Big Data quality issues regarding multi-domain statistical data combining – a survey and case studies

Abstract

Employing Big Data methods and tools to produce statistical data makes the necessity of the revision of the data quality framework for official statistics. Although many different efforts have been made, including UNECE Big Data Quality Framework or different approaches in research papers, there is no unified Big Data quality framework that can be applied for different type of data sets, such as social media or large structured data sets. On the other hand, the variety of Big Data quality frameworks allows creating the set of quality indicators that will assess different aspects of the data source usability. Therefore, the solution is to create different frameworks depending on the data set used. It is rather easy when one dataset is used. More complicated is when different data sets are integrated, including various data types. The aim of the paper is to show how Big Data quality frameworks can be applied to create the set of indicators that will allow assessing the data set quality in three different stages – as input data sources, during processing phase (data sources integration) and when producing the output data (final experimental tables). The paper covers different aspects of Big Data integration. The first is intra-domain when combining data sets within three different statistical domains: population, tourism and agriculture. It includes combining data from traditional surveys and Big Data sources, such as social media data or satellite data. The second aspect is when combining inter-domain data sources. We have tested data integration by combining population and tourism data sets. The case studies and pilot surveys allows creating original conclusions on how to measure the data quality and which quality indicators can be applied to provide reliable assessment of the data sources and results.

Introduction

The paper shows an overview of the results of pilots conducted within ESSNet Big Data WP7. Among pilots conducted, there is a variety of data sources. We have tested several different use cases, including data sources such as business registers, web scrapped data, traffic loops, satellite data as well as Google Trends or Google Traffic. Finally, we decided to give brief description of original and reliable, sustainable data sources.

The pilots are within the following statistical domains:

- Population,
- Tourism,
- Agriculture.

All pilots have already been conducted by partner countries. For instance, two different approaches for Agriculture domain (Crop Types identifying by satellite images) have been prepared by Statistics Ireland and Statistics Poland. The case study on life satisfaction has been prepared separately by ONS UK and Statistics Poland. Moreover, our approaches have been tested by ESSNet partners (Ireland, Netherlands, Poland, Portugal, United Kingdom) with good results.

In this paper you will find information regarding data combining and quality issues – what has been done in the first wave of pilots. Use cases on data combining were divided by us into two approaches:

- intra-domain (e.g., satellite data and data from in-situ surveys),
- inter-domains (e.g., agriculture and agritourism accommodation establishments).

Quality issues

In WP7 three topics were investigated: sentiments revealed in the social media; agriculture; tourism/border crossing. Sentiment analysis is based on the posts from social media. People that are active on social media may differ from the total population and their number and activity can change over time. So the representativeness is an

issue when using this data source. This makes a challenge to compare the results with the current population. Also extracting background characteristics of the users and accurately determining the sentiment of the text not easy tasks. Depending on the social media channel, in some countries Twitter is not so popular and used only by selected group of people. The accuracy is measured with machine learning algorithms and varies between 60 and 90%, depending on the training dataset and country (pilot was conducted by Poland, Portugal and UK).

In the field of agriculture, WP7 combined satellite images with administrative data and survey data to train the machine learning algorithms to recognize different crop types. Sources can be linked accurately, findings can be verified by manual inspection of land lots and the first results are very promising. The accuracy is measured by the number of fields with crop types identified correctly and varies from 75% to 85% depending on the crop type and machine learning algorithm used (KNN and SVM are the most accurate). The pilot was conducted by Poland and Ireland, using different approaches.

In the Tourism/border crossing domain WP7 developed a methodology to measure intensity of border traffic by the number of vehicles/airplanes and passengers. Data about border traffic was obtained from government authorities in Poland and neighbouring Schengen member countries. Moreover, data on air traffic and train trips was scrapped from web portals to see the whole perspective of border traffic movement. This enable comparison of the traffic from both sides of the border. At this moment the results of traffic movement by road sensors can be presented for the border between Poland and Czech Republic, Lithuania, Germany and Slovakia. The estimates show number of vehicles, people/vehicle and expected vehicles. More detailed information is for air traffic – the number of countries is significantly higher because of the specific information available on the aircraft type and source and destination airport. The accuracy is an issue with the road traffic as there are some gaps in the data and some data must be estimated – it leads to the possibility of providing the data under or overestimated.

Interdomain data combining

Rural areas in Poland have significant touristic potential in the form of rich natural resources, culture and infrastructure. In addition, agritouristical farms continue to expand their offer addressed to the most demanding guests - better accommodation, additional attractions such as horse riding, fishing, tennis courts, summer pools, bonfires or regional events.

In the third quarter of 2017 Polish people made 18.7 million of domestic trips. Cottage house or agritouristical lodging was chosen in 2.7% of trips what gives around half million of trips. Sample size does not allow to produce results on NUTS 2 level. For each NUTS 2 region coefficient of variation exceeds 20% (on average it is 47%). One of region even did not appear in a survey. Therefore, we cannot state anything on expenses, trips and nights spent etc. on NUTS 2 level without relevant support of external sources.

We suppose that landscape and natural resources are crucial in choosing the best place for agritouristical activity. Thus, information on landscape structure may be used. Number of agritouristical lodgings is a derivate of landscape and natural resources. Distribution of agritouristical lodgings seems to be also valuable.

The aim of this study is to produce estimates of expenses, trips and nights spent on NUTS 2 level with better precision than from sample survey solely. Following data sources are planned to be used:

- Farm Structure Survey (NUTS 2 level) conducted by Central Statistical Office in Poland
- Integrated Survey of Tourism conducted by Statistical Office in Rzeszów (NUTS 2 level)
- Land and Buildings Register from Central Office of Geodesy and Cartography (NUTS 5 level)
- Lowland / highland structure (NUTS 3 level)
- Data from web scraping of webpages on agritouristical objects

Integrated Survey of Tourism is the source of information on expenses, trips and nights spent. Farm Structure Survey is a sample survey which covers data on location of homesteads, legal status (a natural person, a legal person etc.), presence of economic activity other than agriculture or agritourism. Land and Buildings Register provides data on landscape structure. Number of agritouristical lodgings is obtained from web scraping of relevant websites.

Since direct estimations on NUTS 2 level are not reliable and there are no additional variables in a tourism survey that totals or means are known in general population there is very limited number of methods that can be applied. At the beginning we used synthetic regression estimator. It turned out that OLS regression produces inadmissible outcomes (negative values). We needed a tool that is less sensitive to unreliable dependent variable and may shrink result towards mean. One of these is lasso regression. Details about that methods are described in attachment. Regression results are presented below.

Table. Lasso-based estimates on agritourism in Poland in the third quarter of 2017.

NUTS 2 region	Nights spent	Visitors	Total expenditures	Expenditures on accommodation
dolnośląskie	157299	49154	38063176	14679858
kujawsko-pomorskie	136604	36714	29430374	10747222
lubelskie	70444	22043	16848246	5726209
lubuskie	128360	21880	16659689	5419688
łódzkie	59530	11610	11132149	3784183
małopolskie	287019	93947	78408199	31192865
mazowieckie	8560	13869	2554044	174132

opolskie	98382	13508	16270084	5796491
podkarpackie	236122	73300	59136618	23019659
podlaskie	42345	4429	3984058	443955
pomorskie	187210	51759	38114586	14071677
śląskie	103083	20885	17365269	6587026
świętokrzyskie	100661	16825	18306975	6588852
warmińsko-mazurskie	190729	60776	42748426	15414115
wielkopolskie	91135	32685	18947858	6505026
zachodniopomorskie	226147	68740	47743768	17433312

Source: WP7 Report, ESSNet Big Data, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP7_Multiple_domains, accessed 30th May 2018.

Summary

WP7 prepared and tested 6 intra-domain pilots in three different domains (3 Population, 2 Tourism, 1 Agriculture with 2 pilots implemented by different approaches, first by Statistics Poland and the second by CSO Ireland). For data combining we have two different approaches – intra-domain data combining (combining of different data sources within one domain – e.g., survey data and web data) and inter-domain data combining (combining data sources from two or more domains, e.g., agriculture-tourism). It shows general problems with data quality on different aspects, including representativeness, coverage, missing data, accuracy of machine learning algorithms as well as under vs. over- estimated data.