

## ESSnet Big Data

### Specific Grant Agreement No 2 (SGA-2)

<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata>  
[https://ec.europa.eu/eurostat/cros/content/essnetbigdata\\_en](https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en)

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2016.010-2016.756**

### Work Package 0

#### Co-ordination

#### Deliverable 0.7

### Final Technical Report

Final version 2018-06-30

#### Prepared by:

Martin van Seville (WP 0, CBS, Netherlands)  
Peter Struijs (WP 0, CBS, Netherlands)  
Nigel Swier (WP 1, ONS, United Kingdom)  
Monica Scannapieco (WP 2, ISTAT, Italy)  
Toomas Kirt (WP 3, EE, Estonia)  
Anke Consten (WP 4, WP 8, CBS, Netherlands)  
David Salgado (WP 5, INE, Spain)  
Tomaž Špeh (WP 6, SURS, Slovenia)  
Anna Nowicka (WP 7, GUS, Poland)  
Piet Daas (WP 8, CBS, Netherlands)  
Marc Debusschere (WP 9, SB, Belgium)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)

[p.struijs@cbs.nl](mailto:p.struijs@cbs.nl)

telephone : +31 45 570 7441

mobile phone : +31 6 5248 7775

# Contents

|   | <i>page</i> |
|---|-------------|
| Executive summary                             | 3           |
| 1. Introduction                               | 7           |
| 1.1. Background                               | 7           |
| 1.2. General approach                         | 8           |
| 1.3. Organisation                             | 9           |
| 2. Results of the workpackages                | 10          |
| 2.1. Webscraping / Job Vacancies              | 10          |
| 2.2. Webscraping / Enterprise Characteristics | 18          |
| 2.3. Smart Meters                             | 25          |
| 2.4. AIS Data                                 | 32          |
| 2.5. Mobile Phone Data                        | 39          |
| 2.6. Early Estimates                          | 46          |
| 2.7. Multi Domains                            | 53          |
| 2.8. Methodology                              | 59          |
| 3. Issues encountered                         | 68          |
| 3.1. General issues                           | 68          |
| 3.2. Issues at the level of the workpackages  | 70          |
| Annex 1: Communication and dissemination      | 75          |
| Annex 2: Evaluation of SGA-2                  | 77          |

## Executive Summary

This is deliverable 0.7, the Final Technical Report, of the Specific Grant Agreement No 2 (SGA-2) of the Framework Partnership Agreement (FPA) for the ESSnet Big Data. The FPA, which has 22 partners, covers the period from January 2016 to May 2018. SGA-2 covers the period from January 2017 to May 2018, on which this deliverable reports.

The ESSnet has organised the core of its work around seven workpackages, each workpackage dealing with one pilot and a concrete output. The pilots cover five phases: (1) data access, (2) data handling, (3) methodology and technology, (4) statistical output, and (5) future perspectives. SGA-1 covered only some of the five phases for each of the workpackages, the rest being covered by SGA-2. In addition, an eighth workpackage, which was not part of SGA-1 but started in SGA-2, concerned the horizontal issues of methodology, quality and IT.

These are the main results obtained in SGA-2:

### WP 1 Webscraping / Job Vacancies

This pilot explores the potential of on-line job vacancy (OJV) advertisements as a data source for job vacancy statistics. OJV data contain more variables and detail than existing job vacancy statistics. Various approaches to accessing data were explored including web-scraping of job portals and enterprise web-sites as well as arranged access with owners of on on-line job vacancy data. A key outcome of this pilot was the establishment of a relationship with the European Centre for the Development of Vocational Training (CEDEFOP), who are undertaking a major project to collect OJV data for all EU member states. This is expected to be the main source of OJV data for use within the ESS. A range of data processing, quality and methodological issues were explored including deduplication, classification of data from unstructured text, data linking, validation and time series analysis. However, quality issues around coverage, representativity, definitions and structural differences in on-line labour markets between countries, mean that it cannot replace existing job vacancy statistics. Future challenges will centre on how to produce and present new statistics based on OJV data together with existing statistics.

### WP 2 Webscraping / Enterprise Characteristics

The purpose of this workpackage is to investigate whether web scraping, text mining and inference techniques can be used to collect process and improve general information about enterprises. In particular, the aim is twofold: (i) to demonstrate whether business registers can be improved by using web scraping techniques and by applying model-based approaches in order to predict for each enterprise the values of some key variables; (ii) to verify the possibility to produce statistical outputs with more predictive power combined or not with other sources of data , like the “ICT use by enterprises” survey. SGA-1 work had in scope 4 use cases (URLs retrieval, e-commerce/web sales, social media detection, job advertisement detection). SGA-2 work considered two additional use cases, namely: NACE detection and Sustainable Development Goals (SDGs) detection. Moreover, in SGA-2, all the pilots implemented in the first phase have been enhanced by (i) consolidating adopted techniques and (ii) extending the number of enterprises to which the scraping activity was targeted. In addition, new pilots were implemented not only for new use cases but also for SGA-1 use cases for which additional countries committed to develop pilots. Final methodological and technological

conclusions based on all the work done within the ESSnet were drawn and a set of output indicators were published on the ESSnet wiki as experimental statistics.

### WP 3 Smart Meters

The main goal of workpackage 3 was to demonstrate the potential use of data from smart electricity meters for production of official statistics. The pilot had three goals with regard to expected outputs. First, to assess whether current survey based business statistics can be replaced by statistics produced from electricity smart meter data, second, to produce new household statistics and third, to identify vacant or seasonally vacant dwellings. The aim of the study conducted for SGA-2 was to look beyond the goals set for the workpackage and identify potential statistical products in the domain of energy consumption or in other statistical domains by relaying on the data produced during the previous task. The results of this study are summarized in Report 3 Future perspectives. During the SGA-2 the workpackage focused on classification and providing different use cases of using the smart meter data. The smart meter data was used to produce regional electricity statistics and it was evaluated whether the data can be used to produce tourism statistics. In addition, from the theoretical viewpoint potential uses for different kinds of smart meters (e.g. natural gas, water) were proposed. The impact of aggregation for producing different statistical products from the smart meter data was also evaluated. The second aim was to provide a compact overview of lessons learned during the project and give recommendations to other countries, which start using the smart meter data. As a result, Report 4 Recommendations was produced.

### WP 4 AIS Data

The workpackage investigates whether real-time measurement data of ship positions (measured by the so-called AIS-system) data can be used for improving the quality and internal comparability of existing statistics and for new statistical products relevant to the ESS. Reports were produced (1) on the use of AIS data to determine emissions, (2) on its use to produce possible new statistical output, and (3) summarising the contents and the outcomes of WP 4. The latter report also includes the results on analysing and elaborating scenarios for production of European and national statistics based on one single European data source and a cost-benefit analysis of using AIS-data for official statistics. The original plan included developing a model to calculate emissions and getting access to data from the European Maritime Safety Agency (EMSA). However, at the start of SGA-2 the scope of SGA-2 was adjusted, because good models for calculating emissions already existed and access to AIS data from EMSA was not granted. In this workpackage European AIS data from Dirkzwager (DZ) was used and the quality of this source was compared to satellite data from Luxspace (LS) and national data from Greece. Visualisations were made of transshipment by container giants, and of the average and maximum speed of ships in European seas and oceans. Results from this workpackage show the potential wealth of AIS data to improve current statistics and to generate new statistical products. Although some important elements of current maritime statistics such as type and quantity of goods loaded or unloaded at the port are not part of AIS, AIS still is useful to improve other aspects of maritime statistics and provide new products. In order to use AIS in official statistics, further investigation is needed. Also, an AIS source of good quality should be available at least covering European waters. This should be data from both land based stations and satellites.

## WP 5 Mobile Phone Data

This workpackage has focused on the development of a methodological framework, the analysis of the IT infrastructure and software tools, and the assessment of quality issues regarding the use of mobile phone data in the production of official statistics. Concerning methodology, WP 5 investigated the whole process from data collection to statistical output. Concrete methodological proposals were provided for different elements of this process. Linked to this framework, IT platforms for data access and processing were described and two R packages developed. The first one aims at implementing the Bayesian approach to geolocate network events based on the signal strength and the second one at implementing the statistical model developed to estimate population counts. Concerning quality, WP 5 has focused on two aspects. On the one hand, an analysis was made on how the European Statistics Code of Practice is going to be affected according to the preceding proposals. On the other hand, proposals have been made to deal with the accuracy dimension of quality in the context of the new inference model for the production of official statistics using mobile network data. Finally, the workpackage has made recommendations for future research.

## WP 6 Early Estimates

The aim of WP 6 was to investigate how a combination of multiple big data sources and existing official statistical data can be used in order to create existing or new early estimates for statistics. Several pilots were carried out during the SGA-2. The most promising estimator was GDP, but the pilots were not limited to GDP due to the fact that results of analysing data sources proposed the calculation of estimates of other economic indicators. The main outcome was the calculation of a testing set of early estimates of a concrete economic indicators (GDP, TIO, IPI, ...) together with the defined methodology and process needed for this purpose. Those estimates were calculated with respect to country specifics (availability of data sources) and possibly broken down into economic activity classes and (or) regional domains. Recommendations about the mode of combining (big) data sources, models for estimating economic indicators, quality assessment, IT infrastructure and statistical process of conducting the calculation of early estimates of economic indicators were prepared. The main achievements are calculated concrete estimates for economic indicators with impact and quality assessment of results. It was shown that incomplete early micro-level sources and a real-time big data source such as the traffic loop data can be used to produce early estimates of economic indicators. NSIs can shorten the publication lag in a straightforward way without compromising the quality of results. Concrete advice on how to embrace the opportunities of nowcasting was provided. A range of methodological recommendations related to methodology being investigated for the purposes of nowcasting early economic indicators using traffic loop data was offered. NSIs can address a major quality issue, namely the timeliness, by using a range of micro-level data sources accumulated in the registers well before the official release is made, by employing large dimensional econometric models, to form an initial quick estimate of the target indicator. This does not necessarily lead to too large revisions, but adds significantly to the quality of official statistics through timeliness dimension.

## WP 7 Multi Domains

This workpackage prepared and tested 6 intra-domain pilots in three different domains: three in Population, two in Tourism, one in Agriculture. The best tested pilot and most promising in the Population domain is Life Satisfaction – it uses machine learning algorithm to produce the results of life satisfaction according to the classification from the EU-SILC survey (happy, neutral, calm, upset, depressed, discouraged), based on Twitter data. The other two pilots in the Population domain are related to the selected health status of population and to Peoples opinion/interest by topics based on websites by ONS UK by Facebook. 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 and Traffic Loops data. The agriculture domain has one pilot prepared by two different methodological approaches: the first was prepared by Statistics Poland and the second by CSO Ireland. For combining data two different approaches were carried out – intra-domain data combining (all domains) and inter-domain data combining (agriculture-tourism). The results of the pilots conducted show that the greatest 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.

## WP 8 Methodology

The aim of this workpackage was to lay down a general foundation in the areas of methodology, quality and IT infrastructure when using big data for statistics produced within the European Statistical System. WP 8 therefore started with a workshop in which the most important topics in the area of IT, quality and methodology were identified. Next an overview of important papers, project results, presentations and webpages relevant to the application of big data for official statistics was created. This overview was linked with the findings of the pilots in the first phase (SGA-1) and the input available from the second one (SGA-2) of the ESSnet Big Data. The main outcome of WP 8 was an overview of the methodological, quality and IT findings when using big data for official statistics. Experiences obtained both in- and outside the ESSnet Big Data were used as input for WP 8.

# 1. Introduction

## 1.1 Background

This is deliverable 0.7, the Final Technical Report, of the Specific Grant Agreement No 2 (SGA-2) of the Framework Partnership Agreement (FPA) for the ESSnet Big Data. The FPA covers the period from January 2016 to May 2018. SGA-1 covers the period from February 2016 to July 2017. SGA-2, on which this deliverable reports, covers the period from January 2017 to May 2018. Thus, SGA-1 and SGA-2 have a time overlap from January till July 2017, but this report does not include activities of SGA-1.

This Final Technical Report of SGA-2 builds on deliverable 0.2 of SGA-1, the Final Technical Report of that SGA. Unless needed for reasons of clarity, its contents has not been repeated in the current report. This report describes the results of the action, not does not list the activities carried out by the workpackages, as that will be done in the Final Report on the Implementation of the Action, due 60 days following the closing date of the action. For the same reason, this report does not comprise an evaluation of the budget needed and used.

The overall objective of this ESSnet is to prepare the ESS for integration of big data sources into the production of official statistics. The FPA is founded on a consortium of 22 partners, consisting of 20 National Statistical Institutes (NSIs) and two Statistical Authorities. For SGA-2, all but one NSIs and one of the Statistical Authorities have been involved as beneficiaries of the agreement, so SGA-2 was carried out by 20 partners.

For SGA-1 as well as SGA-2, the consortium has organised the core of its work around a number of workpackages, each workpackage (WP) dealing with one pilot and a concrete output. In SGA-1 there were seven workpackages, focused on specific sources or domains:

1. WP 1 Webscraping / Job Vacancies
2. WP 2 Webscraping / Enterprise Characteristics
3. WP 3 Smart Meters
4. WP 4 AIS Data
5. WP 5 Mobile Phone Data
6. WP 6 Early Estimates
7. WP 7 Multi Domains

A separate workpackage, WP 0, was created for the co-ordination of the ESSnet. For dissemination a separate workpackage was created as well, WP 9. That workpackage is also responsible for facilitating communication. All these workpackages were continued in SGA-2. However, SGA-2 includes an additional workpackage as, given the overall objective, the findings needed to be generalised. For this, the new workpackage WP 8 was added. This concerns Methodology, but also covers other overarching aspects, in particular Quality and IT.

SGA-2 specifies the agreed outputs of the workpackages, and its inputs, both in terms of number of days contributed by partner and workpackage and in terms of material costs. For SGA-2, the total budget available is one million euro, but only 90% of costs, as a maximum, will be reimbursed. (The same budget and percentage applied to SGA-1.

For more specifics on the FPA, SGA-1 and SGA-2 reference is made to the actual agreements. For the current deliverable it is useful to mention that an overview of the milestones and deliverables of SGA-2 can be found on page 10 of the signed version of Annex II of SGA-2, an overview of the distribution of manpower (by partner and work package) is given on page 47, and an overview of the foreseen physical meetings on page 48 of the same document. For each workpackage the document (Annex II of SGA-2) provides a description of tasks, specifying, among other things, the tasks to be carried out, the milestones and deliverables, and the number of days each partner contributes to the workpackage. A specification of the budget is given in Annex III of SGA-2.

The remainder of this chapter describes the approach generally taken to the pilots, and the way the ESSnet has organised itself. The next chapter presents the results obtained so far, for the eight workpackages. The third chapter describes the issues encountered in the action, at a general level and for the workpackages.

## **1.2 General approach**

The pilots, as foreseen in the FPA, have one thing in common: they cover the complete statistical process, from data acquisition to the production of statistical output. In addition, and in accordance with the general objective to prepare the ESS for the integration of big data sources into the production of official statistics, the pilots also consider future perspectives. Thus, all pilots recognise the following five phases:

1. Data access
2. Data handling
3. Methodology and technology
4. Statistical output
5. Future perspectives

The tasks, milestones and deliverables of the workpackages refer to these phases. However, SGA-1 covered only some of the five phases for each of the workpackages, the rest being covered by SGA-2. And the phases covered by SGA-1 are not the same for each pilot (workpackage), as for some areas it was possible to plan ahead further (in time and phases) than for other areas. In particular, WP 5 concentrated on data access problems in SGA-1 and could not plan further ahead, as data processing would depend on the results of the efforts to realise data access. Therefore, WP 5 was planned to end in December 2016, whereas the other workpackages continued into 2017. For WP 6 and WP 7, a longer exploration period was needed for the first two phases, therefore they were planned to end – and did end – in February 2017. This explains the overlap in time of SGA-1 and SGA-2.

At a practical level, this approach required to be facilitated in several respects. First of all, an organisational approach was needed to ensure that the agreed output would be produced with the resources foreseen. This is the subject of the next section, 1.3. In order for the partners of the ESSnet to be able to process big data, some IT facilities were considered necessary, although these were needed at the beginning of the work of the workpackages, when data access had to be arranged first. IT facilities were ensured by subscribing to the so-called Sandbox in Ireland. This is explained further in section 3.1. Facilities were also needed for communication, in order to share and work on documents together and for virtual meetings, among other things. This is the subject of the Annex to this report.



### 1.3 Organisation

The organisational has been carried out as foreseen in the agreement of SGA-2. Each workpackage has a workpackage leader who is in charge of organising the realisation of the milestones and deliverables of the workpackage. This includes the organisation of virtual and physical meetings of the workpackage members. The results of the workpackages are described in chapter 2.

At the level of the ESSnet as a whole, the main instrument for co-ordination is the monthly virtual meeting of the workpackage leaders, including WP 0 and WP 9, supported by the secretary (Martin van Sebille) provided by WP 0. These are called the meetings of the co-ordination group, or CG meetings. Eighteen virtual meetings were held during SGA-2, eight of them taken place in the time of overlap with SGA-1.

A physical co-ordination meeting with the workpackage leaders was held in Brussels, Belgium, in October 2017, for which a report has been made available on the wiki of the ESSnet: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/b3/CG\\_Meeting\\_2017\\_10\\_26-27\\_Brussels\\_20\\_Minutes.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/b3/CG_Meeting_2017_10_26-27_Brussels_20_Minutes.pdf). A dissemination conference, called BDES 2018 (Big Data for European Statistics) was held in Sofia in May 2018 for a wider audience, in which the main results of the ESSnet were presented and discussed. Again, a separate report for the meeting is available on the wiki of the ESSnet: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/BDES\\_2018](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/BDES_2018) (see also the first annex to this report).

The aim of the virtual CG meetings and the physical co-ordination meetings was to stay in control of the realisation of SGA-2. For most CG meetings the partners were asked to provide information on the realisation of the foreseen budget in the form of a spread sheet, which was consolidated by the secretariat (WP 0), thereby enabling the CG to link the progress in producing results to the resources actually spent. The meetings were also used, of course, to discuss cross-cutting issues. In addition to the workpackage leaders, the virtual CG meetings and other co-ordination meetings were also attended by the Eurostat project manager of the ESSnet, Albrecht Wirthmann.

In order to ensure the quality of the deliverables of the ESSnet, a Review Board was created at the beginning of SGA-1. The Review Board continued during SGA-2. During SGA-2, the members were Lilli Japac (chair), Anders Holmberg and Faiz Alsuhail. All workpackage leaders have arranged that their deliverables were reviewed by the Review Board. This has worked well, both in practical terms (planning etc.) as in terms of contents (usefulness of the reviews): all deliverables of SGA-2 have been reviewed and the comments have been taken into account. The members of the Review Board are also invited to the CG meetings.

The organisational arrangements of the ESSnet are considered to be quite adequate. They were the same for SGA-1 and SGA-2. At the end of SGA-2 an internal evaluation was carried out. The conclusions were generally positive. They are summarised in Annex 2.

## **2. Results of the work packages**

### **2.1 Webscraping / Job Vacancies**

The aim of this pilot is to demonstrate by concrete estimates which approaches (techniques, methodology etc.) are most suitable to produce statistical estimates in the domain of job vacancies and under which conditions these approaches can be used in the ESS.

The workpackage produced two reports. The first one describes the strategy for ongoing engagement. Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e0/WP1\\_SGA2\\_Deliverable\\_1\\_1.0docx.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e0/WP1_SGA2_Deliverable_1_1.0docx.pdf)

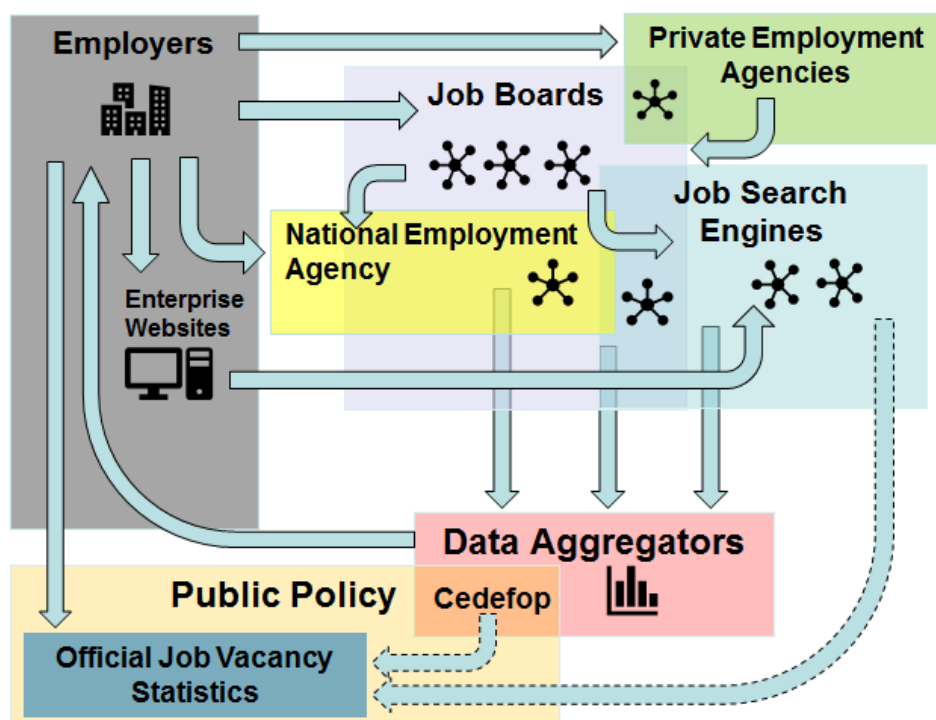
The second one is the final technical report, including a roadmap for moving experimental research into production. Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5e/SGA2\\_WP1\\_Deliverable\\_2\\_2\\_main\\_report\\_with\\_annexes\\_final.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5e/SGA2_WP1_Deliverable_2_2_main_report_with_annexes_final.pdf)

#### **1. Data Access**

A key aim of the pilot was to better understand the OJV data landscape and how it operates within different countries. The landscape is complex consisting of multiple actors with different and interlocking roles (See Figure 1). Employers have various options for advertising job vacancies including posting directly on job boards, advertising vacancies on their own website, or advertising through private employment agencies. Job search engines gather key details of job vacancies and then effectively advertise on behalf of job boards. National Employment Agencies have a pivotal role in matching job seekers with employers and typically have their own job portals. There are also data aggregators who collect data from various parts of this data eco-system, provide added value services and then sell this information back to employers. The role of NSIs in estimating job vacancies to support public policy making both now, and in the future, can be seen as part of this data landscape.

**Figure 1: On-line Job Vacancy data landscape**



The complexity of this data eco-system means that there are numerous routes into accessing OJV data. Broadly, these fall into two categories, direct web scraping (for either job portals or enterprise websites) or arranged access (e.g. with job portals, the National Employment Agency, or data aggregators). All these options for acquiring data have been explored by this pilot to varying degrees. It had been the intention during SGA-2 to focus on applying the techniques developed by WP 2 to capture information about job ads from enterprise websites. However, early on in SGA-2 it became clear these methods were not yet sufficiently well developed to capture information about individual job ads and so this work has for the most not been carried forward.

A key conclusion was that while there may be good reasons for direct web scraping, future efforts should generally be focused on gaining access to data that has already been collected. Apart from the technical and legal issues of web scraping and associated maintenance costs, there is also the issue that it takes time to generate a sufficient time series to properly evaluate the quality of the data. Acquiring data directly from data owners should mostly circumvent these problems.

The pilot has also established a close working relationship with the European Centre for the Development of Vocational Training (CEDEFOP) who are developing a web scraping system for all EU member states. There is broad agreement that this project should also aim to help serve the long-term OJV data needs of the ESS. Therefore, a key conclusion is that NSIs should also avoid investing heavily in developing web scraping and data cleaning approaches as OJV data is expected to become widely available to EU member states via CEDEFOP by the end of 2020.

## 2. Data Processing

The nature of the OJV data landscape and the lack of any definitive source of OJV data has important implications for data quality and the type of processing that is needed to make it usable for analysis. One such issue is duplication, since a single job vacancy is typically advertised multiple times. Some investigations were made into duplication methods, although these were not explored in too much depth due to the expectation that this problem will be addressed by the CEDEFOP web scraping project.

There was a major effort during SGA-2 to investigate methods for classifying job ads by both occupation and industry sector of the employer. This is a key step in adding value to OJV data.

This research involved processing of both semi-structured and unstructured text data (i.e. the full text of the job advertisement) and then applying machine learning techniques using manually labelled data sets as inputs. The aim is to test how well the text of each job advertisement can be used to predict the industry of the employer or occupation sector of each job ad. The results are then compared with different but similarly labelled test data to measure their specificity/precision and recall/sensitivity.

An example of one of these analyses is shown in Figure 2. This relates to the classification of industry sector (i.e. NACE) based on the full text description of Belgian job advertisements. The results give a mixed picture, with some sectors giving much more accurate results than others. However, there is scope for improving these results by using more training data and incorporating other variables, such as the company name, where this is contained in the advertisement. One interesting finding by German colleagues was the observation that the detailed information on job ads collected by the Federal Employment Agency was potentially a very useful source of training data that could be used to classify data from other job portals.

**Figure 2: Confusion Matrix for NACE coding experiment using Belgian job ads**

|                |   | NACE code by Machine Learning |     |     |      |     |      |      |      |      |     |     |      |      |      |      |      |     |     |      |     | Sensitivity |
|----------------|---|-------------------------------|-----|-----|------|-----|------|------|------|------|-----|-----|------|------|------|------|------|-----|-----|------|-----|-------------|
|                |   | A                             | C   | D   | E    | F   | G    | H    | I    | J    | K   | L   | M    | N    | O    | P    | Q    | R   | S   | T    | U   |             |
| Real NACE code | A | 3                             | 0   | 0   | 0    | 1   | 16   | 0    | 0    | 0    | 0   | 1   | 0    | 0    | 0    | 1    | 0    | 0   | 0   | 0    | 14% |             |
|                | C | 0                             | 469 | 1   | 0    | 41  | 287  | 2    | 27   | 16   | 0   | 1   | 44   | 18   | 14   | 2    | 22   | 0   | 0   | 0    | 50% |             |
|                | D | 0                             | 0   | 219 | 0    | 1   | 16   | 0    | 0    | 6    | 0   | 7   | 5    | 7    | 0    | 3    | 0    | 2   | 0   | 0    | 82% |             |
|                | E | 0                             | 0   | 0   | 16   | 3   | 9    | 0    | 1    | 1    | 0   | 0   | 2    | 3    | 8    | 2    | 0    | 0   | 0   | 0    | 36% |             |
|                | F | 0                             | 4   | 0   | 0    | 671 | 200  | 7    | 21   | 6    | 1   | 6   | 45   | 24   | 42   | 3    | 24   | 0   | 3   | 0    | 63% |             |
|                | G | 0                             | 26  | 0   | 0    | 39  | 4320 | 29   | 120  | 85   | 1   | 2   | 118  | 37   | 29   | 10   | 44   | 1   | 18  | 0    | 89% |             |
|                | H | 0                             | 4   | 0   | 0    | 11  | 213  | 1499 | 3    | 18   | 0   | 1   | 25   | 5    | 57   | 0    | 89   | 0   | 4   | 0    | 78% |             |
|                | I | 0                             | 10  | 2   | 0    | 6   | 355  | 1    | 1879 | 3    | 0   | 1   | 23   | 19   | 40   | 3    | 37   | 2   | 4   | 0    | 79% |             |
|                | J | 0                             | 14  | 2   | 0    | 14  | 386  | 6    | 11   | 1096 | 1   | 3   | 152  | 40   | 61   | 4    | 28   | 0   | 5   | 0    | 60% |             |
|                | K | 0                             | 0   | 0   | 0    | 1   | 138  | 0    | 1    | 23   | 524 | 8   | 73   | 22   | 24   | 1    | 5    | 0   | 1   | 0    | 64% |             |
|                | L | 0                             | 2   | 0   | 0    | 12  | 118  | 1    | 21   | 4    | 1   | 249 | 48   | 20   | 32   | 8    | 69   | 1   | 6   | 0    | 42% |             |
|                | M | 0                             | 44  | 0   | 0    | 45  | 876  | 10   | 38   | 134  | 29  | 10  | 2847 | 130  | 103  | 14   | 106  | 2   | 14  | 0    | 65% |             |
|                | N | 0                             | 18  | 1   | 0    | 49  | 699  | 14   | 35   | 106  | 8   | 8   | 313  | 2116 | 108  | 12   | 107  | 2   | 6   | 0    | 59% |             |
|                | O | 0                             | 4   | 1   | 0    | 5   | 51   | 4    | 12   | 6    | 2   | 12  | 33   | 20   | 6950 | 362  | 156  | 1   | 6   | 0    | 91% |             |
|                | P | 0                             | 3   | 0   | 0    | 7   | 59   | 2    | 10   | 7    | 0   | 1   | 57   | 14   | 295  | 5141 | 276  | 8   | 21  | 0    | 87% |             |
|                | Q | 0                             | 8   | 0   | 0    | 10  | 167  | 9    | 56   | 12   | 0   | 4   | 42   | 60   | 375  | 187  | 5499 | 6   | 45  | 0    | 85% |             |
|                | R | 0                             | 3   | 0   | 0    | 2   | 103  | 0    | 7    | 9    | 0   | 2   | 22   | 7    | 91   | 26   | 158  | 213 | 15  | 0    | 32% |             |
|                | S | 0                             | 6   | 0   | 0    | 3   | 193  | 3    | 16   | 13   | 1   | 8   | 86   | 58   | 232  | 39   | 389  | 16  | 601 | 0    | 36% |             |
|                | T | 0                             | 0   | 0   | 0    | 0   | 0    | 0    | 0    | 0    | 0   | 0   | 0    | 4    | 1    | 0    | 1    | 0   | 0   | 2    | 25% |             |
| U              | 0 | 0                             | 0   | 0   | 0    | 12  | 3    | 2    | 4    | 0    | 0   | 4   | 2    | 5    | 0    | 2    | 1    | 2   | 0   | 20   | 35% |             |
| Specificity    |   | 100%                          | 76% | 97% | 100% | 73% | 53%  | 94%  | 83%  | 71%  | 92% | 79% | 72%  | 81%  | 82%  | 88%  | 78%  | 84% | 80% | 100% | 95% |             |

Another area of data handling processing was to account for definitional differences between a job vacancy as defined by the JVS and a live job advertisement. These are not the same because an employer may continue to take active steps to fill a vacancy after a job advertisement has closed. In the case of Slovenia, jobs advertised by the Slovenian Employment agency could be linked to administrative data on when those jobs are actually filled. These data are used to model the OJV data so that it more closely aligns with JVS definitions.

### **3. Methodology**

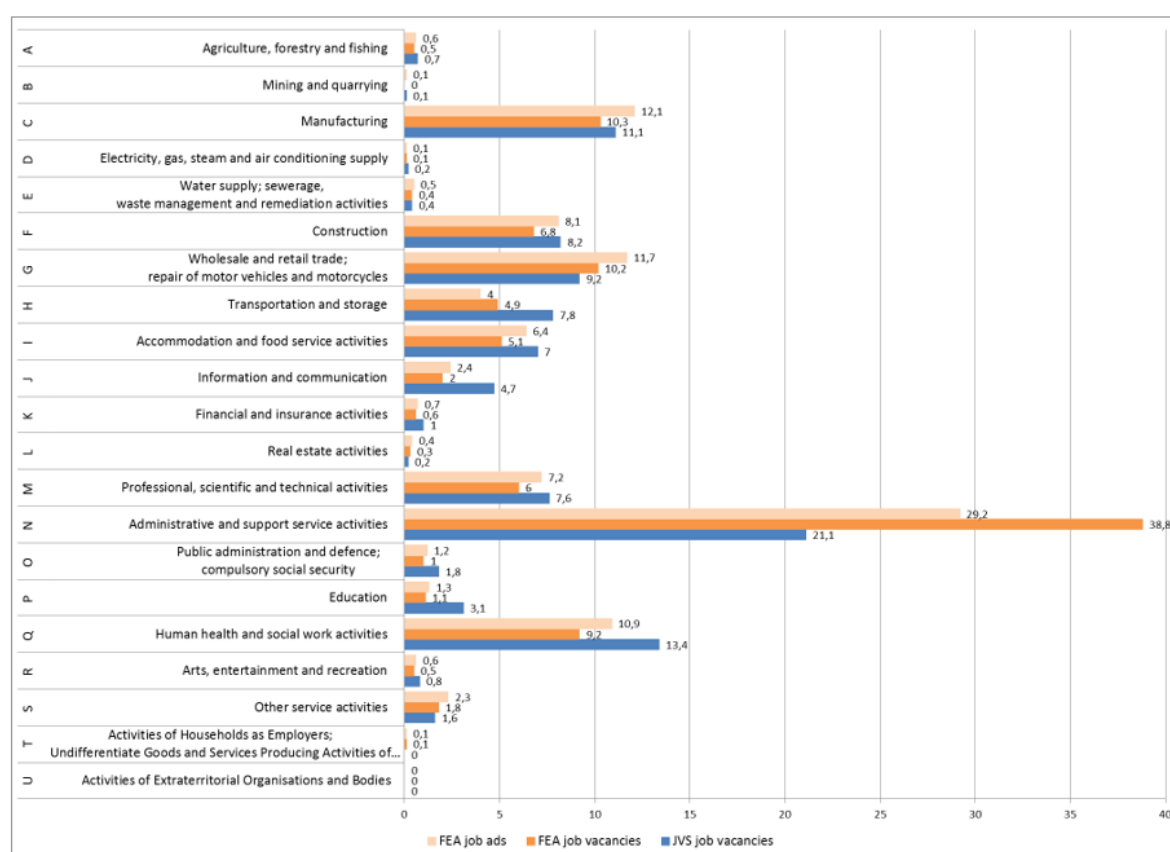
A key area of methodological development was around the matching and linking of statistical units. This is partly related to the problem of duplication and indeed deduplication can be viewed as a special type of matching problem. The more specific application related to the matching of JVS reporting units. This was required to support validation activities. This matching proved to be challenging. Particularly for large enterprises due to complex organisational structures.

Considerable work was completed by WP 1 into assessing and validation different aspects of data quality. Various aspects were explored including:

- Measurement of coverage through additional questions to employers in the JVS about use of recruitment channels.
- Comparisons of the distribution of job vacancies by industry between the OJV data sources and the JVS.
- Time series comparisons of OJV and JVS data both in terms of aggregate data and in terms of vacancy counts for individual enterprises.

Figure 3 shows the distribution of job vacancies by industry sector (NACE) estimated by the German JVS against both the distribution of job advertisements and the actual number of underlying vacancies from the German Federal Employment Agency (FEA). This is a useful exploration of an important definitional issue which is that some job ads may contain more than one vacancy and this is often not explicit in the job advertisement itself. Some sectors, such as manufacturing, compare favourably. Larger differences in the administrative and support services sector may be partially explained by the FEA data being classified by the economic activity rather than the activity of the employing business. However, the differences between the FEA data on job ads versus actual vacancies, show that this definitional difference has a material impact on these distributions.

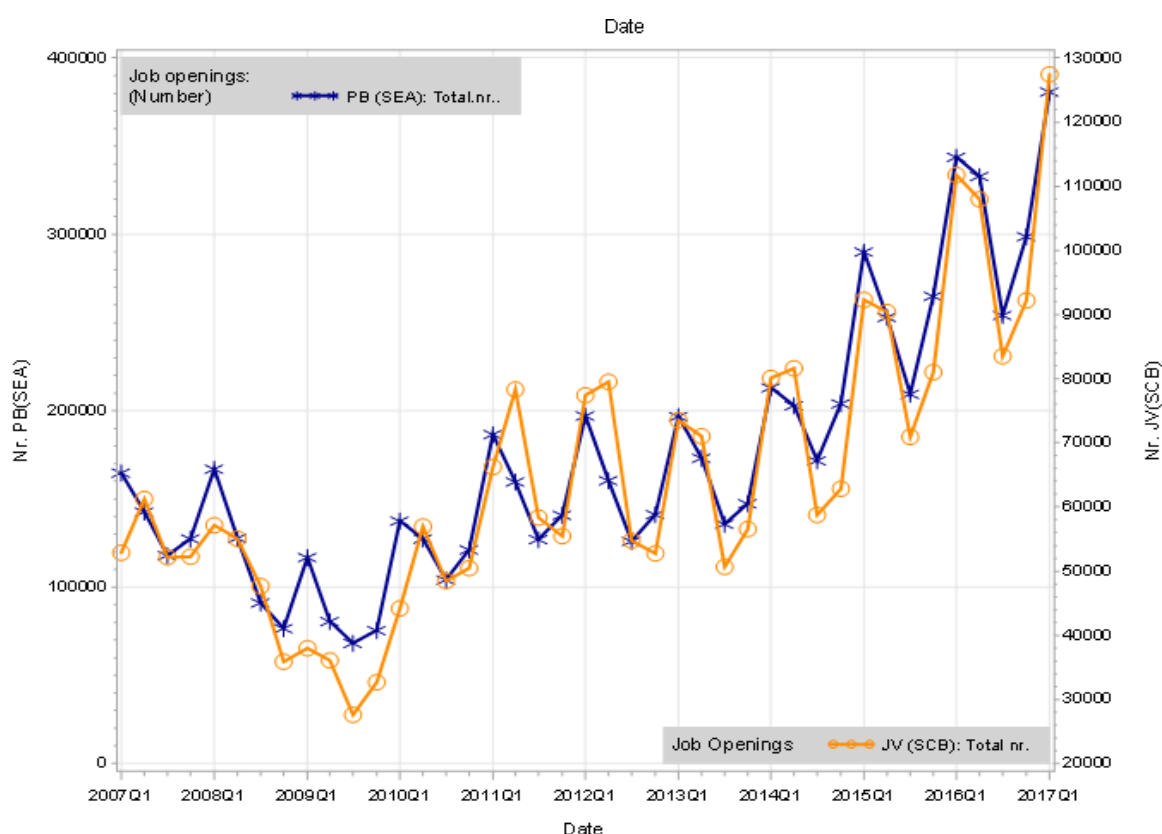
**Figure 3: FEA-OJV versus JVS: job ads and job vacancies by industry sector (NACE)**



Source: Own calculations on FEA data and special analysis on JVS data by IAB.

Sweden has a continuous source of on-line vacancy data from their NES, Platsbaken (PB), going back to 2007. While, the overall levels of job openings are much higher than the Swedish JVS, a standardized view of the data shows that the data has similar time series properties, including a very distinct seasonal pattern. Disaggregation of these series into public and private sector jobs shows greater trend correspondence for the public sector (especially in recent years) with a correspondingly weaker pattern for the private sector. Although this type of analysis shows a relationship between OJV data and existing statistics, these relationships both vary for different types of job and vary across time.

**Figure 4: Job openings from Swedish National Employment Agency (PB) and JVS (2007-2017)**



The work package has attempted to assess OJV data using data quality frameworks, including the proposed UNECE framework for the quality of big data<sup>1</sup>. The main quality issues for OJV data can be summarized as follows:

- Not all job vacancies are advertised on-line and some types of jobs are more likely to be advertised than others.
- There is no definitive source of OJV data. It is generated and managed by various and mostly commercial actors.
- Data about on-line job ads usually contain a mix of structured and non-structured elements, but the specific structure and variables may vary between sources.
- Some job ads are out of scope of the official definition of a job vacancy (e.g. student internships, international jobs, “ghost” vacancies).
- The official definition of a job vacancy does not correspond directly to the concept of a live job ad. Critically, a vacancy will usually persist after the advertisement closes.
- The specific OJV data landscape varies considerably between countries, for example, in terms of the number and type of portals and use of on-line platforms. There may also be differences in terms of the role of the National Employment Agency and what type of information is contained in job ads. There may also be legal difference and finally, processing will often require language specific solutions.

<sup>1</sup> <http://www1.unece.org/stat/platform/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2>

In summary, OJV data is not representative of the overall labour market and there are various definitional issues that make it difficult to compare directly with official statistics.

#### 4. Statistical Outputs

The lack of any definitive source of OJV data and the wide range of quality and definitional issues meant that most countries participating within the ESSnet did not get as far as producing concrete estimates. Slovenia went furthest in developing an approach for producing experimental statistics based on OJV data (Table 1). These figures are the fully deduplicated detected job ads from the National Employment Agency, the two largest job portals and enterprise websites. The “detected job ads for the quarter” include data from before the reference period where a distribution has been applied to adjust for unfilled jobs after the jobs have closed. A comparison with the official job vacancy estimates show that only about 40% of all Slovenian job vacancies can be found on-line.

**Table 1: Experimental on-line job vacancy statistics for Slovenia**

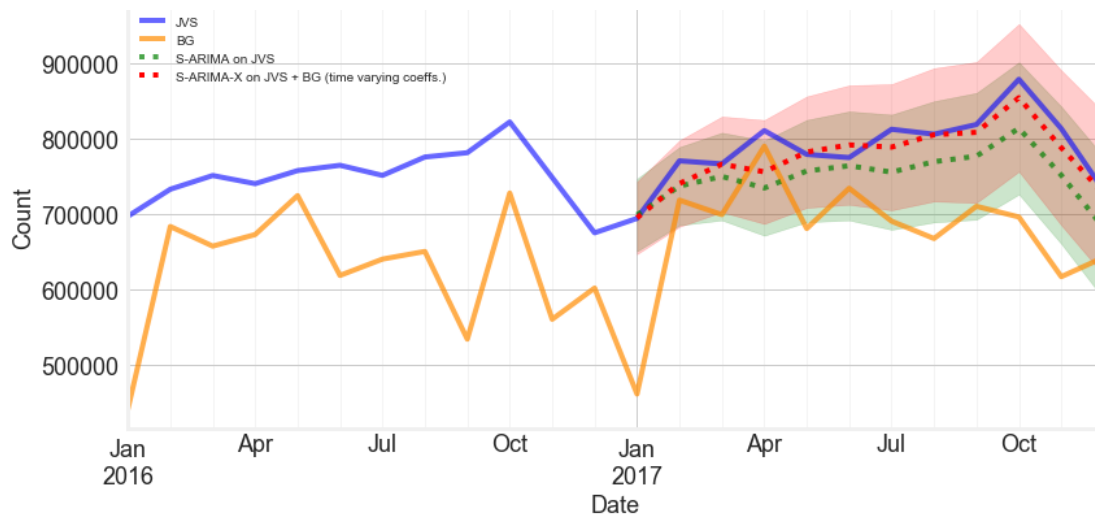
| Estimates                          | Type         | 28 August 2017 (Q3) | 30 November 2017 (Q4) |
|------------------------------------|--------------|---------------------|-----------------------|
| Detected job ads for quarter       | Stock        | 6849                | 6327                  |
| <i>Official JVS estimate</i>       | <i>Stock</i> | <i>17221</i>        | <i>15243</i>          |
| Available in reference month       | Stock        | 3542                | 4493                  |
| Available on reference day         | Stock        | 1368                | 1285                  |
| Newly available on reference month | Flow         | 1984                | 2115                  |
| Newly available on reference day   | Flow         | 123                 | 76                    |

Despite the large difference with the official estimate, the greater frequency of OJV data offers the possibility of variations to these statistics. These include job ads that are available (i.e. advertised) both during the reference month and reference day, as well as new ads for both the reference month and the reference day.

The UK explored the possibility of using the real-time availability of OJV data to produce nowcasts of job vacancy statistics. Initial work focused on developing models for individual enterprises. This found that the relationship between the time series of OJV counts and the reported JVS figures were often highly variable and disjointed with only a loose relationship between them. However, an approach combining aggregate OJV and JVS data gave more promising results (Figure 5). This shows a simple baseline persistence model using only the JVS and then the impact of introducing OJV data with time varying coefficients. The inclusion of the OJV data made a noticeable improvement to the model and so this seems a promising area worthy of further exploration.



**Figure 5: UK nowcasts using an S-ARIMA-X time series model.**



### Concluding Remarks

The results from Slovenia raise some fundamental questions about the OJV data and whether and how it should be used for official statistics. While this has shown that it is feasible to produce estimates of on-line job vacancies, there is a big difference between these and the official job vacancy estimates. It is therefore clear that these could never replace the official estimates. Further, the use of these estimates to support policy making need to be considered carefully as they only give a partial (and not easily defined) view of overall labour market demand.

It therefore seems that the role of OJV data within official statistics is more likely to be as the basis for producing supplementary indicators. These could include indicators of local labour market demand, occupation groups and associated skills. However, rather than measuring absolute levels, these would be more useful for measuring change over time. Using OJV data for nowcasting purposes is another promising application. One also cannot yet rule out the possibility of using OJV data in conjunction with the JVS to reduce the frequency of the survey or reduce the size of survey samples and thereby reduce sampling costs. Another possibility could be using these data for imputing non-response in the JVS.

## 2.2 Webscraping / Enterprise Characteristics

The purpose of this workpackage is to investigate whether web scraping, text mining and inference techniques can be used to collect, process and improve general information about enterprises.

Following up the work done within SGA-1, SGA-2 work was articulated along the following main directions:

- SGA-1 work had in scope 4 use cases (URLs retrieval, e-commerce/web sales, social media detection, job advertisement detection). SGA-2 work considered two additional use cases, namely: NACE detection and Sustainable Development Goals (SDGs) detection.
- Pilots enhancements: all the pilots implemented in the first phase have been enhanced by (i) consolidating adopted techniques and (ii) extending the number of enterprises to which the scraping activity was targeted.
- New pilot development: new pilots were implemented not only for new use cases but also for SGA-1 use cases for which additional countries committed to develop pilots.
- Final methodological and technological conclusions based on all the work done within the ESSnet.
- Production of output indicators and comparison with related survey statistics.

The following sections describe the principal obtained results as reported in the deliverable 2.4 (see link [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/ee/Wp2\\_Del2\\_4.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/ee/Wp2_Del2_4.pdf)).

### 1. Reference framework, Pilots and Software solutions

The whole set of identified use cases is:

1. Enterprise URLs Inventory. This use case is about the generation of a URL inventory of enterprises for the Business register.
2. E-Commerce on enterprises' websites. This use case is about predicting whether an enterprise provides or not web sales facilities on its website.
3. Job advertisements on enterprises' websites. This use case is about investigating how enterprises use their websites to handle the job ads.
4. Social Media Presence on enterprises' webpages, aimed at providing information on existence of enterprises in social media.
5. Sustainability reporting on enterprises' websites. One of the Sustainability Development Goals set up by the UN is to encourage enterprises to produce regular sustainability reports highlighting sustainability actions which the business has taken. In order to measure companies' response to this, this use case will look at what companies publish on their official website and track changes over time.
6. Identifying categories relevant to enterprises' types of activity (NACE). Aimed at identifying relevant categories of enterprises' activity sector from enterprises' web sites to check or complete Business registers.

In order to guide the piloting activities to implement use cases, a generic reference logical architecture was designed, consisting of several building blocks organized into four main layers (see Figure 1), namely:

- « Internet access »,
- « Storage »,
- « Data preparation » and
- « Analysis ».

Each layer consists of building blocks that have been implemented by specific software solutions realized within the piloting activity.

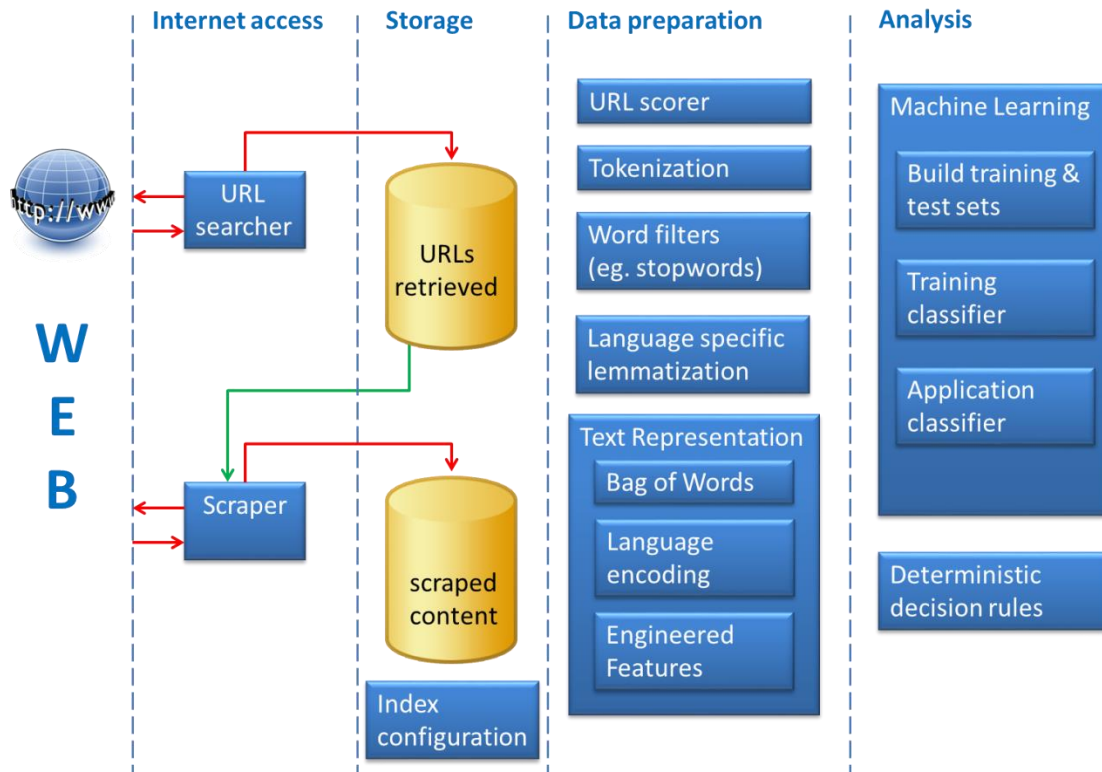


Figure 1: WP2 Logical Reference Architecture

A total number of **24 different pilots** were implemented, namely: **6** for use case 1 (with Bulgaria implementing two pilots with two different technologies), **6** for use case 2 (with Bulgaria implementing two pilots with two different technologies), **3** for use case 3, **5** for use case 4, **3** for use case 5 and **1** for use case 6.

Within the piloting activity, some generalized software solutions were implemented and are available at: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2\\_Links](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_Links)

## 2. Methodological, Quality and IT Recommendations

From the piloting work, some conclusions have been drawn and recommendations from the methodological, quality and IT perspectives derived.

### *Conclusions and recommendations on methodology and quality*

The conclusions and recommendations on methodology can be summarized as answers to the following three main questions.

- “What can we learn methodologically from the four pilots executed in the field of web scraping of enterprises web sites in six countries?”

One thing learned is that it is useful and feasible to apply web scraping techniques in the field of official statistics to compute experimental indicators. However, there is not one preferred way of doing these very different pilots in different countries. Even per pilot the methods being used differ, which may have been caused by different data landscapes per country or other circumstantial differences. However some common machine learning methods have been applied in some of the pilots, especially the URL finding pilot where all participant country applied (almost) the same methodology. In addition, some practical lessons were learned on building training sets and on evaluating the quality of the results.

- “How easy is it to compare the approaches taken in different circumstances and what are the general underlying principles?”

Using the same terminology and describing the work being done in general building blocks has been a big advantage to compare the different approaches, not only from an IT viewpoint, but from a methodological viewpoint as well. The concept of generic versus specific scraping and deterministic versus machine learning approaches (and within machine learning a wide spectrum of different classifiers and text representation methods) form a useful general underlying basis for scraping for official statistics.

- “Can we identify some methodological best practices, common problems and solutions from the pilots that were executed?”

All of the approaches used in the pilots in the different countries produced results that are quite good. Both deterministic as well as machine learning approaches have been successfully applied. In the latter case some practical lesson learned have been described.

### *Conclusions and recommendations on quality*

With respect to quality, it was observed that predicted values can be used for a twofold purpose:

1. **At unit level**, to enrich the information contained in the register of the population of interest. The quality of data pertaining the unit level can be measured by considering the same indicators used to evaluate the trained model. If the test set is representative of the whole population and not used for training the model, the performance measures (like accuracy and F1-score) calculated for the test set can be considered a good estimate for the overall performance. Another way to train a model and measure the performance is to use n-fold cross validation. Assuming that the train/test set is

representative for the whole population, the average performance of the model against the train/test set is a good estimate for the overall performance of the model. Table 1 reports the qualitative judgement on the quality at unit level. For the detailed information, please refer to the methodological notes.

| Use Case             | Quality  |
|----------------------|--|
| 1: URLs Retrieval    | Results are good (accuracy/F1 measure)<br>Coverage: good for Istat, NL, BG, PL, UK partially good<br>Report UK diagram as an example |
| 2: Ecommerce         | Results are good (accuracy/F1 measure) for all participants, partially good for UK   |
| 3: Job advertisement | Results are good (accuracy/F1 measure) for all participants  |
| 4: Social media      | Results are very good (accuracy/F1 measure) for all on social media presence<br>Partially good on social media usage                 |
| 5: NACE              | Results (accuracy/F1 measure) are partially good for all participants  |
| 6: SDG               | Results (accuracy/F1 measure) are partially good   |

Table 1 General quality evaluation per use case

2. **At population level**, to produce estimates. The quality estimation is much more complex comparing with the unit level. Italy proposed a strategy applied in the pilots and is summarized in Table 2.

| Estimator                     | Formula   | Weighting  | Description   |
|-------------------------------|---|--|---|
| Design based / model assisted | $\hat{Y} = \sum_r y_k w_k$  | $\sum_{k=1}^r w_k = N_U$   | $w_k$ weights are obtained by calibration procedure of basic weights (inverse of inclusion probabilities) making use of known totals in the population in order to reduce the bias due to non-response and the variability due to sampling errors   |
| Model based                   | $\hat{Y} = \sum_{U^2} \tilde{y}_k w'_k$   | $\sum_{k=1}^{U^2} w'_k = N_{U^2}$  | The estimate of the total number of enterprises offering web ordering facilities on their websites is given by the count of the predicted values $\tilde{y}_k$ for all units for which it was possible reach their websites (population $U^2$ ), calibrated in order to make them representative of all the population having websites ( $U^1$ ).   |
| Combined                      | $\hat{Y} = \sum_{(U^2-r^1)} \tilde{y}_k + \sum_{r^1} (\tilde{y}_k - y_k) w''_k + \sum_{(r^2)} y_k w'''_k$ | $\sum_{k=1}^{r^1} w''_k = N_{U^2}$<br>and<br>$\sum_{k=1}^{r^2} w'''_k = N_{U^1 - U^2}$ | Estimates are produced by summing three components:<br>1. the counting of predicted values in the subpopulation $U^2$ of units for which it was possible to scrape and process corresponding websites;<br>2. an adjustment based on the consideration of the differences between the $r^1$ reported values and the predicted values (expanded to the same subpopulation $U^2$ );<br>3. the counting of observed values for the $r^2$ respondents that declared a website, that was not found nor scraped, expanded to the whole subpopulation $U^1 - U^2$ . |

Table 2: Estimators

Three Estimators are proposed, namely:

- (i) Design-based/model-assisted;
- (ii) Model based, the main flaws of the model based estimator are in the presence of
  - prediction errors;
  - undercoverage of the population of enterprises owning websites, part of which has not been reached by web scraping.
- (iii) Combined. Looking at coherence as one important dimension of quality, both combined estimates and full model based estimates can be considered as equally acceptable. But two considerations can be made:
  - The second component of the combined estimator is based on an assumption of perfect correctness of reported values, and considers predicted values as errors when they do not coincide with the reported ones. But controls have been carried out when fitting models, and in half of the cases in which predicted values were contradictory with reported ones, this was not due to model fault, but to response errors. So, this assumption does not always hold. In any case it would be advisable to deepen this phase also by returning to the respondents to verify if it is an error in response or if, for example, the model has evaluated the content of a site different from that one considered by the respondent.
  - If a medium-term aim is to make multi-annual frequency of the questions in the survey related to the websites characteristics, then the combined estimator cannot be applied, as it relies on the current availability of reported values from the survey, and the full model based estimators remains the only alternative. In this case, there would be an issue in time series analysis due to problems in comparability between survey estimates and model based ones.

The work done so far could be extended in multiple ways. In particular, if the different use cases are considered:

1. The e-commerce detection algorithms could be refined to distinguish between different levels of e-commerce maturity (for instance, determined by the presence of only an ordering facility, or also payment and deliver tracking ones).
2. The job advertisement spiders could be trained to additionally take the job details and the enterprise characteristics into consideration. The identification of the characteristics of each single job (economic activity, profession) and even the skills required, is a much more ambitious task that implies a different approach, more oriented to “information retrieval” than to “machine learning”.
3. The social media presence detection could be extended to not only observe the enterprise website, but also inspect the social media itself in order to investigate what kind of use of social media is being done in a more detailed way. This is of course subject to ethical and legal considerations.

#### *Conclusions and recommendations on IT techniques*

Several different technologies and programming languages have been used for pilots' implementation.

Python, Java and PHP are the main web scraping languages. R and Python are used for data analysis. Both NoSQL (Apache Solr and MongoDB) and relational databases (MySQL, MS SQL) are used for data storage; in some cases, the data are stored in CSV files. The most convenient programming language can be chosen depending on the in-house competence.

However, libraries for different programming languages provide possibility for parameters setting. Such as scikit-learn of Python, have options for algorithm “tuning”. The varieties of parameters setting can generate different results on the same dataset. This should be interpreted with care. More efforts need to be made to investigate the parameters’ setting and the result comparison.

The differences in methods are those causing the main differences in results, the impact of the specific technology is quite limited.

### 3. Output results as experimental statistics and methodological notes

WP 2 produced some output statistics resulting from the piloting activity. Such statistics are to be published as experimental statistics within the project wiki, alongside comparisons against survey estimates (except for the ‘rate of retrieved URLs’ use-case, where there is no relevant estimate available). Methodological notes will accompany each of these estimates.

The following reports on two relevant examples of the output indicators produced.

The first one is reported in Table 3 and is related to the rate of enterprises engaged in websales on their website like resulting from the piloting activity carried out by the different countries for use case 2. Please notice that the NL result is based on a keyword/feature engineering approach for text representation that could be affected by errors. This could explain at least part of the difference that is reported.

*Table 3: Rate of enterprises engaged in websales on their website*

|                    | Estimate from web data | Survey Estimate                   |
|--------------------|------------------------|-----------------------------------|
| <b>Italy</b>       | 16%                    | 15%                               |
| <b>UK</b>          | 30%                    | 21% (unweighted ICT survey 2015)  |
| <b>Netherlands</b> | 14%                    | 33%                               |
| <b>Bulgaria</b>    | 6,2 %                  | 8,6% (unweighted ICT survey 2017) |

The second example of output indicators is the rate of enterprises that are present on social media. Looking at the Table 4, one can observe that figures related to estimates from web data and survey are quite aligned.

Table 4: Rate of enterprises that are present on social media

|                                   | <b>Estimate from web data</b>    | <b>Survey Estimate</b>           |
|-----------------------------------|----------------------------------|----------------------------------|
| <b>Italy</b>                      | 37%                              | 31%                              |
| <b>UK</b>                         | 80%                              | 66% (unweighted ICT survey 2015) |
| <b>Netherlands</b>                | 65%                              | 69%                              |
| <b>Bulgaria</b> (BNSI software)   | 31%                              | 34%                              |
| <b>Bulgaria</b> (Polish software) | 37%                              | 34%                              |
| <b>Poland</b>                     | 26% (Pomeranian voivodship only) | 25% (Pomeranian voivodship only) |



## 2.3 Smart Meters

The aim of this pilot was to demonstrate by concrete estimates whether buildings equipped with smart meters can be used to produce energy statistics but can also be relevant as a supplement for other statistics e.g. census housing statistics, household costs, impact on environment, statistics about energy production. A smart meter is usually an electronic device that records consumption of electric energy in intervals of an hour or less and communicates that information at least daily back to the utility for monitoring and billing. Challenges ahead with this dataset are: representativity issues, linking to other datasets, privacy concerns. Another challenge with smart meters data is that these are currently available in a few countries only, but will be available in several countries before 2020. Second aim of this workpackage was to relate the results of this pilot to future use in other countries.

Members from six national statistical offices carried out the pilot: Statistics Austria, Statistics Denmark, Statistics Estonia, Statistics Italy, Statistics Portugal, and Statistics Sweden. Statistics Estonia was responsible for coordinating the work in this workpackage.

During the pilot study the following tasks were carried out:

### Task 4. Future perspectives.

Potential usages of smart meters data. Based on the data obtained from the previous phases of the project, other potential statistical products in the domain of energy consumption or in other statistical domains are suggested (e.g. classification of users by their consumption patterns, studying relationship between consumption and different economic indicators).

Other smart meters. In addition to the electricity smart meters, the potential usages for different kinds of smart meters (e.g. natural gas, water) are proposed. No data of other smart meters were available for this pilot, so the discussion is made from theoretical point only.

Feasibility of the use of on different level aggregated data. Overview of the different ways to aggregate the raw data and produce statistics from different level of aggregated data (e.g. yearly, monthly, hourly) is given.

### Task 5. Recommendations

Based on the experiences acquired while carrying out Tasks 1-4, a list of recommendations regarding access, IT-infrastructure, methodology, data processing, potential statistical outputs and output quality were given. The aim is to give recommendations and lessons learned to other countries that could help them to start using smart meters data for the production of statistics.

Based on the results of the two tasks mentioned, two reports were delivered. The first report covered the results of Tasks 4 (Deliverable 3.5). Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3\\_Report\\_3](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Report_3)

The second report covered the results of Task 5 (Deliverable 3.6). Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3\\_Report\\_4](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Report_4)

This pilot study is not linked directly with other pilots in this project, but one can find similar issues with other pilots like the methods, IT-technologies and quality issues. Methodological, quality and technical results of the workpackage, including intermediate findings, have been used as inputs for WP 8 of SGA-2. When carrying out the tasks listed below, care has been taken that these results are stored for later use and communicated to broader audience, by using the facilities described at WP 9.

## 1. Potential usages of smart meters data

The aim of this task was to use linked data obtained from previous tasks and find potential usages for this data. The electricity data has a great potential as a complementary data source for different tasks. The data could be related to several social and economic areas so that it would provide new insight to a phenomenon. Based on this intention the workpackage evaluated different categorization methods and used classification methods to get better insight to behaviour and properties of electricity consumers. Potential of the data was also used to reveal the dynamics of energy use by economic activity and regional consumption patterns.

For categorization and classification machine learning techniques were used. Those methods are classified in broader sense as:

Supervised learning methods - when the training data contain labels (e.g., the corresponding household size is known for a metering point)

Unsupervised learning methods - when the training data do not contain labels (e.g., the corresponding household size is unknown for a metering point and must be identified from the data)

When high quality dependent variables available, which describe households, the supervised learning approach can be used. The unsupervised methods is used if the cluster structure of the data has to be discovered from the data.

In the report a number of methods were tested: Logistic regression, Boosted logistic regression, k nearest neighbour, Bayesian Generalized Linear Model, Support vector machine, Random forest, kmeans and hierarchical clustering.

As an example, a random forest model estimation for the nominal variable urban/rural about 77 percent of the smart meters correctly (Table 1). Clustering methods were used to discriminate between household and business consumption.

Table 1. The random forest model for urban/rural categorizing of households.

| 76.9 % | Rural  | Urban  |
|--------|--------|--------|
| Rural  | 16.5 % | 8.8 %  |
| Urban  | 14.3 % | 60.3 % |

One of the main goal of the project was to identify vacant or seasonally vacant dwellings. On this purpose two approaches can be used. One is to use electricity consumption data and identify zero or close to zero consumption at a certain period of time, another is to apply classification methods to electricity data. By using zero or low consumption data empty dwellings can be identified (e.g., Figure 1). But classification is needed if there is higher than zero consumption due to some device (e.g., heating by electricity) or due to steady consumption it is difficult to discriminate whether the

dwelling is empty or not. Depending on the aggregation level different goals can be achieved. By using yearly data only empty dwellings with zero consumption can be identified and dwellings can be categorized as empty by stating certain threshold for consumption. By using monthly aggregated data seasonal patterns can be identified and by using hourly or more frequent data patterns of living can be identified and there are higher chances to classify households correctly.

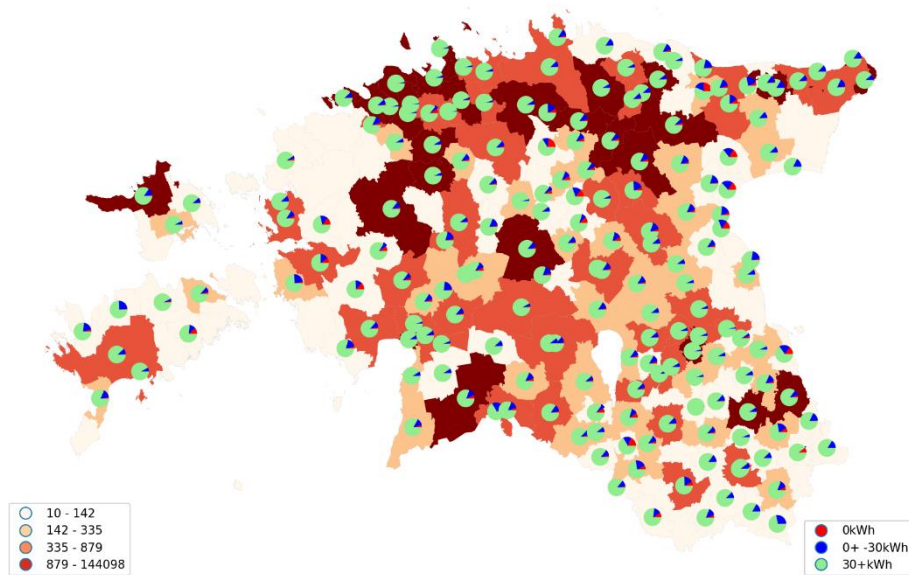


Figure 1. Distribution of low consumption apartments by local municipality. Legend (left) number of apartments, (right) pie chart of share of low consumption.

As cluster centres of dwellings monthly energy consumption were visualized (Figure 2), there was possible to see seasonal patterns, and dwellings could be identified that are used in winter- or summertime.

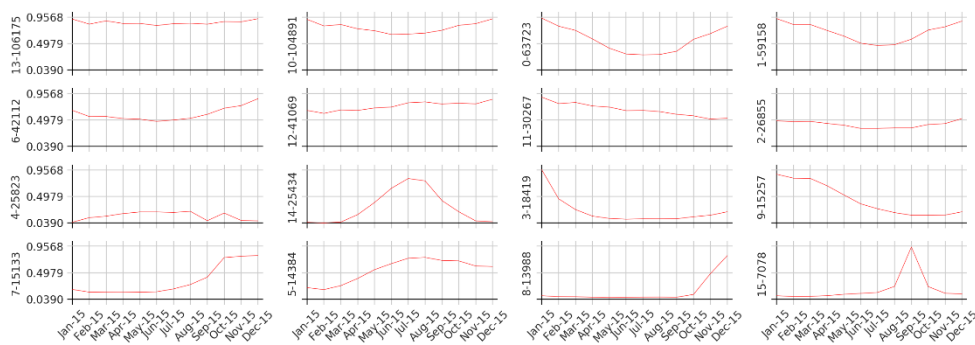


Figure 2. 16 cluster centres of normalized monthly data.

The electricity data has a great potential to produce regional statistics and to see what are the main business activates in regions and identify the main centres of different activity areas (e.g., Figure 3).

The total consumption of electricity of agriculture per municipality, given in 1000 kwh

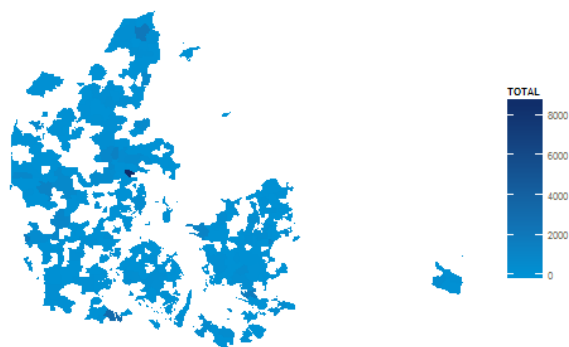


Figure 3. Regional statistics of mean annual electricity consumption in agriculture

In the report it is shown that the smart meter data could be used for identifying consumption patterns and give some background knowledge to some everyday phenomena. As energy consumption around switching time between summer- and wintertime was studied, no very clear evidence was found of saving electricity by applying daylight saving time policy (Figure 4).

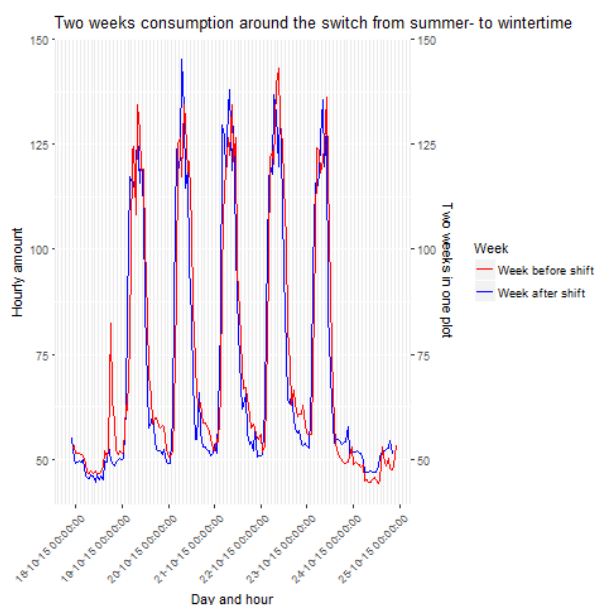


Figure 4. Two weeks consumption around the time switch from summer- to wintertime

## 2. Other smart meters

In addition to the electricity smart meters, the potential usage of other smart meters was discussed. Due to the fine granularity of the smart meter data new possibilities of data usage are possible.

There are several different smart meters in use which have potential to provide data for producing statistics or to be a supplementary data source:

- The water and gas meters can be used for identifying vacant dwellings and doing energy statistics,
- The heating meters could be used in the energy statistics,
- The weather stations information can be used as a supplementary source to evaluate impact of weather to energy consumption,
- The data of power monitors can be used for tracking energy usage, and allow customers to better understand their energy consumption and wisely choose billing rates.
- The waste bins trackers can be used for environmental statistics.

The amount of IoT (internet of things) devices is increasing rapidly and probably there are much more potential devices, which can provide data for replacing existing data sources or be a new or additional data source.

### 3. Feasibility of the use of on different level aggregated data

From the study of the aggregated data it was concluded that aggregation level determines what are the statistical products that can be produced from the smart meter data. Smart meter data is a time series of recordings of consumption of electric energy in intervals of an hour or less. Raw smart meter data contains usually a metering point id, timestamp of recordings and recording of consumption in a fixed time period. In addition there is information about the location of the smart meter and customer information who has signed contract with an energy provider. A smart meter is usually located in a building or in a part of the building, which has an address. Address information is relevant for identifying the end consumer as the contract owner may not be the actual end consumer. Aggregation level and additional metadata is relevant question when a NSI gets access to the smart metering data. Depending on what is the aggregation level of data and what kind of additional metadata about a metering point is available determines what are the possible statistical products (Table 3) and analysis that can be performed. Data can be aggregated in time scale or in space scale or when the end consumer is identified in some other scale (e.g., area of activity of businesses). Using a proper aggregation level is also relevant in processing the data as it is much efficient to produce monthly statistics from monthly level aggregated data than from the raw hourly data.

Table 3. Aggregation levels of smart metering data and possible statistical outputs

|                           | Metering point | Address             | Region  |
|---------------------------|----------------|---------------------|---------|
| Raw data (hourly or less) | 1, 2, 7        | 1, 2, 3, 4, 5, 6, 7 | 1, 6, 7 |
| Daily                     | 1, 7           | 1, 3, 4, 5, 6, 7    | 1, 6, 7 |
| Monthly                   | 1, 7           | 1, 4, 5, 6, 7       | 1, 6, 7 |
| Yearly                    | 1              | 1, 4, 5, 6          | 1, 6    |

Table 4. Aggregation levels

|                | Table Size (KB) | File Size (KB) | Processing time (s) |
|----------------|-----------------|----------------|---------------------|
| Quarter minute | 1.447.035       | 1.138.681      | 4                   |
| Hourly         | 335.872         | 265.985        | 1                   |
| Daily          | 14.336          | 10.683         | 0,178               |
| Monthly        | 448             | 311            | 0,094               |
| Yearly         | 128             | 18             | 0,089               |

The aggregation level and availability of metadata of a metering point determines what kind of products can be produced from the electricity data. The more detailed information is available, the more products can be produced, but at the same time the more storage space and computing power is needed to handle the data (Table 4). To speed up the calculations it is recommended to use temporary tables of aggregated data even in the case the raw metering data is available.

#### 4. Recommendations

The aim of Task 5 was to summarize the findings of country-specific studies and give advice on how to work with smart meter data, including access to data, methodology for processing and analysing data, IT-architecture, validating the quality of data, and possible outputs. The results of the task are published in Report 4. The report contains detailed description from acquiring access to the data until producing potential statistical outputs. Standard SQL- or R-code is provided where appropriate. The report includes detailed diagrams how the data sources can be linked to produce relevant output (Figure 5).

The main contribution of the report is to demonstrate the use of data from smart electricity meters for production of official statistics. The pilot had three goals with regard to expected outputs. First, to assess whether current survey based business statistics can be replaced by statistics produced from electricity smart meter data, second, to produce new household statistics and third, to identify vacant or seasonally vacant dwellings. A main challenge is how the observed units - metering points - can be mapped to statistical units - businesses, households and dwellings - and how existing statistics can be produced based on the results. The methodology used depends on what kind of data are available for the NSI and the aggregation level (on the time scale). Availability of raw (not aggregated) data and additional information about the metering points (e.g. address information) can widen the scale of possible statistical outputs.

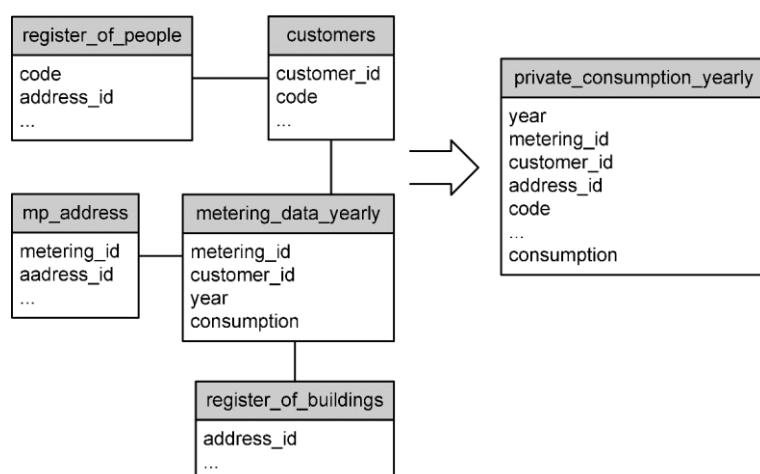


Figure 5. Linking tables to establish aggregated private consumption

## **5. Outlook**

The results obtained from the project indicate that the source, smart meter data, has a great potential to become a source for producing official statistics. During the project, the linking quality was continuously improved and new knowledge about the data was obtained, which improved the ability to find better match between measured units, metering points, and statistical units, businesses, households and dwellings. Therefore, it is becoming clear that after improving the linking quality further the source can be used for producing official statistics.

Further activities should be devoted to improving address quality and other sources for finding proper address information might be used. There is also need for finding adequate sources for categorization labels, as the quality of categorization depends on the quality of the training data. Thereafter the data can be used to train more complex models of artificial neural networks and to make predictions that are more adequate.

## 2.4 AIS Data

Aim of this workpackage was to investigate whether real-time measurement data of ship positions (measured by the so-called AIS-system) can be used 1) to improve the quality and internal comparability of existing statistics and 2) for new statistical products relevant for the ESS.

The added value of running a pilot with AIS-data at an European level is that the source data are generic worldwide and data can be obtained at European level.

Methodological, quality and technical results of the workpackage, including intermediate findings, were used as inputs for WP 8 of SGA-2. When carrying out the tasks listed below, care was taken that these results will be stored for later use, by using the facilities described for WP 9.

The original plan for SGA-2 (August 2017-May 2018) was to develop a model to calculate emissions, investigate AIS data from the European Maritime Safety Agency (EMSA) and investigate future perspectives for AIS data as source data for new statistical output. However, at the start of SGA-2 the scope of SGA-2 was adjusted, because access to this AIS data could not be secured and good models for calculating emissions already existed.

### *Task 3 – Methodology and Techniques*

#### *Estimate emissions*

Aim of this task was to investigate existing models for deriving emissions and describe how these models could be implemented in statistical processes. Other possibilities for improving current emission statistics by using AIS data were also investigated.

### *Task 4 – Access to and analysing AIS data from EMSA*

AIS data are available for national territories and the entire European territory. In this workpackage commercial European AIS data from Dirk Zwager (DZ) was used. But for the future it would be preferable to gain free European AIS data. Therefore, gaining access to data with European coverage from EMSA was tried. Unfortunately, this could not be realized. However, satellite data from Luxspace (LS) covering the Mediterranean Sea for the same period as the DZ data was obtained. So the quality of satellite data was assessed by comparing it to the DZ data and the national AIS data from Greece.

### *Task 5 – New statistical output*

Aim of this task was to explore possibilities of new statistical products (for example intra port statistics) by using AIS data.

### *Task 6 – Future perspectives*

Aim of this task was to produce a *consolidated* report summarising the contents and the outcomes of WP 4. This task also included analysing and elaborating scenarios for production of European and national statistics based on one single European data source and a cost-benefit analysis of using AIS data for official statistics.

Concrete results of the tasks mentioned above are deliverables on:



- 1 [Determining emissions using AIS](#)
- 2 [Possible new statistics using AIS](#)
- 3 [Consolidated report on project results](#)

## **1. Results on determining emissions**

It was investigated how emissions from maritime ships can be determined using AIS data. Aim of this task was to investigate existing models for deriving emissions from AIS and describe how these models could be implemented in the statistical processes. Possibilities for improving existing emission statistics by using AIS data were also investigated.

One of the ways shipping impacts the environment is through atmospheric pollution. Exhaust gases result in both conventional pollutants and greenhouse gas emissions. To maintain Kyoto Protocol (1997), it is necessary to monitor chemical pollutants such as Carbon Dioxide(CO<sub>2</sub>), Methane (CH<sub>4</sub>), Nitrous Oxide, Nitrogen Dioxide (NO<sub>x</sub>), Sulphur Dioxide (SO<sub>2</sub>,) and Particulate Matter (PM). AIS provides a method to determine the route of a ship. Combined with a model to estimate the emission of vessels (which depends on factors like travel distance, speed, draught, weather conditions and characteristics of the vessel itself), emissions of e.g. CO<sub>2</sub> and NO<sub>x</sub> can be estimated per ship and per (national) territory.

To identify goals for this task, the needs of stakeholders/ experts on this topic were first investigated. Then, various papers were described and it was explained how existing methodology would be useful for our purposes. Finally, recommendations were provided on how stakeholders' needs on a national and European level can be met.

### *Stakeholder's needs*

There are two approaches to determine and publish about emissions: using the territorial principle or the residence principle. In the territorial approach, emissions are calculated for confined geographical areas, such as a country or a sea. In the residence approach emissions are calculated for ships that are owned by companies from a certain nationality for the whole world. The latter approach is necessary for the National Account System.

The needs of stakeholders with regard to information on emissions from AIS were investigated. It was considered how data on emissions are reported internationally and then national agencies from the countries in this workpackage were approached. The comments received from stakeholders were mainly geared towards the National Accounts approach. For example, Statistics Netherlands expressed interest in receiving travelled distances by Dutch owned ships. Hardly any requests on the territorial approach were received. This might be because it was harder to contact the institutions that work on this as these were usually not the statistical offices themselves. Therefore, it was not possible to conclude from this investigation that there is no interest in work on the territorial approach.

### *Literature review*

To investigate work that has been done on AIS and emissions on a national level, and on a more global level, three papers were discussed. First, papers focusing on two geographical areas of the members of this workpackage (Norway and the Netherlands) were discussed, and then a paper

widening the geographical area of interest using the so-called Ship Traffic Emission Assessment 3 Model (STEAM3).

The models presented here all build on the same principle of using actual speeds compared to maximum speeds as a proxy for propulsion power used. The methods applied become more advanced in the STEAM3 model. The method described in the first and second paper are not sufficient for a European analysis because they only cover a smaller geographical area. For a larger scale analysis two aspects are needed, which are included in the STEAM3 model. The first aspect is an algorithm dealing with missing AIS data due to scarcity of AIS in some areas. Second, a method to obtain detailed vessel information is needed. To this end, an extended/automatic means of combining multiple sources has to be incorporated. Another benefit of the STEAM3 model is that it takes into account aspects like control areas, scrubbers and dual fuels.

It was concluded that the STEAM3 model would best serve the purpose of analysing emissions of maritime shipping in large geographical (European) areas or even worldwide shipping. However, local models also do quite well for their own area and are much simpler. Starting off with the local models might thus be a good first step to calculate emissions, then turning to the more complex global model (of course also depending on the needs).

#### *Recommendations for calculating emissions using AIS*

When constructing emission statistics on maritime (and inland waterway) transport, a number of data sets are indispensable: worldwide AIS data, a route generation algorithm, ship's characteristics (IHS Fairplay-based registry, web scraping and a similar vessel algorithm), emission factors, an emission calculation algorithm and storage. Some opportunities for efficiency gains and other opportunities have been described, like using "average" paths and speeds, the latter being presented in a visualisation. All in all, it is thought that the analysis of AIS data for statistics eminently lends itself to a process on a European level.

As developing such a method would take a big investment in terms of getting data, algorithms and storage, an intermediate step might already be possible. Stakeholders have expressed interest in two specific aspects that might be easier to obtain. Firstly, the Dutch department of National Accounts has expressed interest in getting the development of the distance travelled for Dutch-owned ships. Even European AIS data would be an enrichment of the current method. Secondly, the Dutch Pollutant Release and Transfer Register would be interested in getting information on average speeds over 100 meter-intervals for Dutch territory (mainly for inland waterway ships).

## **2. Results on accessing and analysing AIS data from EMSA**

AIS-data are available for national territories, the entire European territory, and covering the whole world. In this workpackage European AIS data from DZ has been used. As this costs money, there is interest in obtaining free European AIS data. That is why it was tried together with Eurostat to obtain AIS data at the European level from the European Maritime Safety Agency (EMSA). Unfortunately, getting access to data from EMSA proved not to be possible during this ESSnet .

However, satellite AIS data from LuxSpace was obtained. Satellite data is needed when ships travel across the oceans, as ships here are positioned too far from land so that land receivers are not able

to pick up the signal. Getting this data provided the workpackage with the first possibility to examine satellite AIS data. The data covered the Mediterranean Sea for the same period as the DZ data already available and also covered the area of the Greek national data. This gave the opportunity to assess the quality of satellite data by comparing it to the DZ data and the Greek national AIS data.

As expected, satellite data is indispensable when it comes to following ships across the Mediterranean Sea. Also, in some port areas it did pick up signals from ships that were not picked up in the DZ data. However in some ports, satellite AIS contained signals from a lower number of ships than the Greek national data. In addition, the interval between signals sometimes is too long to derive detailed enough information to linearly model shipping routes in more complex areas. That is, for areas where the geography is more intricate, e.g. due to isles or different water depths, frequency of reporting by satellites can be too low to just linearly interpolate successive points. As satellite AIS has significantly higher latency and lower frequency of receiving, the number of messages for linear interpolation can result in a ship's route crossing land. When AIS data is needed to cover both coastal and oceanic regions, a combination of land and satellite receivers is necessary to track all (routes of) maritime ships.

The results of this task were added to deliverable 4.3 of SGA-1.

### **3. Results on possible new statistical output**

Aim of this task was to explore possibilities of new statistical products from using AIS data. In the workpackage a questionnaire had already been administered to maritime experts at all European NSI's. Results from this expert questionnaire were also taken into account.

#### *Intra port distances*

Port authorities do not always have a complete insight in the activities within their port. For example, they do not always know whether ships visit one or multiple terminals during one visit, rendering a higher intra port journey distance. The expert questionnaire shows that most of the NSI's are interested in this. Journeys within the port, and with that intra-port distances can be derived from AIS, as shown in the PoC (see deliverable 4.3).

#### *Fluvio-maritime statistics*

If transport is partly performed on inland waterways (IWW) and partly on sea, it is classified as fluvio-maritime transport. For Eurostat, this has created discussion about whether fluvio-maritime transport should be reported as maritime transport, IWW transport, or both. Countries have non-harmonised approaches to report fluvio-maritime data in the maritime and IWW statistics. Deliverable 4.3 describes the results of the PoC performed on investigating the extent of fluvio-maritime transport in the Netherlands. AIS data could help getting insight in the relationship between maritime and IWW transport.

#### *Port-to-port distance matrix*

(Dynamics between) Different factors can influence the route a ship travels. If one does not want to calculate the distance for every single journey, this requires a flexible way of calculating the distance a ship travels. AIS provides the opportunity to follow ships' routes and thus calculate the distance

travelled. In addition, it can also provide more insight into patterns and factors determining these travel patterns.

The quality of Eurostat's average port-to-port distance matrix was compared to Marine Traffic's model for estimating routes of ships and the calculated distance per single journey (obviously the most precise measurement). Only a very limited number of journeys was examined, so a strong preferred averaging method was not established. Yet, the method of calculating each journey separately requires an enormous amount of processing. It is thought that the fixed Eurostat's distance matrix at the moment is too limited. Reasons for this are that it only uses a short period of historical AIS data, it does not take into account important determining factors (e.g. type of ship, weather conditions) and it does not use worldwide AIS data. Using a historically richer model like Marine Traffic's, and adding factors like type of ship and weather conditions, could further improve the calculation of travelled average distances of ships.

#### *AIS as an economic indicator*

The EU's international trade market highly depends on seaports. To get up-to-date information on the development of international trade, and with that economic development, an analysis of activity in the seaports would be valuable information. Part of this information could be based on AIS data, so AIS could provide a rapid indicator of economic development, where international comparisons can also be made quickly.

#### *Insight in disruptions or regulations*

Multiple parties, such as experts in the questionnaire and port authorities, have shown interest in gaining insight into the effects of disruptions or (changed) regulations by means of AIS. For example a shift in traffic patterns caused by port closure could be found by using AIS. Another example is speed regulations in for example emissions zones. To examine and visualize this, a speed grid was developed.

#### *Visualizations*

To gain insight in data, particularly big data, visualizations are indispensable. Although visualizations are not a statistic as such, it does ask for a new type of skill, requiring an investment for most statistical offices. As such, it is important to be aware of the opportunities opened up by this investment. An example of a visualization published by Statistics Netherlands was presented which showed sea routes taken by container giants (which have a cargo capacity of more than 10 thousand TEU) performing transshipment. See [http://research.cbs.nl/AIS\\_transshipment](http://research.cbs.nl/AIS_transshipment).

### **4. Future perspectives**

#### *Cost-benefit analysis of using AIS for official statistics*

Regarding sustainability, it is foreseen that AIS data, on both the transmitting side (ships) and receiving side (land-based stations and satellites), will be available for quite some time into the future, as it is part of the general security system for both maritime and inland shipping. Even if the mode of transmission might change (e.g. wifi instead of GPS), most of the algorithms developed will still be useful. National data will probably remain freely available, as they are also used for other

governmental ends. When using commercial data, which can provide a larger area coverage, new arrangements would have to be made once in a while, asking for new negotiations that might pose some risk

#### *Scenarios for producing national/European statistics based on AIS data*

The analysis of AIS data for statistics eminently lends itself to a process on a European level. Having every country determine the route of every ship (also entering other countries) would result in redundant processing. Therefore, to optimize efficiency on a European level, it would be advisable to perform calculations per (European-owned) ship and then aggregate this either to nationality of owner or to the territory a ship traverses.

For future use, it is advised to at least obtain European AIS data from EMSA. This has the big advantage that European statistics could be made and that all national statistics are comparable. The advantage of EMSA data compared to other European sources is that the EMSA data is freely available and the coverage and frequency will be on the same level as the national AIS data. Also data access will be more sustainable compared to data from other, commercial, companies.

The only disadvantage of this scenario would be that there is no world coverage, so travel between European and non-European countries cannot be followed. If there is sufficient funding, it is advised to use worldwide data, because then all ship journeys can be followed and a more complete picture of worldwide patterns can be analyzed.

Eurostat could have an important role in decoding, cleaning and storing the European data in one central database and could enable each NSI to query this database for their own national statistics. Another alternative could be that one of the NSI's will take this role and provide a central European database with AIS data. Still another alternative would be asking an external organization (like EMSA) to build an AIS reference architecture for ship transport from AIS data, based on a standardized journey approach with time stamps and location information. This architecture would represent microdata, from which it should be possible to generate all statistics that are currently being produced (and intended to be produced in the future).

#### *Sketch of a possible statistical process and needed infrastructure*

For collecting streaming data a failsafe infrastructure with backup functionality is needed. Batch data collection will stress the infrastructure less. However, extending collection periods increases the size of the data and computational power needed for processing the data.

For AIS data it was concluded that decoding the messages and encrypting the crucial fields can be done on a normal infrastructure. More complicated analysis should be done on distributed environments like Spark. Storage of the data should be done on high performance big data file systems. During analysis, it is important to have a performance environment available, like Spark. The chosen environment must be able to integrate with analysis tools such as R and elastic search. After analysis, dissemination is comparable with disseminating traditional statistics. The same tools and infrastructure can be used.

Using AIS in statistical processes asks for new skills, not always available in the Statistical offices yet, so training for employees or new employees is probably needed.

### *Future work on AIS*

Although there is a lot of work done in this workpackage, further investigation on this promising data source is still needed. A new ESSnet Big Data has been planned already (end 2018-2020). One of the principal aims of this ESSnet will be to develop functional production prototypes and promote and support their implementation in a limited number of participating National Statistical Offices. The final description for the work to be done in the new ESSnet has to be worked out in the next couple of months. The ideas for further work based on AIS data could be part of this new ESSnet if Eurostat thinks they are of importance to the ESS.

### **5. Overall conclusion**

Although the workpackage is not satisfied yet with the quality of the currently used European AIS source, the results from this workpackage show the potential wealth of AIS data to improve current statistics and to generate new statistical products.

Although some important elements of current maritime statistics such as type and quantity of goods loaded or unloaded at the port are not part of AIS, AIS is still useful to improve other aspects of maritime statistics and provide new products. There are also concerns about not having suitable hardware and software tools for the exploration and exploitation of the huge amount of AIS data and the lack of documentation and guidance for using AIS. Of course, there is a role here for Eurostat or an experienced NSI.

For using AIS in official statistics more investigation is certainly needed and at least a European AIS source of good quality will be needed. It is also important for the future to have data from both land based stations as well as data from satellites. By having new data sources like EMSA available in the future the possibilities of AIS data seem to be even more promising.

## 2.5 Mobile Phone Data

### 1. Introduction

The specific grant agreement 2 (SGA-2) for the WP 5 on mobile phone data has focused on the development of a methodological framework, the analysis of the IT infrastructure and software tools, and the assessment of quality issues regarding the use of mobile phone data in the production of official statistics. The results in these three aspects are described below. These are the links to the three deliverables:

Methodology:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/4d/WP5\\_Deliverable\\_5.3\\_Final.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/4d/WP5_Deliverable_5.3_Final.pdf)

Technology:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/c/ce/WP5\\_Deliverable\\_5.4\\_Final.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/c/ce/WP5_Deliverable_5.4_Final.pdf)

Quality:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/71/WP5\\_Deliverable\\_5.5\\_Final.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/71/WP5_Deliverable_5.5_Final.pdf)

### 2. Methodology

Based upon the experiences in accessing data during the SGA-1 and looking for a starting point to find and propose statistical methods to process mobile phone data, WP 5 started its deliverable on methodology by revisiting the definition of big data for official statistics. Instead of the three Vs widely known and cited to introduce big data, it was concluded that for official statistics there exist more important features to be taken into account:

- Data refer to **third people** and not to data holders;
- Data are **central in data holders' economic activity**;
- Data **lack statistical metadata** (since they are generated for very different purposes).

The first two characteristics clearly lie behind some of the aforementioned issues to access data. The third feature is an essential trait for methodological considerations. Notice how administrative data also share these characteristics. Thus, it can be claimed that existing tools for the use of admin data in official statistical production, with due modifications, are still valid for mobile phone data.

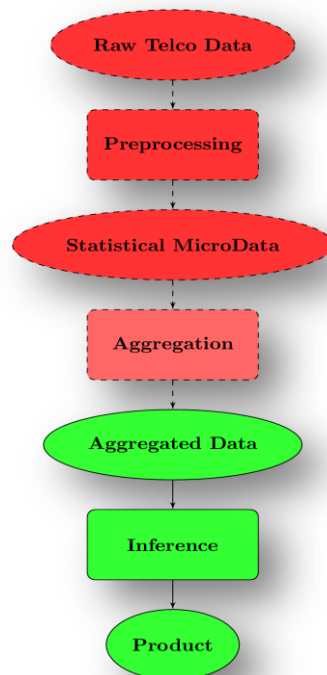


Figure 1. Sequence of large processing steps of mobile phone data

Although WP 5 has not been able to produce a concrete statistical output based on an end-to-end process, it does provide a set of elements for the construction of a production framework of official statistics based on these data. A schematic representation of the whole process is depicted in Figure 1. With oval elements different forms of data are represented, whereas with rectangular shapes process steps are represented. In red those elements and phases of the process under no control by NSIs are depicted, whereas in green those elements and phases under control by NSIs in the project are shown. Data accessed in the current process are in aggregated form (exceptions under limited conditions are INSEE, ISTAT, and CBS).

The process starts with the raw telecommunication data generated by the mobile device-antennae interaction in the telecommunication network. This form of data cannot be exploited for statistical purposes and thus needs pre-processing to turn them into statistical microdata. These are data at the mobile device level basically gathering information about their identification, time attributes, and spatial attributes together with some extra variables depending on the network (event types, duration, ...). They can be aggregated to produce the number of detected individuals at each territorial cell in a sequence of time instants. The final step infers from these aggregated data the estimates for the target population at stake. This is the ultimate goal of the whole process: to produce high-quality official statistics for a target population.

Concrete methodological proposals were provided for different elements of this process:

- The generation of data and the identification of error sources can be undertaken using the two-phase life-cycle model by Zhang (2012) (see also Reid et al. (2017) for a wider adaptation to admin data). It is claimed that the Total Survey Error paradigm, duly adjusted, is still valid for mobile phone data. Especially, WP 5 underlines the increasing relevance of statistical



methods to go from events to statistical units (from network events such as calls, SMS/MMS, Internet connections, pings,... to individuals in the case of mobile phone data).

- The assignment of spatial attributes to network events is a key step in the generation of statistical microdata. This step strongly depends on data availability. Three scenarios of increasing complexity were explored:
  - Using only the location of the antennae, a Voronoi tessellation of the territory was computed to assign a geolocation to each event. Since this tessellation technique does not take into account both the directional character of antennae and the overlapping nature of the covered territorial cells, a first conclusion is the need for more sophisticated methods.
  - Taking into account the directional character of antennae and depending on the morphology of the territory and the estimated population density of subscribers, so-called Best Service Areas were computed and then used to locate the events (hence the individuals). Although offering much better results than Voronoi techniques, the non-overlapping assumption still poses some limitations. These areas are to be computed by MNOs since not all information is under NSIs' control.
  - Finally, considering both the directional and overlapping character of the antennae cells, a Bayesian approach to estimate the probability that an event (hence a mobile device) is located in a given territorial cell was followed. The probabilities are constructed using the prior information of each antennae and the geographical partition of the territory at stake. The likelihoods are computed using the signal strength. Should access be obtained to more variables, more sophisticated computations could be undertaken with this same structure.
- Once pseudonymised ID, spatial, and time attributes have been computed and assigned to each unit, a data model must be constructed containing diverse information elements for further producing statistical outputs of interest. These range from stay and movement sections (conceptually this is indeed observing sequences of stay and movement periods) to country of residence, anchor points (work/home/...), usual environment, trips, ... These are to be complemented with official data such as geographical administrative units, cell grids, etc.
- The inference exercise (possibly including the aggregation of microdata as an initial step) connecting aggregated data at each cell with the target population cannot follow the traditional probability sampling scheme, since no probabilistic sample selection can be undertaken. An alternative inference model must be used. Important points to consider are:
  - It can be argued that the concept of representativity must be duly understood and not to request from new data sources something which is not already present in traditional sources and probability sampling. Representativity is not a mathematical concept. Unbiased estimates with as low a variance as possible should be pursued. Certainly, there will be selection biases and new elements such as model checking and model assessment will be needed.
  - hierarchical model was adapted which was already used by ecologists to solve the so-called species abundance problem to estimate population counts. The main working assumptions are:
    - At  $t_0$  individuals are assumed to be physically in the territorial cell of auxiliary admin/survey data.

- Mobility patterns of individuals do not depend on the concrete MNO they are subscribed to.

The key parts in the specification of the model to estimate the number of individuals  $N_i(t_n)$  at each cell  $i$  and time period  $t_n$  are:

$$N_i(t_n) = \left[ N_i(t_0) + \sum_{\substack{j=1 \\ j \neq i}}^I p_{ji}(t_0, t_n) N_j(t_0) - \sum_{\substack{j=1 \\ j \neq i}}^I p_{ji}(t_0, t_n) N_i(t_0) \right], \quad i = 1, \dots, I,$$

$$N_i^{MNO}(t_0) \simeq \text{Binomial}(N_i(t_0), p_i(t_0)), \quad i = 1, \dots, I,$$

where  $p_{ij}(t_0, t_n)$  are detection probabilities of individuals moving from cell  $i$  to cell  $j$ . The random variables  $N_i(t_0)$  and  $p_{ij}(t_0, t_n)$  are further specified according to prior probability distributions with their corresponding (hyper-) parameters modelled using our available data (from official population registers, survey data, and mobile network data).

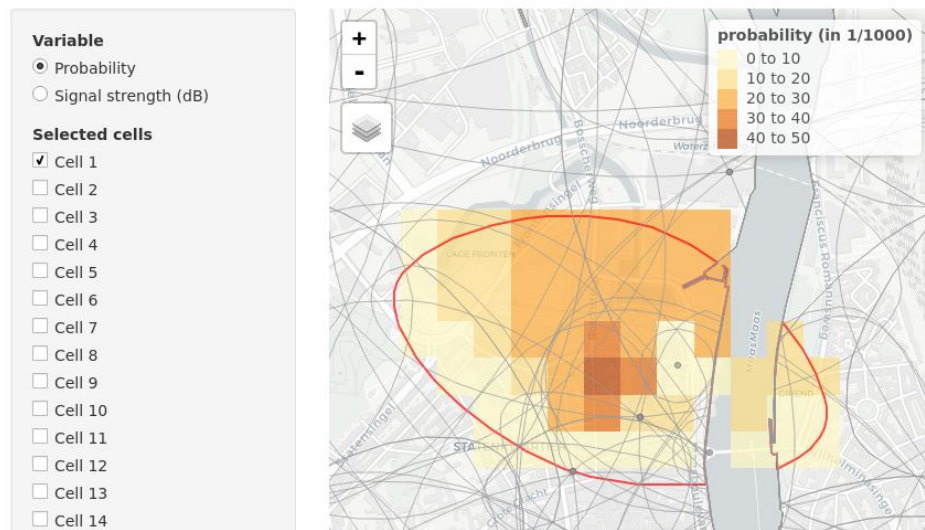
This is not intended to provide a definitive solution for the inference stage of the process, but to set up the first elements for an inferential framework in which the official statistician can adapt the model to the concrete inference exercise at stake. A Bayesian approach to fit the model was followed.

### 3. IT tools and infrastructure

Not having full access to data to undergo an end-to-end statistical process to produce a concrete output, the workpackage has concentrated on the IT part of the aforementioned elements for a production framework. The following main outputs have been provided:

- A general description was provided of IT platforms to process data at NSI's premises and at MNO's premises. In the first case, advantage was taken of the dataset of Call Detail Records collected by ISTAT and transmitted from TIM to the Italian NSI. A cluster with 8 nodes with 6 drives of 1.2Tb capacity, 128Gb of RAM, 32 or 16 CPU in each node has been used to process the data. This has been exemplified as a potential IT infrastructure when data are transmitted to the NSI.
- Complementarily, an IT platform to access mobile phone data in situ at an MNO's premises was also described. Taking advantage of the situation of access by the French INSEE to Call Detail Records of Orange at Orange Labs, WP 5 has provided a description of this platform. It is basically also a cluster providing an overall storage capacity of 3.5 Pb with 24Gb of RAM memory on average.
- An R package was developed called mobloc implementing the Bayesian approach to geolocate network events based on the signal strength. The package makes use of the antennae position and provides algorithms and tools to translate antennae properties to geospatial distributions. Depending on the availability of the antenna properties, such as coordinates, direction, and height, different algorithms can be used. Based on the Bayesian approach, and using these input data, the user will obtain the probability of geolocation for each cell.

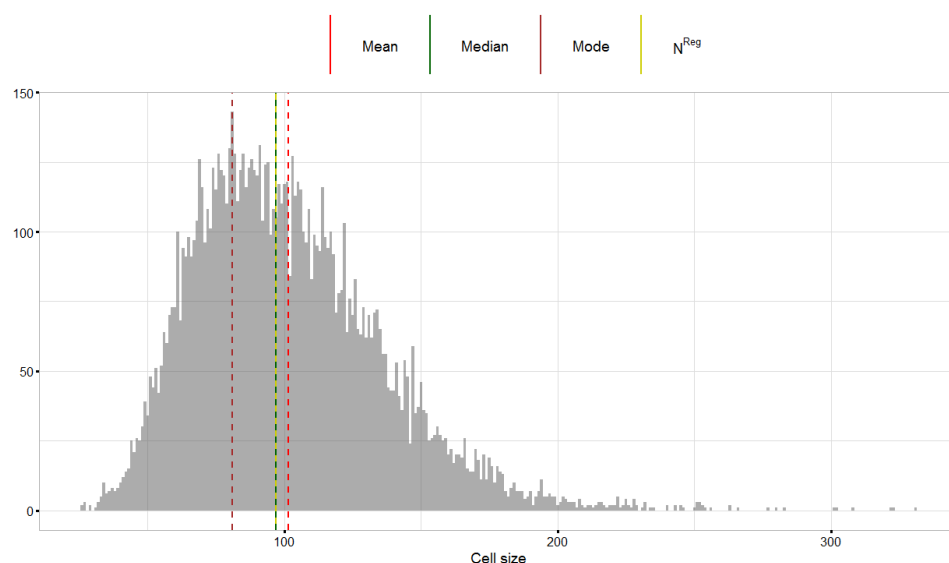
## Cell Inspection Tool



A snapshot of the geospatial and computing capabilities of the mobloc package.

- An R package was developed called pestim implementing the aforementioned statistical model to estimate population counts. To compute estimates a Bayesian approach was followed mainly for two reasons. Firstly, by the specification of priors in this approach this gave the possibility to incorporate in a natural way all available information (from survey data, from administrative registers, ...) into the estimation process. Secondly, the Bayesian approach invites also in a natural way to parallelize the computation. As a result of the application of this approach, the user, after specifying both MNO and admin data for each territorial cell, will obtain a posterior distribution for the number of individuals in each of these cells. This distribution enables the user to provide both point estimates, credible intervals, and a number of indicators associated to the estimation process.

Both packages are freely available at WP5's GitHub page.



Posterior distribution for the number of individuals at a cell.

#### 4. Quality

Quality must be an ultimate goal in the production of official statistics and also for new data sources. Challenges arise when using mobile phone data which can be derived from the new changes in the production process depicted above. WP 5 has focused basically on two aspects of quality in the project. On the one hand, a first incursion was made on how the European Statistics Code of Practice (CoP hereafter) is going to be affected according to the preceding proposals. On the other hand, proposals have been made to deal with the accuracy dimension of quality in the context of the new inference model for the production of official statistics using mobile network data.

Regarding the CoP, the workpackage has briefly analysed principle by principle suggesting how each one will potentially be affected by the use of mobile phone data for the production of official statistics. In summary, the three main factors as sources of change have been identified: (i) MNOs will be an **active part of the production process**; (ii) a change will be needed of **inferential paradigm** from design-based to model-based (even Bayesian), and (iii) there will be an unprecedented **higher** degree of spatial and temporal **breakdown** in outputs. These three factors will affect the CoP in a cutting-cross way.

As the accuracy dimension is the most traditional measure of quality in Statistics, the workpackage has focused on producing accuracy measures in the context of the inference stage depicted above. Having followed a Bayesian approach, since the output of the modelling exercise is a posterior distribution for the number of individuals  $N_i(t_n)$  in each cell  $i$  and time period  $t_n$ , one has the conditions to produce any statistical indicator at will. In this line, the traditional confidence intervals and coefficients of variation can now be replaced by credible intervals (at least three alternative versions can be computed) and posterior coefficients of variation (which now can be also computed using such robust measures as the posterior interquartile range, the posterior median and similar robust indicators).

As a novel element, the goodness of fit of the model now needs to be checked and assessed to make sure that final estimates are not provided upon a useless model (hence starting from inappropriate prior hypotheses). The core element here is the posterior predictive distribution by which it is checked whether the input data (mobile network data) can be reproduced firstly estimating the hyperparameters and then using the model to generate replicated mobile phone data.

#### 5. The future research

A whole chapter in the last deliverable has been devoted to reflect on the future and provide some insights. The results in this workpackage constitute a first step and more research is needed. Different recommendations have been provided for this future research:

- The linear structure access-methodology-IT-quality has proven not to be the most efficient strategy to conduct this research, since the access to mobile phone data is currently blocked. It is recommended to follow a parallel double-track with access issues on one track and methodology-IT-quality on the other track. They must advance in parallel working on simulated data as much as possible until real data can be used in optimal conditions.

- A track of research on how to simulate mobile phone data must be initiated, including simulations of the whole population. This will enable NSIs to advance on the second track and to test both hypotheses and statistical models.
- The geolocation of network events must be further investigated including accuracy issues to build a full data model providing service for the production of any kind of statistics (population, tourism, transport,...). The construction of this model must contain the spatiotemporal interpolation of data.
- The inference framework must be enriched with more hypotheses and more sophisticated and realistic models.
- The Quality Assurance Framework must be revised in terms of potential new indicators. The Total Survey Error paradigm must be adapted to this new data source in the search for the identification of all error sources.

Hopefully, in the forthcoming second ESSnet on Big Data the lines of work initiated here will be continued.







## 2.6 Early Estimates

The main goal of WP 6 was to explore how a combination of (early available) multiple big data sources, administrative and existing official statistical data could be used in producing existing or new early estimates for official statistics.

Several pilots were carried out during the SGA-2. The aim of the pilots was to investigate big data and other existing sources for calculating flash and (or) intermediate estimates of economic indicators. They describe in particular:

- Big data sources and statistical areas where used
- Other statistical data sources combined with investigated big data sources
- Methods used and impact on the quality of results
- Data treatment and related methodological and IT issues

*Picture 1: SGA-2 WP6 Pilots*

|   |  |
|---|--|
|    | Use of electronic transactions of System of payments and of the Anti-Money Laundering Reports data on estimating private household consumption |
|   | Machine learning approaches for <u>nowcasting</u> GDP and TIO using firm-level traffic loops data  |
|  | Estimating early GDP and IPI using traffic loops data  |
|  | The use of high frequency indicators for predicting macroeconomic variables  |
|  | Using internet data sources about the property market and job offers to forecast coincident and leading indicators                             |
|  | Using Monte Carlo Markov Chain (MCMC) to clean the data, remove noise and solve the problem of missing data                                    |

During the conducting of the pilots the correlation of the data sources and early economic indicators was explored and according to the results (detected combining sources and testing early economic indicators), various models for flash and (or) intermediate estimates were used. The most promising estimator was GDP, but the pilots were not limited to GDP due to the fact that results of analysing data sources proposed the calculation of estimates of other economic indicators.

The main outcome was the calculation of a testing set of early estimates of a concrete economic indicator together with the defined methodology and process needed for this purpose. Those estimates were calculated with respect to country specifics (availability of data sources) and possibly broken down into economic activity classes and (or) regional domains. Recommendations about the mode of combining (big) data sources, models for estimating economic indicators, quality assessment, IT infrastructure and statistical process of conducting the calculation of early estimates of economic indicators were prepared.

The main achievements are calculated concrete estimates for economic indicators with impact and quality assessment of results from the pilots carried out during the SGA-2.

## **1. Deliverables**

Four deliverables were produced during the SGA-2 (completed May 2018):

6.6 Report about the impact of one (or more) big data (and other) sources on economic indicators (correlation, time lag, selectivity, etc.). Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/27/WP6\\_SGA2\\_Deliverable\\_6\\_6\\_L.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/27/WP6_SGA2_Deliverable_6_6_L.pdf)

6.7 Recommendation about the methodology and process of calculating estimates for at last one early economic indicator. Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/ef/WP6\\_SGA2\\_Deliverable\\_6\\_7\\_L.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/ef/WP6_SGA2_Deliverable_6_7_L.pdf)

6.8 At least one example of calculated concrete estimates for one of the economic indicators with quality assessment of the input, throughput and output phase of the process. Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/ae/WP6\\_SGA2\\_Deliverable\\_6\\_8\\_L.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/ae/WP6_SGA2_Deliverable_6_8_L.pdf)

6.9 Report and recommendations about the IT infrastructure needed for the storage, analysing combining data sources and the process of calculating early economic indicators. Link:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/a0/WP6\\_SGA2\\_Deliverable\\_6\\_9\\_L.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/a0/WP6_SGA2_Deliverable_6_9_L.pdf)

## **2. Main findings**

During the ESSnet project in the SGA-1 period some nowcasting methods for the purpose of estimating turnover indices were discovered and tested. The original purpose was to explore the possibilities of estimating the consumer confidence index and (or) turnover indices. Due to the early findings about the inaccessibility of social media data, which is crucial for assessing the consumer confidence index, the WP 6 team focused on turnover indices.

Statistic Slovenia (SURS) and Statistics Finland both explored the usage of traffic loop data in the process of estimating early economic indicators.

Based on the investigation of various big data sources, SURS had the idea to use the data acquired from traffic sensors and use them as primary and secondary regressors in a linear regression method for nowcasting GDP 45 days after the reference period.

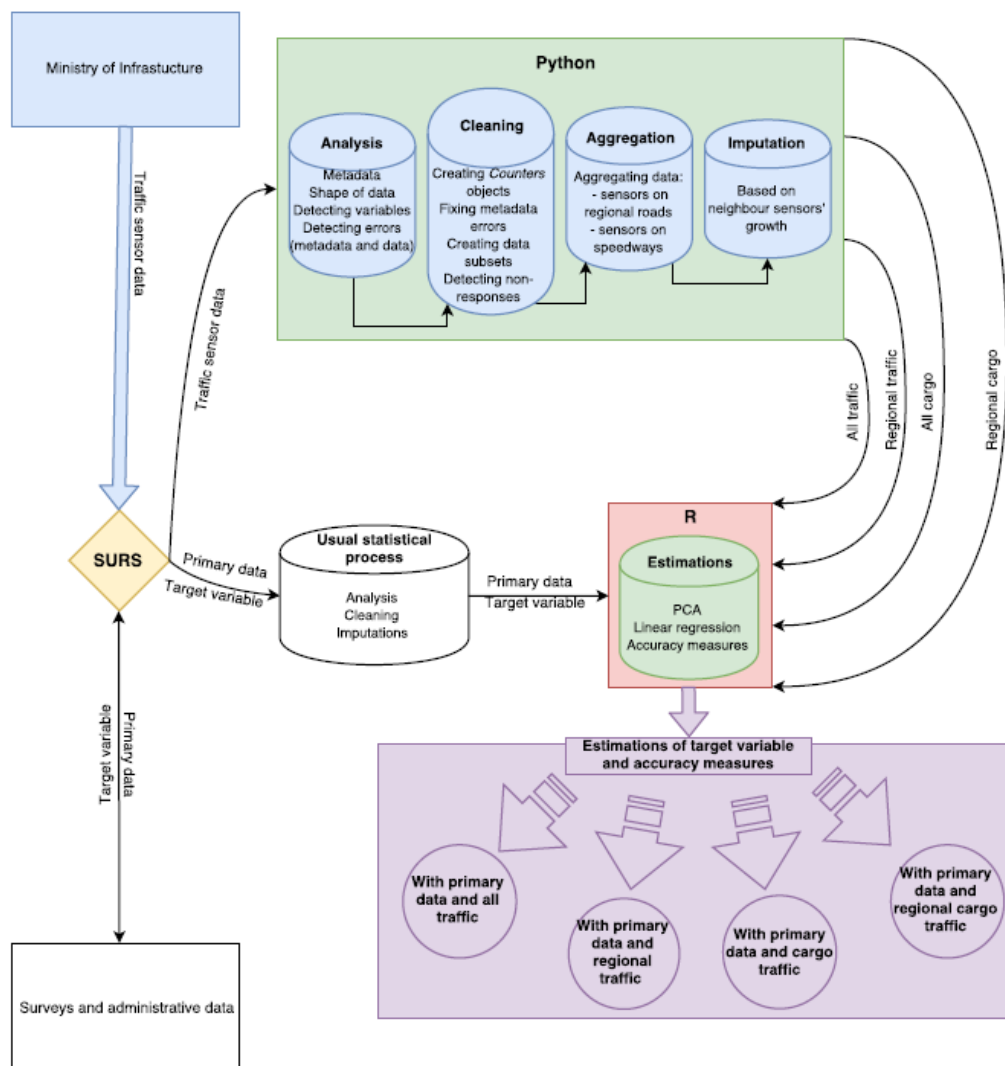
Table 1: Estimates and errors of no-traffic data and traffic data sets

| Period | Official values of GDP (in million EUR) | PCA method | No traffic data estimates | Traffic data as secondary regressor estimates | Absolute values of relative errors of the 1 <sup>st</sup> est. (in %) | Absolute values of relative errors of the 2 <sup>nd</sup> est. (in %) |
|--------|---|------------|---------------------------|---|---|---|
| 2016Q2 | 9755.848669                             | 75%        | 9641.769                  | 9620.851                                      | 1.17  | 1.38  |
|        |   | 80%        | 9655.510                  | 9601.865                                      | 1.03  | 1.58  |
|        |   | 85%        | 9627.988                  | 9660.667                                      | 1.31  | 0.98  |
|        |   | 90%        | 9689.009                  | 9531.461                                      | 0.69  | 2.30  |
|        |   | zadnja5    | 9531.461                  | 9541.131                                      | 2.30  | 2.20  |
| 2016Q3 | 9748.845959                             | 75%        | 9642.126                  | 9699.763                                      | 1.10  | 0.50  |
|        |   | 80%        | 9681.626                  | 9637.750                                      | 0.69  | 1.14  |
|        |   | 85%        | 9656.819                  | 9658.668                                      | 0.94  | 0.93  |
|        |   | 90%        | 9716.181                  | 9464.554                                      | 0.34  | 2.92  |
|        |   | zadnja5    | 9704.672                  | 9530.760                                      | 0.45  | 2.24  |
| 2016Q4 | 9686.031852                             | 75%        | 9570.723                  | 9700.913                                      | 1.19  | 0.15  |
|        |   | 80%        | 9629.284                  | 9582.602                                      | 0.59  | 1.07  |
|        |   | 85%        | 9722.534                  | 9553.433                                      | 0.38  | 1.37  |
|        |   | 90%        | 9593.896                  | 9600.487                                      | 0.95  | 0.88  |
|        |   | zadnja5    | 9646.320                  | 9646.767                                      | 0.41  | 0.41  |
| 2017Q1 | 9395.205718                             | 75%        | 9355.285                  | 9317.238                                      | 0.43  | 0.83  |
|        |   | 80%        | 9419.886                  | 9336.230                                      | 0.26  | 0.63  |
|        |   | 85%        | 9285.980                  | 9305.568                                      | 1.16  | 0.95  |
|        |   | 90%        | 9133.499                  | 9275.574                                      | 2.79  | 1.27  |
|        |   | zadnja5    | 9300.724                  | 9308.459                                      | 1.01  | 0.92  |
| 2017Q2 | 10197.88759                             | 75%        | 10137.822                 | 10103.581                                     | 0.59  | 0.92  |
|        |   | 80%        | 10201.657                 | 10118.330                                     | 0.04  | 0.78  |
|        |   | 85%        | 10111.837                 | 10096.056                                     | 0.84  | 1.00  |
|        |   | 90%        | 10130.813                 | 10178.251                                     | 0.66  | 0.19  |
|        |   | zadnja5    | 10182.237                 | 10248.388                                     | 0.15  | 0.50  |
| 2017Q3 | 10187.24812                             | 75%        | 10151.038                 | 10077.054                                     | 0.36  | 1.08  |
|        |   | 80%        | 10164.924                 | 10045.334                                     | 0.22  | 1.39  |
|        |   | 85%        | 10148.311                 | 10002.641                                     | 0.38  | 1.81  |
|        |   | 90%        | 10152.910                 | 10505.859                                     | 0.34  | 3.13  |
|        |   | zadnja5    | 10273.918                 | 10347.609                                     | 0.85  | 1.57  |
| 2017Q4 | 10265.53999                             | 75%        | 10110.472                 | 10224.349                                     | 1.51  | 0.40  |
|        |   | 80%        | 10065.457                 | 10099.572                                     | 1.95  | 1.62  |
|        |   | 85%        | 10346.287                 | 9998.1344                                     | 0.79  | 2.61  |
|        |   | 90%        | 10188.307                 | 10232.035                                     | 0.75  | 0.33  |
|        |   | zadnja5    | 10277.841                 | 10339.279                                     | 0.12  | 0.72  |

The pre-treated traffic loop data were used as a secondary regressor in a nowcasting method. The method consisted of a linear regression model with principal component analysis (PCA) that was used to find the best fit of quarterly enterprise production data onto GDP values.



Picture 2: Data processing and nowcasting phases

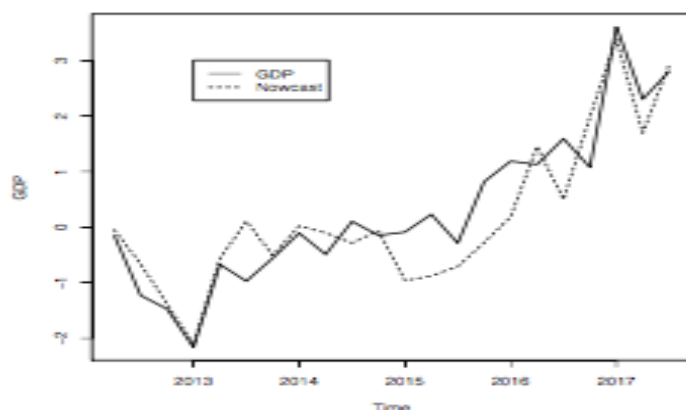


With the help of Statistics Finland SURS prepared the application (together with instructions how to use it) which allows inputting different kinds of data, testing various conditions for choosing principal components, producing quality indicators which compare results of different nowcasting methods and producing quality indicators which compare results with disseminated official statistics. This work describes the methodology and process of nowcasting indicators from the point of data acquisition to the end results on GDP and also on a known GDP correlator, the Industry Turnover Index. Imputations of missing count spots data received a big consideration in this work.

Statistics Finland carried out extensive testing of methodologies capable of dealing with large dimensional data, which are relevant when one tries to extract information from big data and disaggregated microdata that can be used in producing early estimates. Built upon the results obtained from the SGA-1, they explored machine learning approaches to nowcasting Finnish economic indicators (TIO and GDP) using firm-level data and traffic loops. The study describes the process of data preparation and analysis of the calculated estimates. It was concluded that incomplete early micro-level sources and a real-time big data source such as the traffic loop data can be used to produce early estimates of economic indicators. Statistics Finland regularly produces turnover index nowcasts based on the methodology presented here. Therefore, as a result of this

work, it is expected that turnover indicators will become timelier and (what is remarkable) more accurate, which in turn makes it possible to produce the aggregate output indicators (TIO and GDP) faster and more accurately by the existing national accounts procedures. A new methodology is provided by which an early estimate of GDP can be produced reliably (quantitatively with similar accuracy as the official first estimates) by employing micro-level turnovers and traffic loop data together with various machine learning methods designed to handle such large dimensional data. They used more than one hundred models in order to compute early estimates of the indicators of interest. In this study they discuss the generalities of the model classes explored, and then provide a list of the models which were incorporated in the forecast combinations, an approach which provides the best performance and is therefore the preferred methodological choice in nowcasting with the type of data employed. Some practical considerations related to the overall methodological framework are discussed. All the models used are easily implementable, in terms of software used and coding efforts. This study also evaluates the quality in terms of timeliness and accuracy in capturing the target indicators' next values. Statistics Finland manages to produce early estimates with similar accuracy as the official release 1.5 months earlier.

*Picture 3: Early Estimate of GDP vs the Official Release*



The above figure shows results which correspond to 0.49 of a percentage point MAE in relation to the t+60 publication by the statistical office. The early estimate is obtained by aggregating the monthly TIO nowcasts to the quarterly level. The process can be automated almost fully, and is applicable in many countries interested in producing early estimates in this way, using disaggregated micro-level sources.

In the Netherlands, a different approach was developed. Due to the large number of traffic loops (20,000 on the Dutch highways), another data cleaning strategy had to be developed. To enhance the speed of the data cleaning process, a Monte Carlo Markov Chain (MCMC) algorithm was used to clean the data, remove noise and solve the problem of missing data. The neural network is able to deal with missing and erroneous values, which makes it ideal for data cleaning and dimension reduction at the same time. By training subsets of road sensors in different neural networks, the dimensionality problem as described above is solved. The reduced dataset can then be used for regression purposes.

Through an institutional collaboration agreement with the Bank of Italy, Istat had the possibility to access new data sources concerning electronic transactions of the exchange circuits and interbank

settlement of the System of Payments and others deriving from the Anti-Money Laundering Aggregate Reports (SARA) that financial intermediaries file monthly to the Financial Intelligence Unit (UIF), which is a separate branch of the Bank of Italy. A joint Bank of Italy-Istat working group was set up to produce new series of indicators from these big data sources, to study the characteristics of these series and to test their possible use – besides traditional economic series – for the purposes of nowcasting or forecasting macroeconomic aggregates. This report is aimed at presenting the actions taken to produce the new series and the first results of the insertion of the new series in Istat models. There are two case studies: the nowcasting of the value added of Services and the forecasting of Consumption. The results are to be considered still preliminary and experimental.

Statistics Portugal aims at assessing early estimates for macroeconomic variables of interest over the most recent reference period based on big data in the regular production of official statistics. Regression and Spatial Panel Data Models are used as the methodological approaches.

Statistics Poland nowcasts the ILO unemployment rate, which is calculated on a quarterly basis and released with a delay of four months. The structural time series (STS) model has been used to produce flash estimates of the ILO unemployment rate, i.e. trend and seasonality. Two variables were considered: (1) registered unemployment rate based on data from District Labour Offices and (2) job vacancy barometer based on online job offers. In the project the following data sources were used: the Labour Force Survey, the registered unemployment rate and online job offers. It was concluded that the job vacancy index actually worsens the ILO unemployment rate. The model without any variables seems to be sufficient but this is mainly because of a clear falling trend in the unemployment rate.

### **3. Outlook**

It was shown that incomplete early micro-level sources and a real-time big data source such as the traffic loop data can be used to produce early estimates of economic indicators. NSIs can shorten the publication lag in a straightforward way without compromising. Concrete advice on how to embrace the opportunities of nowcasting was provided. A range of methodological recommendations related to methodology being investigated for the purposes of nowcasting early economic indicators using traffic loop data was offered. NSIs can address a major quality issue, namely the timeliness, by using a range of micro-level data sources accumulated in the registers well before the official release is made, by employing large dimensional econometric models, to form an initial quick estimate of the target indicator. This does not necessarily lead to too large revisions, but adds significantly to the quality of official statistics through timeliness dimension.

What was gathered during the processing of data and the errors is that quite a lot of errors arise from specific characteristics of given datasets (metadata errors, weighting circumstances, editing) and cannot be generalized to the whole or at least to a big part of the big data field. In the case of traffic sensors, measurement errors mainly arise from sensors malfunctions and/or switching off. The solutions to most of these individual problems need to be specifically made for them. In the case of dealing with a large number of traffic loops, a more novel approach as an alternative for PCA should be considered for dimension reduction. The neural network is able to deal with missing and erroneous values, which makes it ideal for data cleaning and dimension reduction at the same time. By training subsets of road sensors in different neural networks, the dimensionality problem as described above is solved. The reduced dataset can then be used for regression purposes.

In some years enough quarters will pass to efficiently use RMSFE criteria for optimal model selecting on GDP. At that moment one will be able to accurately assess whether it is better to flash estimate directly GDP or if one should focus on rapid estimations of its less-timely components and then use them as regular data in the normal estimation process for GDP. In addition, it has to be investigated what precision is required in order to produce reliable earlier estimates, explore the possibilities to nowcast only some components of quarterly GDP (for which all data are not available on time), include other big data (and other) sources and also test some additional methods for nowcasting.

Given that the aim is to form a strategy by which one can forecast the near past and the present evolution of the target indicators, the quality in terms of timeliness and accuracy in capturing the target indicators' next values has to be evaluated. The standard measures of accuracy in these types of exercises are the Mean Absolute Error (MAE), and the Root Mean Squared Error (RMSE), but the workpackage also considered the Mean Error (ME), and the Maximum Error (MaxE). ME is an important measure because it indicates the unbiasedness of the estimate vs the target indicator. MaxE can be important to the NSIs because very large errors in a widely followed GDP series can be a cause for negative publicity. MAE and RMSE are the standard measures of deviations from the target (forecast errors), and RMSE gives relatively more weight to very large errors. A very useful point of assessing the accuracy dimension in the quality of nowcast is to compare the nowcasting accuracy to the revisions of the statistical office; in other words, to the standard that is acceptable by the NSI for regular dissemination. This is done by looking at the revisions in the published series (the final available value of the series growth rates vs the initial estimates by the statistical office).

It is expected that turnover indicators will become timelier and (what is remarkable) more accurate, which in turn makes it possible to produce the aggregate output indicators faster and more accurately by the existing national accounts procedures. WP 6 also provided a new methodology by which an early estimate of economic indicators can be produced reliably (quantitatively with similar accuracy as the official first estimates) by employing micro-level turnovers and traffic loop data together with various machine learning methods designed to handle such large dimensional data. These results have several implications. Firstly, the statistical office can easily meet requirements for timelier delivery of the flash GDP. Secondly, one can potentially construct real time and possibly high frequency (e.g. daily) estimates of the economic output. Thirdly, the traffic loop data should be inspected further, for instance by exploring clusters of the measurement points around designated areas (borders, manufacturing clusters, mining fields, etc.). Further, the explanatory power of traffic data should be examined in relation to the turnover indicators or other specific sectors.

Last but not least, this line of work can proceed in multiple directions:

- Other data sources can be explored with the methodologies presented and possibly in relation to other indicators.
- Other modelling frameworks are possible.
- A real-time application can be programmed, especially relying on the traffic loop data.
- These methodologies can be implemented into the production systems of official statistics, and their added value is not limited to nowcasting the GDP or some aggregate final indicators, but could be explored in order to impute some missing components of the aggregated figures. Such approaches can be easily implemented across the entire European Statistical System.

## 2.7 Multi Domains

These were the general goals of WP7 Multi domains:

- The aim was to investigate how a combination of big data sources and existing official statistical data can be used to improve current statistics and create new statistics in statistical domains.
- The workpackage focussed on the statistical domains: Population, Tourism/border crossings and Agriculture.
- The workpackage team described the data collection, data linking, data processing and methodological aspects when combining data in statistical domains.

Under SGA-2 WP 7 carried out experimental work on combining various sources in the statistical domain and between. Among all WP 7 pilots, there is a variety of data sources. Several different use cases have been tested, including data sources such as business registers, web scrapped data, traffic loops, satellite data as well as Google Trends or Google Traffic.

The pilots are within the following statistical domains:

- Population:
  - Life Satisfaction by Twitter(GitHub)
  - Life Satisfaction by Facebook(GitHub)
  - Morbidity areas and personal well-being with Google Trends by ONS UK (wiki)
  - Estimating daily and night population with Google Traffic(wiki)
- Tourism:
  - Tourism accommodation establishments (data sources—various portals)
  - Border movement (road traffic, air traffic, train traffic)
- Agriculture:
  - Crop types identification by PL
  - Crop types identification by IE

The results of the pilots can be summarized as follows:

- There are not so many cases in Population domain that can be implemented according to the rules of the data quality and accessibility.
- Two use cases in Tourism domain can be implemented with success – one of them can be treated as a replacement for the traditional survey – tourism accommodation establishments, the second can be treated as a supplement for traditional survey – border movement.
- The implementation of the border movement pilots relies on the data availability on road traffic sensors. The data on air traffic is easily available.
- The implementation of tourism accommodation establishments is based on the availability of web data on popular accommodation portals.
- The agriculture domain has one successful project – crop types identification that is compliant with in-situ survey on crop types. It can be used to replace the surveys on crop types in agricultural region. It is based on satellite data.

- The population domain in big data is strongly related to the social statistics – this is about estimating life satisfaction according to the human sourced information – comments, posts and blogs in a specific term.
- The pilots on Life satisfaction within Population domain have been shared on GitHub repositories and are easy to implement by other countries – they are using Machine Learning algorithms to identify the sentiment.
- Pilot on Tourism accommodation establishments can be implemented with the use of web scraping methods.

WP 7 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 there were 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). The report for this workpackage can be found here:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/0/04/WP7\\_Deliverable\\_7\\_7\\_2018\\_05\\_31.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/0/04/WP7_Deliverable_7_7_2018_05_31.pdf)

## 1. Population use cases

In the **Population** domain the goal of use cases was to show the structure of population in different regions according to the specific facts – e.g., public opinion on a topic (in the pilot Brexit) and life satisfaction in different regions.

Therefore, three use cases have been conducted. One of them was to identify the scale of the depression in different countries based on the Google Trends. The second was to find social mood according to public events or facts (e.g., Brexit). The goal of the third use case was to identify life satisfaction in population according to their comments/posts/tweets.

Preparing the use cases in Population and conducting pilots provide the following conclusions:

- There is still little information available on the Internet to provide reliable data in population domain,
- The structure of population to be analysed is strictly related to social media users or website users,
- The only way to conduct pilots of population is to select specific group of users (e.g., Twitter users) and make analysis of its structure by regions/gender and different aspects of life satisfaction,
- Although the data source is not representative for the whole population, the results are very promising in terms of accuracy of machine learning algorithms,
- With current data sources, social surveys can be supported, such as EU-SILC to provide information on selected life satisfaction,
- The value added of the pilots in population domain is that one can do it every day to analyse the causes of changing moods/opinions/life satisfaction,

- The use cases on population/life satisfaction are easily replicable to other countries as the source code is publicly available on the Internet – the most time-consuming work is to prepare a training dataset,
- EU-SILC results are different than the ones received by social media sentiment analysis, probably due to the coverage and representativeness issues,
- Negative things appear faster in social media than positive – the fake news spread very fast,
- People are more willing to comment on news when they are angry.

Future perspectives:

- Social media is very useful to find the sentiment of the population – this use case should be developed rather to provide information on whether people are angry or not, than to estimate the number of population.
- One should not concentrate just of Facebook or Twitter because they may not exist in the future.
- EU-SILC classification must be revised and for the social media sentiment analysis one should try to find the best to identify the sentiment in reliable way.

## 2. Tourism use cases

In the **Tourism** domain as a result of the inclusion of Poland in the Schengen Information System, on 21 December 2007 the Border Guard discontinued the registration of movement of persons and means of transport at the borders with the countries of the European Union. New regulations on customs clearances resulted in the loss of data on border traffic at the Polish borders with EU member states including land, air and maritime border. Therefore, methodological work has been undertaken in order to develop methods for collecting the missing data on border traffic.

Border traffic and tourism seem to be perfect statistical domains to be combined with big data. The preliminary research shows huge potential and usefulness of big data.

Under WP 7, the team from Statistics Portugal decided to extract reviews from tourism related web portals and compare the results with official statistics from tourism surveys. The main benefits and added value for official statistics would be to provide timely insights of the phenomena that are difficult to measure through traditional surveys.

Conclusions:

- Data on road traffic intensity may decrease the burden of interviewers because data on border crossings can be estimated,
- Data obtained has a better spatial granularity than from sample survey,
- Great care needs to be taken in choosing sensors with respect to their location, elsewhere data may be heavily biased,
- Missing data imputation needs a completely non-standard approach. A cross-entropy-based approach seems to be relevant,
- Big data on air traffic may increase countries coverage for which statistics are produced,
- Estimates are much more stable and distributions are smoother than in sample survey,

- Nevertheless, aircraft schedules cannot be used solely or even with administrative data. Some crucial statistics still need to be obtained from sample survey.
- Big data produces much more Europeanised trips distribution rather than country-specific. Therefore, distributions derived from a big data approach need to be integrated with distributions derived from sample survey which requires relevant methods,
- Train traffic data has a potential to be used to enhance domestic tourism statistics but it is a much more complex task than other presented cases.

Big Data can be implemented in the Integrated Survey of Tourism conducted by Statistical Office in Rzeszów. Road sensor data from Lithuania has been already implemented in border traffic estimation since 2018.

Next steps can be undertaken in Big Data usage research. From the experience of this study future direction of research are:

- Full integration of big data and sample survey. Estimation of impact of big data on all variables in sample survey – not only number of trips,
- Develop methods for combining distribution received from big data and distribution received from sample survey.

### 3. Agriculture use case

In the **Agriculture** domain, the purpose of the work was to use big data in the mechanism for recognizing selected crops in Poland on the basis of satellite images, administrative data and field surveys (in-situ). The remote sensing data came from Sentinel satellites operating under the European Copernicus program. A time series of satellite data representing the development of crops in 2017 were used, which were processed and then classified. The reference data were databases on crops that were collected during in-situ studies and information from administrative systems, including: handling the direct payments to farmers - Land Parcel Identification System (LPIS).

During the project integration of administrative data, field surveys and satellite images were used in the scope of:

- Source data - preparation of a training set - combining georeferenced vector data with satellite and in-situ data to prepare machine learning algorithms;
- Result data - the ability to publish results from satellite images and combine them with data from administrative data, e.g. in areas where there is no information in administrative sources.

This is not a registration number or other descriptive attribute of the parcel but georeference (spatial reference) of these data is the main attribute (data linkage) for combining administrative data with satellite, while combining spatial data from different sources is called as data fusion. Unfortunately, there may be even several different crops on one record plot, which prevents their full reference mapping with remote sensing data, hence the use of the image segmentation method. The biggest problem encountered during the analysis of optical data from the Sentinel 2 satellite was cloudiness. In the context of the optical imaging of the terrain surface, it is important to use every observation opportunity. In principle, this should translate into an increase in the number of cloudless observations.



## Results:

- The general distinction between classes is good, but in some cases the separation of morphologically similar plants is a problem (spring wheat vs. spring triticale).
- The use of the multitemporal NDVI mosaic for the autumn 2016 - spring 2017 period allowed a better distinction of winter and spring cereals.
- The machine learning accuracy varies from 80% to 97%. A detailed analysis of the causes of differences in the machine's learning accuracy is required, most of which result from the differences in field sizes in individual regions of Poland, which affects the accuracy of machine learning of the system.
- Segmentation carried out in both radar and optical images is better than that carried out only in radar images.
- When starting work related to the recognition of crops using satellite data, each type of crop should be represented by about 50 points in the field survey and have a relational mapping in the administrative data.
- Due to the specificity of radar data processing, the field of field surveys should be defined in accordance with the distribution of picture frames, not administrative borders.
- The addition of previous autumn images to the series of data allows for a better distinction between spring and winter crops and improves the overall accuracy of the classification.
- Due to the spatial resolution of the Sentinel-1 data, it is not possible to correctly classify arable fields smaller than 60x60 m at this time.

## Data combining:

- Connecting data at the source stage - administrative and in-situ data as well as result - recognition of crops, requires time-consuming and advanced system work, which is associated with a significant load of IT systems with big data.
- The use of data fusion satellite and administrative data allows to increase the accuracy of published results.
- The combination of various remote sensing and administration data allows in individual cases to eliminate errors in administrative data and improve their quality.
- The fusion of satellite and administrative data creates the possibility of preparing and developing the estimated and resulting data much earlier than in comparison with the results of statistical surveys carried out in the field of cropping areas.

## 4. Conclusions

The pilots conducted in WP 7 resulted in many benefits to NSIs. These include:

- It is possible to have accurate data from satellite images with the high accuracy,
- The rules of using API's and processing the data from social media, such as Facebook and Twitter, are known,
- It is possible to get the current data from road government authorities from road sensors for different countries and it is known how to process them when there is no data,
- One can get data by web scraping alternative data sources, such as air and train traffic,

- The participants have better skills on machine learning regarding sentiment analysis and satellite data processing.

There are still a few challenges:

- The results show that it is not possible to support population use cases (e.g., counting population) because the data sources available at this moment are not representative,
- Web scraping tourism accommodation establishments makes it possible to identify all the objects, but the classification used by websites for tourism is not unified (e.g., hotels vs. pensions),
- Constantly working on improving algorithms is necessary as the data sources are not stable and their structure and format may change (e.g., resolution of satellite images).

To summarize:

- WP 7 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).
- 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.
- Other two pilots in Population domain are related to the selected health status of population (Morbidity areas use case on depression (Google Trends) by ONS UK) and to Peoples opinion/interest by topics based on websites by ONS UK by Facebook.
- 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 and Traffic Loops data.
- The agriculture domain has one pilot prepared by two different methodological approaches: first was prepared by Statistics Poland and the second was prepared by CSO Ireland.
- For data combining two different approaches were carried out – intra-domain data combining (each domain) and inter-domain data combining (agriculture-tourism).
- The results of machine learning algorithms for agriculture domain have an average accuracy of 80% (the percentage of positively identified crop types), while in pilots for population domain also 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.
- 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.

## **1.8 Methodology**

The aim of this workpackage was laying down a general foundation in the areas of methodology, quality and IT infrastructure when using big data for statistics produced within the European Statistical System. WP 8 therefore started with a workshop in which the most important topics in the area of IT, quality and methodology were identified. Next an overview of important papers, project results, presentations and webpages relevant to the application of big data for official statistics was created. This overview was linked with the findings of the pilots in the first phase (SGA-1) and the input available from the second one (SGA-2) of the ESSnet Big Data. The main outcome of WP 8 was an overview of the methodological, quality and IT findings when using big data for official statistics. Experiences obtained both in- and outside the ESSnet Big Data were used as input for WP 8.

### *Task 1 - Literature overview*

The aim of this task was creating an overview of the findings described in the products of WP 1 to 7 of SGA-1 to be used as input. Any SGA-2 findings and findings of other projects were also included. In addition, a literature study in which the most important papers, project results, presentations and webpages (such as blogs) relevant to the application of big data in the context of official statistics were performed. Also a workshop was organized for which Big Data / Data Science experts, one for each partner involved in SGA-2, were invited. In the workshop, the important methodological, quality and IT related topics in the area of using big data for official statistics were identified, discussed and prioritized.

### *Task 2 - Quality of Big Data*

Since various big data sources will be used to produce official statistics it is important to determine which quality aspects of big data are essential for this use. This part of WP 8 focused on the quality related observation of the actual big data studies identified in Task 1 of WP 8. The product of this task was a document in which the important quality considerations, findings, detection methods (such as visualizations) and solutions (if available) are described.

### *Task 3 - Big Data and IT*

The link with IT is described in a report providing an overview of the infrastructures and processes applied in SGA-1 and SGA-2 of the project and those identified in Task 1. By focusing on the overall process, it was assured that big data specific IT-infrastructure were not discussed on their own but always in the context of the applied production process. This task included an inventory of the technical skills needed.

### *Task 4 - Big Data Methodology*

This task focussed on the production of a report in which the methodology and methodological challenges when using big data for official statistics were described. In this report the major challenges for future research were also included. The results of Tasks 1 and 2 were important input for this deliverable.

This resulted in the following products:

1. [Results of the workshop](#)
2. [Literature overview](#)
3. [Quality aspects of Big Data for official statistics](#)
4. [IT-infrastructure used and the accompanying processes developed and skills needed to study or produce Big Data based official statistics](#)
5. [Methodology of using Big Data for official statistics and the most important questions for future studies](#)

## 1. Results of the workshop

On the 25th and 26th of April 2017, WP 8 organized a workshop at the Statistics Netherlands location in Heerlen. During these two days, a group of eighteen people identified the main topics in the areas of Methodology, Quality and IT when using big data for official statistics in the context of WP 1-7 of the ESSnet on Big Data. The results of the workshop were three related lists. The list of Methodology and IT contain eleven identified issues and the list of Quality contains seven issues for Quality. The work of the remainder of WP 8 focused on these identified issues. The lists and their interrelations are shown in the table below.

| IT  | Quality                      | Methodology                              |
|---|------------------------------|--|
| Big Data processing Life Cycle ↔          | Comparability over time ↔    | Changes in Data Sources                  |
| Data source integration ↔                 | Linkability ↔                | Data linkage                             |
| Metadata management                       | Coverage ↔                   | Unit identification problem              |
| Format of Big Data processing             | Process chain control ↔      | Secure multi-party computation           |
| Datahub                                   |                              | Data process architecture                |
| Choosing the right infrastructure         | Model errors and precision ↔ | Inference                                |
| List of secure and tested API's           | Measurement error            | Assessing accuracy                       |
| Shared libraries and documented standards | Processing errors            | What should our final product look like? |
| Data-lakes                                |                              | Deal with spatial dimension              |
| Training/skills/knowledge                 |                              | Machine learning in official statistics  |
| Speed of algorithms                       |                              | Sampling                                 |

*Table 1: overview of the issues identified during the workshop for each of the three areas and their interrelations*

## 2. Results on literature overview

The aim of the literature review was to classify and provide an overview of papers relevant for the application of big data in official statistics. The variety of the papers included reveal different statistical domains for which big data sources can be used.

The literature overview includes a short characteristic of each paper with respect to the following categories: data sources, domains, keywords. Each paper has full bibliographic data and a link (if possible) enabling the reader interested in the paper to quickly access the full content. However, some of the papers included are restricted and cannot be accessed without a subscription. Since many official statistical offices have a subscription to the most important digital libraries, it will not be an issue to access these papers.

| SPECIFICATION                                 | DESCRIPTION  |
|---|--|
| <b>Bibliographic data</b>                     | AAPOR (2013): Report of the Task Force on Non-probability sampling, June.  |
| <b>Link</b>                                   | <a href="https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_T_F_Report_Final_7_revised_FNL_6_22_13.pdf">https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_T_F_Report_Final_7_revised_FNL_6_22_13.pdf</a>  |
| <b>Short overview (strengths, weaknesses)</b> | The report shows different aspects on non-probability sampling, including sampling matching, network sampling, estimation and weight adjustment methods or measures of quality. Therefore, the paper is a good start to work on data sources with non-probability sampling. However the paper does not provide a complete framework to prepare non-probability sampling, it is rather a discussion on the topic. |
| <b>Data sources</b>                           | Social networks, surveys   |
| <b>Domains</b>                                | Population   |
| <b>Keywords</b>                               | Non-probability sampling, data quality   |
| <b>Classification (A – very relevant, B</b>   | C  |

*Table 2: example of the information in the literature overview*

The target audience of the literature overview are official statistics employees interested in developing a big data project or data scientists looking for new opportunities and exciting projects. The literature overview is available on the wiki since December 2017 as a formal deliverable. However, the literature overview is a living document, so new articles were added during the remainder of the ESSnet Big Data project.

### **3. Results on Quality of Big Data**

In this report quality is discussed in the context of big data. Topics included were identified in a WP 8 workshop during which seven quality aspects were listed as the most important ones when using big data for official statistics (in the context of WP 1-7 of the ESSnet on Big Data). The quality aspects identified are: Coverage, Comparability over time, Processing errors, Process chain control, Linkability, Measurement errors and Model errors and precision.

It is worth noting that the WP 8 discussion of quality aspects is based on the experiences of pilots, not on products already declared as official statistics. As a result, the approach might differ a bit here: with official statistical products, it is clear from the beginning which quality dimensions have to be met. Working with pilots involving big data sources, the approach differs here as a more data driven approach is followed. The workpackages explored new data sources as well as possibilities to use them. The consideration of quality aspects (at the output side) was not the main focus from the beginning, but happened more in the course of exploring and using the new data sources, or happened even only retrospectively.

The UNECE quality framework is the most well-known quality framework so far, focussing specifically on the quality of statistics based on Big Data sources, is the Big Data Framework by the UNECE from December 2014<sup>2</sup>.

In Table 3, the quality aspects in relation to the structure (the three phases as well as the three hyperdimensions) of the UNECE's quality report are contextualized:

- The quality aspects of this report were assigned to one (or more) phases of the business process.
- With the help of colours it was illustrated to which hyperdimension the considered quality aspect (mostly) corresponds to.

Please note that "Coverage" as well as "Comparability over time" are listed both in the Input as well as in the Output phase. Further, "Comparability over time" affects all three hyperdimensions, because the availability of the source, the metadata as well as the data itself can change over time.

| Input                   | Throughput                 | Output                  |
|-------------------------|----------------------------|-------------------------|
| Coverage                | Processing errors          | Comparability over time |
| Measurement error       | Process chain control      | Coverage                |
|                         | Model errors and precision |                         |
| Comparability over time |                            |                         |
|                         |                            |                         |
| Linkability             |                            |                         |

|                          |
|--------------------------|
| Hyperdimension: Source   |
| Hyperdimension: Data     |
| Hyperdimension: Metadata |

*Table 3: The seven quality aspects in the business process and with relation to the hyperdimensions "Source", "Data" and "Metadata"*

What the 7 quality aspects have in common is a clear relation with either one or more causes of error (Coverage, Processing errors, Linkability, Measurement errors and Model errors and precision) or the need to detect and deal with changes in the composition of the source (Comparability over time and Process chain control). In the various WP's in the ESSnet Big Data many causes of error were found. Some of them are unique, such as the effects of the scrambling of the Automated Identification Signal of ships in WP 4 and the coverage issues for job portal vacancies in WP 1. These clearly indicate the need for new (big data specific) checks and correction methods. These findings also indicate the need to develop or update a quality framework for big data sources. Regarding quality frameworks it should be mentioned that the Code of Practice for European Statistics (CoP) is currently reviewed to be fitter for modern production of statistics. One of the main aspects of the revision is the potential use of so called new data sources including big data.

<sup>2</sup>

<https://statswiki.unece.org/display/bigdata/2014+Project?preview=%2F108102944%2F108298642%2FBig+Data+Quality+Framework+-+final+-+Jan08-2015.pdf>, accessed March 3rd, 2018.

In addition, big data sources may also change over time for which a number of causes were identified; mainly related to changes in the composition of the data source. These causes affect the comparability over time and require the need to track those changes. Causes identified were technological changes, changes in the policy of the data holder and changes in the population composition and/or amount included. The need to check and control these causes and their effect on the entire chain is part of Process chain control. Because of the large volumes of data involved, it is important to use efficiently implemented quality indicators or predictors. For a data driven process this introduces the need to not only focus on the quality of the output but also on the quality of each individual step in the process chain.

Overall it can be concluded that some familiar and some new quality aspects have to be considered for big data sources which urges the need for the development of an extended quality framework. The experience gathered in the ESSnet Big Data, and especially in this WP 8, shows that on an abstract level, the same quality aspects can be considered for different big data sources, but the diverse nature of big data sources can make it very difficult to apply standardized quality measures for different big data projects. It is also important to emphasize that for new big data projects and for new big data sources, other quality aspects than the ones listed in this report can be decisive.

#### **4. Results on Big Data and IT**

The results of this report are a description of the main topics identified in the area of IT when using big data for official statistics. Looking upon the issue of using big data for official statistics it is clear that the available IT infrastructure determines the way by which quality and methodological issues will be dealt with. The goal of this report was to expand on the topics identified: big data processing life cycle, metadata management (ontology), format of big data processing, data-hub and data-lake, data source integration, choosing the right infrastructure, list of secure and tested API's, shared libraries and documented standards, speed of algorithms and finally training/skills and knowledge. These topics were linked to the work performed in WP1 -7 of the ESSnet Big Data. It is also a valuable resource of information for anyone, statisticians in particular, interested in the essential IT-aspects of Big Data.

The most important finding on big data and IT are:

1. A big data process for official statistics can be characterized at four different stages: collect, process, analyze and disseminate, this sequence was mostly used by big data pilots such as traffic intensity, web scraping enterprise characteristics or AIS data.
2. There is no unified framework for Metadata Management. One can rely on the Big Data Quality Framework by UNECE but for metadata purposes, a better way is to apply the Common Metadata Framework by GSBPM and GAMSQ.
3. The format of big data sources in official statistics is various but some common data files can be listed, such as semi-structured CSV files (the most common format) and JSON files (for storing processed data), structured relational databases and unstructured NoSQL-like databases.
4. Data-hubs and Data-lakes mostly refer to the environment created in Apache Hadoop cluster or NoSQL database. A good example is the Sandbox created by UNECE for big data purposes. None of the ESSnet Big Data workpackages used a data-lake. Data-hubs were used to access smart meter data by WP 3.

5. The Variety of big data sources makes it difficult to integrate them directly. Different data sources (e.g., structured vs. unstructured) use different processing techniques and as a result, it may be difficult to find an attribute to link them during processing phase.
6. The right infrastructure for big data processing must be characterized by the following attributes: linear scalability, high throughput, fault tolerance, auto recovery, programming language interfaces, high degree of parallelism and distributed data processing. In the ESSnet Big Data the variety of platforms used included Linux and Windows as a landing zone and Java, Python, R, Spark and SAS for processing purposes.
7. Several API's (Application Programming Interfaces) were used by ESSnet Big Data workpackages: Twitter API, Facebook Graph API, Google Maps API, Google Custom Search API, Bing API, Guardian API and Copernicus Open Access Hub to access public available data.
8. There is a list of GitHub repositories that contain software to start with big data in official statistics. It includes: Awesome Official Statistics software, ONS (Office for National Statistics) UK Big Data team, ONS (Office for National Statistics) UK Data Science Campus and ESTP Big Data course software.
9. Using the right algorithm will allow to increase the speed of data processing. It is especially important when processing streaming data.
10. The best way to become a data scientist is to start with improving the skills in the area of negotiating with data owners, set and maintain infrastructure, data combining, checking and editing large amounts of data, analyzing large amounts of data, using statistical methods including machine learning and visualize the data. Most of these skills can be improved with ESTP courses – European Statistical Training Program.

The general conclusion of the report is that there is a common set of tools, methods and libraries used by ESSnet countries and shared, e.g., Python and R language, various API's and machine learning algorithms. Moreover, several GitHub repositories with statistical software exist where countries can use and give feedback on different tools, such as software to detect the presence of an enterprise in social media or its ecommerce activity.

Exchanging these experiences allows us avoiding common problems regarding planning and developing further big data projects. These results will be a good support for any further big data projects that can be conducted in official statistics.

## **5. Results on Big Data Methodology**

At the start of the workpackage the most important topics related to big data methodology were identified. These topics are: What should the final product look like?, Data process architecture, Changes in data sources, Deal with spatial dimension, Unit identification problem, Sampling, Data linkage, Secure multi-party computation, Machine learning in official statistics, Assessing accuracy and Inference. This diverse set of topics either aims to create the best results achievable from the data available (Deal with spatial dimension, Unit identification problem, Sampling Data linkage, Machine learning in official statistics, Assessing accuracy and Inference,)), aim to deal with changes as good as possible (What should the final product look like?, Data process architecture, Changes in data sources) or want to exchange data in the most secure way (Secure multi-party computation). For each of these tasks, methods need to be available or developed to achieve them. The overview of the methodological work performed in each workpackage of the ESSnet revealed a number of



examples belonging to each topic with the exception of multi-party computation. The examples of methods for the other topics are listed in Table 4 to which other relevant applications in other NSI related big data research are added. As such, this table provides an overview on familiar and new methodology in this exiting area of statistics.

Data process architecture is not included in Table 4 because it lays the foundation of the comparison of the other topics and relates big data processes to that of other processes in an NSI. Be aware that this process view is mainly used to aid the reader here. Therefore, to help the reader to get more grip on the methodological topics identified, the topics are related to the various steps in a big data based statistical process. For this process, the Generic Statistical Business Process Model (GSBPM) is used. According to this model, a Big Data process is composed of 4 steps: Collect, Process, Analyse and Disseminate. The Collect step is all about obtaining data, the Process step is about processing data and quality checking, the Analyse step is about estimation and the Disseminate step is about producing output. Based on the experiences in WP1-7 of the ESSnet, the topics included in Table 4 are assigned to one or more of these four steps. The assigned topic of each step is indicated in Table 1 between brackets. Below each topic, different kind of uses are listed, for which various methods are needed. Many of these methods are new for official statistics. However, beware that new in this context may simply indicate methods not yet familiar to official statistics but available (and developed) in other areas of research.

No workpackage reached the full cycle of big data statistical production yet, i.e. from data access through data processing to data dissemination. However, some workpackages are close. WP 2, WP 3 and WP 4 are examples where big data based methods emerge and start to become fully developed. However, these methods still have to be flexible enough, since big data is a product of technological development. Technology development underlies the variety and dynamic of data sources. To be in line with technology NSIs using big data need to cope with many dynamic data sources, like for example Facebook.

It is important to realize that big data sources are scattered across many private and government agencies. To ensure access to data and produce big data based statistics NSIs have to build partnerships network with both public and privately owned organizations. It is essential to ensure stable, long lasting, access to data to enable the production of big data based statistics that also enables the development of generic big data methodology.

**What should our final product look like? (*Disseminate*)**

Especially important in data-driven studies  
Relevant for all WP's (such as WP6, WP7, ...)

**Changes in data sources (*Collect, Process*)**

- Especially important for relative new data sources (such as social media)
- Less relevant for data sources with (high quality) international standards (e.g. CDR, AIS, ...)

**Deal with spatial dimension (*Process*)**

- Used to identify populations (WP3, WP7, social media NL)
- Used to derive routs of ships in WP4
- Basis of satellite study in WP7
- Work of FlowMinder

**Unit identification problem (*Process*)**

- Identifying which part of the target population is included (WP1, social media NL)
- To distinguish mixed populations (WP3 business's and persons, social media NL)
- Deriving background characteristics of units (WP7 social media, social media NL)

**Sampling (*Collect*)**

- Draw samples of Big Data in exploratory studies
- Considering Big Data as a non-probability sample (PhD NL)
- To compare Big Data variable values with those in target population (WP1)
- To compare population composition in Big Data and target population (WP5)

**Data linkage (*Process*)**

Combining at three different levels:

- Location/area: Geolocation data, address and buildings (WP3, FlowMinder)
- Unit: Companies + URL's (WP2), Companies + job adds.(WP1)
- Period: GDP and traffic intensity (WP6), Consumer Conf. + Social media sent. NL)

**Machine learning in official statistics (*Process, Analyse*)**

Can be applied for processing and for estimation:

- Processing (social media NL)
- Estimation (WP1, WP2, WP3, WP6, WP7)

**Assessing accuracy (*Analyse*)**

- Deal with bias (WP2, WP4, ...)
- Deal with variance (WP6, ...)

**Inference (*Analyse*)**

Can be Survey based, BD census like (complete coverage), or partly complete BD population

- Survey based (~WP2, WP7 Social media, Consumer Conf. + Social media sent. NL)
- Census like (WP4, road sensors NL) - Partly complete (WP1, WP3, WP5, ...)
- Partly complete WP1, WP3, ...

*Table 4: Overview of the methodological topics and there application in WP1-7 and other Big Data areas. Secure multi-party computation and Data process architecture are not included here.*

Another nice result of the report of Big Data Methodology is an overview of the general approach followed by an NSI that wants to include Big Data in official statistics. These steps are:

1. Get access to Big Data (BD)
2. Perform an BD exploratory data analysis study (including a privacy assessment)
3. Study the objects (units/events) in BD and check if events need to be converted to units for the foreseen application
4. Compare the coverage of the objects in BD to those of the target population of the NSI
5. Study the variable(s) of interest in BD and compare these with those needed by the NSI (variables may be combined and/or processed here; e.g. creating features)
6. Compare the development over time and/or per area of the variable(s) of interest in BD with similar variables included in any other survey or register based results (if available)
7. Check the performance of various models and/or machine learning based applications on improving the relation described in the previous step
8. Determine the effect of any assumptions, short-cuts made, and/or quality issues and corrections on the comparison described in the previous step (may need to restart at step 3-6)
9. In case of any positive findings, check the reproducibility and stability over time of those results
10. Produce a first (beta-)product

In principle, this is an overview of (data-driven) big data methodology. After each step a go/no go decision can be made whether to proceed to the next step. The list is a starting point and will undoubtable form the foundation of new and exciting future developments in the area of big data research.

### **3. Issues encountered**

#### **3.1 General issues**

##### **1. Planning and related issues**

The planning of the deliverables and milestones of SGA-2 is given in the form of a Gantt chart on page 10 of Annex II of the agreement of SGA-2.

The planning proved to be realistic. All deliverables and milestones were realised and there were only minor delays. Moreover, without exception all deliverables were realised in their final version before the end of the project (end of May 2018)<sup>3</sup>. There were a few changes in the foreseen contents of deliverables and in the foreseen contribution by the partners, especially concerning WP 4 and the contribution of Denmark, but this could be solved without imperilling the planning, and in full agreement with Eurostat and the partners concerned. (These changes were described in the previous chapter.)

More details on the realisation of the planning of the specific workpackages will be provided in the Final Report on the Implementation of the Action, due 60 days following the closing date of the action.

The issues that occurred in carrying out the action of SGA-2 for the specific workpackages are discussed in section 3.2. There were no major cross-cutting issues, and the CG meetings proved to be effective in solving all matters that affected multiple workpackages.

##### **2. The Review Board**

As mentioned in section 1.3, the Review Board was very effective. However, there were two issues. The first is the bottleneck in workload that naturally developed towards the end of the project, as all workpackages finalized their last deliverables during the last months of the project, and these all had to be reviewed. This issue was solved by agreeing on an overall project review planning for the last half year of the project – and by the flexibility and efforts of the Review Board. The second issue is the fact that the Review Board worked on an unpaid, volunteer basis, whereas the required efforts were considerable. On the longer term, this construct is not desirable and sustainable, especially since the work of the Review Board in the ESSnet proved to be an important success factor.

##### **3. Use of the Sandbox<sup>4</sup>**

The Sandbox is a shared platform for storage and computation of big data, hosted and managed by ICHEC (Irish Centre for High-End Computing). The Sandbox is one of the outcomes of the HLG Big Data project, carried out in 2014 and 2015 and facilitated by UNECE. The use of the Sandbox as a training and collaboration platform was successfully tested during the project and after the end of the project, an agreement with ICHEC to grant the use of the Sandbox to organizations on a

---

<sup>3</sup> Only the current report, i.e. the Final Technical Report, was finalized after the end of May, since this could only be written after the BDES 2018 conference in Sofia, May 2018, and the finalization of more than ten deliverables and milestones in that same month.

<sup>4</sup> This section was kindly written by Antonio Virgillito, who is the Sandbox officer for the ESSnet.

subscription basis. The main characteristic of the Sandbox is the possibility to share datasets and work with big data tools without any software installation and configuration. The tools included in the Sandbox are accessible remotely from any computer simply through a web browser and do not need special software or hardware requirements.

A subscription to ICHEC for the use of the Sandbox was acquired by the ESSnet project. The subscription gives the possibility to all the project participants to create an account in the Sandbox to upload/download datasets and use the tools for analysis.

A special section in the project wiki is dedicated to the Sandbox, with instructions on access and use and documentation for all the tools:

<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Sandbox>

Whereas four workpackages made use of the Sandbox in SGA-1, only WP 4 used it in SGA-2. For WP 4 a sample of the European dataset concerning ship positioning data from land-based stations was uploaded in the Sandbox. There is a sample of satellite data from the Mediterranean Sea available on the Sandbox as well. Both datasets are available for all the users and they represent six months of observations of AIS messages from ships in European waters.

### **3.2 Issues at the level of the work packages**

#### **1. Webscraping / Job Vacancies**

##### *Workpackage*

WP 1 had six partners during SGA-1. A further four countries joined for SGA-2, although Denmark pulled out very early on. Although it was good to have a large group of interested countries, it was often difficult to coordinate the contributions from so many countries. This was exacerbated by the fact that the OJV landscape in each country was different and so the approaches taken and lessons learned from one country were not always easy to transfer to others.

In addition, the level of skills and experience between different partners was somewhat uneven. For example, several partners were learning about and applying machine learning approaches for the first time. Although there was a deliberate strategy of the ESSnet to widen participation as far as possible, this possibly came at the expense of making more progress addressing the methodological challenges of using OJV data. For future ESSnets it might be useful to have a clearer separation between the aims around methodological development and aims of developing capability.

##### *Staffing*

Apart from Denmark who had to withdraw from WP 1 due to the departure of a key member of staff, other countries also experienced the loss of key staff during the pilot including Germany, Sweden, UK, and Slovenia. In some instances, it took many months for key staff members to be replaced and in some cases staff were not replaced at all.

In some cases, this meant that a lot of the development work occurred in a very concentrated period or was rushed and under developed. In some cases, key staff seemed to have too many commitments, including working on other ESSnet workpackages. In general it seems that the specific data science skills required to work with big data are still quite scarce within the ESS and that this is impacting on the ability to make the progress expected.

#### **2. Webscraping / Enterprise Characteristics**

More and more data pertaining to individuals and enterprises are being placed on Internet. This is a significant opportunity; however, challenges exist in the several areas, i.e., ethics and legal framework, Internet as the data source, methodology for integrating big data for Official Statistics and the complete IT platform of handling the Internet data life cycle.

##### *Ethics and legal framework*

Extracting knowledge from online data draws attention and public concerns how governments are utilizing online data, including relatively uncontroversial cases such as NSIs utilizing textual data on company websites. GDPR also adds further requirements to the web scraping scenario. As a response to this, NSIs and Eurostat need to develop transparent web-scraping policies in order to allay public concerns about the data collected and the usage of them. The 'netiquette' developed as a part of this work-package is an important first step. It is remarkable that NSIs adopt different interpretations on the question. Sharing good experiences among NSIs need to continue.

### *Challenge of adopting Internet data*

Websites are created with various techniques and different standards. Some are even behind firewalls. Some are built with JavaScript, which make the content unseen to an ordinary scraper. To develop efficient scrapers, in the case of generic web scraping, intense web technique is necessary in order to handle the webpages' varieties. As the internet evolves, data in forms of audio or video files are increasing in volume; and the use of interactive or user-specific content increases. The new data formats set even higher requirements on the scraper and the storage. To make the web scraping approach lasting in the longer term, it is necessary to build up new competence, both for web scraping and for data storage at NSIs. Sharing software and knowledge is very important for competence building.

### *Methodological challenge*

As remarked in section 2.2, a methodological challenge is to finalize a quality framework of the statistics generated from web data.

### *Data management platform*

The web scraping system is separated from the ordinary systems at some NSIs for security reasons. The gap is inconvenient when transferring the scraped data from the system on one network to the system on another network, especially for big data. The data handling process is different from other types of data. The data are stored, not in the traditional relational database, but in a NoSQL data storage; then they are analyzed and extracted for loading into databases. It is a challenge to build up the pipeline to manage the entire data life circle i.e., storing data scraped, extracting, transferring and loading process and data analysis.

## **3. Smart Meters**

The biggest issues the partners have faced during the project are related to getting access to data, delays caused by the lack of knowledge and experience with big data and big data infrastructure

### *Management*

There was change of the workpackage leader in the middle of the project, which caused some discontinuity in the management style of the project. It also put some extra load to the Estonian team, which was responsible for coordinating the project.

### *Personnel*

There were no changes in teams, but due to vacancies in Sweden the progress of the project was disturbed. However, this could be overcome.

### *Data access*

During the project, only two countries had access to the data – Estonia and Denmark. The Danish dataset was fully updated during the second half of the project. A small set of anonymized Estonian data was made available for Austria. Sweden and Portugal got access to the sample data and Italy did not have access to the data at all.

## *IT Infrastructure*

To complete one of the main task of the WP in Estonia – classification – the software of Spark had to be updated by the IT department. Unfortunately, the process of update took more than three months and disturbed the progress of achieving the goals of the project. Due to the lack of proper infrastructure the preparation of 2016 data in Estonia took more than a year and therefore it was not possible to use the updated dataset during the project and the data from 2013 - 2015 was used.

## **4. AIS Data**

### *Staffing issues*

Very shortly after the start of SGA-2 in August 2017 the team member from DST (Olav Grøndal) left DST. Unfortunately, DST could not replace him because of staffing problems, leading to a withdrawal from DST from SGA-2. This issue was solved by distributing the available 51 working days and related tasks from DST in SGA-2 among the other participants in this WP. So, the promised deliverables could still be developed in time.

### *Getting AIS data by EMSA*

Free European AIS data from EMSA was not obtained. On this topic actions were carried out together with Eurostat. That is why the originally planned deliverable 4.7 could not be delivered, concerning the results of comparing the quality and coverage of the European AIS data from Dirkzwager and EMSA. There were no further consequences for the other deliverables in this workpackage.

### *Adjusted scope of SGA-2*

The original plan included developing a model to calculate emissions and getting access to data from the European Maritime Safety Agency (EMSA). However, at the start of SGA-2 the scope of SGA-2 was adjusted, in agreement with Eurostat, because good models for calculating emissions already existed and access to AIS data from EMSA could not be obtained. Instead of developing a model to calculate emissions existing models for deriving emissions were investigated and it was described how to implement these models in statistical processes. Other possibilities for improving current emission statistics by using AIS data were also investigated. Instead of comparing EMSA data with the European AIS source from Dirkzwager (DZ) in this WP, the quality of satellite AIS data from Luxspace (LS) to the DZ data and the national data from Greece was compared.

## **5. Mobile Phone Data**

WP 5 has found no major issues in the execution of the research plan on mobile phone data. However, two aspects can be mentioned to be taken into account both for the assessment of the current ESSnet and the execution of similar projects in the ESS in the future regarding new data sources (big data, smart statistics,...).

A lot of lessons have been learnt as a consequence of the present ESSnet regarding the use of mobile phone data in the production of official statistics. The original research plans were composed by the workpackage with noticeable lack of knowledge about concrete details regarding this data source, details now perfectly present in the conclusions and the prospects for the future of these activities.



### *Phasing issues*

Firstly, the original strategy based on the sequence of access, methodology, IT aspects, and quality issues has been proven to be not as efficient as originally expected. The access to this sort of data has been proven to be extremely hard not just for research purposes but especially for production (indeed, currently blocked). In the sequence, this puts the rest of phases into danger since they are proposed with access to real data as the basis. A change of strategy using simulated data in parallel to the work upon the access conditions and agreements has been proposed without forgetting about the hands-on bottom-up approach.

### *Production framework*

Another important lesson for this kind of project within the ESS is that it cannot be understood as a compendium of national initiatives somewhat amalgamated in the deliverable documents of the project. A novel production framework is needed with different elements and parts (new metadata, new methodology, new software tools, new quality indicators,...). The workpackage concluded from this experience that with these new data sources (at least with mobile phone data) the statistical production system needs the whole ESS as the actor, since the efforts by a single NSI would demand astounding resources.

## **6. Early Estimates**

### *Staffing issues*

Some of the issues encountered within this pilot are country specific. SURS has experienced difficulties in retaining staff, in particular data science specialists, who have been difficult to replace. SURS has lost two key team members: Boro Nikić, who was also WP 6 coordinator, and Vesna Horvat. Thus, the experience is that it is difficult to recruit and retain staff with strong data science skills.

### *Ambitions*

In the proposal of WP 6 ambitions were high in the sense that it was planned to carry out two pilots which would yield "quick wins". Besides the pilot related to nowcasting the turnover indices, the WP 6 team tried to work also on the pilot aimed at testing the possibility to estimate the sentiment indicators such is the Consumer Confidence Index (CCI) using data from social media (Facebook, Twitter, etc.). However, due to early findings where it was found out that access to a sufficient amount of data is impossible, this idea was abandoned. The problem was also access to the historical data.

Some countries, e.g. Italy and Portugal, which joined the project only in the second phase (i.e., SGA-2), did not have enough time to produce final results. Their results are to be considered still preliminary and experimental.

### *Communication*

WebEx for internal workpackage meetings is a good instrument but there is a dependency on two people. It would be helpful if workpackage leaders can arrange them to.

## **7. Multi Domains**

### *Staffing issues*

Due to the huge scope of work (three wide domains: Population, Tourism, Agriculture and work on many BD sources) at various stages of the work WP 7 needed a lot of people with specific skills and expert knowledge. Therefore, the number of substantive and organizational problems was incomparable with other WPs.

### *Use of the Sandbox*

Access to the Sandbox was not necessary at any stage of work. For experimental purposes the team were using their own server with dedicated software running on Apache Spark and Apache Hadoop. They were using Python language for pilot surveys.

### *Data access*

During this project WP 7 had to tackle with many various challenges such as: ethical and legal aspects of working on different sources. Analysis of these issues were described in the documents of WP 7 in the context of both: specific countries and various Big Data sources.

## **8. Methodology**

### *Staffing issues*

After having the starting workshop of WP 8, the contribution of the workpackage leader (Piet Daas) to this workpackage was under pressure, because of all the work he had to do for the Centre of Big Data Statistics (CBDS) at Statistics Netherlands. This caused a delay in producing the deliverables and milestones of WP 8. After rescheduling the deliverables and getting another WP leader, the workpackage managed to deliver all reports on time according to the new deadlines. Piet was still involved in and responsible for the contents in WP 8.

### *Timing issue*

The aim of WP 8 was laying down a general foundation in the areas of methodology, quality and IT infrastructure when using big data for statistics produced within the ESS. The results of the workpackages 1 to 7 were therefore important input for WP 8. As identified during the Co-ordination Group meeting in Brussels on 26 and 27 of October, there was a timing issue for the deliverables on WP 8. Workpackages 1 to 7 were all producing input for WP 8. The problem was that all workpackages delivered their results near the end of SGA-2. This meant there was little time for WP 8 to gather all information and produce the final deliverables for WP 8. Therefore, it was decided to interview all workpackage leaders of the ESSnet Big Data programme to know in an earlier stage what they were working on.

## Annex 1: Communication and Dissemination

WP 9 Dissemination had as overall objectives to facilitate communication and information exchange between project participants and to disseminate ESSnet Big Data results to all stakeholders and to the general public, mainly through a suite of websites and a high-profile dissemination conference.

### 1. Websites

#### Objectives and overall architecture

The suite of ESSnet Big Data websites has to serve several purposes simultaneously:

- **intranet** for coordinators, workpackage leaders and project participants, to exchange project management and financial information, contact information; calendars, meeting agendas, documents and minutes, reports under construction, background articles and references, software and other useful resources;
- **extranet** to disseminate the project's progress and results to all stakeholders: Eurostat (funding ESSnet Big Data), non-participants from the European Statistical System (ESS), anyone else in official statistics involved with big data and data analytics, international organizations, academia and private enterprises owning data or providing services;
- **one-stop shop**, beyond the ESSnet Big Data project. for resources and information about projects, events and results focusing on big data in official statistics.

To these ends, three seamlessly integrated websites were created and are continuously being managed and expanded:

- **ESSnetbigdata** (<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata>), a Mediawiki-based wiki which all project participants can edit and which anyone else may freely consult;
- Its **mirror site** ([https://ec.europa.eu/eurostat/cros/content/essnetbigdata\\_en](https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en)) on the Eurostat Cros Portal, the official dissemination channel for ESS projects, consisting only of the upper navigation layers and redirecting for content to Essnetbigdata;
- **ESSnet Big Data** (<https://webgate.ec.europa.eu/fpfis/wikis/display/EstatBigData/ESSnet+Big+Data>), a restricted Confluence wiki for personal, financial and confidential information, only accessible to project participants.

These are complemented by special-purpose satellites:

- the **Sandbox** (<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Sandbox>) for testing;
- a **GitHub repository** (<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/GitHub>) for storing and sharing code.

#### Actions during the course of SGA-II

The Essnetbigdata wiki was expanded considerably and now consists of some 1200 web pages and 800 uploaded files (see <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Special:Statistics>). It is intensively used, both in terms of production (with 55 'editors' allowed to make changes, 10-15 of whom are active at any given time, and almost 8000 saved edits until now) and consultation (more than 200,000 page views). The main actions during SGA-II were:

- Continuous **maintenance and expansion** of webpages, uploaded files and user accounts, imposing and improving consistency, optimizing navigation and further integrating via links, redirects, overview and category pages.
- **Assistance** to editors, by creating new tutorials or improving existing ones (<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Category:Tutorial>), providing instruction and support, solving problems and taking care of more specialized editing and uploading.
- **Adding new sections and features:** an overview of all results and deliverables, the GitHub repository, a section on experimental big data statistics ([https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Experimental\\_big\\_data\\_statistics](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Experimental_big_data_statistics)), creating and filling documentation and external links overview pages for all workpackages.

## Preparing for continuity beyond ESSnet Big Data I

Ideally, the Essnetbigdata website should remain useful and used beyond the present ESSnet Big Data project. Care was taken to facilitate this through a clear and well-labelled structure, easy-to-use and intuitive navigation and the insertion of resources and tools which are also useful for non-participants. The success of this, even now, is demonstrated by the high number of page views, more than could be generated from the project itself, for the extensive list of linked mobile phone articles (11,500, highest of all) and the big data event calendar (5200), clearly indicating interest from outside the ESSnet Big Data.

### 2. Dissemination conference: BDES 2018

The conference **Big Data for European Statistics (BDES 2018)**, co-organized by Statistics Belgium (Statbel) and Statistics Bulgaria (NSI), took place in **Sofia (BG)** on **14 and 15 May 2018**. A complete overview of the conference arrangements, programme and presentations (downloadable as PDF) can be found at the conference website:

[https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/BDES\\_2018](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/BDES_2018)

Speakers included Mariana Kotzeva (director general Eurostat), Dominik Rozkrut (director general Statistics Poland), Sergey Tsvetarsky (director general Statistics Bulgaria) and Diego Kuonen (Statoo, Switzerland).

Interest for the conference was high: compared to the 2017 Dissemination workshop, with some 70 attendees, all 120 available places were now taken and a number of applicants had to be refused, regrettably, and added to a waiting list. Apart from project participants making up about half of the audience, representatives from academia, international organizations (ILO, CEDEFOP, OECD) and several non-European statistical institutes (Algeria, Brazil, Egypt, Israel, Saudi Arabia, South Korea) were also present.

## Annex 2: Evaluation Results

For the evaluation 4 questions were asked to the partners and response was received from 13 countries with no item non-response. The evaluation was not only aimed at what was experienced but also what lessons were learned. The evaluation will be used as input to write a better proposal for the ESSnet Big Data II.

### *First question:*

Have the objectives of the ESSnet been realized for your NSI? Have your expectations been met? If not, please explain.

Summary of answers: Yes, largely, fully or beyond expectations. Good way to obtain knowledge, also from more experienced countries. Positive knowledge (e.g. methods and IT) was gained, and also negative (what not to do). Results are highly relevant and capabilities increased. However, in some cases data access was less than anticipated, leading one respondent to answer that expectations were “partially” met. Use of the Sandbox was less than expected, as was the development of IT infrastructures.

### *Second question:*

To what extent can the outputs of the ESSnet be used by your NSI? Have they opened possibilities for implementation, for new domains, etc.?

Summary of answers: Generally the outputs are considered useful. Examples mentioned: job vacancies (4 times), web scraping enterprises (3), mobile phone (2), early estimates (2), WP 7, methodology of WP 8, exchange of IT tools (2), semi-simulated data, start of experimental statistics. One country mentioned that on their own they would never have been able to develop the ESSnet results. Some remarked that it is too early to answer this question. One remarked that results can also be applied in other areas. Another remarked that results are not ready to be incorporated in production – but are still useful. Moreover, the results point directly to colleagues with relevant experience.

### *Third question:*

In your opinion, do the benefits surpass the costs, for your NSI and for the ESS as a whole?

Summary of answers: Yes, without exception. The answers all seem to refer to at least the NSI level; five countries said explicitly that this was also true for the ESS level. The knowledge developed is seen as very important, and this way of acquiring the knowledge is seen as better than any alternative way. This includes IT. One country mentioned that the cost of NOT doing this would be high, and another mentioned that the level of budget and activity should be much higher, saying that the budget “is probably nowhere near enough to achieve the transformation required”.

### *Fourth question:*

Other lessons learned, e.g. on the organization of the ESSnet, meetings, communication, ....

Summary of answers: There were very many lessons learned. The positive include: being part of a network of experts is very good (3), the knowledge-sharing (wiki, etc.) is very cost-efficient (2), virtual meetings were effective and in good balance with physical meetings, general management was good and finances well organized, it all worked because the participants felt their responsibility.

Suggestions for improvement included: reimburse more persons for meetings (2) and reimburse external academics, reduce the number of requests about resources spent, better and more detailed division of roles in workpackages needed (2), more small physical meetings needed (2), too much coordination burden for workpackage leaders, the wiki could be improved, the tools for sharing results are insufficient, there are too many communication channels, the project required more time than budgeted, too many partners in one of the workpackages (and not always really committed), there was a conflict between involving as many countries as possible and delivering quality outcomes.