Wikipedia online activity data for temporal disaggregation of tourism indicators

Serena Signorelli [1]
Fernando Reis [2]

[1] Independent researcher, Italy
[2] Eurostat, Luxembourg

March 13, 2019

Introduction

Introduction

> *In 2015, 45% of individuals with 16 to 74 years old living in EU "consulted wikis to obtain knowledge (e.g. Wikipedia)"*[1]

People leave **digital traces** of their interactions with Wikipedia in several forms (contributed content of the articles, history of editions of the articles, discussions, history of access to the articles)

Assessment of the potential use of these digital traces for the production of relevant statistics

Our research: use of Wikipedia page views data for the **temporal disaggregation of tourism indicators**

[1] Community survey on ICT usage by individuals and households

Policy need

Tourism indicators:

- Arrivals
- Overnight stays

Available at Eurostat in tourism statistics broken down by:

- **NUTS 2 region** at **annual level**
- **Whole country** at **monthly level**

At the moment not available **NUTS 2 region** at **monthly level**

Possible methodologies

The spatio-temporal disaggregation could be done:

- Assuming independence between space and time (NOT a reasonable assumption!)
- Using an **auxiliary variable** which captures different temporal profiles between regions

Data on the consultation of Wikipedia articles related to **tourism points of interest** has the potential to capture the differing temporal profiles of the various regions and can be useful to perform the spatio-temporal disaggregation of tourism indicators

Sources of data

Official tourism data

Data on arrivals and overnight stays are publicly available in Eurostat online database[2]:

- by region & by month
- in absolute numbers and in % change on previous period
- broken down by NACE and for resident tourists

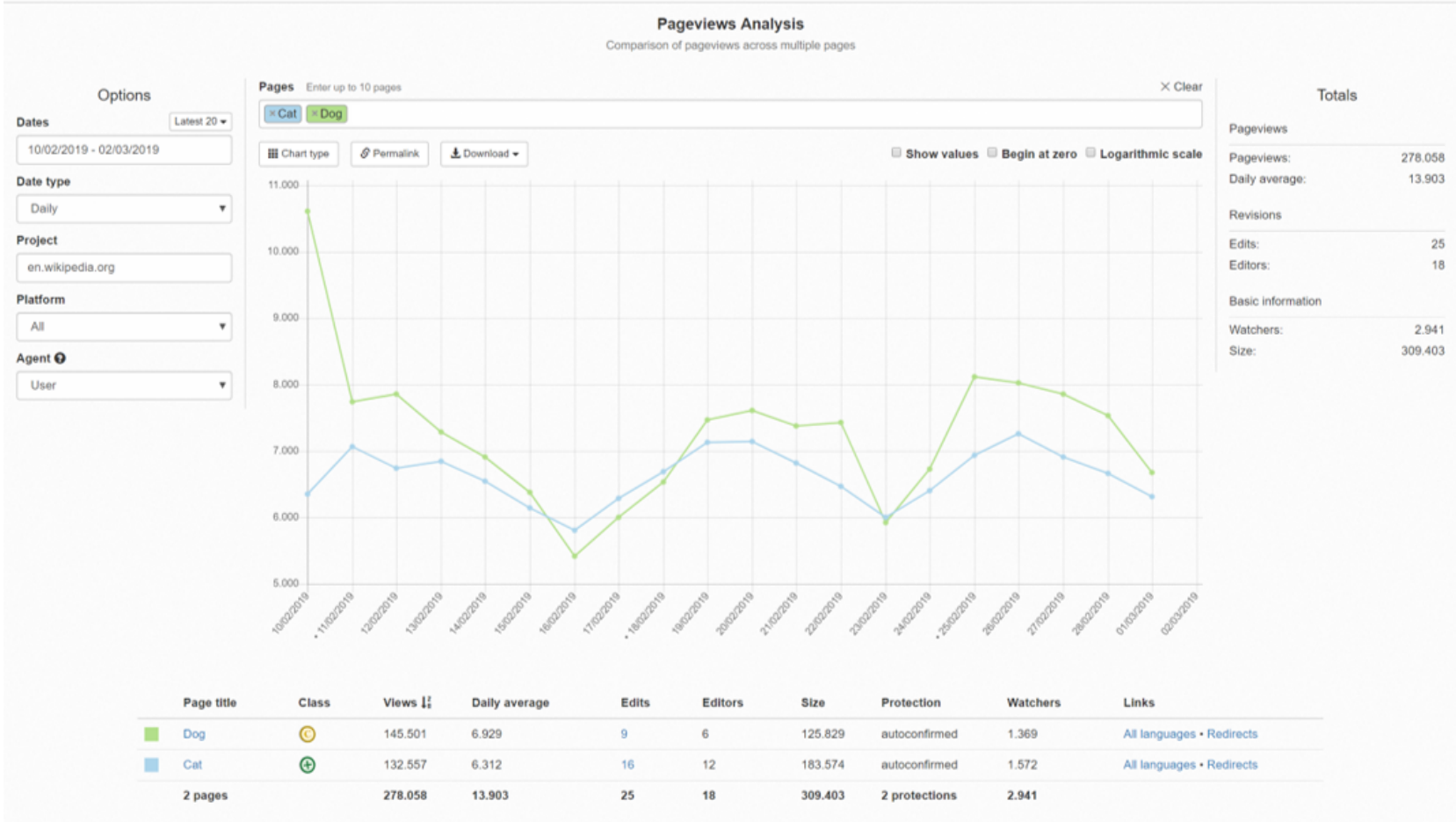
Indicator used in this work: **arrivals** at tourist accomodation establishments in **absolute numbers** for **both resident tourists and non-resident tourists**

Time & space: time scope for which Wikipedia page views data was available and all countries for which data on both the tourism indicator and Wikipedia page views data was available

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

Wikipedia data

Page view statistics: tool available for Wikipedia pages



Go to the tool

Wikipedia data

Wikipedia page views data require some pre-processing

Selection of articles:

1. starting from Wikidata, the linked data source of the Wikimedia foundation, we selected all Wikidata items with geo-coordinates
2. after the identification of Wikidata items, we got all the Wikipedia articles related to them in 31 languages[3]
3. we downloaded the pageviews related to those articles (considering also redirect articles) from January 2012 to December 2015

[3] 24 official EU languages + Icelandic, Macedonian, Norwegian, Russian, Albanian, Serbian, Turskish

Data used in this work

To test the methodology we chose the Italian NUTS 2 region **Lombardia** as monthly tourism data are made available from the Italian Statistical Institute website[4]

Official monthly data allowed to compare the temporal disaggregation with ground truth data

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

Methodology and results

Methodology

Use of Wikipedia online activity for the temporal disaggregation of tourism indicators has TWO STAGES:

1. Synthesise the relevant signal in Wikipedia views data into one (or a few) indicator(s)
2. Use of the previously identified proxy indicator to perform the temporal disaggregation

Stage 1

The creation of the indicator passes first through the selection of a list of tourism points of interest (TPOI)

How?

1. Querying Wikidata for all items that have geo-coordinates within the geographical area of interest (identification of points of interest - POI)
2. Filtering only those POIs with touristic relevance

Point 2: Attempt with a curated list of POI from TomTom: **limited success**, only a few matched POIs

Then: matched POIs only used for quality control purposes, all POIs from Wikidata were considered for the extraction of the synthetic indicator

Stage 2

The creation of the indicator passes then through the use of **Principal Component Analysis** (PCA)

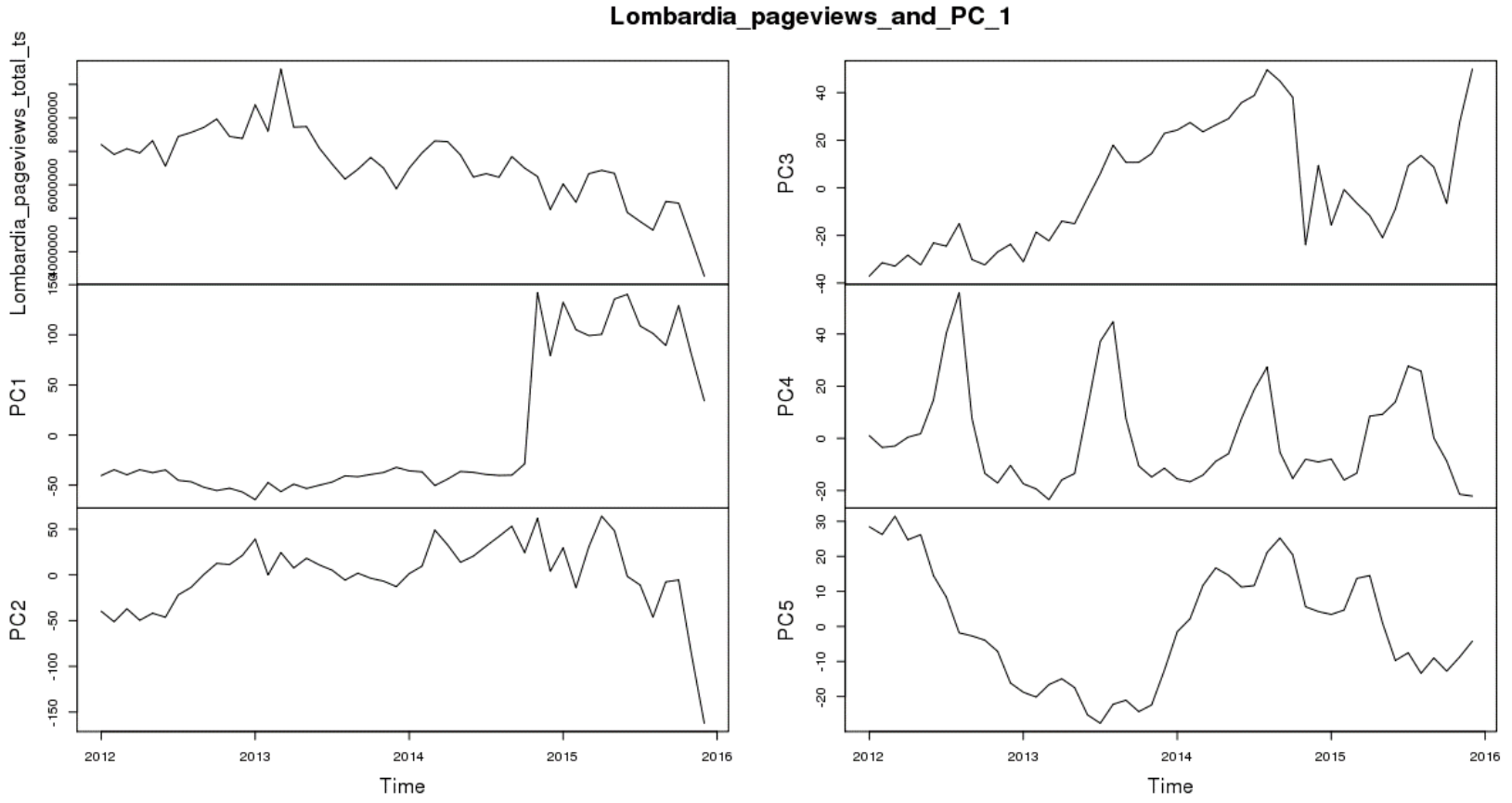
Why?

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

Aim: identify intra-annual profiles (i.e. **seasonality**) that can be used for the temporal disaggregation

Results

First 5 components of the PCA decomposition of the monthly page views of all POIs selected from Wikidata, as well as the total number of page views for all POIs



Results

- Components 3 and 5 extract **trend/cyclical movement**
- Components 1 and 2 extract **outliers**
- Component 4 presents an **infra-annual movement**

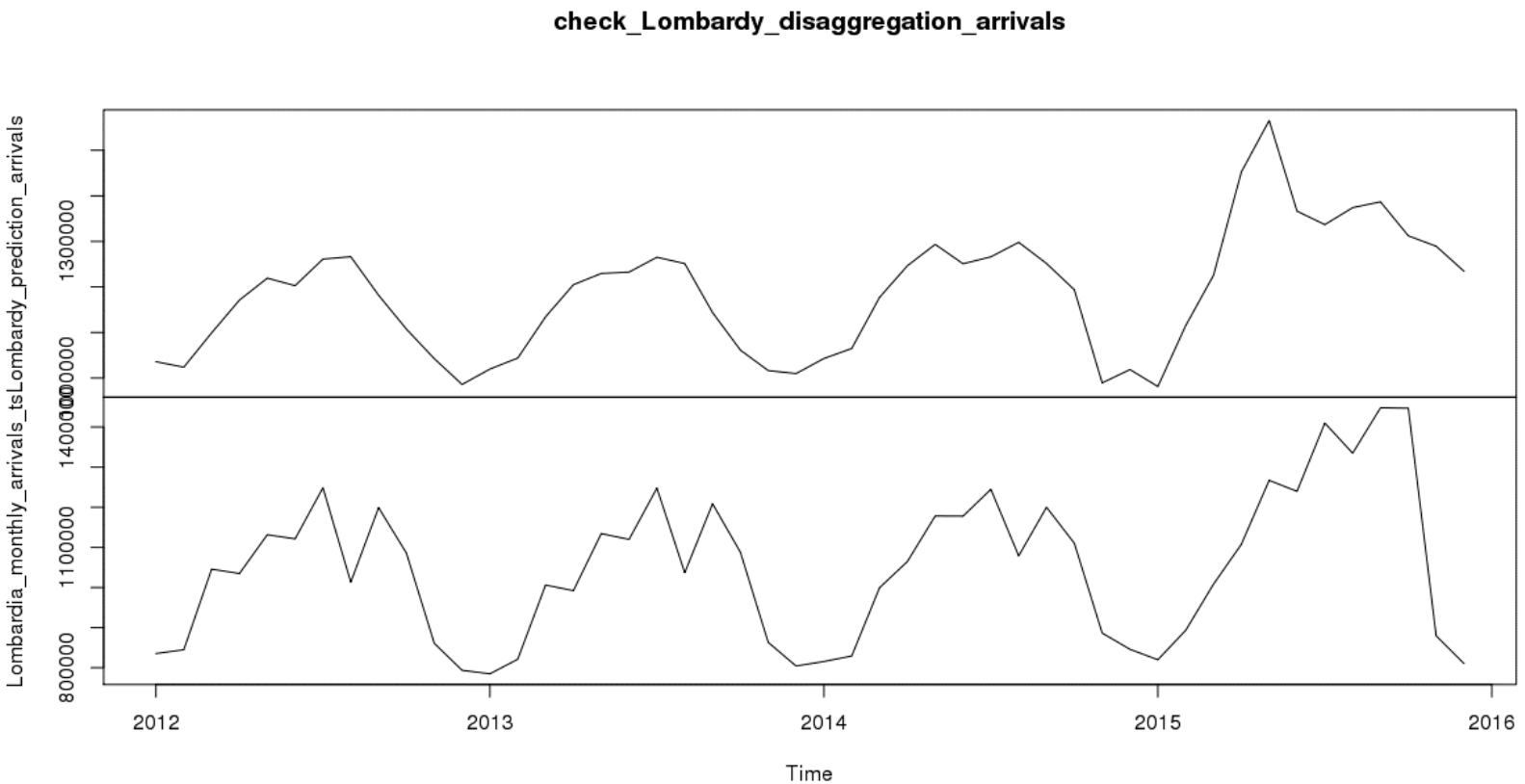
We analysed the first 10 components and found that the component 6 presented an infra-annual regular movement

Note: no preprocessing was performed on page views series

Results

Component 4 + Component 6 = auxiliary indicators

Disaggregation of annual data for Lombardy region (top) vs real monthly data (bottom)



Test on results (1)

Model real monthly data using all the principal components as regressors:

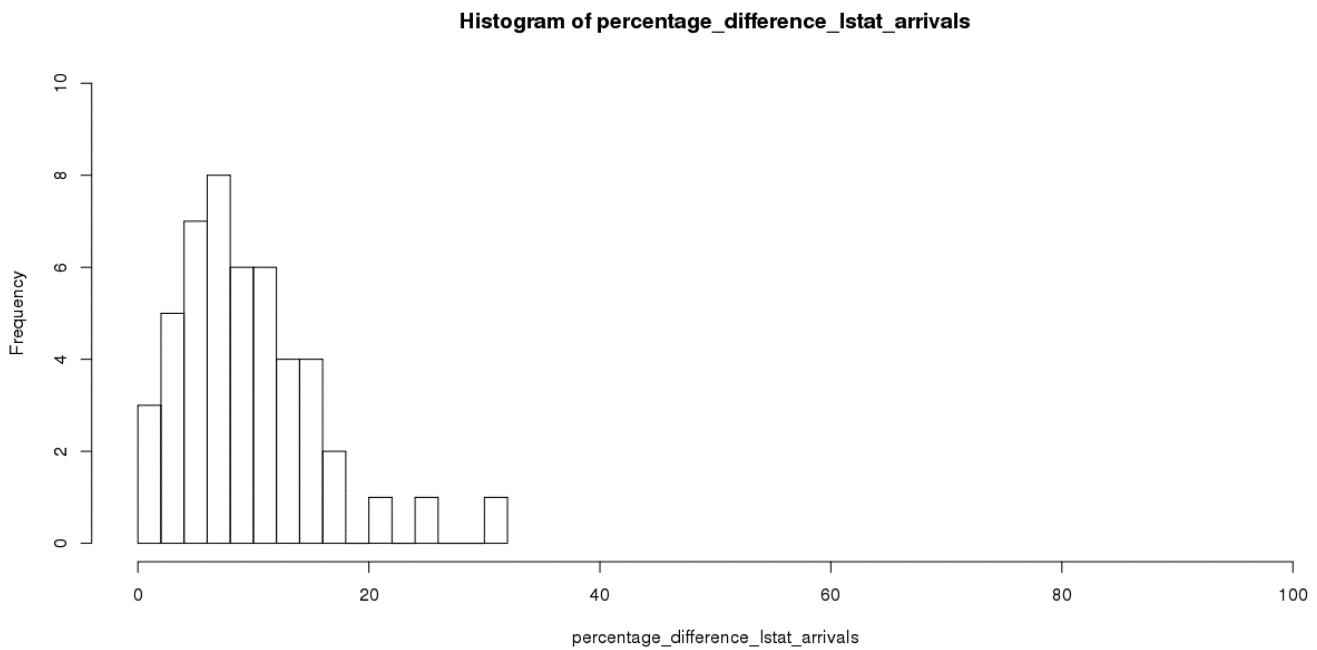
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1050413.5    14707.3  71.421  < 2e-16 ***
PC1             593.0      210.7   2.815  0.00777 **
PC2             579.5      371.9   1.558  0.12776
PC3             766.2      580.3   1.320  0.19481
PC4            4986.7      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
```

Test on results (2)

We computed the differences between:

- disaggregated series computed using components 4 and 6
- official monthly data in absolute value

Then computed the percentage of those differences on official monthly data:



Conclusions and further research

Conclusions and further research

- the use of Wikipedia page views as auxiliary indicator could **successfully capture the seasonal profile**
- the accuracy of the methodology still relatively low
- some preprocessing on page views data is needed before the temporal disaggregation

THANK YOU!

serena.signorelli.87@gmail.com

Fernando.REIS@ec.europa.eu