Statistics Poland

Implementation of flash estimates in the field of tourism

Jarosław Napora, Sebastian Wójcik

Statistical Office in Rzeszów, Poland

Face-to-face Meeting
Rzeszów, 8-10 July 2019

Task 4 – Flash estimates

Motivation

- Flash estimates in the field of tourism would respond to the growing demand from stakeholders regarding the rapidly changing situation on the tourism market.
- Data in the field of tourism statistics are provided with a relatively long delay (in Poland monthly data on the accommodation base at T + 45).

Aim

- Flash estimates of monthly data on nights spent at tourist accommodation establishments

Task 4 – Flash estimates

Procedure

- Merging outputs tables from Survey On Tourist Accommodation Establishments (10 or more bed places) into one database of time series
- Merging outputs files from webscraping
- Cleaning and formatting data
- Trimming outliers
- Generating set of monthly statistics
- Selecting crucial variables for flash estimates
- Time-series model for crucial variables from Survey On Tourist Accommodation Establishments

Task 4 – Flash estimates

Merging outputs tables from Survey On Tourist Accommodation Establishments (10 or more bed places) into one database of time series

Year	Month	Tourists - residents	Tourists - non residents	Nights spent - residents	Nights spent- non residents
2018	01	240553	96037	400287	194805
2018	02	253066	90922	412485	181207
2018	03	283611	120979	468270	243800
2018	04	308799	126390	499931	252984
2018	05	331807	140764	545853	267939
2018	06	364805	152876	579633	288049
2018	07	309757	167022	561992	324325
2018	08	313204	176323	580325	340409
2018	09	361310	149996	599181	291070
2018	10	363969	136928	597114	273904
2018	11	339078	124608	558790	259784
2018	12	278285	105874	451189	225556
2019	01	262143	90523	439041	183032
2019	02	260927	86145	430539	168564
2019	03	312061	114048	504240	228588

Task 4 – Flash estimates

Web scraping aim

- scrap day-by-day all offers to get micro level data for each NUTS region (NUTS 2 in the case of Poland)

Web scraping query setup for Hotels.com

- 1 adult, 0 children
- Destination: NUTS 2 region
- Dates: check-in is *t+1*, check-out is *t+2* where *t* stands for today. For instance, on 19.06. we set check-in date: 20.06., check-out date: 21.06.

Task 4 – Flash estimates

Using R-script:

- Merging outputs files from webscraping
- Cleaning data and trimming outliers with respect to prices

In selection | Match case | Whole word | Regex | Wrap

```
#----old format----

file_names=list.files(path = ".", pattern = "Lok*", all.files = FALSE,full.names = FALSE)

baza=as.data.frame(matrix(0,1,16))
colnames(baza)[]=c("offerId","hotelId","hotelName","accType","locality","postalCode","region","street","price","starRating","guestReviews

for(i in 1:length(file_names))
{
  file=read.csv2(file_names[[i]]) #,encoding ="UTF-8"

  if(dim(file)[1]>0)
  {
    year=as.numeric(substr(file_names[[i]],start=13,stop=16))
    month=as.numeric(substr(file_names[[i]],start=18,stop=19))
    day=as.numeric(substr(file_names[[i]],start=21,stop=22))
    file=cbind(file,year,month,day)
    #file=file[,-1]
    colnames(file)[]=c("offerId","hotelId","hotelName","accType","locality","postalCode","region","street","price","starRating","guestRev
    baza=rbind(baza,file)
  }
}
```

Task 4 – Flash estimates

Generating set of monthly statistics

- Scraped data contains information about:
  - locality,
  - accommodation type,
  - price,
  - star rating, number of guest reviews, review rating
- Mean, median, deciles, minimum and maximum price and the number of offers

with respect fo accommodation type were calculated

- Star rating, the number of guest reviews, review rating seem to be quasi-constant or irrelevenat for time series model

Task 4 – Flash estimates

Selecting crucial variables for flash estimates

- Variables highly correlated with number of nights spent should be chosen
- In our case it is:
  - The median price
  - The number of offers

Year	Month	Number of offers	Mean price	Median price
2018	11	1660	280,46	263,00
2018	12	6712	261,76	245,00
2019	1	8474	260,63	239,00
2019	2	4193	253,26	240,00
2019	3	8980	261,68	254,32
2019	4	3562	271,61	260,00
2019	5	1193	270,60	257,00

Task 4 – Flash estimates

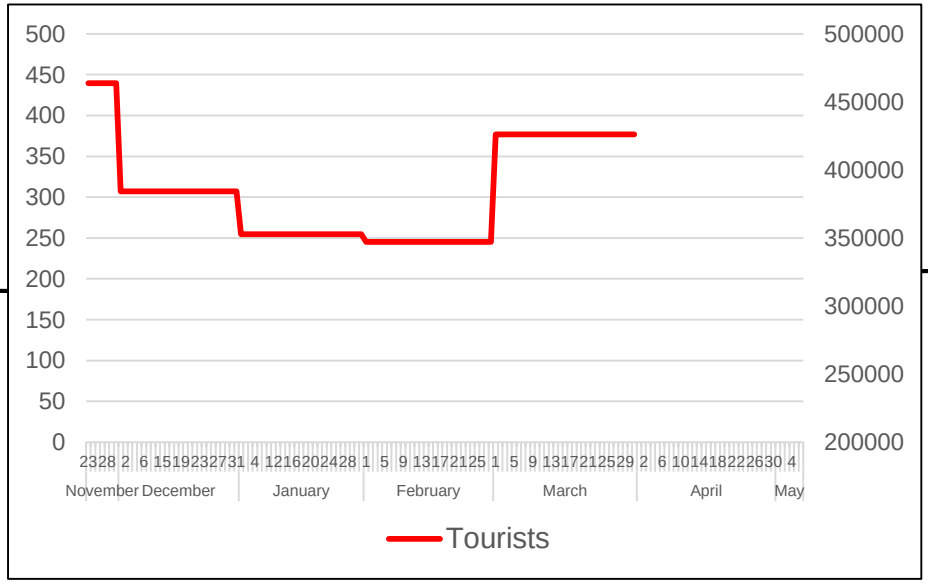
Time-series model.

- Y – the nights spent
- X’s – selected monthly statistics from web scraping e.g. median price
- If works the simpler the better e.g. LM over ARIMAX

Task 4 – Flash estimates

Web scraping of tourist accommodation portal

Survey On Tourist Accommodation Establishments

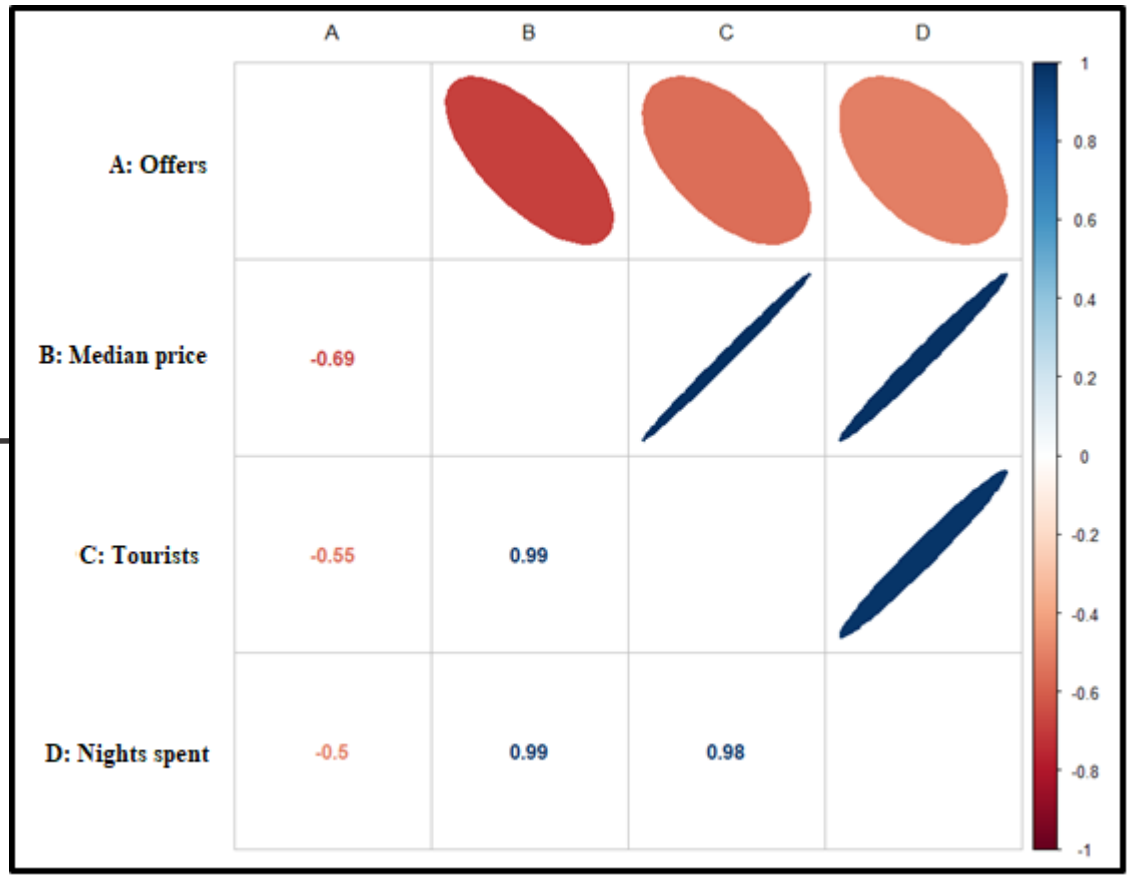


Flash estimates model

Statistics Poland

Task 4 – Flash estimates

Flash estimates model



Task 4 – Flash estimates

Flash estimates model

Statistics Poland

Thank you for your attention!

Sebastian Wójcik, Jarosław Napora