ESSnet Big Data

Specific Grant Agreement No 2 (SGA-2)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
http://www.cros-portal.eu/

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2016.010-2016.756**

Work Package 7

Multi domains

Milestone 7.9

List of potential pilots and domains with successful implementation potential for further elaboration in the second wave of pilots in 2018

Prepared by: WP7 team

PL (leaders): Anna Nowicka, Jacek Maślankowski, Łukasz Błaszczyk, Sebastian Wójcik
IE: John Sheridan
NL: Piet J.H. Daas
PT: Rui Alves, Maria José Fernandes, Sónia Quaresma
UK: Alessandra Sozzi

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775

Table of contents

1. Introduction

The main goal of this milestone is to list pilots that were implemented with success by WP7 and can be conducted at a European level by other countries.

Among all WP7 pilots, 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 (chapter 3),
- Tourism (chapter 4),
- Agriculture (chapter 5).

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 report you will find also information regarding data combining – 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).

Intra-domain pilots have been included in chapters 3-5, according to the domain. Inter-domain data combining pilots were described in chapter 6.

Please note, that WP7 is still working intensively on additional pilots so the final list of the positively implemented pilots ("List of potential pilots and domains with successful implementation potential for further elaboration in the second wave of pilots in 2018") can be extended and published in Annex 1 the final SGA-2 report of WP7: "Deliverable 7.7. The General Report for Each Case Study/Domain including Recommendations on Legal Aspects, Availability, Sustainability, Methodology, Quality and Technical Requirements".

2. Executive Summary

WP7 prepared and tested 6 intra-domain pilots in three different domains (3 Population, 2 Tourism, 1 Agriculture with two 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).

Main findings from suggested list of pilots are as follows:

- The most tested pilot and most promising in **Population** domain is Life Satisfaction – it uses machine learning algorithm to produce the results of life satisfaction according to classification from EU-SILC survey (happy, neutral, calm, upset, depressed, discouraged), based on Twitter data, the accuracy varies between ~55% and 80% (the percentage of positively identified life satisfaction class – e.g., upset, calm, happy, depressed, …), depending on the algorithm and training dataset.
- Other two data sources that can be explored in **Population** domain are related to the selected health status of population (Morbidity areas use case on depression (Google Trends) by ONS UK) – the results showed us that there are differences in the official data and Big Data and to Peoples opinion/interestingness by topics based on websites by ONS UK by Facebook – this use case is very promising and it has a potential to be used to monitor everyday people's opinion on the specific event/news.
- Alternative use cases on **Population** include the possibility of estimating the density of population by BTS location (Base Transceiver Stations used by mobile phone operators) and daily and night population by Google Traffic, but they were not implemented at international level in the first wave of pilots.
- In **Tourism** there are two different pilots: Tourism accommodation establishments and Internal EU Border Crossing, including data sources by Air Traffic – Flight Movement web scraping (the data are available online and estimation of number of passengers is possible as well) and Traffic Loops data (data were obtained from road authorities but they are ready to be used after preprocessing and estimating the number of passengers with entropy econometrics method).
- **Agriculture** domain has one pilot prepared by two different methodological approach: first was prepared by Statistics Poland and the second was prepared by CSO Ireland – it was successful with the accuracy depending on the crop type and algorithm used, an average accuracy was about 80% (the percentage of positively identified crop types).
- There were two approaches used for intra-domain data combining – in the input stage for agricultural domain, for tourism domain is was combined in the output phase.
- The results of the pilots conducted shows that the great potential is in the Agriculture domain – to identify crop types. The case is ready to use with the open data that can be accessed on the Internet. Pilots related to Population domain

2.1. Life Satisfaction by Twitter

<table>
<tr><td colspan="2">USE CASE NAME Life Satisfaction by Twitter
(countries that can provide results as of February 2018: Poland, Portugal)</td></tr>
<tr><td>LEGAL ASPECTS
Twitter data may be collected via Twitter API. Personal data is limited to name (screen name), location.</td><td>TESTING/USEFUL REFERENCES
POLAND: The 6.187 tweets were used in implementation in Polish language, processed with stemming, lemmatization and stop words issues. Classification was made manually to prepare the training dataset, according to the list of classes presented below.

PORTUGAL: The 2.370 tweets were submitted to the following pre-process: Portuguese PorterStemmer; In CountVectorizer; Custom portuguese stopwords list (combined from several available Portuguese stopwords lists); Preserved accentuation with strip_accents = 'unicode'; Parameter max_features wasn’t subject to any restriction; the whole corpus was considered.</td></tr>
<tr><td>QUALITY (1-5)
availability – 5
sustainability – 4</td><td rowspan="3">IMPLEMENTATION
We decided to prepare training and testing datasets and used supervised learning in Python 3 language. INSEE Portugal modified Polish Software to work with Python 2.

The Training dataset for Portugal had 2.370 tweets (Portugal) and 6.187 tweets (Poland), (UK)for the 6 categories (1.happy, 2.neutral, 3.calm, 4.upset, 5.depressed and 6.discouraged).

In Portugal an additional twitter dataset with 600 tweets (100 for each category) was used for category prediction.
In Poland the testing dataset was 15% of the training population – it is a common way of providing information about the training dataset accuracy.</td></tr>
<tr><td>METHODOLOGY
Machine learning – supervised learning
Web scraping – we use Twitter API to gather and process the data.</td></tr>
<tr><td>TECHNICAL REQUIREMENTS
Python
MongoDB
Apache Spark for efficient processing</td></tr>
<tr><td colspan="2">RESULTS / REMARKS / MAIN FINDINGS
Statistics Poland findings:
1. Due to the rules in Polish language, the and testing datas had to be prepared very precisely and pre-processed using ad-hoc stemming and lemmatization .
2. In the first phase of the implementation we had to increase the number of tweets used in the training dataset due to the problems with low accuracy in the results of the training dataset.
3. Increasing the number of tweets added a significant boost to accuracy.
4. A small training dataset and the disproportion on the number of cases in different classes (e.g., 70% of positive comments compared to 10% of negative) resulted in almost accidental classification of tweets.

Statistics Portugal findings:
Each sentiment should be accurately defined to avoid imprecision both in the extraction step as well in the manual validation step. Since they are not mutually exclusive some of the sentiments</td></tr>
</table>

may have overlap meanings, e.g., happy/calm or discouraged/depressed. There are limitations to the classification when you don't properly conceptualised what you are trying to measure.

1. These sentiments can be polarized e.g., happy/sad or discouraged/encouraged, but others cannot (depressed).
2. Specify the domain: We are measuring sentiments regarding to what? Sentiment expression on politics (e.g., EU Commission) or a social event (e.g., Olympic Games) rely on different vocabulary and therefore it should be used a different approach. If the domain is purposely vague and undetermined then the sample should be much larger.
3. Too many categories
4. Larger sample training dataset
5. Twitter is not the most popular social media platform in Portugal. This is not a problem in an experiment, but it is relevant issue for further developments.
6. As in opinions, we should consider 5 elements when classifying sentiments:
    a. Target entity/object: "I'm calm" VS "She's calm"
    b. Aspect/feature of the entity: "She has a calm voice" VS "She is calm"
    c. Sentiment value of the opinion from the opinion holder "She's barely calm" VS "She's completely calm"
    d. Opinion holder/opinion source: " I think she's calm" VS "She thinks she's calm"
    e. Time when the opinion is expressed: "I was depressed" VS "I am depressed"

SHORT DESCRIPTION

The goal of the use case is to deliver data on life satisfaction - 1.happy, 2.neutral, 3.calm, 4.upset, 5.depressed and 6.discouraged. The goal is to support the data from EU-SILC survey with more recent data. The major drawback from this case study is that the dataset may not be representative, which is an issue for the future pilots.

2.2. Life Satisfaction by Facebook

USE CASE NAME **Life Satisfaction by Facebook** (ONS UK)

LEGAL ASPECTS

Facebook comments to public news pages can be collected up to a week before from the Facebook Graph API. Personal information is limited to screen name

TESTING/USEFUL REFERENCES

Hundreds of thousands of comments were collected in the time frame of a month from a single news page. Positive/Negative sentiment was derived from comments comparing 4 different Sentiment lexicons and the Vader library

QUALITY (1-5)

availability – 4
sustainability – 4

METHODOLOGY

Sentiment lexicons used and score calculated based on sum of scores of individual words, averaged across sentences
Web scraping – we use Facebook API to gather and the data.

TECHNICAL REQUIREMENTS

Python
MongoDB

IMPLEMENTATION

We decided to compute the sentiment score (from -1 to 1) by extracting lexicon words from the comments and combining the individual word scores first by sentence, then for the whole comment.

Clerical review of comments sentiment
Approx. 1000 comments for the 3 categories (1. positive, 2. neutral, 3. negative)

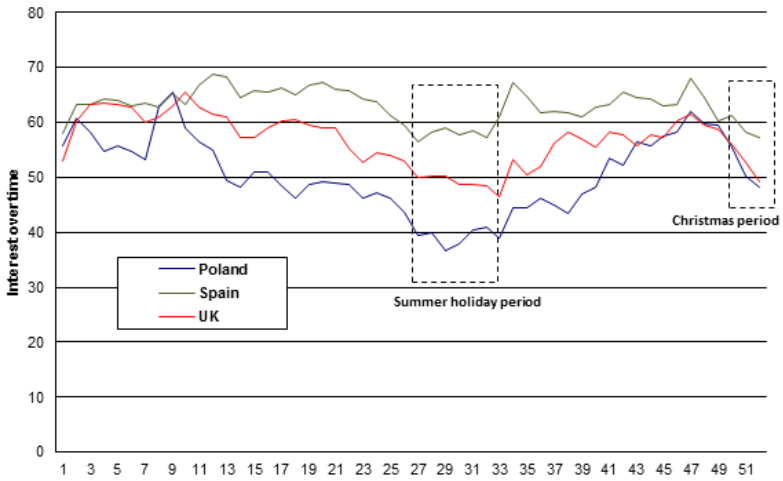
RESULTS / REMARKS / MAIN FINDINGS

ONS findings:

There are several limitations that arose from using lexicon-based sentiment analysis methods for analysing Facebook comments. Some of them include:

1. Long text 2. Noisy comments: many comments with just a name in it 3. Context relevant 4. Keyword-based approach is totally based on the set of keywords. Sentences without any keyword would imply that they do not carry any sentiment at all. 5. Meanings of keywords could be multiple and vague, as most words could change their meanings according to different usages and contexts. 6. The Facebook API does not provide any other information than person's full name and ID. Additionally, users' IDs are meaningful only within the same API/app. This limits the understanding of representativeness issues 7. Case study was limited to a single news agency, whereas should include multiple of them to be as representative as possible 8. Using an emotion lexicon (where each word is rated according to 8 different emotional states: joy, sadness, fear, anger, disgust, surprise, anticipation, trust), rather than just positive/negative, didn't provide any robust finding
SHORT DESCRIPTION The purpose of the case study was to examine the level of daily satisfaction by analyzing the content of messages for the presence of defined expressions describing emotional states, e.g., positive and negative sentiments and understand how they varies over time.

2.3. Alternative pilots

Name of the pilot/data source	Description
Morbidity areas and personal well-being by ONS UK/Google Trends	Short overview: The goal is to find personal well-being related attributes and morbidity areas using Google Trends. Implementation: Google Trends Analysis tools Main findings: The results may provide information related to the number of queries regarding personal well-being or morbidity areas attributes by countries. The results can show how people are affected by different issues related to health status or any personal issues, like depression by countries shown below.  Source: Nigel Swier, WP7 document, Google Trends as a source for measuring sentiment and personal well-being

<table>
<tr><td></td><td>The key issue is to prepare a set of queries in different languages regarding specific attribute or phenomenon that can be used to explain research topic (e.g., morbidity areas or personal well-being).</td></tr>
<tr><td>Estimating daily and night population /Google Traffic</td><td>Short overview:
The goal was to scrap information on people's commuting to estimate daily and night population. The data source was a Google Traffic. It was possible to scrap the data by API and receive information as a JSON file about the current situation on road (e.g, heavy traffic).

Implementation:
Python 3, Google Traffic API, JSON files

Main findings:
The results provided the data to construct the matrix on the road situation, like presented below:
<table>
<tr><td></td><td>Point X</td><td>Point Y</td><td>Point Z</td></tr>
<tr><td>Point X</td><td>X</td><td>40 min.</td><td>55 min.</td></tr>
<tr><td>Point Y</td><td>30 min.</td><td>X</td><td>15 min.</td></tr>
<tr><td>Point Z</td><td>45 min.</td><td></td><td>X</td></tr>
</table>
According to the matrix we can estimate the people commuting in every direction with the frequency of every minute.
This table can be generated every half an hour to provide information about direction of population movement (e.g., from the city centre to suburbs – point X to Z) and as a result can be used to estimate day and night population in the specific area/city.</td></tr>
</table>

2.4. Future data sources

Since the valuable data sources, such as mobile call records, are not available now, we decided to list data sources that are very promising but not ready to take them because of legal or quality issues. Call Detail Records may provide information on the location of the person, what is important both in Population (e.g., day and night population, number of commuters) as well as in Tourism domain (e.g., number of tourists).

Figure 1. Call Detail Records as a data source in population domain

3. Pilots related to Tourism domain and Border crossing movement

3.1. Tourism accommodation establishments

<table>
<tr><td colspan="2">USE CASE NAME: Tourism accommodation establishments (Statistics Poland)
(the result below shows issues related to boarding houses in one of the most popular touristic place in Poland – Kołobrzeg city)</td></tr>
<tr><td>LEGAL ASPECTS
Legal aspects are related to the possibilities of web scraping information on tourism accommodation establishments from websites such as booking.com, tripadvisor.com and related.</td><td>TESTING/USEFUL REFERENCES
We have tested using different data sources – web portals, including the most popular but also very specified oriented to tourism accommodation establishments (boarding houses) – booking.com, tripadvisor.com, agroturystyka.pl.</td></tr>
<tr><td>QUALITY (1-5)
availability - 5
sustainability - 3
METHODOLOGY
Web scraping – extracting data from tourism web portals and loading them into database (currently csv files).
Information from the Internet was used and compared with information from tourism register used for statistical survey.
TECHNICAL REQUIREMENTS
Python 3</td><td>IMPLEMENTATION
We have used the data from various Internet Portals to find tourism accommodation establishments (boarding houses).
The code was written in Python 3 and data was stored in CSV files.</td></tr>
<tr><td colspan="2">RESULTS / REMARKS / MAIN FINDINGS
The following description shows a case of wrong classification in accommodation offering web portals, according to official data:
1. As part of the work, 17 objects that used the name of the boarding house (in Polish pensjonat – boarding house) were detected.
2. Based on web scraping, it is difficult to determine if a given object meets the mandatory requirements related to the official classification of an object for a given type of facility.
3. Based on the analysis carried out and a comparison of the list of objects obtained from scraping with the list of units in the tourist accommodation survey card, it was established that only one object submits a statistical report as a boarding house. Others are classified as objects of other types.
4. On the basis of direct contact with representatives of the city of Kołobrzeg having information on the number of tourist accommodation establishments classified as a boarding house, it turned out that there is only one such facility that has obtained the appropriate status, meeting the requirements imposed by law.
5. The web scraping method is not applicable to searching for tourist accommodation facilities by type of facility. Therefore, we have to combine this data with official register of Tourism accommodation establishments</td></tr>
<tr><td colspan="2">SHORT DESCRIPTION
The goal of the use case is to web scrape all information related to tourism accommodation establishments. The aim of the work described in remarks was to identify tourism accommodation establishments providing services as boarding houses in the town of Kołobrzeg.</td></tr>
</table>

3.2. Border movement

USE CASE NAME**: EU Internal Porter Crossing by air, maritime and land border traffic estimation** (Statistics Poland)

LEGAL ASPECTS

Data from Polish Authorities, sample survey and websites with web scraping allowed. No legal obstacles.

TESTING/USEFUL REFERENCES

We have tested machine generated data from road authorities and air traffic

QUALITY

availability 4
sustainability 5

METHODOLOGY

Non-extensive Cross Entropy Econometrics, GLS and benchmarking

TECHNICAL REQUIREMENTS

Python, RStudio

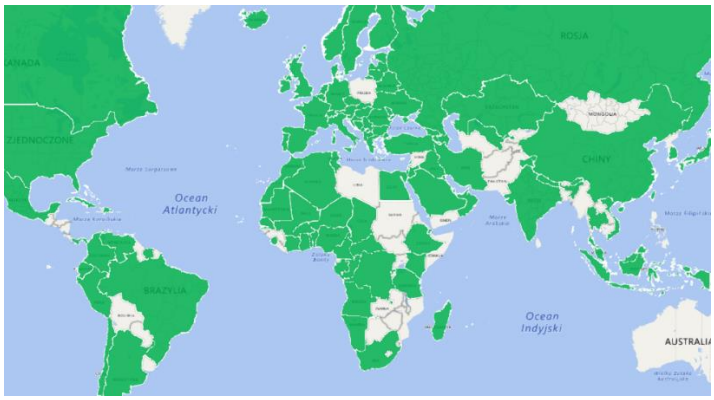
IMPLEMENTATION

Obtained data concerns 29 airports: 15 Polish airports and 14 hub airports to 484 airports in 145 countries. It contains information on around 270 thousand flights. Aircraft technical information on number of seats with respect to class pertain to 105 types of aircrafts. Base of IATA, ICAO and FAA codes covers 10259 airports. Data from traffic sensors comes from countries like Poland, Germany and more listed below. Data on maritime traffic comes from government authorities.

RESULTS / REMARKS / MAIN FINDINGS

1. Big Data on air traffic may increase countries coverage for which statistics on trips are produced. We obtained estimates for 145 countries while sample survey covers up to 80 countries.

   Graph. Countries with non-zero estimate from Big Data and sample survey.



2. Estimates are much more stable in time and distributions are smoother than in sample survey. For instance, with sample survey solely we reached only 6 destinations in Latin America and the distribution seems to be distorted and unreliable. At the same time with Big Data we reached 22 destinations out of around 30 possible destinations.
3. Big Data needs assistance of sample survey with respect to travel behaviours such as frequency of using hub airports, share of residents in total traffic etc.
4. Distribution of flights reflects preferences of many nations – not Polish directly. Therefore, obtained results may be more European than Polish
5. Data on road traffic intensity from traffic loops may partially replace manual counting.
6. With traffic loops only, we were able to cover 75.7% of land border traffic. Increasing density of traffic loop's net may improve the coverage.
7. Data sources and methods are completely different for air traffic and land border traffic.

Table. Border traffic with respect to country and dimension (land, sea and air) in 2015

	Number of crossings		
	Total	Polish residents	Non-residents
Total	217 175 084	88 121 949	129 053 135
land border	187 972 038	70 960 914	117 011 124
external EU's border	35 031 906	6 371 720	28 660 186
Russian-Polish border	6 098 860	3 369 138	2 729 722
Belarusian-Polish border	7 817 371	863 830	6 953 541
Ukrainian-Polish border	21 115 675	2 138 752	18 976 923
internal EU's border	152 940 132	64 589 194	88 350 938
Lithuanian-Polish border	4 946 268	1 528 605	3 417 663
Slovakian-Polish border	15 271 001	6 864 510	8 406 492
Czech-Polish border	50 870 006	24 917 817	25 952 189
German-Polish border	81 852 856	31 278 263	50 574 593
maritime border	1 851 524	1 088 040	763 485
air border	27 351 522	16 072 995	11 278 527
external EU's border	1 037 256	611 388	425 867
Russian-Polish border	315 118	185 740	129 378
Belarusian-Polish border	49 997	29 469	20 527
Ukrainian-Polish border	672 141	396 179	275 962
internal EU's border	3 947 888	2 326 998	1 620 890
Lithuanian-Polish border	90 604	53 405	37 199
Slovakian-Polish border	54 588	32 176	22 412
Czech-Polish border	179 694	105 917	73 777
German-Polish border	3 623 002	2 135 501	1 487 501
other air borders	22 366 379	13 134 608	9 231 770

SHORT DESCRIPTION

The aim of this study is to estimate air, maritime and land border traffic with respect to destination country in a case of air traffic.

For land border traffic data from traffic sensor in Poland, Germany, Slovakia and Lithuania was collected. Traffic sensors relevant to border traffic estimation were selected. For each source coefficient of variation was assigned basing on technical parameters of loops. Data biasedness was partially reduced with historical time series. Mirror statistics were combined with GLS method. Missing data was imputed basing on cross entropy measure. It enabled to estimate land border traffic with respect to country, vehicle and month.

For air traffic we gathered data on flights schedules from web scraping of flight movement webpages, data on technical information on aircrafts, base of IATA, ICAO and FAA codes, administrative data from Civil Aviation Authority of Poland, sample survey on tourism. Web scraping covered all Polish airports and crucial airports in Europe. Databases were linked into one database with country code (IACO or IATA), airport code and aircraft type as a key. Conditional distributions of passenger from each country were calculated and combined. Some statistics such as frequency of using hub airports, share of residents in total traffic were taken from sample survey on tourism. Data from Civil Aviation Authority of Poland were used to benchmark number of passengers to each hub country. Irrelevant routes were removed from database. It enabled to estimate number of passenger with respect to country using direct flights from Poland and using hub airports.

3.3. Alternative pilots

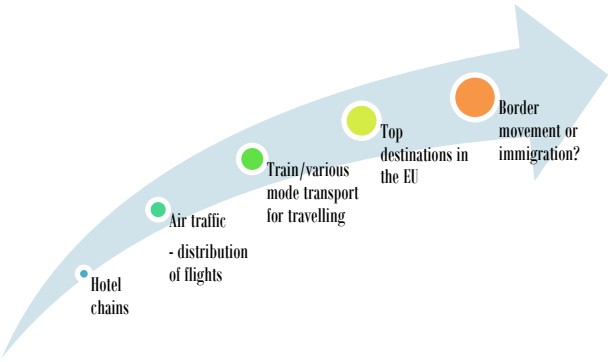
Name of the pilot/data source	**Description**
Road traffic/Google Traffic	Short overview: The goal was to scrap information on road traffic between countries by Google Traffic. It was possible to scrap the data by API and receive information as a JSON file about the current situation on road (e.g., heavy traffic). Implementation: Python 3, Google Traffic API, JSON files Main findings:

	As mentioned in 4.2 (population), the results provided the data to construct the matrix on the road situation, like presented below: <table><tr><td></td><td>Point X (country A)</td><td>Point Y (country B)</td></tr><tr><td>Point X (country A)</td><td>X</td><td>20 min.</td></tr><tr><td>Point Y (country B)</td><td>10 min.</td><td>X</td></tr></table> According to the matrix we can estimate the people crossing the border in every direction with the frequency of every minute. Long time of travelling will indicate in most cases the heavy traffic on the road.

3.4. Future data sources

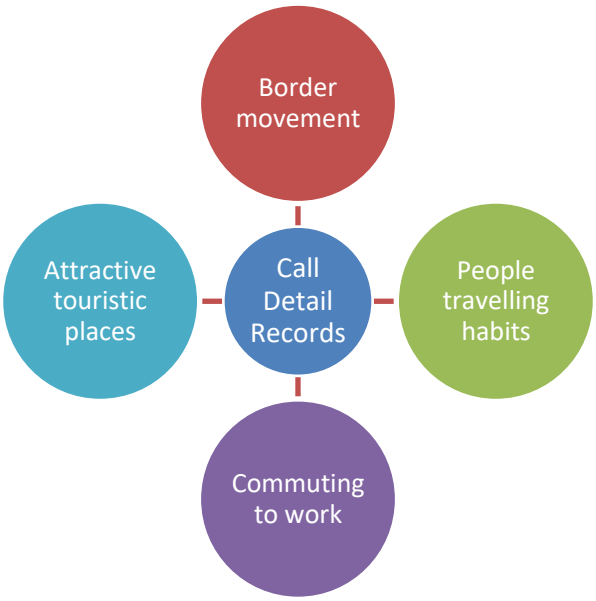
The goal of this chapter is to provide additional data sources giving a potential for new pilots on tourism. The variety of data sources allows to apply several different cases according to Figure 2. As mentioned in previous sections, we have tested the pilot for Tourism accommodation establishments (section 4.1, e.g., hotel chains), border movement (section 4.2 – air traffic, various mode transport for travelling) and as a result we can provide applications according to Figure 2.

Figure 2. Potential applications of various data source in tourism domain



As it was mentioned in section 3.4, Call Detail Records can provide information on the location of population. This is very important from Tourism point of view, where movement of the people may also be related to tourism – e.g., border movement, attractive touristic places, people travelling habits and commuting to work – see Figure 3.

Figure 3. Call Detail Records or antennas/network registrations of sim-cards as a data source in tourism domain



4. Pilots related to Agriculture domain

4.1. Crop types identification (1st methodological approach) by Statistics Poland

<table>
<tr><td colspan="2">USE CASE NAME Crop Types Identification (1) using satellite imagery Sentinel-1 and Sentinel-2 for crops classification (Statistics Poland)</td></tr>
<tr><td>LEGAL ASPECTS
Sentinel data may be downloaded via ESA open access hub. Sentinel data are free and available for anyone.</td><td>TESTING/USEFUL REFERENCES
One-year time series raw satellite data for Poland.
Over 5.000 in situ samples were collected as a ground truth for satellite images classification.</td></tr>
<tr><td>QUALITY (1-5)
1. Accuracy – 4 (on average 80-97%)
2. Completeness – 5
3. Sustainability – 5
4. Availability – 5
Measured also by classification results compared to control in situ dataset (confusion matrix) and classification results compared to official administrative data (CSO, ARMA).</td><td rowspan="3">IMPLEMENTATION
• Sentinel-1 A & B From the beginning of the year 2017 more than 2200 scenes were downloaded for the whole country (3.75 TB). Data were downloaded in a fully automatic way. The procedure for image automatic downloading was developed in Earth Observation Group of Space Research Centre of Polish Academy of Sciences. Surface of Poland was covered by 6 paths of Sentinel-1 satellites.
• Sentinel-2 Images were acquired in order to support the segmentation process and to improve winter and summer crops separation. More than 1160 images were downloaded, around 1.2 TB of data.
• NDVI Basing on a series of optical Sentinel-2 images from the period autumn 2016 and</td></tr>
<tr><td>METHODOLOGY
1. Raw satellite image processing
2. Segmentation of processed data.
3. Object based image classification (machine learning algorithms).</td></tr>
<tr><td>TECHNICAL REQUIREMENTS</td></tr>
</table>

Workstation with high computing power (min. Core i7 processor, 64GB of RAM and 2TB SSD drive).
ESA SNAP and PolSARPro software for satellite data processing.
E-cognition Developer software for object-based image classification.

spring 2017 NDVI index was calculated from each image.

- **Segmentation** It was done in the eCognition software. In the first phase, LPIS vector was used for the exclusion of the areas that are outside of the arable lands (e.g., forests, water, built-up areas). First level of the segmentation was done based on the LPIS vector. Next, multitemporal radar images (lambda) were used for the creation of another level of segments. In this level, good separation of inhomogeneous arable lands (that occupy one field) was obtained.
- **Crop recognition** Based on the previous experience, the time series of the coherence matrix and H / L decomposition were classified using the Support Vector Machine (SVM) algorithm.

RESULTS / REMARKS / MAIN FINDINGS

Obtained results allow us to formulate following conclusions:

1. New method of in-situ samples gathering ensured their proper spatial distribution. This allowed proper learning of the classifier
2. Utilization of multi temporal NDVI mosaic obtained using data acquired between fall 2016 and spring 2017 allowed for more precise distinction of winter and summer crops.
3. Accuracy of classifier learning varies from 80% for path 3 to 97% for path 2. More detailed analysis of this phenomenon is needed. It means that the algorithms manages to correctly classify a crop from the images in 80% of the times.

Area [ha] of classified crops (preliminary results):

	winter wheat	corn	spring cereal mixes	winter rape	spring barley	oat	winter triticale	winter barley	winter cereal mixes	rye	spring wheat	spring triticale	spring rape
	2	3	4	5	6	7	8	9	10	11	12	13	14
Dolnośląskie	173253	145881	47073	154136	59687	41720	47594	39244	14211	42753	49929	11192	18243
Kujawsko-pomorskie	106582	221238	99912	136147	62746	37729	67247	56993	19979	90421	36283	19340	18053
Lubelskie	80226	134882	54770	104419	61947	54242	127243	16081	53798	87439	79212	68082	51356
Lubuskie	37762	53809	44258	52875	25441	21838	44464	19008	9486	32579	21665	12868	11471
Łódzkie	82661	127058	117794	49828	81216	86792	68448	36679	34115	87062	43037	19569	21436
Małopolskie	73606	55655	18170	30125	39699	14592	26300	21019	19546	14241	31585	7411	13414
Mazowieckie	128534	285548	93481	79358	106266	94890	115691	98596	61298	123378	72642	62513	28074
Opolskie	79370	60109	19870	66634	24564	11434	28393	20131	7224	23074	19036	5672	9106
Podkarpackie	43720	48561	12391	27083	21252	25955	37957	27630	13067	33414	43715	28357	40337
Podlaskie	57053	148827	72773	22294	44103	64786	59429	23586	31177	78835	51738	29961	12161
Pomorskie	94457	80850	62211	108989	49716	56912	65305	21061	16989	55091	49370	14028	9681

Śląskie	63784	42242	27681	31258	19671	21060	27327	9276	9928	18534	28313	7033	7325
Świętokrzyskie	71479	55883	28618	21855	44362	29265	23556	19767	10970	15941	39932	6231	9594
Warmińsko-mazurskie	160519	121593	62298	114711	48849	31884	95813	34177	29862	78641	74727	36332	17037
Wielkopolskie	130171	323677	149718	152008	120466	88570	135491	94627	27284	161912	50577	28547	29495
Zachodniopomorskie	90595	86818	66665	136496	56379	45216	64045	26356	11254	82772	78270	22716	26544

Confusion matrix for classified crops (1-st compilation):

Klasyfikacja \ Próbka	winter wheat	corn	spring cereal mixes	winter rape	spring barley	oat	winter triticale	winter barley	winter cereal mixes	rye	spring wheat	spring triticale	spring rape	Sum
winter wheat	27	0	0	0	0	0	1	0	1	2	1	0	0	32
corn	0	11	0	0	0	0	0	0	0	0	0	0	0	11
spring cereal mixes	0	0	17	0	0	0	0	0	0	0	0	0	0	17
winter rape	0	0	0	9	0	0	0	0	0	0	0	0	0	9
spring barley	0	0	0	0	20	0	0	1	0	0	0	0	0	21
oat	0	0	0	0	0	29	1	0	0	0	0	0	0	30
winter triticale	0	0	0	0	0	0	28	0	0	0	0	0	0	28
winter barley	0	0	0	0	0	0	0	12	0	0	0	0	0	12
winter cereal mixes	0	0	0	0	0	0	0	0	9	0	0	0	0	9
rye	0	0	0	0	0	0	0	0	0	24	0	0	0	24
spring wheat	0	0	0	0	0	0	0	0	0	0	17	0	0	17
spring triticale	0	0	0	0	0	0	0	0	0	0	0	12	0	12
spring rape	0	0	0	0	0	0	0	0	0	0	0	0	4	4
unclassified	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sum	27	11	17	9	20	29	30	13	10	26	18	12	4	
Producer	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.92	0.90	0.92	0.94	1.00	1.00	
User	0.84	1.00	1.00	1.00	0.95	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Hellden	0.92	1.00	1.00	1.00	0.98	0.98	0.97	0.96	0.95	0.96	0.97	1.00	1.00	
Short	0.84	1.00	1.00	1.00	0.95	0.97	0.93	0.92	0.90	0.92	0.94	1.00	1.00	
KIA Per Class	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.92	0.90	0.91	0.94	1.00	1.00	
Overall Accuracy	0.97													
KIA	0.97													

SHORT DESCRIPTION

The main goal of the use case is to deliver methodology for calculating area of main crops in Poland using free remote sensing data.

5.2 Crop Types Identification (2nd methodological approach) by CSO Ireland

<table>
<tr><td colspan="2">USE CASE NAME Crop Types Identification (2) by the use of satellite imagery for crop classification (CSO Ireland)</td></tr>
<tr><td>LEGAL ASPECTS
Sentinel data may be downloaded via ESA open access hub. Sentinel data are free and available for anyone.</td><td>TESTING/USEFUL REFERENCES
Used a sample area of Irish satellite data.
LPIS data was used for training.
Data covered a sample time period.</td></tr>
<tr><td>QUALITY (1-5)
Availability – 5
Sustainability – 4</td><td rowspan="3">IMPLEMENTATION
Sentinel 1 and 2 images for May to July 2016 were downloaded for the whole of Ireland. Due to processing constraints a smaller area was used. This was selected based on variety of land cover types and minimizing cloud cover for the period of interest.
LPIS data used was restricted to the same period and the same area to ensure the crops that would be in evidence in that time frame were being used in training.
Classification was performed using the QGIS software using a variety of algorithms</td></tr>
<tr><td>METHODOLOGY
Raw satellite image acquisition and processing
Preparation of LPIS data
Classification using various machine learning algorithms</td></tr>
<tr><td>TECHNICAL REQUIREMENTS:
Workstation with higher than average computing power (min. Core i7 processor, 8Gb RAM, powerful graphics card).
ESA SNAP software for satellite data processing.
QGIS and the Orfeo toolbox for image processing and classification.</td></tr>
<tr><td colspan="2">RESULTS / REMARKS / MAIN FINDINGS
Conclusion:
1. The methodology did successfully classify data however the accuracy rate would need to be improved, with the limited sample and training set it was too low.

Issues:
1. Usable optical data was difficult to find for Ireland due to cloud cover.
2. Software was slow to pre-process, classify and post-process.
3. Spectral signatures of vegetative land cover types are quite similar which resulted in misclassified pixels.
4. Some classifiers often found Irish field sizes too small to generate spectral signature.</td></tr>
<tr><td colspan="2">SHORT DESCRIPTION
The goal is to determine the feasibility and deliver the methodology for determining crop types in Ireland using free satellite imagery.</td></tr>
</table>

Administrative sources (1) LPIS/IACS, (2) Tax register, (3) Weather sources, (4) Maps, (5) import/export data, (6) IUNG, (7) data related to the industry: food, (8) in situ surveys

6. Data combining

6.1. Agriculture and Tourism data combining

<table>
<tr><td colspan="2">USE CASE NAME Impact of landscape resources in rural areas on agritourism intensity (Statistics Poland)</td></tr>
<tr><td>LEGAL ASPECTS
Data from registers, sample surveys and websites with web scraping allowed. No legal obstacles.</td><td>TESTING/USEFUL REFERENCES
None</td></tr>
<tr><td>QUALITY
availability 5
sustainability 5</td><td rowspan="3">IMPLEMENTATION
With web scraping we collected data from 1950 agritourist lodgings. Sample survey on homesteads covered over 1600 farmers while survey on tourism covered 150 records pertaining to agritourism: expenditures, nights spends etc. Land and Buildings Register is available on NUTS 5 level and contains detailed information on land structure and usage.</td></tr>
<tr><td>METHODOLOGY
Spatial disaggregation methods, shrinkage estimator.</td></tr>
<tr><td>TECHNICAL REQUIREMENTS
Python, RStudio</td></tr>
<tr><td colspan="2">RESULTS / REMARKS / MAIN FINDINGS
1. Preliminary statistical analysis reveals that agritourist lodgings are mostly located on hilly areas. Number of them is moderately correlated with area of forest land, meadows, and pastures, while weakly correlated with area of arable land, lakes and rivers.
2. Number of agritourists is mostly related to area of lakes and rivers, then to forest land. It is negatively correlated with area of urban areas and arable areas.
3. With points 1-2 we may conclude that conditions of agritourism development are not the same as for agriculture.
4. Raw estimates on agritourism on NUTS 2 level are unreliable. It is ill-posed inference problem therefore OLS models seem to be irrelevant (mediocre R-squared while adjusted R-squared is low)</td></tr>
<tr><td colspan="2">SHORT DESCRIPTION
The aim of this study is to disaggregate data on agritourism intensity from national level to NUTS 2 level.
Data collected in sample survey on agritourism does not allow to produce statistics on NUTS 2 level. We assumed that agritourist intensity is related to landscape resources of the region. Such a relation would help us to disaggregate data from national level to lower level of aggregation. We gathered data on landscape structure on NUTS 5 level from Land and Buildings Register from Central Office of Geodesy and Cartography, data on lowland / highland structure, data from sample survey on homesteads and survey on tourism on NUTS 2 level. Web scraping of webpages on agritourist objects was also performed to collect information on prices and locations.
The basic idea is to use spatial disaggregation methods to estimate statistics on NUTS 2 level. These methods consist of regression and benchmarking stage. Since we have very small amount of data on variables of interest, we may encounter ill-posed inference problem – raw estimates are unreliable. Therefore shrinkage estimators may be useful.</td></tr>
</table>

6.2. List of potential inter-domain pilots

The list of potential inter-domain pilots was identified during the WP7 meeting in Warsaw on 12-13 June 2017. We decided to explore the potential of the following inter-domain data combining pilots:

(1) Agriculture and others: key geolocation; data from census/registers;
(2) Agriculture and tourism: web scraping on agricultural hotels/tourism; Instagram - geolocated photos - how often is different area photographed; transportation and data from mobile phones - potential tourism in the agricultural regions; data on flights
(3) Workers in summer-season; border crossing; estimate the number of people working in agriculture;
(4) Agriculture-Population - satisfaction of living in agricultural regions vs. other regions - census data; can be measured by Twitter or Instagram;
(5) Agriculture-Tourism - model assisted data, modelling the relationship between (the impact of agriculture cheap local food on touristic movement)
(6) Migration - Border traffic: two directions - it may indicate the migrations;
(7) Facebook images and Tourism - mark each image from Facebook by location to improve tourism statistics;
(8) Check the correlation between state of population and agriculture (depopulation process) - connection with unemployment rate;
(9) Gather data and divide data into group profiles - grouping data by aged 40-50, married, with children, check if they are football fans - cross with data from air traffic.