Wikipedia online activity data for temporal disaggregation of tourism indicators

Keywords: Temporal disaggregation, tourism, online data, Wikipedia

1. INTRODUCTION

Wikipedia is a widely known on-line encyclopedia used by people all over the world. According to the Community survey on ICT usage by individuals and households, in 2015, 45% of individuals with 16 to 74 years old living in EU "consulted wikis to obtain knowledge (e.g. Wikipedia)". This was 66% for individuals of between 16 and 24 years old. People leave digital traces of their interactions with Wikipedia in several forms, such as contributed content to the articles, the history of editions of articles (when and what type of editions), discussions (each article has a discussion page) and the history of access to the articles.

There has been previous work on the assessment of the potential use of these digital traces for the production of relevant statistics [1]. The research presented in this paper builds on those previous attempts and consists of the use of Wikipedia page views data for the temporal disaggregation of tourism indicators.

These indicators are available at Eurostat in tourism statistics broken down by NUTS 2 region at annual level and at monthly level for the whole country. There is a policy need for having these indicators broken down simultaneously in space (NUTS 2) and in time (monthly), in order to obtain values at monthly level for each NUTS 2 region.

This spatio-temporal disaggregation could be done simply assuming independence between space and time, i.e. assuming that the infra-annual temporal profile of these indicators is the same in every region. However, that is not a reasonable assumption as regions tend to have different tourism profiles with some attracting more visits in the summer (beaches) while others attracting them also in winter (mountains). One way to do a more accurate spatio-temporal disaggregation of tourism indicators is to use an auxiliary variable which captures these different temporal profiles between regions.

Data on the consultation of Wikipedia articles related to tourism points of interest (e.g. monuments, cultural sites) by tourists when planning their trips has the potential to capture the differing temporal profiles of the various regions and be useful to perform the spatio-temporal disaggregation of tourism indicators.

2. DATA

Data on the two tourism indicators (arrivals and overnight stays) broken down by region and by month are publicly available in Eurostat online database[1], in absolute numbers and in percentage change on previous period, broken down by NACE and for resident tourists. The indicator considered in this work is "Arrivals at tourist accommodation establishments" in absolute numbers, for both resident tourists and non-resident tourists. The time scope considered was the one for which Wikipedia page views data is available

[1] http://ec.europa.eu/eurostat/web/tourism/data/database

and all countries for which data on both the tourism indicator and Wikipedia page views data was available were considered.

Wikipedia pageviews represent the source of data we use in our analysis, but they are not immediately available, and they require pre-processing before it can be used. Page view statistics is a tool available for Wikipedia pages, which allows to know how many people visited an article during a given period (hourly counts, at the highest level of detail).

We first selected the articles to include in the study. To do so, we decided not to start directly on the selection of articles on Wikipedia, but to use the Wikimedia Foundation linked data source, Wikidata. In our study, we need to get all Wikidata items with geo-coordinates. This is made possible through the Wikidata Query Service, an interface that allows to query its database using the SPARQL language. After having identified the Wikidata items (from now onwards simply points of interest), we got all the Wikipedia articles related to them in the 31 languages considered in the analysis[2]. Afterwards we downloaded the page views related to those articles (considering also redirect articles) from January 2012 to December 2016.

Lombardia is the Italian NUTS 2 region that was chosen as a test as monthly tourism data are made available from the Italian Statistical Institute (ISTAT) website.[3] Having official monthly data allowed us to compare the results of the temporal disaggregation with ground truth data.

3. METHODS

The use of the Wikipedia online activity for the temporal disaggregation of the tourism indicators consisted in two stages. The first stage was to synthesise the relevant signal in Wikipedia page views data into one (or a few) indicator(s). In a second stage, this physical visits proxy indicator would then be used in the temporal disaggregation of the existing tourism indicators.

In order to create the synthetic indicator, the first step was to select a list of tourism points of interest (TPOI). Points of interest (POI) were selected by firstly querying Wikidata for all items which have geo-coordinates within the geographical area of interest. These POI should then be filtered to those with touristic relevance. In order to do that, a match was attempted with a curated list of POI from TomTom where they were classified (e.g. as monuments, museums, churches). However, that match was of limited success with few TomTom points being matched to Wikidata POI. Therefore, the matched points were only used for quality control purposes and all POI obtained from Wikidata were considered for the extraction of the synthetic indicator.

The second step of the creation of the synthetic indicator was to use Principal Components Analysis (PCA) to disentangle the several temporal signals included in the time-series of the number of page views of the Wikipedia articles associated to each POI. The idea is to identify intra-annual profiles, such as the seasonality, which can be of use for the temporal disaggregation.

[2] 31 Language versions cover 24 official EU languages and as well Icelandic, Macedonian, Norwegian, Russian, Albanian, Serbian and Turkish.

[3] http://dati.istat.it/

4. RESULTS

Figure 1 shows the first 5 components of the PCA decomposition of the monthly page views from 2012 to 2015 of all POI selected from Wikidata, as well as the total number of pageviews for all POIs.

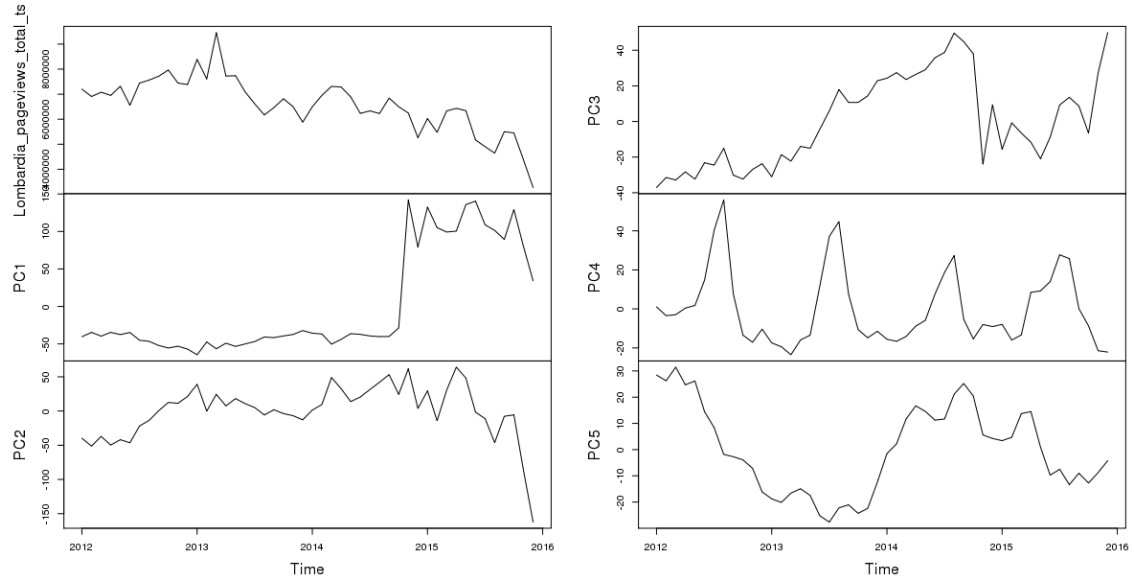

Figure 1. Wikipedia pageviews of articles associated to all POI selected from Wikidata and the corresponding first 5 PC

Most components seem to extract trend / cyclical movement in the time-series (components 3 and 5) or outliers (components 1 and 2), which are expected to be already present in the original annual time-series. Component 4, on the other hand presents an infra-annual movement.

An analysis of the first 10 components revealed that also a sixth component presented infra-annual regular movement which could be used for the temporal disaggregation.

Using the two components identified previously as auxiliary indicators, we disaggregated [2] the annual data for Lombardy region and plotted it against the real monthly data on the number of arrivals. The results can be seen in Figure 2.

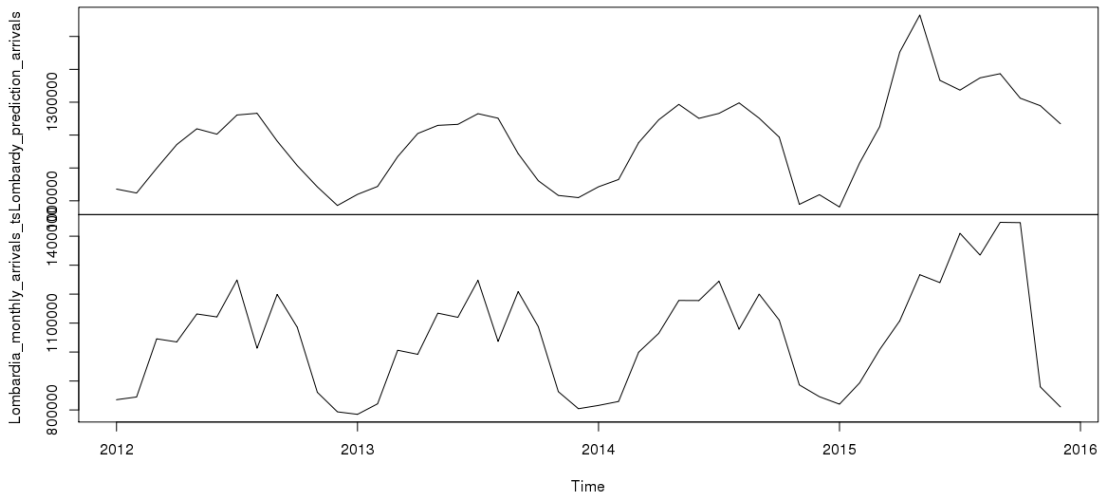

Figure 1. Monthly number of tourism arrivals (top) and temporally disaggregated annual number of arrivals (bottom)

In order to confirm that all the relevant signal was captured by the components showing infra-annual regular movement, we modelled the real data using all the principal components as regressors. The results confirmed that components 4 and 6 were the most significant ones. However, other components still included relevant signal.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1050413.5    14707.3  71.421  < 2e-16 ***
PC1             593.0      210.7   2.815  0.00777 **
PC2             579.5      371.9   1.558  0.12776
PC3             766.2      580.3   1.320  0.19481
PC4            4986.7      781.0   6.385 1.89e-07 ***
PC5            -229.2      890.4  -0.257  0.79830
PC6            6678.6     1007.6   6.628 8.92e-08 ***
PC7           -2247.6     1067.1  -2.106  0.04203 *
PC8           -2747.8     1120.9  -2.451  0.01907 *
PC9            3307.8     1224.0   2.702  0.01033 *
PC10          -2584.7     1250.3  -2.067  0.04576 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101900 on 37 degrees of freedom
Multiple R-squared:  0.7626,   Adjusted R-squared:  0.6985
F-statistic: 11.89 on 10 and 37 DF,  p-value: 8.099e-09
```

We computed the differences between the disaggregated series (computed using only PC4 and PC6) and the official monthly data in absolute value. Then we computed the percentage of those differences on official monthly data. The histogram in Figure 3 shows that although for most of the months the difference was below 10%, there were 3 instances where the estimated value was wrong by more than 20% of the real value.

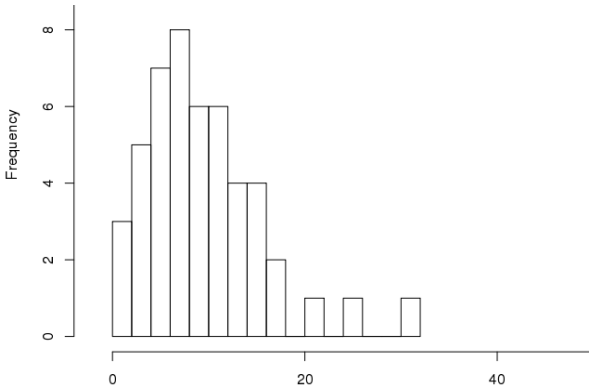

Figure 3. Histogram of the percentage difference between disaggregated and real values

5. CONCLUSIONS

The use of Wikipedia page views as an auxiliary indicator to perform temporal disaggregation on tourism indicators could successfully capture the seasonal profile that should be present in the infra-annual indicator. However, the accuracy of the methodology is still relatively low and it requires further development, in particular the inclusion of monthly information at country level.

REFERENCES

[1] S. Signorelli, F. Reis and S. Biffignandi. Virtual vs. real visits: an analysis of three cities through Wikipedia page views and tourism data. In Proceedings of NTTS2017 (2017).

[2] C. Sax and P. Steiner. Temporal Disaggregation of Time Series, The R Journal 5(2):80-87 (2013)