top 6 languages
- Through interactive plots of the standardized time series.

Again, these visualizations are available online<sup>19</sup>, you can see a preview of them in the following screenshots (Figures 39 to 44).

Figure 39 – Points of interest per category in Barcelona (Urban Audit Level C)



<sup>19</sup> Maps:

[http://serenasignorelli.altervista.org/Barcelona\\_categories\\_map/Barcelona\\_categories\\_map.html](http://serenasignorelli.altervista.org/Barcelona_categories_map/Barcelona_categories_map.html)

[http://serenasignorelli.altervista.org/Bruges\\_categories\\_map/Bruges\\_categories\\_map.html](http://serenasignorelli.altervista.org/Bruges_categories_map/Bruges_categories_map.html)

[http://serenasignorelli.altervista.org/Vienna\\_categories\\_map/Vienna\\_categories\\_map.html](http://serenasignorelli.altervista.org/Vienna_categories_map/Vienna_categories_map.html)

Plots:

[http://serenasignorelli.altervista.org/Barcelona\\_categories/Barcelona\\_categories.html](http://serenasignorelli.altervista.org/Barcelona_categories/Barcelona_categories.html)

[http://serenasignorelli.altervista.org/Bruges\\_categories/Bruges\\_categories.html](http://serenasignorelli.altervista.org/Bruges_categories/Bruges_categories.html)

[http://serenasignorelli.altervista.org/Vienna\\_categories/Vienna\\_categories.html](http://serenasignorelli.altervista.org/Vienna_categories/Vienna_categories.html)

Figure 40 – Points of interest per category in Bruges (Urban Audit Level C)

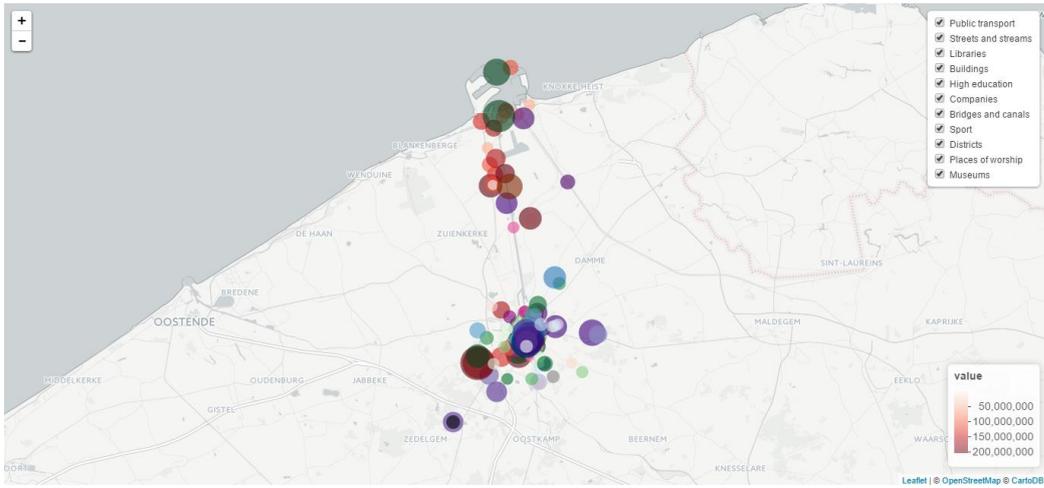


Figure 41 – Points of interest per category in Vienna (Urban Audit Level C)

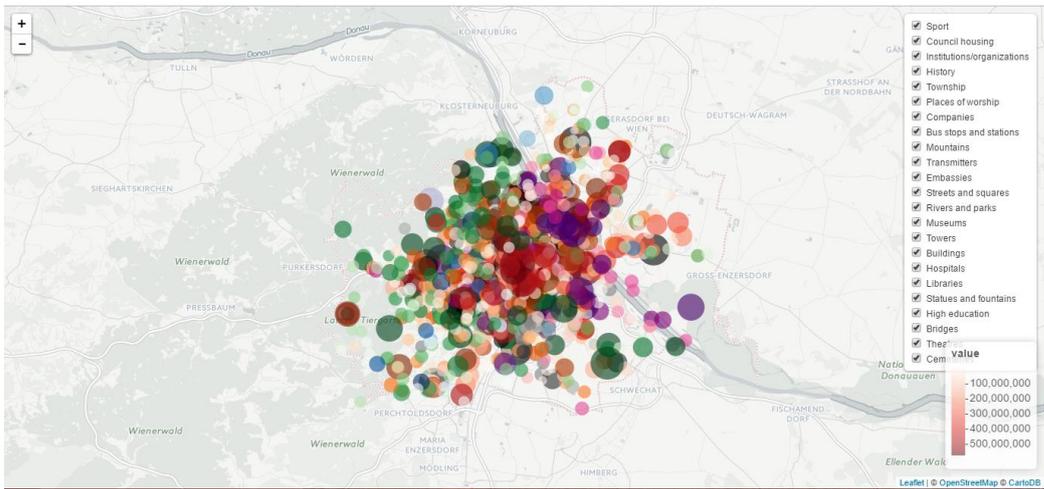


Figure 42 – Standardized time series of categories in Barcelona (Urban Audit Level C)

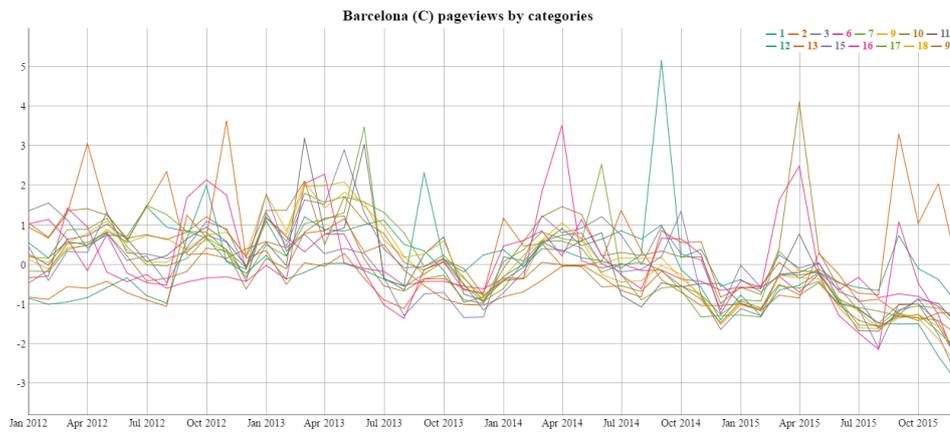


Figure 43 – Standardized time series of categories in Bruges (Urban Audit Level C)

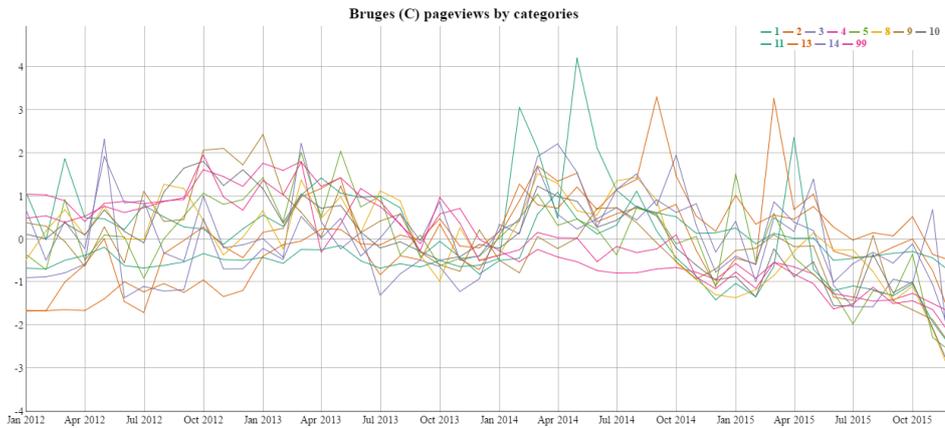
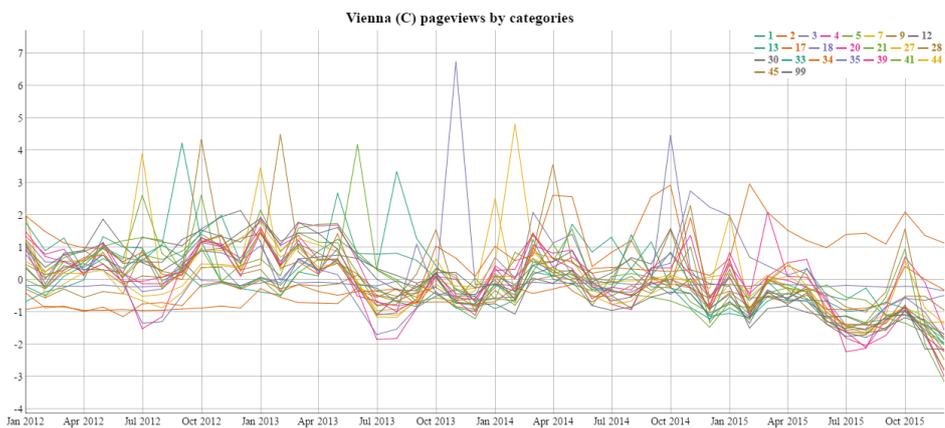


Figure 44 – Standardized time series of categories in Vienna (Urban Audit Level C)



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.

#### 4.3.2 Combined data sources

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

The model that we thought was suitable, at least for a first analysis, is an Autoregressive Integrated Moving Average with Explanatory Variables (ARIMAX). The choice was made considering that the official tourism series we are trying to model could be in some way linked to the values assumed in previous periods (that's why the AR part) and that the regression errors could represent a linear combination of error

terms whose values occurred contemporaneously and at various times in the past (the MA part). The 'I' stays for 'Integrated', because, as we will see, the series are non-stationary and we had to difference them before. We also decided to standardize the series. But let's see in detail how we proceed.

First of all, the variables that we consider in this phase are:

- For tourism data: number of passengers and number of bookings, which they respectively represent arrivals and overnight stays.
- For the big data source, we consider all the classified series that we identified in the previous paragraph (14 series for Barcelona, 11 for Bruges and 23 for Vienna, not considering the unclassified category).

Before trying to model the series, it is necessary to look at them to identify some issues. First of all, we decompose the series and find that all the series present seasonality, so we remove it in order to deal with seasonal adjusted data. A trend component is also present in all the series, so we have to difference the series before trying to model them.

#### 4.3.2.1 Initial models

Once the series are ready, we load the *forecast*<sup>20</sup> R package. It allows us to use an ARIMA function with external regressors (represented by the Wikipedia page views series), resulting in an ARIMAX model. The results appear in Table 12.

Table 12 – ARIMAX results of official tourism data

City	Arrivals	Overnight stays
Barcelona	Random walk	Random walk
Bruges	AR(1)	AR(1)
Vienna	Random walk	Random walk

In the following tables (13 to 18) only the significant parameters that arise from each model are shown.

<sup>20</sup> <https://cran.r-project.org/web/packages/forecast/index.html>

*Table 13 – Significance of parameters in the model for Barcelona arrivals*

<b>BARCELONA Arrivals</b>	Estimate	Std. Error	z value	Pr(> z )
sport	-0.30998697221615362	0.12877206229534527	-2.4073	0.01607 *
theatres	-0.42136590963869014	0.19755381089363347	-2.1329	0.03293 *
institutions_organis	0.32947151907455818	0.18891403770962925	1.7440	0.08115 .
parks	0.68299039610063406	0.22554398903643211	3.0282	0.00246 **
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

*Table 14 – Significance of parameters in the model for Barcelona overnight stays*

<b>BARCELONA Overnight stays</b>	Estimate	Std. Error	z value	Pr(> z )
sport	-0.361016079056739825	0.133185691118201111	-2.7106	0.006716 **
parks	0.711587149642935346	0.233274457772622318	3.0504	0.002285 **
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

*Table 15 – Significance of parameters in the model for Bruges arrivals*

<b>BRUGES Arrivals</b>	Estimate	Std. Error	z value	Pr(> z )
ar1	-0.7727996	0.0984480	-7.8498	4.166e-15 ***
buildings	0.2662761	0.1028732	2.5884	0.009643 **
bridges_and_canals	0.1733007	0.0919555	1.8846	0.059482 .
sport	-0.2186508	0.1012729	-2.1590	0.030848 *
districts	-0.6598638	0.2991065	-2.2061	0.027376 *
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

*Table 16 – Significance of parameters in the model for Bruges overnight stays*

<b>BRUGES Overnight stays</b>	Estimate	Std. Error	z value	Pr(> z )
ar1	-0.77879474	0.09699727	-8.0290	9.824e-16 ***
streets_and_streams	-0.16882002	0.09583642	-1.7615	0.078146 .
buildings	0.30994291	0.09668776	3.2056	0.001348 **
bridges_and_canals	0.14445082	0.08619322	1.6759	0.093759 .
districts	-0.62176140	0.28124965	-2.2107	0.027056 *
companies	0.48045249	0.27553858	1.7437	0.081214 .
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Looking at the sign of the significant parameters, a 'positive' in the models seems reasonable, as that means that an increase in the number of page views of the articles in that category leads to an increase in the physical visits to the city. The negative sign of

other parameters, indeed, seems a little bit uncertain. So before taking any conclusions, we decided to make another step ahead in the analysis.

*Table 17 – Significance of parameters in the model for Vienna arrivals*

VIENNA Arrivals	Estimate	Std. Error	z value	Pr(> z )	
council_housi	-0.21150479025133434	0.11687612441086716	-1.8096	0.070350	.
history	-0.21689593819554567	0.10389593986207492	-2.0876	0.036832	*
institutions_	-0.27068649574237968	0.11013290280547726	-2.4578	0.013978	*
companies	0.21930883645215180	0.10433625184222449	2.1019	0.035558	*
places_of_wors	0.52169366154791974	0.12822147370991749	4.0687	0.0000472778	***
mountains	0.61366345226427943	0.14653629386270214	4.1878	0.0000281682	***
transmitters	-0.16721008364738582	0.09517255720810915	-1.7569	0.078932	.
streets_and_s	-0.36434819945770541	0.09203982426056539	-3.9586	0.0000753926	***
rivers_and_par	0.42523625080431249	0.13447311507870041	3.1622	0.001566	**
museums	-0.30183600731860266	0.13450972715519283	-2.2440	0.024834	*
buildings	-0.70549071830122045	0.13918660297496876	-5.0687	0.000004006	***
libraries	-0.28103575299898470	0.16370068560374301	-1.7168	0.086022	.
statues_and_f	-0.23788281174234691	0.12831294590770873	-1.8539	0.063750	.
high_education	-0.31259171551245335	0.11827955062999722	-2.6428	0.008222	**
bridges	-0.21546620525578508	0.11965678645584629	-1.8007	0.071750	.
theatres	0.19538887642323732	0.10426671478648680	1.8739	0.060940	.
cemeteries	0.26994789836933519	0.12755473973663606	2.1163	0.034317	*
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

The approach we decided to follow is general-to-specific, in the sense that through the models we just showed, we identify the significant categories, and these are the only ones 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 each of them as a unique category.

We want to verify if this significant effect is caused just by one single article or not.

Table 18 – Significance of parameters in the model for Vienna overnight stays

VIENNA Overnight stay	Estimate	Std. Error	z value	Pr(> z )	
institu-	0.23909924113269234724	0.11665870723048821223	-2.0496	0.040407 *	
history-	0.29080966343579456312	0.11005217856315256675	-2.6425	0.008230 **	
township	0.52741045489362692322	0.18312495920444493702	2.8801	0.003976 **	
places_	0.63446662616271076729	0.13581910004981564977	4.6714	0.0000299140 ***	
company	0.28386921464500086687	0.11051858084025541207	2.5685	0.010213 *	
mountain	0.72992056603182664531	0.15521914790050073130	4.7025	0.0000256975 ***	
embassi-	0.27973158526066743690	0.13732596290927245875	-2.0370	0.041651 *	
streets-	0.17445779706511521656	0.09749353904927396397	-1.7894	0.073546 .	
museums-	0.38241441767586953349	0.14247995816202979613	-2.6840	0.007275 **	
rivers_	0.44982111779051214828	0.14244117666777417197	3.1579	0.001589 **	
buildin-	0.82588606473613934700	0.14743395792716282311	-5.6017	0.0000002122 ***	
librari-	0.55873696260414451586	0.17340059953052652797	-3.2222	0.001272 **	
bridges-	0.28150223481248132229	0.12674692005907631231	-2.2210	0.026352 *	
theatre	0.25684970154197328540	0.11044492343104718446	2.3256	0.020040 *	
cemeteri	0.35340679201710567536	0.13511285937010902858	2.6156	0.008906 **	
high_ed-	0.27752304350836104474	0.12528807736668814976	-2.2151	0.026755 *	
---					
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

#### 4.3.2.2 Inspection on significant series

For **Barcelona**, the category *sport* has the article 'Futbol Club Barcelona' that leads the page views rankings with 75.8% of the total amount of the whole category. At the same time, a similar situation appears in the *theatres* category, with the article 'Gran Teatro del Liceo' leading with 75.3% of page views. We then take out these two articles from the categories and re-run the model with only the significant categories, plus the corrected categories and the articles alone, in order to verify if something changes. The updated results are shown in Tables 19 and 20 (only significant parameters).

Table 19 – Significance of parameters in the corrected model for Barcelona arrivals

BARCELONA Arrivals	Estimate	Std. Error	z value	Pr(> z )	
sport	-0.482018	0.130877	-3.6830	0.0002305 ***	
gran_teatro_del_liceo	-0.372502	0.126306	-2.9492	0.0031861 **	
parks	0.362687	0.147159	2.4646	0.0137172 *	
---					
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' ' 1

Table 20 – Significance of parameters in the corrected model for Barcelona overnight stays

BARCELONA Overnight stays	Estimate	Std. Error	z value	Pr(> z )
sport	-0.38893801833560382	0.13689066946161060	-2.8412	0.0044940 **
parks	0.45129409089986566	0.13210024266218320	3.4163	0.0006348 ***
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

As you can easily see, for arrivals, it is the single Wikipedia article related to the theatres that is driving the negative effect on number of passenger, while the *theatres* category itself does not seem to have an importance. The opposite situation appears for *sport*, where the Futbol Club Barcelona article seems not important and the negative effects remain. The situation for the number of bookings, instead, does not change.

For **Bruges**, we investigate three categories: *sport*, *districts* and *streets and streams*. The first one has the article about the soccer team 'Cercle Brugge' that collects 57.3% of the page views. For *districts*, the article about the port 'Zeebrugge' has 82.1% of page views, while 'Belfort Van Brugge' for the last category collects 45.3% of page views. Following the same procedure as for Barcelona, we re-run the model, obtaining the results in Tables 21 and 22.

Table 21 – Significance of parameters in the corrected model for Bruges arrivals

BRUGES Arrivals	Estimate	Std. Error	z value	Pr(> z )
ar1	-0.7456239	0.1016159	-7.3377	2.173e-13 ***
buildings	0.1976131	0.0864569	2.2857	0.02227 *
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 22 – Significance of parameters in the corrected model for Bruges overnight stay

BRUGES Overnight stays	Estimate	Std. Error	z value	Pr(> z )
ar1	-0.8040564	0.0919090	-8.7484	< 2.2e-16 ***
buildings	0.3175176	0.0822512	3.8603	0.0001132 ***
districts	-0.5548691	0.2045750	-2.7123	0.0066818 **
companies	0.5243503	0.2342962	2.2380	0.0252223 *
zeebrugge	-0.2993752	0.1460667	-2.0496	0.0404056 *
belfort_van_brugge	-0.2037880	0.0975403	-2.0893	0.0366835 *
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Concerning the number of passengers, the negative effects of *sport* and *districts* disappears, leaving only the positive effect of *buildings*.

If we look at the number of bookings, the negative effect of *streets and streams* disappears, leaving just the negative effect of the main article (in number of page views) of the category, 'Belfort Van Brugge'. The negative effect of *districts* remains and gets reinforced by the article that was taken out from the category, 'Zeebrugge'. On the positive side, as for number of passengers, we still see the positive effects of *buildings* and *companies*, while *bridges and canals* disappeared.

To perform the same kind of analysis on the city of **Vienna**, we had to consider only the parameters with a level of significance of 5% or less, otherwise we would have more regressors than observations in our model. As for the previous cases, we decided to split the categories that have a negative effect on number of passengers and number of bookings, taking out the article that have the highest number of page views. Here it is a recap of the articles per category:

- for *history*, the article 'Österreich-Ungarn';
- for *institutions/organizations*, 'Organisation erdölexportierender Länder' (OPEC);
- for *streets and squares*, 'Wiener Ringstraße';
- for *museums*, the article 'Der Kuss (Klimt)';
- for *buildings*, 'Schloss Schönbrunn';
- for *high education*, 'Universität Wien'.

The output of the new models is available in Tables 23 and 24.

For Vienna number of passengers, the negative effect of *history* category disappears, leaving only the effect of the article about the Austro-Ungarian empire. The same happens for *high education*, in fact only the article about the University of Vienna appears among the significant parameters. The opposite situation appears for *institutions*, that continue to show a significant negative impact, while the article about OPEC does not appear. Concerning the *buildings* category, a double effect shows up: the category itself continues to have a negative effect, and it is amplified by the article about Schloss Schönbrunn, a castle in Vienna.

Table 23 – Significance of parameters in the corrected model for Vienna arrivals

VIENNA Arrivals	Estimate	Std. Error	z value	Pr(> z )	
institutions_org	-0.36298965941271039	0.09855037654282442	-3.6833	0.0002302	***
places_of_worship	0.32820297327481268	0.12721534205886809	2.5799	0.0098829	**
companies	0.19521932214393181	0.10402101248952993	1.8767	0.0605552	.
mountains	0.28961368727461345	0.09116441673164651	3.1768	0.0014890	**
buildings	-0.31661461801440505	0.10008425899850198	-3.1635	0.0015589	**
cemeteries	0.24054310627959383	0.10647146566921614	2.2592	0.0238693	*
schloss_schonbrun	-0.27755183898650693	0.09935496000118831	-2.7935	0.0052135	**
universitat_wien	-0.30124015245408076	0.11063706750558214	-2.7228	0.0064736	**
osterreich_ungarn	-0.24502760409333993	0.09724452586309917	-2.5197	0.0117453	*
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Regarding the number of bookings, some results that we just saw for arrivals remain, while others change completely.

Table 24 – Significance of parameters in the corrected model for Vienna overnight stays

VIENNA Overnight stays	Estimate	Std. Error	z value	Pr(> z )	
places_of	0.344184620137694797	0.125318300549748907	2.7465	0.0060238	**
companies	0.219496196483167205	0.090272504096270204	2.4315	0.0150371	*
mountains	0.141981304176390072	0.075859911025589907	1.8716	0.0612585	.
embassies	-0.529876307822059633	0.133429722547206003	-3.9712	0.0000715110764	***
rivers_an	-0.936682864173964291	0.254396200297594188	-3.6820	0.0002314	***
museums	0.638847732524532974	0.236769384418305223	2.6982	0.0069719	**
buildings	-0.432914712782285160	0.156215129904921757	-2.7713	0.0055838	**
high_educ	-0.335415359429994675	0.137315237544817559	-2.4427	0.0145792	*
theatres	0.322709276944901258	0.117829570406176615	2.7388	0.0061668	**
cemeterie	0.896416392969378806	0.161753163363808128	5.5419	0.000000299244	***
schloss_s	-0.409401437305662363	0.114108891777931279	-3.5878	0.0003335	***
universit	-0.271926024897625851	0.086755145839510842	-3.1344	0.0017220	**
osterreic	-0.571975443428321850	0.089908252638242675	-6.3618	0.000000001994	***
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Considering *history* and *buildings*, the situation is the same as before, with the Austro-Ungarian empire article that remains instead of the category, and the other category that shows a double negative effect (amplified by the castle article). The *high education* category in this case behave in the same way as *buildings*, showing a double negative effect. Other two categories appear, *embassies* and *rivers and parks*, showing

that the articles that collected the most page views do not appear to influence the number of bookings.

#### 4.3.2.3 Final models

Considering only the significant parameters, the model that best describes the number of passengers and number of bookings in **Barcelona** is a random walk composed as follows:

*Number of passengers*

$$\begin{aligned} &= \text{Number of passengers}_{t-1} - 0.482018\text{sport} \\ &- 0.372502\text{Gran Teatro del Liceo} + 0.362687\text{parks} + \epsilon_t \end{aligned}$$

*Number of bookings*

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

As the models suggest, the number of passengers arriving in Barcelona each month is positively affected by the number of page views concerning *parks*, while a negative effect is related to page views about *sport* and the specific Gran Teatro del Liceo. The same negative effect of *sport* category is visible also in the model that explains the number of bookings in the city, plus the positive effects of the page views of articles about *parks*.

We decide also to display the significant parameters in a confidence intervals chart, in order to see not only the parameters with the highest values, but also their variance, to better understand the uncertainty associated with them. In these charts, the parameters are ordered according to their values, and coloured in green or red depending on the positive or negative sign of the effect. In the middle of the chart we highlighted in dark grey the zero value, so that it is possible to quickly identify the parameters with the highest effect, which are the most distant from that line.

Figures 45 and 46 display the results for Barcelona number of passengers and number of bookings.

Figure 45 – Significant parameters in the model for Barcelona arrivals

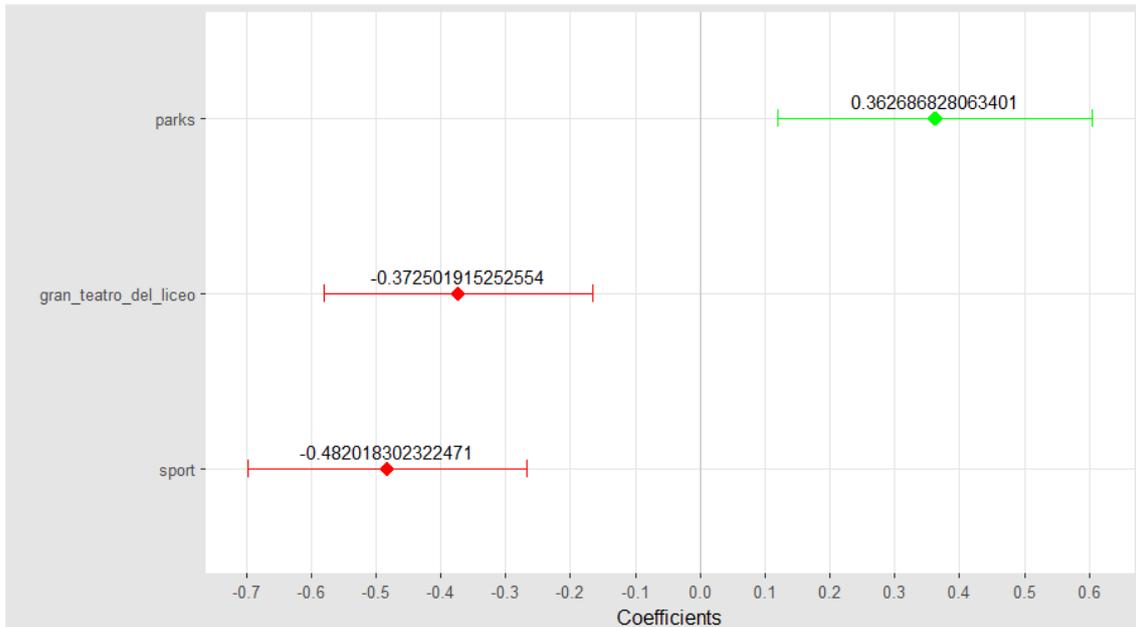
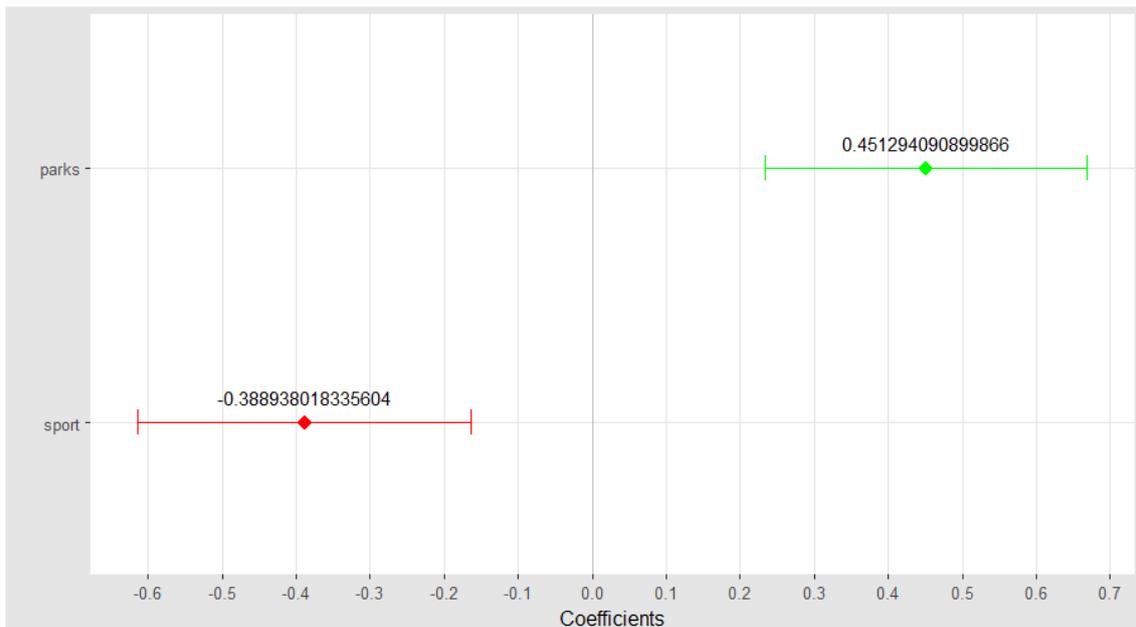


Figure 46 – Significant parameters in the model for Barcelona overnight stays



For **Bruges**, an autoregressive model of order one best describes arrivals and overnight stays as follow:

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

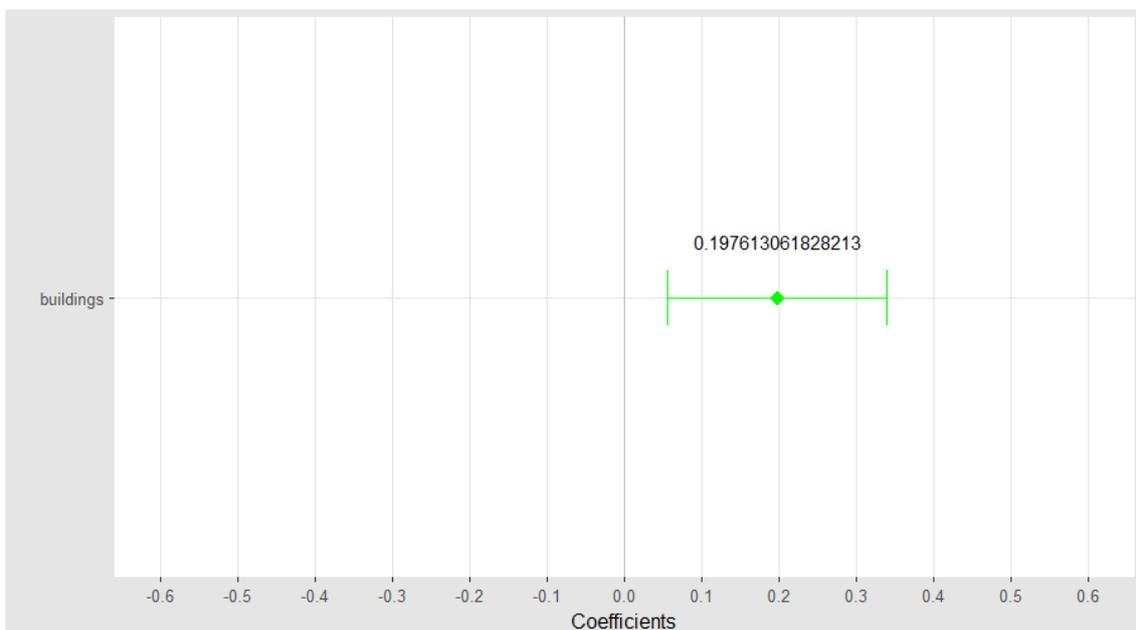
$$\begin{aligned} \text{Number of bookings} \\ &= -0.8040564\text{Number of bookings}_{t-1} + 0.3175176\text{buildings} \\ &\quad - 0.5548691\text{districts} + 0.5243503\text{companies} - 0.2993752\text{Zeebrugge} \\ &\quad - 0.2037880\text{Belfort Van Brugge} + \epsilon_t \end{aligned}$$

The number of passengers is positively affected only by *buildings*. This category also shows a positive effect on number of bookings, combined with another positive effect from *companies*.

The overnight stays are affected also by negative effects from articles about *districts* and two specific articles, Zeebrugge (the port of the city) and Belfort Van Brugge (a medieval bell tower in the city center).

Figures 47 and 48 show the confidence intervals for the parameters.

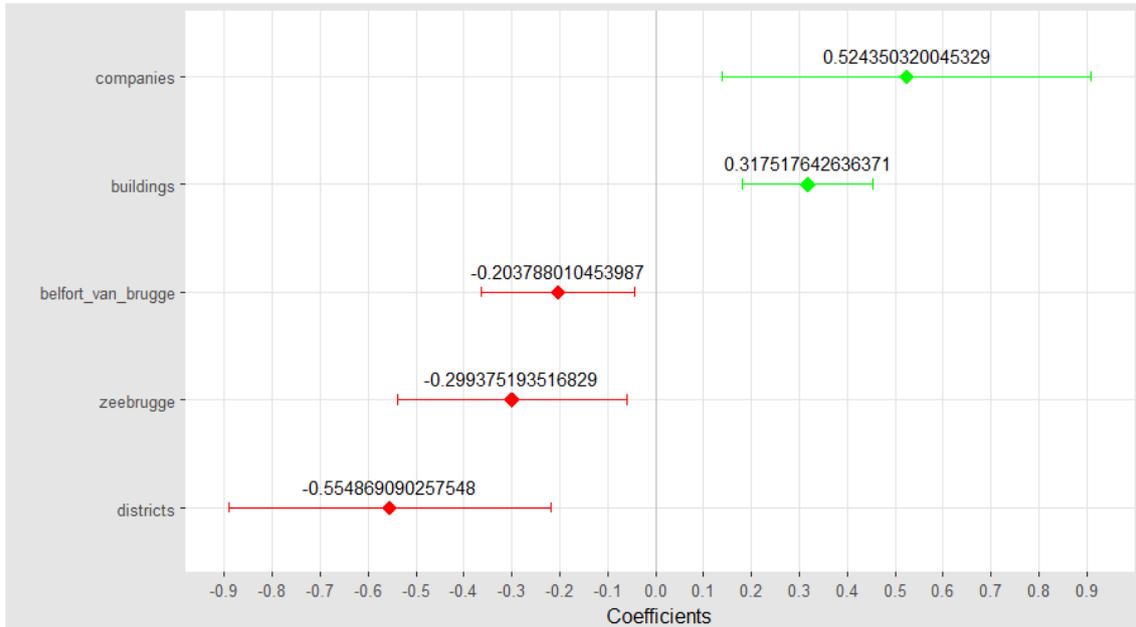
Figure 47 – Significant parameters in the model for Bruges arrivals



The parameters that affect Bruges number of bookings show different levels of uncertainty. It is sufficient to look at the two parameters with positive effects: *companies* and *buildings*. The value of the former is higher than the latter, but the chart well displays how uncertainty plays a role in this case. The value for the parameter

*companies* could be in a range that goes from 0.1 to 0.9, while the range for *buildings* is less than a half of the previous one.

Figure 48 – Significant parameters in the model for Bruges overnight stays



Again, the random walk is the model that best describes the situation for **Vienna**:

*Number of passengers*

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

*Number of bookings*

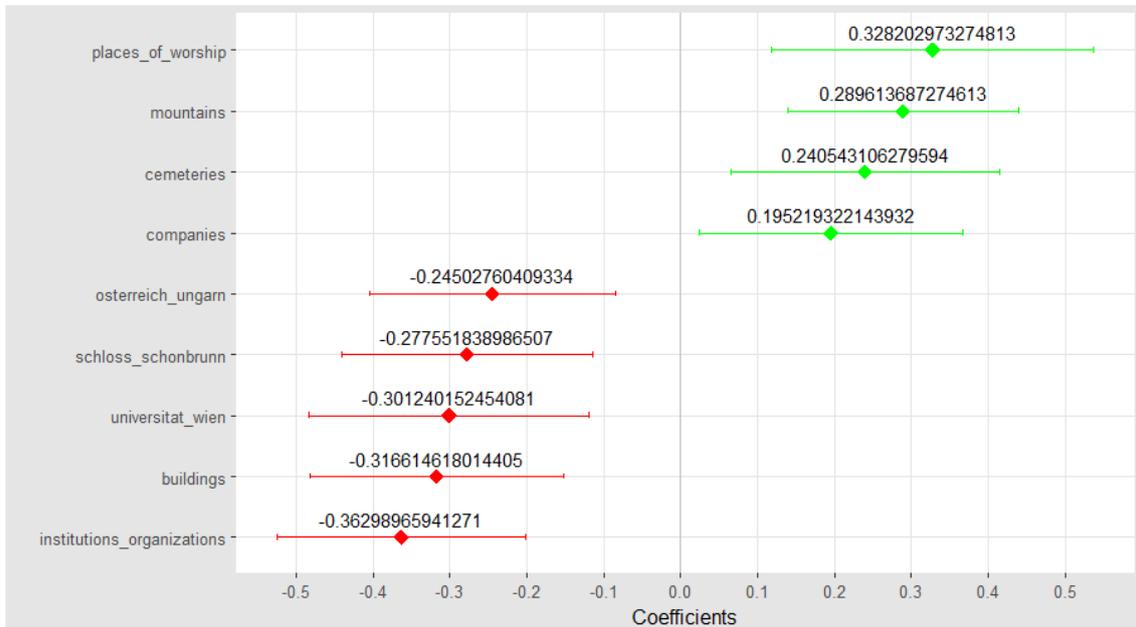
$$\begin{aligned}
 &= \text{Number of bookings}_{t-1} + 0.344185 \text{places of worship} \\
 &+ 0.219496 \text{companies} - 0.529876 \text{embassies} \\
 &- 0.936683 \text{rivers and parks} + 0.638848 \text{museums} - 0.432915 \text{buildings} \\
 &- 0.335415 \text{high education} + 0.322709 \text{theatres}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

Vienna shows a higher number of categories that affect number of passenger and bookings.

In particular, arrivals seem positively affected by page views concerning *places of worship*, *mountains* and *cemeteries*, while articles about *institutions/organizations* and *buildings* (plus three specific articles about the University of Vienna, the Schonbrunn castle and the Austro-Hungarian Empire) show a negative effect.

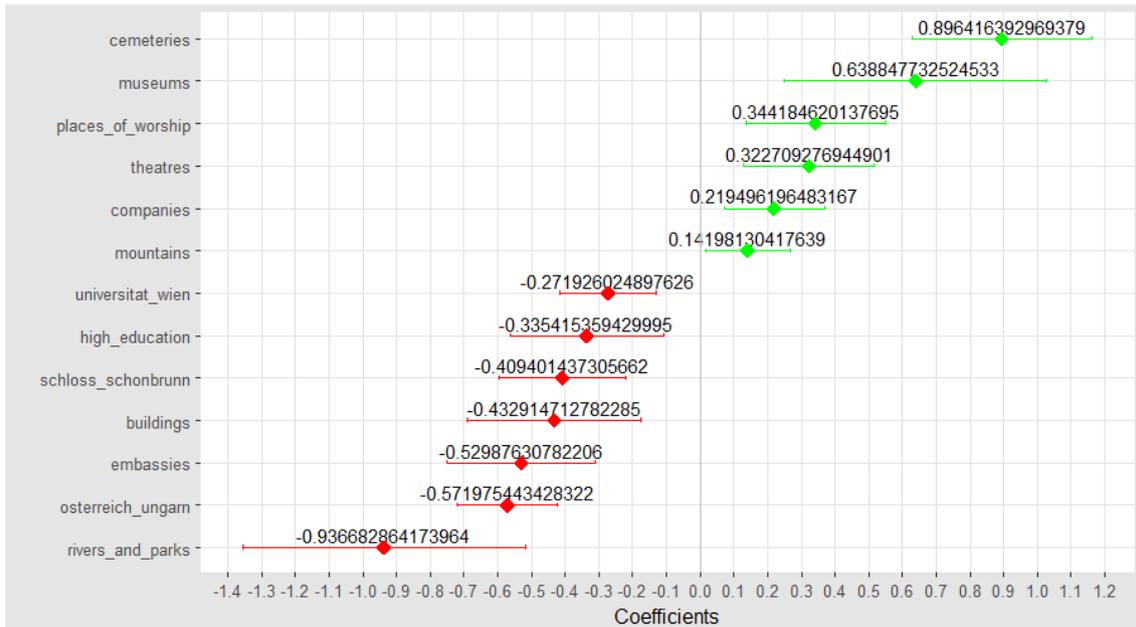
The parameters that affect overnight stays are a little bit different, with a positive effect from *places of worship*, *companies*, *museums* and *theatres*. A negative effect seems provoked by the visits to articles about *embassies*, *rivers and parks*, *buildings* and *high education* (plus the abovementioned three specific articles).

Figure 49 – Significant parameters in the model for Vienna arrivals



For Vienna arrivals, you can notice how the uncertainty of the parameters is more or less the same, while for overnight stays it assumes different ranges, being higher for *rivers and parks*, *museums* and *cemeteries*. For the other parameters, the uncertainty decreases, reaching the lowest values for *companies*, *mountains* and the Austro-Ungarian empire article.

Figure 50 – Significant parameters in the model for Vienna overnight stays



## 4.4 Conclusions and future research

The analysis seems promising, with further analysis to perform, but a lot of work still has to be done and some issues have to be taken into account.

First of all, we should consider the edit history of a Wikipedia article. Plotting the series, we could detect some strange peaks in some categories, followed by a return to their previous trend. This could represent an edit of the article (contributors make some changes and keep opening the article to check if it is ok) or even the phenomenon of *edit warring* that occurs when editors who disagree about the content of a page repeatedly override each other's contributions. The solution to this could be detecting the edits in the time series and adjust the page views number accordingly (by reduce them in some way or adding a dummy variable to the model).

Second, 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.

Third, it would be interesting to split the tourism flow into residents' tourism and foreigners' tourism, in order to identify if something changes in the patterns.

Finally, in our analysis we just considered the page views from desktop connections (due to technology constraints)<sup>21</sup>, while it would be interesting to consider also mobile connections. We can easily think about tourists visiting a city and, when in front of a monument or whatever point of interest, taking their smartphone to look for additional information. Usually, the first source that appears among the results, and we suspect the most visited, is Wikipedia.

A further point that would be interesting to develop, could be the development of a collaboration with the Wikimedia Foundation, in the hypothesis of getting IP addresses of the page views. This variable would incredibly improve the analysis, giving the opportunity to know not only the language of the consulted Wikipedia article, but also the country when the connection happens.

## **4.5 Appendix**

During the analysis, we had to perform some quality checks on the data we were getting. In particular, we checked:

- the quality of Wikidata classification through properties;
- the quality of the points of interest we identified in Vienna, comparing them with some official points from the municipality of Vienna;
- the change in the number of Wikidata items queried every month;
- the boundaries of official tourism data.

### ***4.5.1 Quality of Wikidata classification***

When we were trying to use the already given Wikidata categorization, we needed to check first for the quality of it. To check the quality of properties and/or classes assigned by Wikidata, we decided to take a sample and manually check it. We decided

---

<sup>21</sup> The way the Wikimedia Foundation structures mobile connections data is different from the desktop one, we had to decide which of the two to develop. When I arrived at Eurostat, the technology was already implemented for desktop connections, so we just continue using the same tools.

to take a 10% sample (of property and classes data frame) for each city. This sample has been joined to the Wikipedia articles' list in order to check for the quality. (For each city, we considered only the biggest Urban Audit level, so K for Barcelona, F for Bruges and C for Vienna).

In Table 25 a brief summary of the quality check.

*Table 25– Significance of parameters in the corrected model for Vienna overnight*

<i>City</i>	<i>Level</i>	<i>Size of sample</i>	<i>No of missing categories</i>	<i>No of wrong categories</i>
Barcelona	K	145	30 (20.69%)	14 (12.17%)
Bruges	F	65	38 (58.46%)	4 (14.81%)
Vienna	C	266	73 (27.44%)	15 (7.77%)

We can notice that for the smallest city, Bruges, we have a highest number of missing categories (nearly 60%) as well as the worst quality, that by the way does not exceed 15% of the sample.

In Barcelona, the missing categories represent one fifth of the sample, and the quality is better than Bruges, being around 12%.

In terms of quality, Vienna showed the best results, with just nearly 8% of wrong categories, with 27% of missing categories. It is worth noting that only three categories in Vienna were completely wrong, the other twelve were wrong in the sense that they were more generic (i.e. train station categorized just as 'building', or park as 'geographical object').

#### ***4.5.2 Quality of official points from municipality of Vienna***

For the city of Vienna, the municipality of Vienna provides a shapefile with points in the city, included their names. We decided that it would be nice to compare the points of interest we identified through our analysis with these 'official' points. The municipality of Vienna identifies 4.331 points; the file has been loaded on the Open Government Website on February 27th, 2016.

After a brief meeting with a geostatistician colleague, we decided to join the points in a two steps procedure:

- in ArcGIS, using the Near function. Starting from each of the official points, we computed the 20 shortest distance to the Wikidata points of interest (building a new table with all the variables from the two original datasets and the distances in meters;
- in RStudio, using a string match, to see whether the name of the official point appears in the Wikipedia article's name. Some attention will be paid to this step, as the procedure could lead to false positives.

The idea that we decided to follow is a sort of combination of the previous points, dividing the dataset in two parts, one for training and one for testing, and building a random forest algorithm to categorize items. The random forest algorithm requires features, so we decided to include both a geographic feature (first point) and several text features (second point). Going more into the details, the features we used are the following (the names represent the name of the variables in the dataset):

1. geographic distance:

- NEAR\_DIST; it is the distance that we built in ArcGIS, so the physical distance between one official point of Vienna and one Wikidata point of interest (for each official point we consider just the closest 20);

2. text distance:

- String distance, using *stringdist* function in *stringdist*<sup>22</sup> R package:
  - Stringdist; it computes the Levenshtein distance<sup>23</sup> between the article's title (article) and the official point's name (FEATURENAME);
  - Cat\_stringdist; it computes the Hamming distance<sup>24</sup> between the article and the category of the official point (NAMECAT\_NAME);
  - Cat\_synonym\_stringdist; it computes the Hamming distance between the article and the expanded version of the category

---

<sup>22</sup> <https://cran.r-project.org/web/packages/stringdist/index.html>

<sup>23</sup> [https://en.wikipedia.org/wiki/Levenshtein\\_distance](https://en.wikipedia.org/wiki/Levenshtein_distance)

<sup>24</sup> [https://en.wikipedia.org/wiki/Hamming\\_distance](https://en.wikipedia.org/wiki/Hamming_distance)

(NAMECAT\_NAME\_SYNONYM). We had to add a couple of synonyms for hospital and bus stop.

- String distance (Wikipedia article without brackets), same as before, but taking away the words that some articles had in brackets (usually it was the name of the city itself):
  - Stringdist\_no\_brackets;
  - Cat\_stringist\_no\_brackets;
  - Cat\_synonym\_stringdist\_no\_brackets.
- Check if string appears, it verifies if a string appears in another string, having taken out blank spaces and putting upper letters before:
  - Check\_article; it checks if article string appears in FEATURENAME;
  - Check\_name; it checks if FEATURENAME string appears in article;
  - Check\_category; it checks if NAMECAT\_NAME string appears in article;
  - Check\_cat\_synonym; it checks if NAMECAT\_NAME\_SYNONYM string appears in article.
- Match between single words, it makes a string match between every single word in a string with any other single word in another string; then, it computes the sum of the number of positive matches:
  - Multiple\_match; it checks for matches between article and FEATURENAME;
  - Multiple\_match\_cat; it checks for matches between article and NAMECAT\_NAME;
  - Multiple\_match\_cat\_synonym; it checks for matches between article and NAMECAT\_NAME\_SYNONYM.
- Stem, we decided to look for the stem<sup>25</sup> of a word, using the function stemDocument in *tm* R package; the stemming is made on German words and then a *stringdist* function is applied to compute again the Levenshtein distance. If the Wikipedia article was not in German, the *stringdist* computed before was put (unfortunately, stemDocument function is not available in all the languages

---

<sup>25</sup> <https://en.wikipedia.org/wiki/Stemming>

that we are considering). The kind of string distances are the same as before, with the differences that we are performing them considering the roots of the words:

- Stem\_stringdist;
- Stem\_cat\_stringdist;
- Stem\_cat\_synonym\_stringdist.
- Stem and then match, so we first split the string to single words, find the stem of each word and then make the comparison between each word of the two strings. It then computes the sum of the number of positive matches:
  - Stem\_multiple\_match;
  - Stem\_multiple\_match\_cat;
  - Stem\_multiple\_match\_cat\_synonym.
- Summary variables that summarize the results of the above-mentioned features:
  - Summary\_stringdist, which takes the minimum distance among the three;
  - Summary\_stringdist\_no\_brackets, the same but on the no brackets features;
  - Summary\_check, which sums the number of TRUE that we had in the four checks;
  - Summary\_multiple\_match, which sums the number of matches in the three features;
  - Summary\_stem\_stringdist, which takes again the minimum distance;
  - Summary\_stem\_multiple\_match, which sums up the number of matches.

The total dataset is composed by the matches between official points and Wikidata points of interest and has 88.620 observations. It has been divided as following:

- 880 observations have been used to perform manual training; the number has been chosen so that each category is represented into the dataset (44 categories \* 20 nearest distances each);
- 87.740 observations remain in the dataset and for these variables the variable match will have to be predicted through the algorithm.

The training dataset has been again divided into three parts:

- 187 observations were used to compose the final test dataset, which will not be tested until the final version of the model;
- the remaining 704 were used to train and validate the model, dividing them each time into two subgroups:
  - 70% of the observations were used for training;
  - 30% of the observations were used as a validation test to make changes to the algorithm .

In the final model, only some of these features have been used, in particular:

1. NEAR\_DIST;
2. Summary\_stringdist;
3. Summary\_stringdist\_no\_brackets;
4. Summary\_check;
5. Summary\_multiple\_match;
6. Summary\_stem\_stringdist;
7. Summary\_stem\_multiple\_match.

After having selected only the Wikidata points of interest that have a match with official points, we found out that their total amount is 796. They represent nearly 30% of the total amount of Wiki points (2663). Some maps for comparison have been built. You can see some screenshots in Figures 51 to 53 (they are again interactive maps available online<sup>26</sup>).

---

<sup>26</sup> [http://serenasignorelli.altervista.org/Vienna\\_official/Vienna\\_official.html](http://serenasignorelli.altervista.org/Vienna_official/Vienna_official.html)  
[http://serenasignorelli.altervista.org/Vienna\\_wiki/Vienna\\_wiki.html](http://serenasignorelli.altervista.org/Vienna_wiki/Vienna_wiki.html)  
[http://serenasignorelli.altervista.org/Vienna\\_match/Vienna\\_match.html](http://serenasignorelli.altervista.org/Vienna_match/Vienna_match.html)

Figure 51 – Official points in Vienna according to municipality

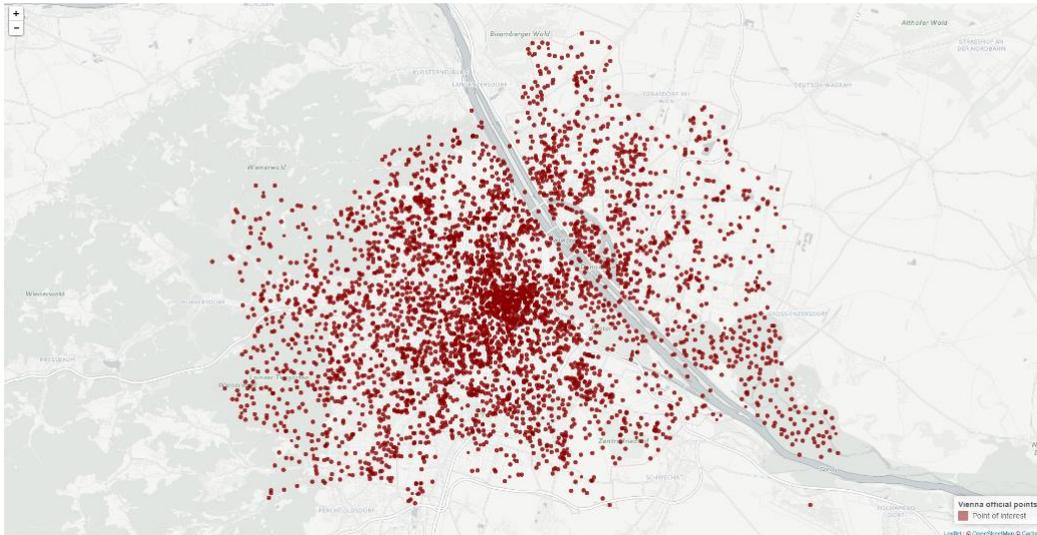
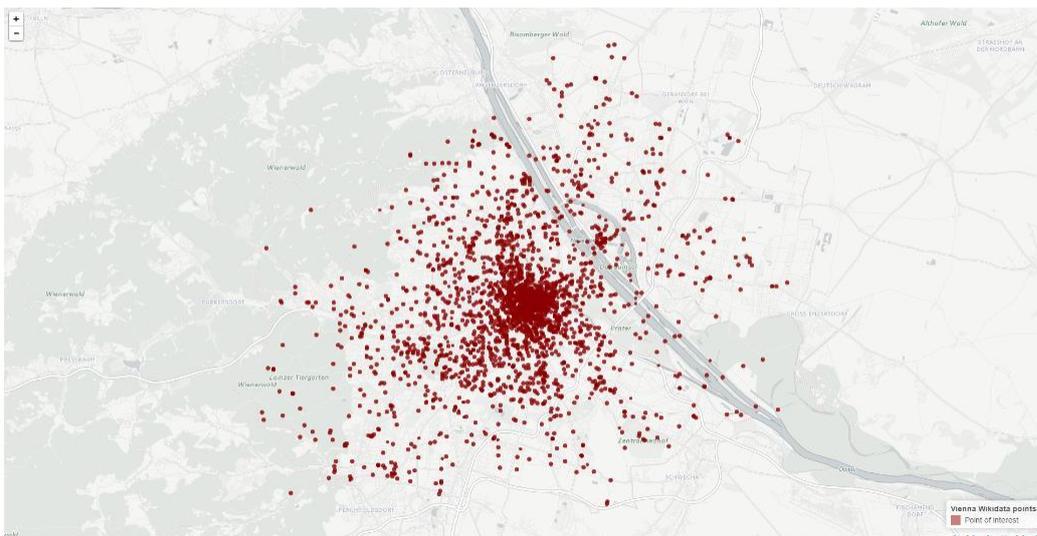


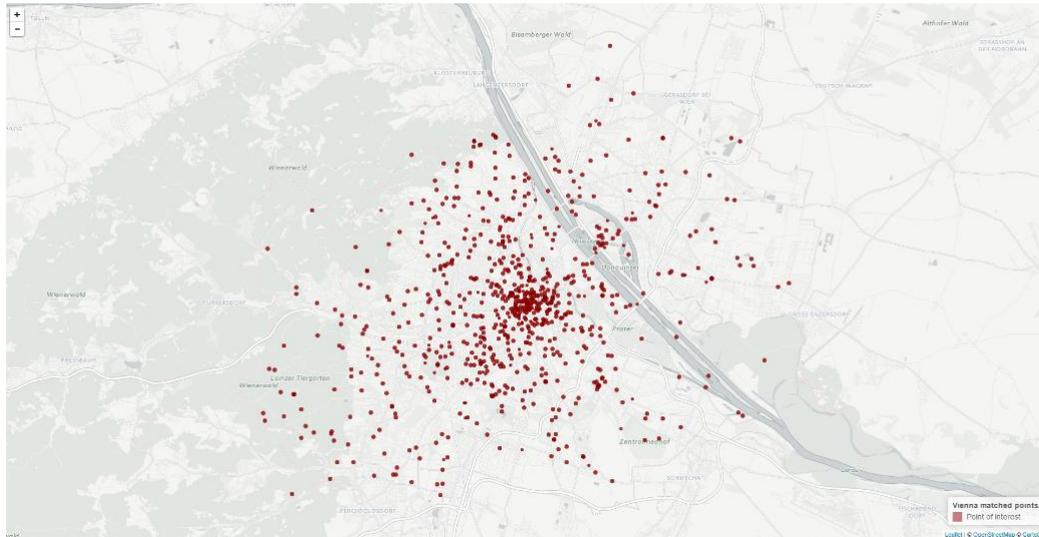
Figure 52 – Points of interest in Vienna according to Wikidata



The interesting fact is that if we look at the variable NAMECAT\_NAME (their category), we have 42 levels (in the official dataframe we had 44 levels). We made two maps to visualize points with the official category (again available as interactive maps online<sup>27</sup>).

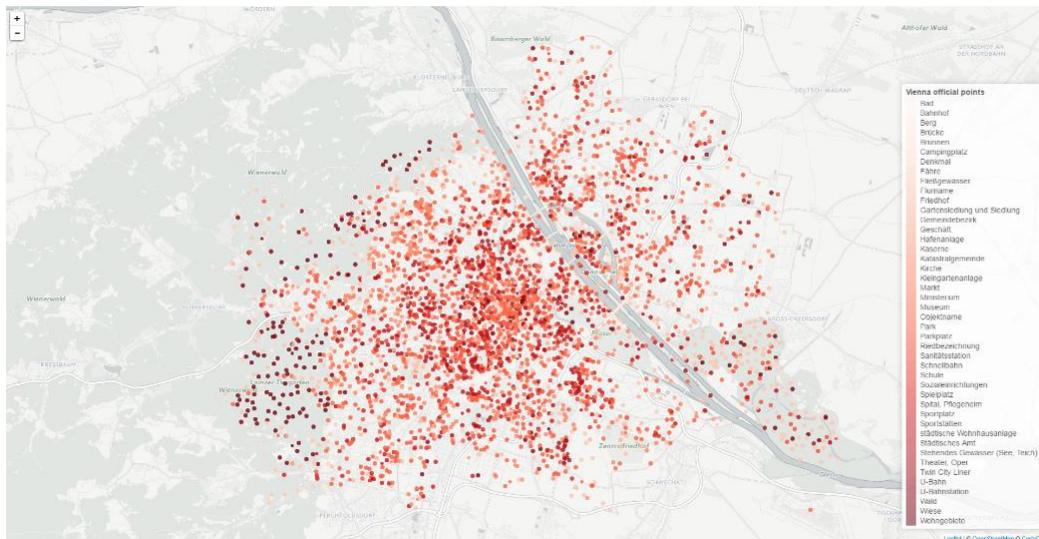
<sup>27</sup> [http://serenasignorelli.altervista.org/Vienna\\_official\\_with\\_cat/Vienna\\_official\\_with\\_cat.html](http://serenasignorelli.altervista.org/Vienna_official_with_cat/Vienna_official_with_cat.html)  
[http://serenasignorelli.altervista.org/Vienna\\_match\\_with\\_cat/Vienna\\_match\\_with\\_cat.html](http://serenasignorelli.altervista.org/Vienna_match_with_cat/Vienna_match_with_cat.html)

Figure 53 – Matched points of interest in Vienna



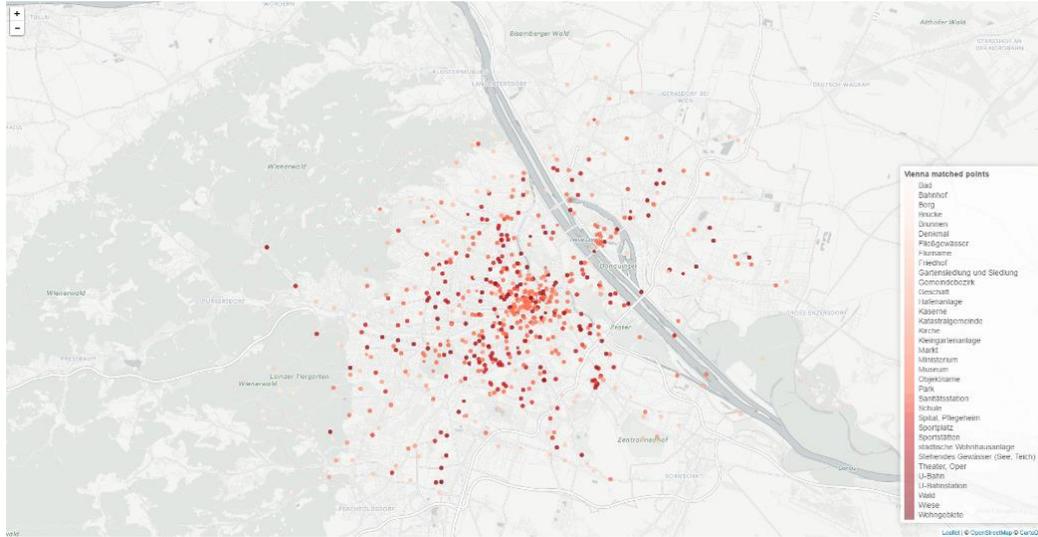
In Figures 54 and 55 the maps show the categorization of official Vienna points according to the municipality and of the matched points.

Figure 54 – Categorization of official points in Vienna according to municipality



After this match, we would have liked to extend this official categorization to the whole Wikidata points dataset. But after a little inspection on the official points, we decided that the categorization does not fit for our purpose, it suits for municipality purposes, but not for tourism.

Figure 55 – Categorization of matched points in Vienna



### 4.5.3 Change in the number of items

The first extraction of Wikidata items and related Wikipedia articles was made on May 24th. One month later (June 24<sup>th</sup>) one error has been identified in the formula that got the Wikipedia articles out from Wikidata items. Once corrected, we performed the extraction again (all the results in the analysis are corrected and updated at June 24<sup>th</sup>). It was nice to see how many new Wikidata items with geo-coordinates were included in the new results. In Table 26 a summary of the changes in one month. In Figures 56 to 58 you can see some screenshots of how the maps looked like in May and in June.

Table 26 – Summary of changes in Wikidata from May 2016 to June 2016

<i>City</i>	<i>Level</i>	<i>No of points</i>	<i>No of points with property</i>	<i>No of properties</i>	<i>No of classes</i>	<i>No of categories</i>
Barcelona	C	+148%	+135%	+47%	+35%	+36%
	K	+152%	+140%	+50%	+35%	+36%
Bruges	C	+162%	+156%	+51%	+52%	+53%
	F	+152%	+153%	+57%	+58%	+56%
Vienna	C	+174%	+155%	+51%	+33%	+40%

Figure 56 – Difference in maps of Barcelona (Urban Audit Level C) from May 2016 to June 2016

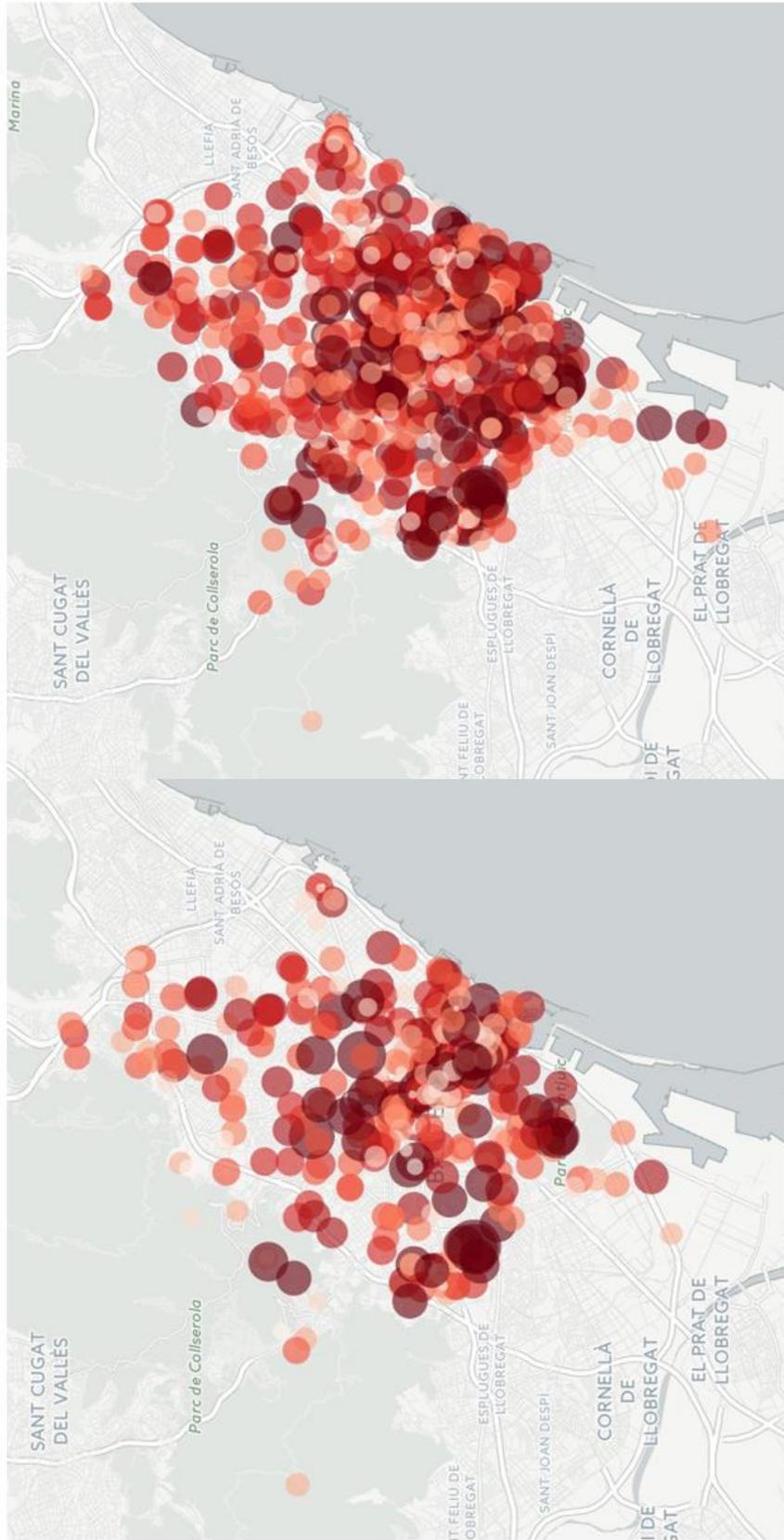


Figure 57 – Difference in maps of Bruges (Urban Audit Level C) from May 2016 to June 2016

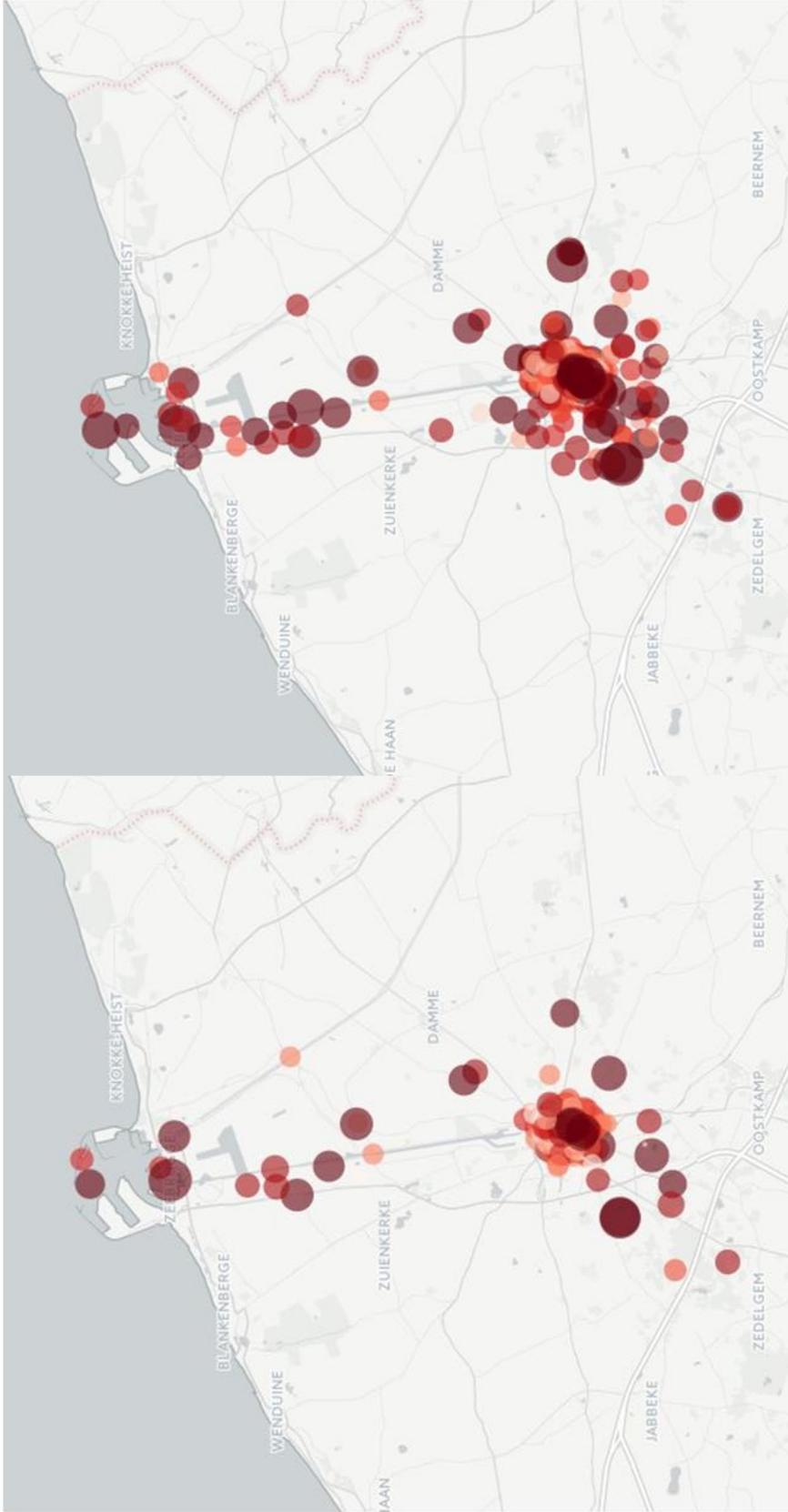
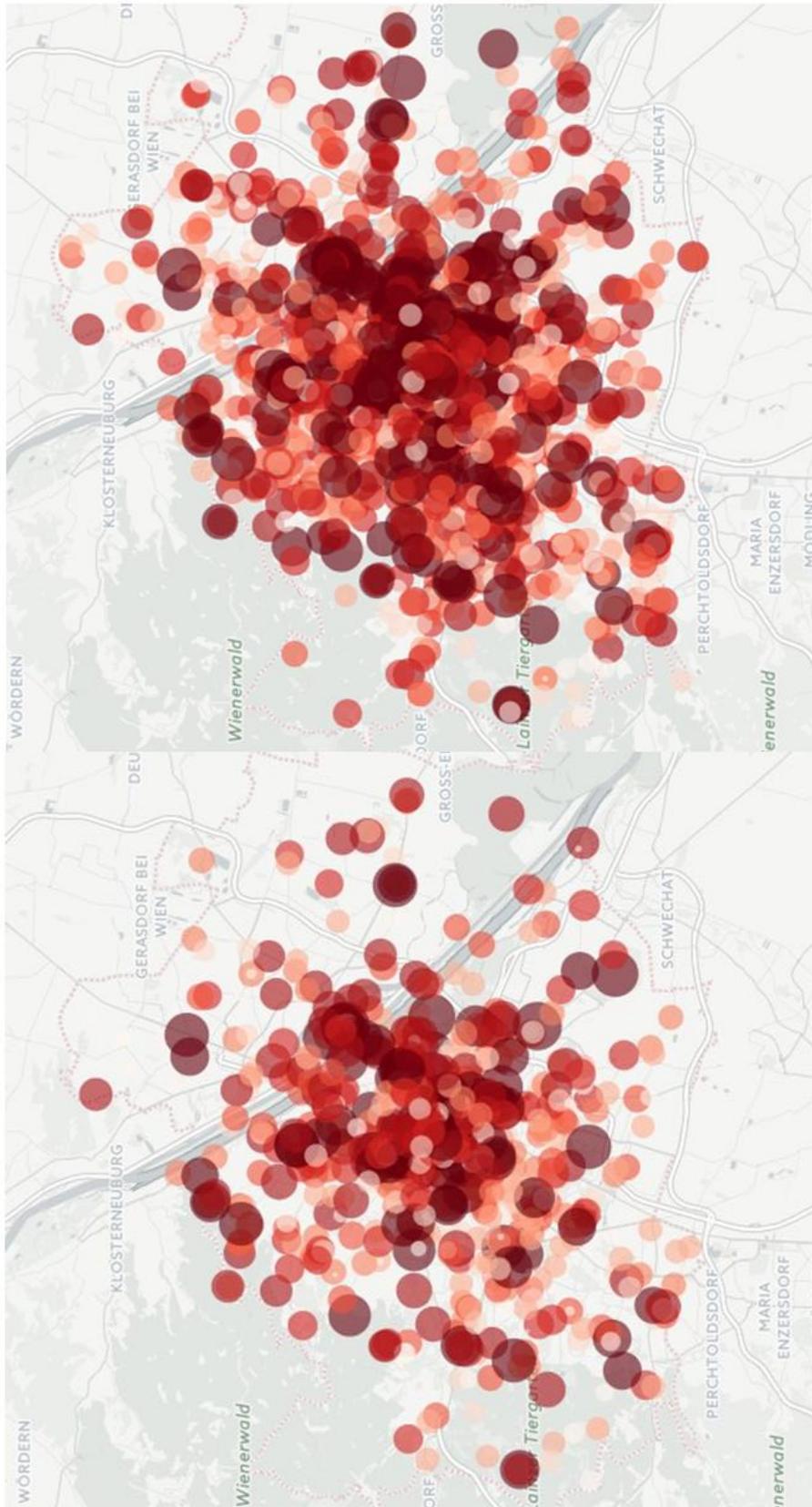


Figure 58 – Difference in maps of Vienna (Urban Audit Level C) from May 2016 to June 2016



We decided to perform the check of number of items each month on the 24<sup>th</sup> without changing our data, just to verify how Wikidata is evolving in time. On July 24<sup>th</sup> these were the changing in number Wikidata items and their properties, with respect to June 24<sup>th</sup> (Table 27).

Table 27 – Summary of changes in Wikidata from June 2016 to July 2016

<i>City</i>	<i>Level</i>	<i>No of points</i>	<i>No of points with property</i>	<i>No of properties</i>	<i>No of classes</i>	<i>No of categories</i>
Barcelona	C	-0,27%	+0,59%	+1,53%	+74,77%	+88,35%
	K	-0,07%	+0,43%	+0,47%	+69,17%	+79,65%
Bruges	C	+0,00%	+0,00%	-1,54%	+109,09%	+100,00%
	F	+0,31%	+0,60%	-1,25%	+111,54%	+107,55%
Vienna	C	+1,16%	+1,22%	-4,96%	+80,99%	+82,50%

There were no big changes in the number of items and property, while we can see that the number of classes increased a lot in one month. It became more than double for Bruges, and grow between 70 and 80% for the other two cities. This means that the Wikidata community has been very active in reviewing especially the property P279 (which is our definition of class). The situation on August 24<sup>th</sup> was as in Table 28.

Table 28 – Summary of changes in Wikidata from July 2016 to August 2016

<i>City</i>	<i>Level</i>	<i>No. of points of interest</i>	<i>No. of points with property</i>	<i>No. of properties</i>	<i>No. of classes</i>	<i>No. of categories</i>
Barcelona	C	+0,28%	+1,17%	+2,01%	+3,61%	+3,61%
	K	+0,28%	+0,94%	+1,40%	+2,96%	+2,96%
Bruges	C	+0,00%	+0,38%	+1,56%	+3,26%	+3,26%
	F	+0,31%	+0,89%	+2,53%	+2,73%	+2,73%
Vienna	C	+0,26%	+0,47%	+0,87%	+1,37%	+1,37%

We can see some small changes concerting property P31 and a stable situation on the number of items. On September 24<sup>th</sup>, it is interesting to note the big change in the number of items with property P31 that happened in Vienna, nearly 25% increasing (Table 29).

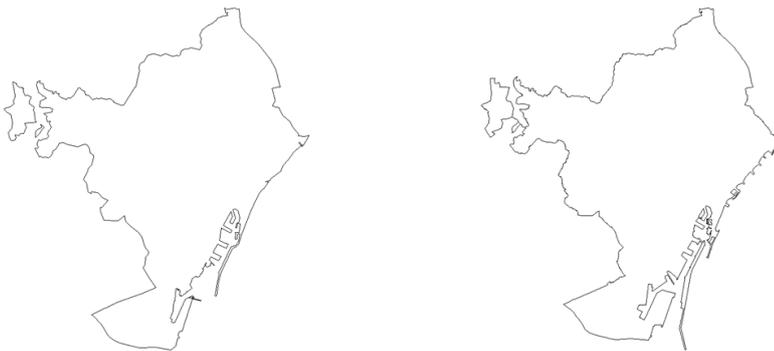
Table 29 – Summary of changes in Wikidata from August 2016 to September 2016

<i>City</i>	<i>Level</i>	<i>No. of points of interest</i>	<i>No. of points with property</i>	<i>No. of properties</i>	<i>No. of classes</i>	<i>No. of categories</i>
Barcelona	C	+0,18%	+1,39%	+0,49%	-1,00%	-1,00%
	K	+0,28%	+1,18%	+0,92%	+0,96%	+0,96%
Bruges	C	+0,00%	+0,00%	+0,00%	-1,05%	-1,05%
	F	+0,00%	+0,00%	+0,00%	-0,88%	-0,88%
Vienna	C	-1,52%	+24,66%	+2,16%	+1,80%	+1,80%

#### 4.5.4 Check for cities boundaries

When we found the Official Statistics data, the first operation to do was to check if their "definition" of the city was the same as ours. In order to do so, we downloaded their city shapefiles and compared with the ones from Urban Audit. In Figures 59 to 62 you can see the comparison.

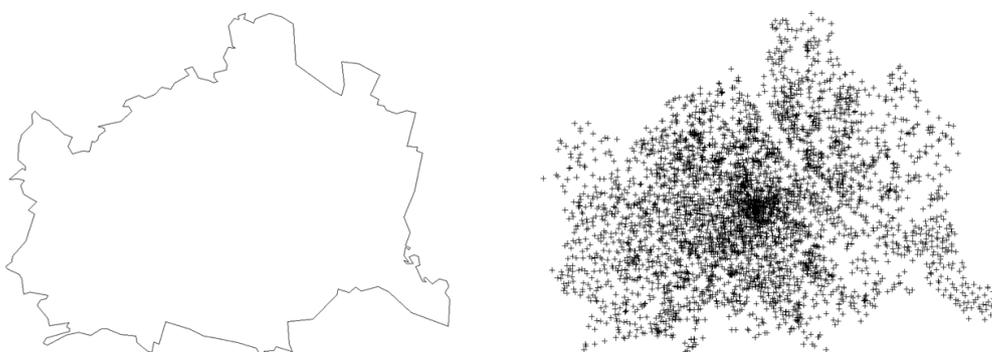
Figure 59 – Comparison of shapefiles of Barcelona: Urban Audit (left) vs. Ajuntament de Barcelona (right)



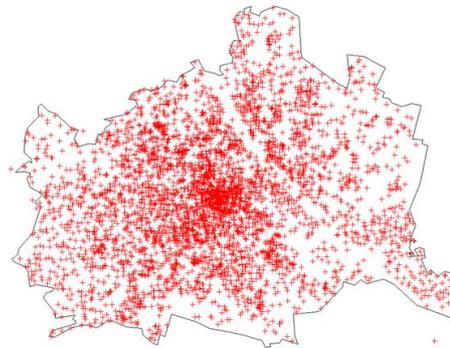
*Figure 60 – Comparison of shapefiles of Bruges: Urban Audit (left) vs. Toerisme Vlaanderen (right)*



*Figure 61 – Comparison of shapefiles of Vienna: Urban Audit (left) vs. Statistik Austria (right)*



*Figure 62 – Comparison of shapefiles of Vienna: overlap of the two sources*



After this comparison, we checked that the Urban Audit level that we must consider for the three cities is level C. In fact, starting from the classification based on the content of the Wikipedia articles, the analysis considers only the Urban Audit level C, as other levels appear to be useless.

## 4.6 Disclaimer

The views expressed here are those of the authors and do not necessarily reflect the official views of the European Commission (Eurostat).



## 5 CONCLUSIONS

Big data is a data source that is becoming more and more popular these days. Official Statistics bodies are trying to deal with this new source to identify possible uses, as a support to existing sources or to produce new statistics.

Through this thesis, we tried to bring new studies in literature, analysing two different sources of big data and their possible uses in Official Statistics. We think we brought some new ideas, in particular concerning Wikipedia. In fact, the latter analysis was quite advanced, bringing some innovation in visualizations and trying to identify factors that drive tourism to an area.

The analysis on mobile phone data was a little bit more limited, due to the data available. The datasets came from the 1<sup>st</sup> Telecom Italia Big Data Challenge, they are freely available on the Web and they are aggregated in 10 minutes' slots. Moreover, they are composed by Call Detail Records, a type of data that registers an observation in the data only when an event occurs (such as a phone call). To perform a better analysis, we should have the possibility to have individual data, better if in the form of pings and not CDRs (pings register the position of the mobile phone even if no event occurs). The methodology of data aggregation is not released by Telecom Italia, adding more uncertainty on the analysis. Moreover, in the second application on mobility we used an updated version of the Origin/Destination matrix, integrated with surveys on commuters, but again we don't have further details on the methodology.

Unfortunately, having access to other sources of data is not easy, especially in academia. Mobile phone providers often try to sell their data, and this represents a purchase not affordable from a university. We tried to do our best with data openly available, of course having to deal with the above-mentioned lacks in methodology clearness. This is the reason why the Wikipedia study seems more promising. The Wikimedia foundation releases its data, and if any change occurs, a methodology note is released.

Collaboration with data providers is always better, as it brings clarity and transparency to the study. Of course, this is not easy, especially for privacy constraints. For Wikipedia page views this problem does not exist, but it is a very important issue, for example, considering mobile phone data. The period I spent at Eurostat allowed me to face this problem very closely, participating in discussions with mobile phone providers. The discussion about access and protection of the data is open, but there's the willingness from both sides (National Statistical Institutes and data providers) to find solutions in this sense.

The exploration of big data sources in Official Statistics is in progress, but still needs a lot of improvements and in-depth analysis before going into production, especially regarding quality aspects.

## Bibliography

- ANTENUCCI, D., CAFARELLA, M., LEVENSTEIN, M.C., RE, C. & SHAPIRO, M.D. (2014) Using Social Media to Measure Labor Market Flows. *NBER Working Papers 20010*, National Bureau of Economic Research, Inc.
- BARCAROLI, G. (2015) Use of Big Data in Official Statistics, July, 2015.
- BARCAROLI, G., NURRA, A., SCARNÒ, M., SUMMA, D., NAZIONALE, I. (2014) Use of web scraping and text mining techniques in the Istat survey on Information and Communication Technology in enterprises. *In: European Conference on Quality in Official Statistics*, Wien, Austria.
- BIEMER, P. (2010) Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, **74(5)**, 817-848.
- BLEI, D. M., NG, A. Y., JORDAN, M. I. (2003) Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3:993–1022, January, 2003.
- BOGOMOLOV, A., LEPRI, B., STAIANO, J., OLIVER, N., PIANESI, F., & PENTLAND, A. (2014) Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data. *Proceedings of the 16th International Conference on Multimodal Interaction*, 427-434.
- BRAAKSMA, B., ZEELENBER, K. (2015) “Re-make/re-model”: should Big Data change the modelling paradigm in Official Statistics?, *Statistical Journal of the IAOS*, IOS Press, **31 (2015)**, 193-202, DOI 10.3233/SJI-150892.
- CICI, B., MARKOPOULOU, A., FRIAS-MARTINEZ, E., LAOTARIS, N. (2014) Assessing the potential of ride-sharing using mobile and social data: a tale of four cities, *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle, September 13-17, 2014.
- DAAS, P.J.H., PUTS, M.J., BUELENS, B., VAN DEN HURK, P.A.M. (2013) Big Data and Official Statistics. *Paper for the 2013 New Techniques and Technologies for Statistics conference*. Brussels, Belgium.
- DAVENPORT, T.H. (2013) At the Big Data Crossroads: turning towards a smarter travel experience, Amadeus.

- DAY, H.R., PARKER, J.D. (2013) Self-report of Diabetes and Claims-based Identification of Diabetes Among Medicare Beneficiaries. *National Health Statistics Reports*, **69**.
- DE MAURO, A., GRECO, M., GRIMALDI, M. (2015) What is Big Data? A consensual definition and a review of key research topics, *AIP Conference Proceedings*, 1644, 97, <http://dx.doi.org/10.1063/1.4907823>.
- DE MONTJOYE, Y., QUIDBACH, J., ROBIC, F., PENTLAND, A. (2013) Predicting people personality using novel mobile phone-based metrics, *Social computing, behavioral-cultural modeling and prediction*.
- DE PRATO, G., SIMON, J. P. (2015) The Next Wave: 'Big Data'?, *COMMUNICATIONS & STRATEGIES*, no.97, 1st quarter 2015, p. 15-39. Available at SSRN: <http://ssrn.com/abstract=2674005>.
- DEL VECCHIO, P., PASSIANTE, G., VITULANO, F., ZAMBETTI, L. (2014) Big Data and Knowledge-intensive entrepreneurship: trends and opportunities in the tourism sector, *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*, North America, 5, December, 2014.
- DEVILLE, P., LINARD, C., MARTIN, S., GILBERT, M., STEVENS, F.R., GAUGHAN, A.E., BLONDEL, V.D., TATEM, A.J. (2014) Dynamic population mapping using mobile phone data, *PNAS* doi: 10.1073/pnas.1408439111.
- DUFTY, D., BÉRARD, H., REEDMAN, L., LEFRANC, S., SIGNORE, M., MUNOZ, J., ORDAZ, E., STRUIJS, P., MAŚLANKOWSKI, J., MAROZKRUT, D., NIKIC, B., JANSEN, R., KOVACS, K., JUG, M. (2014) A Suggested Framework for National Statistical Offices for assessing the Quality of Big Data. *Paper for the 2015 New Techniques and Technologies for Statistics conference*. Brussels, Belgium.
- DUMBILL, E. (2013) Making Sense of Big Data. *Big Data*, Feb 2013, **1(1)**, 1-2.
- DUONG, T., MILLMAN, S. (2014) Behavioral Data as a Complement to Mobile Survey Data in Measuring Effectiveness of Mobile Ad Campaign. *Presented at the CASRO Digital Research Conference*.
- ELLIOTT, M.R. (2009) Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights, *Survey Practice*, **2(6)**.
- EUROPEAN STATISTICAL SYSTEM COMMITTEE (2014) ESS Big Data action plan and roadmap 1.0, *22<sup>nd</sup> Meeting of the European Statistical System Committee*, Riga (Latvia), September 26, 2014.

- FAN, J., HAN, F., & LIU, H. (2014) Challenges of Big Data analysis. *National Science Review*, **1(2)**, 293-314.
- FRIAS-MARTINEZ, V., SOGUERO, C., JOSEPHIDOU, M., FRIAS-MARTINEZ, E. (2013) Forecasting Socioeconomic Trends With Cell Phone Records, *Proceedings of the 3rd ACM Symposium on Computing for Development*.
- FRIAS-MARTINEZ, V., FRIAS-MARTINEZ, E., OLIVER, N. (2010) A Gender-centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records, *AAAI 2010 Spring Symposia Artificial Intelligence for Development*.
- GYLLSTROM, K., MOENS, M-F. (2012) Surfin' Wikipedia: an analysis of the Wikipedia (non-random) surfer's behavior from aggregate access data, *Proceedings of the fourth information interaction in context symposium (IiX 2012)*, pages:155-163.
- GÖBEL, S. AND MUNZERT, S. (2016) Political Advertising on the Wikipedia Market Place of Information (April 4, 2016). Available at SSRN: <http://ssrn.com/abstract=2720141>.
- HAIRE, A. J., MAYER-SCHÖNBERGER, V. (2014) Big Data - Opportunity or Threat, *ITU GSR discussion paper*.
- HEERSCHAP, N., ORTEGA, S., PRIEM, A., OFFERMANS, M. (2014) Innovation of tourism statistics through the use of new big data sources, *12<sup>th</sup> Global Forum on Tourism Statistics*, Prague 15-16 May, 2014.
- HENDERSON, J.V., STOREYGARD, A., WEIL, D.N. (2012) Measuring economic growth from outer space, *American Economic Review*, **102(2)**, 994-1028, DOI: 10.1257/aer.102.2.994.
- JAPEC, L., KREUTER, F. (2015) Unlocking the full potential of Big Data, *JOS anniversary conference*, June, 2015.
- JAPEC, L., KREUTER, F., BERG, M., BIEMER, P., DECKER, P., LAMPE, C., LANE, J., O'NEIL, C., USHER, A. (2015) AAPOR Report on Big Data, *AAPOR (American Association For Public Opinion Research) Big Data Task Force*, February 12, 2015.
- KÄMPF, M., TESSENOW, E., KENETT, D.Y., KANTELHARDT, J.W. (2015) The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks. *PLoS ONE* 10(12): e0141892. doi:10.1371/journal.pone.0141892.

- KITCHIN, R. (2015) Big Data and Official Statistics: Opportunities, Challenges and Risks, *Statistical Journal of the IAOS*, IOS Press, **31(3)**, 471-481.
- LANEY, D. (2001) 3-D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note*.
- LIANG, T., FRIAS-MARTINEZ, V. (2015) Cars and Calls: Using CDR Data to Approximate Official Traffic Counts, *D4D Challenge*, Netmob, to appear.
- MCIVER, D.J., BROWNSTEIN, J.S. (2014) Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time, *PLoS Computational Biology* 10 (4): e1003581. doi:10.1371/journal.pcbi.1003581. <http://dx.doi.org/10.1371/journal.pcbi.1003581>.
- MICROSOFT (2013) The Big Bang: How the Big Data Explosion Is Changing the World.
- MOAT, H. S., CURME, C., AVAKIAN, A., KENETT, D. Y., STANLEY, H. E. AND PREIS, T. (2013) Quantifying Wikipedia Usage Patterns Before Stock Market Moves (May 8, 2013). *Scientific Reports*, Vol. 3, pp. 1801; DOI:10.1038/srep01801 (2013). Available at SSRN: <http://ssrn.com/abstract=2263897>.
- MUNZERT, S. (2015) Using Wikipedia Article Traffic Volume to Measure Public Issue Attention, Version 0.3, October 2015 – work in progress.
- NATHAN, M., ROSSO, A., GATTEN, T., MAJMUDAR, P., MITCHELL, A. (2013) Measuring the UK's digital economy with Big Data. *National Institute of Economic and Social Research (NIESR)*.
- NIST (2014) Big Data Public Working Group, Big Data Interoperability Framework: Definitions (draft).
- NOULAS, A., MASCOLO, C., SCELLATO, S., PONTIL, M. (2011) An empirical study of geographic user activity patterns in Foursquare, In: Adamic, LA and Baeza-Yates, RA and Counts, S, (eds.) *ICWSM*. The AAAI Press.
- OECD (2012) Quality Framework and Guidelines for OECD Statistical Activities.
- PAN, B., YANG, Y. (2014) Monitoring and forecasting tourist activities with Big Data, in *Management science in hospitality and tourism: theory, practice and applications*, Muzaffer Uysal, Zvi Schwartz, Ercan Sirakaya (eds.), Apple Academic Press.
- PAPPALARDO, L., SIMINI, F., RINZIVILLO, S., PEDRESCHI, D., GIANNOTTI, F., BARABÁSI, A.-L. (2015) Returners and explorers dichotomy in human mobility, *Nature Communications*, **6**, doi:10.1038/ncomms9166.

- POLIDORO, F., GIANNINI, R., LO CONTE, R., MOSCA, S., ROSSETTI, F. (2015) Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS* 31:165{176, DOI 10.3233/sji-150901.
- PORTER, S., LAZARO, C.G. (2014) Adding Big Data Booster Packs to Survey Data. *Presented at the CASRO Digital Research Conference.*
- REINOSO, A. J., GONZALEZ-BARAHONA, J. M., ROBLES, G., ORTEGA, F. (2009) A quantitative approach to the use of the Wikipedia, *Computers and Communications, 2009. ISCC 2009. IEEE Symposium on, Sousse, 2009*, pp. 56-61. doi: 10.1109/ISCC.2009.5202401, URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5202401&isnumber=5202209>.
- REINOSO, A. J., MUÑOZ-MANSILLA, R., HERRAIZ TABERNERO, I., ORTEGA, F. (2012) Characterization of the Wikipedia Traffic, *Proceedings of the Seventh International Conference on Internet and Web Applications and Services, ICIW, 2012.*
- REIS, F., DI CONSIGLIO, L., KOVACHEV, B., WIRTHMANN, A., SKALIOTIS, M. (2016) 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.*
- ROCASALVATELLA AND TELEFÓNICA (2014) Big Data and tourism: new indicators for tourism management, *Conclusions and recommendations for the hotel industry, Barcelona, May, 2014.*
- SAHA, B., SRIVASTAVA, D. (2014) Data quality: the other face of Big Data, *2014 IEEE 30th International Conference on Data Engineering (ICDE).*
- SOTO, V., FRIAS-MARTINEZ, V., VIRSEDA, J., FRIAS-MARTINEZ, E. (2011) Prediction of Socioeconomic Levels using Cell Phone Records, *Proceedings of the 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011.*
- TIAN, T., AGRAWAL, A. (2015) Quantifying the Relationship between Hit Count Estimates and Wikipedia Article Traffic, *International Journal of Advanced*

- Computer Science and Applications (IJACSA), 6(5), 2015.  
<http://dx.doi.org/10.14569/IJACSA.2015.060504>.
- TOSTES, A.I.J., SILVA, T.H., DUARTE-FIGUEIREDO, F., LOUREIRO, A.A.F. (2014) Studying traffic conditions by analyzing Foursquare and Instagram data, *Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks*, Montreal, QC, Canada, September 21-26, 2014,
- UNITED NATIONS, UNPulse (2015) Harnessing big data for development and humanitarian action, <http://www.unglobalpulse.org/about-new>
- W3C Semantic Web Activity (2011) World Wide Web Consortium (W3C), November 7, 2011.
- YASSERI, T.; BRIGHT, J. (2015) Wikipedia traffic data and electoral prediction: towards theoretically informed models, Submitted to EPJ Data Science on 5 May, 2015.
- YUCESOY, B., BARABÁSI, A-L. (2016) Untangling Performance from Success, EPJ Data Science, 2016, 5:17, DOI: 10.1140/epjds/s13688-016-0079-z.

### ***Acknowledgements***

*I simply would like to thank all the people that I met during this three-years journey,  
both new and long-term encounters.*

*Each one of you contributed in the success of this adventure.*