MAKSWELL

MAKing Sustainable development and WELL-being frameworks work for policy analysis

www.makswell.eu

Horizon 2020 - Research and Innovation Framework Programme

Call: H2020-SC6-CO-CREATION-2017

Coordination and support actions (Coordinating actions)

Grant Agreement Number 770643

Work Package 2

Methodological aspects of measuring SDG indicators with traditional and non-traditional data sources

Deliverable 2.1

Aspects of existing databases, traditional and non-traditional data sources and collection of good practices

April 2019

Leading partner: Statistics Netherlands (CBS)

Authors:

J.A. van den Brakel, B. Buelens, R.L. Curier, P. Daas, Y. Gootzen, T. de Jong, M. Puts, M. Tennekes, R. Willems (CBS), A. Brunetti, S. Fatello, F. Polidoro, A. Simone, A. Ferruzza, A. L. Palma, G. Tagliacozzo (Istat), N. Rosinski, K. Wichmann, T. Zimmermann (Destatis), F. Ertz, L. Güdemann, and R. Münnich (UT)

This project has received funding from the European Union's Horizon 2020 research and innovation programme.

Deliverable D2.1

Methodological aspects of measuring SDG indicators with traditional and non-traditional data sources;

Aspects of existing databases, traditional and non-traditional data sources and collection of good practices

Summary

The MAKSWELL project was set up to help strengthening the use of evidence and information on well-being and sustainability for policy-making in the EU, as also the political attention to well-being and sustainability indicators has been increasing in recent years. Traditionally sample surveys are the data source used for measurement frameworks for well-being and sustainability. Over the last decades more and more new, alternative data sources become available. Examples are administrative data like tax registers, or other large data sets - so called big data - that are generated as a by-product of processes not directly related to statistical production purposes. This report explores the possibilities to use non-traditional data sources for measurement frameworks for well-being and sustainability. An overview of data sources used for sustainable development goal indicators as well as potential alternative non-traditional data sources is given for three European countries. An extended list of examples are presented where scanner data, mobile phone network data, data obtained through webscraping and social media platforms, data downloaded from the web, satellite images, aerial images, and road sensor data are used for constructing official statistics and measuring sustainable development goal indicators.

1. Introduction

The MAKSWELL project (MAKing Sustainable development and WELL-being frameworks work for policy) was set up to help strengthen the use of evidence and information on well-being and sustainability for policy-making in the EU. During the last decades several initiatives have been developed to propose measurement frameworks to measure well-being in a broader scope than just GDP as well as sustainable development. In the first work package of the MAKSWELL-project the frameworks that are currently in place to measure well-being and sustainable development are evaluated (Tinto et al., 2018, Tinto and Baldazzi, 2018).

National statistical institutes play a central role in providing data for measuring these frameworks. Traditionally, relevant statistical information is obtained from sample surveys, also called traditional data sources. Over the last decades more and more new, alternative data sources become available. Examples are administrative data like tax registers, or other large data sets - so called big data - that are generated as a by-product of processes not directly related to statistical production purposes. Such data sources are further referred to as non-traditional data sources. Examples of these include time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook, internet search behavior from Google Trends, satellite and aerial images and data found on the Web.

These non-traditional data sources can provide useful information for the measurement frameworks for well-being and sustainable development. The purpose of Work Package 2 is to study the usefulness of non-traditional data sources for measuring well-being and sustainability. In this report an overview of data sources that are currently used and potential alternative non-traditional data sources for measuring sustainable development goal indicators is provided for the Netherlands, Italy and Germany. In addition a list of examples how non-traditional data sources are applied in the context of official statistics and measuring sustainable development goal indicators is provided. In deliverable 2.2 the methods related to the use of non-traditional data sources in measurement frame works for well-being and sustainability are treated in more general way (van den Brakel et al., 2019).

The paper is structured as follows. Chapter 2 contains an overview of existing and potential data sources for SDG indicators in the Netherlands, Italy and Germany and is based on contributions from M. Puts, R. Willems, A. Ferruza, J.A. van den Brakel, N. Rosinski, K. Wichmann, and T. Zimmermann.

Chapter 3 describes a project conducted by ISTAT to improve the Italian CPI using scanner data of grocery products. This contribution comes from A. Brunetti, S. Fatello, F. Polidoro and A. Simone and describes how scanner data can be used to calculate price indices.

Chapter 4 contains applications of mobile phone data and is based on contributions from M. Ten-

nekes, Y. Gootzen, A. Ferruzza, A. Laureti Palma, G.Tagliacozzo and N. Rosinski. The methodology developed at Statistics Netherlands' Big data Centre to produce day time population statistics using call detail records is described as well as a correction method for selection bias. Other potential applications where these data are used directly are statistics for mobility, tourism and migration. Alternatively mobile phone data can be used to construct covariates in prediction models for measuring unemployment, poverty and economic growth. Feasibility studies on the use of mobile phone data in Germany are also summarized as well as a project recently initiated at ISTAT to use mobile phone data for measuring the impact of natural disasters.

Chapter 5, by P. Daas, describes the methodology of webscraping using an application to count the number of innovative companies in the Netherlands, which is a project of Statistics Netherlands' Big data Centre.

Chapter 6, by P. Daas and J.A. van den Brakel, contains applications of data obtained from social media platforms from Statistics Netherlands' Big data Centre. The methodology is described how to use social media platforms as a direct data source to construct indicators for social tension and sentiment. The sentiment index is also used as a covariate in a time series model to improve accuracy and timeliness of the Dutch consumer confidence index.

Chapter 7, by B. Buelens and J.A. van den Brakel, contains an application of data found on the Web. Google trends is probably the first application one might think of in this context. In this chapter, however, meteorological time series on solar irradiance that can be downloaded from the Royal Netherlands Meteorological Institute and time series on electricity exchange from the Dutch high power grid, which can be freely downloaded from the website of the Dutch Transmission System Operator are combined in a time series model to estimate the unmetered photovoltaic power production by domestic photovoltaic installations. This is another application of Statistics Netherlands' Big data Centre.

Chapter 8, by F. Ertz, L. Güdemann, R. Münnich and T. Zimmermann, describes the possibilities of using satellite data for official statistics and measuring sustainable development goal indicators. Possibilities and issues with the use of satellite data are discussed. Applications to measure poverty, quality, and different sustainable development goal indicators with satellite images are described. A detailed literature overview of applications using satellite images in the context of measuring sustainability is provided.

In chapter 9 the use of remote sensing data for measuring sustainability related indicators is continued. Three projects conducted at Statistics Netherlands' Big data Centre are presented. The first project, by R.L. Curier, is an application of using satellite data to measure urban extension. The second project, by T. de Jong, is an application to measure the amount of photovoltaic solar panels using aerial images. The third project, by M. Tennekes, describes how traffic intensity and transportation capacity is estimated using road sensors.

The report concludes with a discussion in Chapter 10.

2. Existing and potential data sources for SDG indicators

2.1. Introduction

In 2015 the members of the United Nations adopted an agenda for sustainable development. All 193 members of the UN signed up to an ambitious package of goals: the Sustainable Development Goals (SDGs). This agreement commits the UN members to make greater efforts to end poverty and hunger, protect the Earth, defend human rights, and promote equality between men and women. The package contains a total of 17 goals that are to be achieved by 2030. To monitor progress made towards this ambitious goal, the UN drafted a list of SDG indicators.

In order to identify a common statistical framework as a tool for monitoring and assessing progress towards the objectives of the Agenda, an Inter-Agency Expert Group on SDGs (IAEG-SDGs) was set up by the UN Statistical Commission. To facilitate the implementation of the global indicator framework, all indicators are classified by the IAEG-SDGs into three tiers based on their level of methodological development and the availability of data at the global level, according to the following definitions:

- Tier 1: Indicator is conceptually clear, has an internationally established methodology and standards are available, and data are regularly produced by countries for at least 50 per cent of countries and of the population in every region where the indicator is relevant.

- Tier 2: Indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries.

- Tier 3: No internationally established methodology or standards are yet available for the indicator, but methodology and standards are being (or will be) developed or tested.

As of December 2017, there are 93 Tier I indicators, classified as such because they are conceptually clear, with established methodology and standards and data regularly produced by 50 percent of countries and of population in every region where the indicator is relevant; 66 Tier II indicators, which are conceptually clear, with established methodology and standards, but data not regularly produced by countries; and 68 Tier III indicators, for which there are no internationally established methodology or standards yet available but are to be developed and tested.

2.2. Current and new data sources for measuring sustainability

The implementation process is still in progress and involves more updating steps to ensure a thorough review of the indicators, their correct classification by the different Tiers, and the preparation of the necessary metadata. See the deliverables of Work Package 1 of the MAKSWELL-project for an extended analysis of the measurement frameworks of well-being and sustainability that are currently in place (Tinto et al., 2018, Tinto and Baldazzi, 2018). In addition to this work, an inventory has been made in Work Package 2 for Italy, the Netherlands and Germany, for all SDG indicators whether

they are currently published or not, which data source is used to compile an indicator, its publication frequency, the regional publication level (in terms of NUTS levels), future plans for implementation and which potential alternative data sources can be identified to measure these indicators. The results are presented in the accompanying Excel file and are summarized in Tables 2.1, 2.2, and 2.3.

Italy									
Chapter	Number of indicators		Frequency		Regional level		Data source		
	Total	Published	Anually	Less	National	Regional	Survey	Register	Other
1	10	5	5	0	2	0	4	1	0
2	9	4	4	0	3	0	4	0	0
3	21	13	13	0	4	0	2	19	0
4	8	6	6	0	1	0	4	1	1
5	10	6	6	0	3	0	3	3	0
6	9	5	5	0	0	0	4	1	0
7	4	4	4	0	1	0	2	0	2
8	15	12	12	0	5	0	3	7	0
9	9	6	6	0	2	0	2	3	1
10	8	4	4	0	0	0	1	3	0
11	11	9	9	0	3	0	3	3	3
12	10	4	4	0	2	0	0	4	0
13	6	1	1	0	1	0	0	1	0
14	7	2	2	0	1	0	0	1	1
15	11	6	6	0	-	-	3	1	2
16	21	8	8	0	3	0	6	2	0
17	25	4	4	0	3	0	2	2	0

Table 2.1: SDG indicators published by Italy (- means no information provided).

Some SDG indicators are already produced on a regular basis by national statistical institutes or other government agencies. A first observation is that most indicators are derived from traditional data sources like sample surveys or registers, while alternative non-traditional data sources are rarely used. From the overview in the Excel file it follows that several indicators seem to be ideally suited to be measured by non-traditional data sources using new statistical methods. The most likely indicators that may benefit from (or even being derived from) non-traditional data sources are the tier 3 indicators (although all three tiers may contain such indicators).

In almost all cases the internationally established methodology are design-based procedures based on regular surveys or registers. In these cases the survey (or register) is held under the responsibility of either the statistical office itself or a government agency or some other trustworthy entity. Extraction of the relevant data is almost always straight forward. For example SDG indicator 1.2.1 is based on a survey for all three countries under consideration (for a complete overview consult the accompanying Excel file).

Another observation that can be derived from Tables 2.1, 2.2, and 2.3 is that indicators are predominately published on the national level on an annual or lower frequency. For policy making, however, low regional information and timely information is more relevant than information at national level

Germany									
Chapter	Number of indicators		Frequency		Regional level		Data source		
	Total	Published	Anually	Less	National	Regional	Survey	Register	Other
1	10	4	-	-	-	-	2	0	2
2	9	6	-	-	-	-	3	2	1
3	21	13	-	-	-	-	8	3	2
4	8	6	-	-	-	-	5	0	1
5	10	7	-	-	-	-	2	2	3
6	9	8	-	-	-	-	5	0	3
7	4	4	-	-	-	-	0	0	4
8	15	13	-	-	-	-	6	1	6
9	9	7	-	-	-	-	4	0	3
10	8	4	-	-	-	-	3	0	1
11	11	4	-	-	-	-	2	0	2
12	10	6	-	-	-	-	2	1	3
13	6	1	-	-	-	-	0	0	1
14	7	2	-	-	-	-	1	0	1
15	11	5	-	-	-	-	1	1	3
16	21	12	-	-	-	-	6	3	3
17	25	9	-	-	-	-	5	0	4

Table 2.2: SDG indicators published by Germany (- means no information provided).

The Netherlands									
Chapter	Number of indicators		Frequency		Regional level		Data source		
	Total	Published	Anually	Less	National	Regional	Survey	Register	Other
1	10	4	4	0	4	0	3	0	1
2	9	6	6	0	4	0	1	2	1
3	21	17	17	0	17	0	12	3	2
4	8	6	6	0	6	0	5	0	1
5	10	7	7	0	6	0	1	2	3
6	9	5	3	2	3	0	0	0	3
7	4	2	2	0	2	0	2	0	0
8	15	12	12	0	11	0	4	1	6
9	9	8	5	3	8	0	5	0	3
10	8	4	4	0	4	0	3	0	1
11	11	7	7	0	5	0	3	0	2
12	10	7	7	0	7	0	3	1	3
13	6	1	1	0	1	0	0	0	1
14	7	4	4	0	4	0	3	0	1
15	11	5	5	0	4	0	0	1	3
16	21	10	10	0	10	0	4	3	3
17	25	8	8	0	8	0	4	0	4

Table 2.3: SDG indicators published by the Netherlands.

at a low frequency. So besides using non-traditional data sources directly to compile SDG indicators, another potential contribution of non-traditional data sources is to use them as covariates in small area prediction models in combination with sample surveys, to make reliable regional estimates. Another possibility is to utilize their timeliness and high frequency to obtain more timely nowcasts for sample estimates. In the remaining chapters of this report this is illustrated with a range of examples. In Delivarable 2 of Work Package 2, the methodology for using non-traditional data sources in this context is described more generally (van den Brakel et al., 2019).

The accompanying Excel file summarizes the current situation in Italy, Germany and the Netherlands regarding the implementation of the indicators and future plans for implementation. It also points to possible alternative data sources for many indicators. Some potential alternative sources and methodologies have been identified from this inventory and can be categorized as follows:

- Data found on the web, like google trends and information obtained from websites. Google trends are time series observed on a high frequency and are therefore particularly useful to improve timeliness of sample estimates in nowcast methods as well as the accuracy of sample estimates after finalizing the data collection. In case google trends can be observed on the domain level it also provides relevant information for small area estimation methods. A nowcast example in this context is given in Subsection 6.2. Data found on the web can basically be anything. An example in the context of gathering statistical information on the energy transition is presented in Chapter 7.

- Data from social media platforms. This are typically high frequency time series observed at the national level and can be used to improve timeliness and accuracy of sample estimates. Applications for measuring SDG indicators appear to be minor.

- Data digitally collected by companies, like e.g. scanner data. These data come at a low regional level and are useful in small area estimation models for predicting poverty but also as a direct source for price indices. These data are therefore very relevant for well-being indicators and poverty indicators. An application in this context is provided in Chapter 3.

- Mobile phone data provide a wealth of information about the actual location of people. From this point of view it provides a direct data source for real time population statistics, mobility and tourism statistics. More related to SDG indicators, mobile phone data can be used to produce statistics on migration flows and natural disasters. Besides that, mobile phone data provide relevant auxiliary information for estimating unemployment, economic growth and poverty. This is further explored in Chapter 4.

- Remote sensing data, like e.g. satellite images, aerial images, and geodata. These data have many applications for SDGs like: 1) No poverty, 6) Clean water and sanitation, 11) Sustainable cities and communities, 13) Climate action and 15) Life on land. Chapters 8 and 9 further explore these data sources and illustrate that they are potentially useful to directly estimate SDG related information, like urban sprawl and the amount photovoltaic solar panels, but also

provide auxiliary information for example for small area poverty prediction models.

3. Scanner data: Improvements in Italian CPI/HICP deriving from the use of scanner data

3.1. The state of art

[1] Starting from January 2018 ISTAT introduced scanner data of grocery products (thus excluding fresh food) in the production process of estimation of inflation. This innovation concerns 79 aggregates of product belonging to 5 ECOICOP Divisions (01, 02, 05, 09, 12). Since the end of 2013 a stable cooperation was established among ISTAT, Association of modern distribution, retail trade chains (RTCs) and Nielsen. Scanner data of grocery products have been collected by ISTAT through Nielsen for years 2014, 2015 and 2016 for about 1400 outlets of the main six RTCs for 37 provinces.

Afterwards, in view of the inclusion of scanner data into price indices calculations, a probabilistic design has been implemented for the selection of the sample of outlets, for which Nielsen provided ISTAT from December 2016. Scanner data for 1,781 outlets (510 hypermarkets and 1,271 supermarkets) of the main 16 RTCs covering the entire national territory are monthly collected by ISTAT on a weekly basis at item code level. Outlets have been stratified according to provinces (107), chains (16) and outlet-types (hypermarket, supermarket) for a total of 867 strata, taking into account only the strata with at least one outlet. Probabilities of selection were assigned to each outlet based on the corresponding turnover value. Figure 3.1 shows the number of the strata, the number of the outlets and the coverage in terms of turnover, at regional and national levels for years 2018. The coverage for the year 2017 is slightly lower because a small RTC has been excluded from the analysis.

Concerning the selection of the sample of items, a static approach that mimics traditional price collection method has been adopted[2]. Specifically, a cut off sample of barcodes (GTINs) has been selected within each outlet/aggregate of products (coveraing 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover). The products selected in December are kept fixed during the following year. A 'thank' of potentially replacing outlets (258) and GTINs (until a coverage of 60% of turnover within each outlet/aggregate) has been detected in order to better manage the possible replacements during 2018.

About 1,370,000 price quotes are collected each week to estimate inflation. For each GTIN, prices are calculated taking into account turnover and quantities (weekly price=weekly turnover/weekly quantities). Monthly prices are calculated with arithmetic mean of weekly prices weighted with quantities.

Scanner data indices of aggregate of products are calculated at outlet level as unweighted Jevons index (geometric mean) of GTINs elementary indices. Provincial scanner data indices of aggregate of products are calculated with weighted arithmetic mean of outlet indices using sampling weights.

[1] Alessandro Brunetti, Stefania Fatello, Federico Polidoro and Antonella Simone, Istat, Integrated system for household economic conditions and consumer prices

[2] The static approach to sampling is discussed in Eurostat (2017)

Finally, for each aggregate of products, scanner data indices and indices referred to other channels of retail trade distribution are aggregated with weighted arithmetic mean using expenditure weights.

To calculate weights for the integration of regional indices of modern and traditional distribution at regional level, data are broken down using regional estimates from National Account (at ECOICOP sub-class level), regional expenditure by type of distribution from Ministry of Economic Development and qualitative information on the shopping habits of consumers coming from HBS.

	North								Centre				South								
Region	Piemonte	Valle d'aosta	Liguria	Lombardia	Trentino Alto Adige	Veneto	Friuli Venezia Giulia	Emilia Romagna	Toscana	Umbria	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna	ITALY
Num. of strata	78	4	31	148	12	85	45	84	65	16	43	38	34	11	35	34	6	25	44	29	867
Num. of outlets	152	6	62	286	35	161	77	163	142	32	83	106	58	18	88	85	10	51	111	55	1781
% market shares (hyper + super)	95,9	76,0	99,8	94,8	99,1	87,0	94,3	98,5	99,9	92,2	97,4	93,2	92,9	96,6	77,5	91,3	67,9	87,2	86,4	98,0	93,7

Figure 3.1: Sample size: number of strata, number of outlets and coverage in terms of turnover -Year 2018

3.2. The improvements of the territorial coverage of indices and its effect on the accuracy of inflation estimates

Scanner data allow calculating the indexes for the entire national territory using data from outlets of all Italian provinces and located both in the municipal area and outside. With the aim of evaluating the benefits in terms of accuracy of inflation estimation coming from the improvement of the coverage in territorial terms, price indices are calculated by taking into account the outlet location (i.e. inside municipal area of the provincial chief towns: HICP SD MA) and by distinguishing the 80 provincial chief towns previously involved in the consumer price survey from the rest of the provinces whose data now are made available by scanner data (HICP SD 80P).

Figure 3.2 shows the number of outlets used for the calculation of the different indices at national and macro-regional level for years 2018.

Macroregion	All outlets		Outlets in 80 provinces		Outlets in municipal area	
	N° outlets	*% outlets*	*N° outlets*	*% outlets*	*N° outlets*	*% outlets*
North	942	52,9	857	58,8	304	50,6
Centre	363	20,4	286	19,6	133	22,1
South	476	26,7	315	21,6	164	27,3
Italy	1781	100,0	1458	100,0	601	100,0

Figure 3.2: Number of outlets at national and macro-regional level â€“ Year 2018

In order to point out the methodology used for this analysis, it is necessary to start with a short description of the procedure for the aggregation of indices[3]

[3] For a detailed description of the procedures adopted by Istat for the calculation of the consumer price indices, see Istat (2017)

Let us introduce the following notation[4]:

- n denotes the n-th product aggregate[5] $(n = 1, \ldots, N)$
- g denotes the g-th region $(g = 1, \ldots, G = 20)$
- j denotes the j-th province $(j = 1, \ldots, J(r))$
- h denotes the h-th outlet $(h = 1, \ldots, H(j))$

Let

$$P_{ng,j} = \sum_{h \in j} w_{ngj,h} P_{ngj,h} \textit{ be the provincial index of the product aggregate } n;$$

$$P_{n,g} = \sum_{j \in g} w_{ng,j} P_{ng,j} \textit{ be the regional index of the product aggregate } n;$$

$$P_n = \sum_{g} w_{n,g} P_{n,g} \textit{ be the national index of the product aggregate } n;$$

$$P = \sum_{n} w_n P_n \textit{ be the general index at the national level}$$

where:

$$w_{ngj,h} = \frac{e_{ngj,h}}{\sum_{h \in j} e_{ngj,h}} \; ; w_{ng,j} = \frac{e_{ng,j}}{\sum_{h \in g} e_{ng,j}} \; ; w_{n,g} = \frac{e_{n,g}}{\sum_{r} e_{n,g}} \; ; w_n = \frac{e_n}{\sum_n e_n}$$

and $e_{ngj,h}$ is the expenditure estimate for product aggregate n in the outlet h of province j in region g[6].

For the scope of the present analysis, the general index P is to be compared with the index $\hat{P}$ which is calculated using:

- Transaction prices of the outlets (h') situated inside municipal borders of the provincial chief towns (HICP SD MA);
- Transaction prices of the outlets of the 80 provincial chief towns previously involved in the consumer price survey (HICP SD 80P)

[4] The notation used is adapted from that one suggested in Biggeri and Giommi (1987)

[5] Product aggregates indices are indices calculated at the lower level of aggregation of product-offers.

[6] $e_{ngj,h}$ incorporates the sampling coefficient attached to the outlet h

Concerning the first case, the general index P can be usefully expressed as the weighted arithmetic mean of provincial product aggregate indices:

$$P = \sum_{nj} \frac{e_{nj}}{\sum_n e_n} P_{nj} = \sum_{nj} \pi_{nj} P_{nj}$$

Accordingly, the impact of the improvement of territorial coverage is calculated as follows:

$$P - \hat{P} = \sum_{nj} \pi_{nj} (P_{nj} - \hat{P_{nj}})$$

where

$$\hat{P_{nj}} = \sum_{h^j \in j} \frac{e_{nj,h^j}}{\sum_{h^j \in j} e_{nj,h^j}} P_{nj,h^j}$$

The impact can also be decomposed as suggested in Biggeri et al. (2008). By indicating with k the product NxJ, with δ_{nj} the difference between sub-indices $(P_{nj} - \hat{P_{nj}})$ with $s_{\pi_{nj}}$ and $s_{\delta_{nj}}$ the standard deviation of π_{nj} and δ_{nj} with $R_{\pi_{nj},\delta_{nj}}$ the linear correlation coefficient between π_{nj} and δ_{nj} with $\bar{\delta}_{nj}$ the arithmetic mean of δ_{nj}, we have:

$$P - \hat{P} = k s_{\pi_{nj}} s_{\delta_{nj}} R_{\pi_{nj},\delta_{nj}} + \delta_{nj}$$

As for the second case, it is convenient to express the general index P as the weighted arithmetic mean of regional product aggregate indices:

$$P = \sum_{ng} \frac{e_{ng}}{\sum_n e_n} P_{ng} = \sum_{ng} \pi_{ng} P_{ng}$$

Consequently, it is possible to write:

$$P - \hat{P} = \sum_{ng} \pi_{ng} (P_{ng} - \hat{P_{ng}})$$

where

$$\hat{P_{n,g}} = \sum_{j^j \in g} \frac{e_{ng,j^j}}{\sum_{j^j \in g} e_{ng,j^j}} P_{ng,j^j}$$

and, with similar notation:

$$P - \hat{P} = k s_{\pi_{ng}} s_{\delta_{ng}} R_{\pi_{ng},\delta_{ng}} + \delta_{ng}$$

3.3. Results

By comparing indices calculated on the whole national territory and the corresponding indicators compiled taking into account only the outlets in the municipal area of the provinces (3.3) moderate differences emerge in the first months of 2017 and 2018 and in the middle of the first year. However, when the geographical breakdown is considered, the divergences tend to be relatively larger and persistent, especially in the South of Italy (islands included) (Figure 3.4). For example, the difference of the indices, calculated on March 2018, shows that the HICP SD MA index of the South is about 0.3 percentage points below the corresponding HICP SD index, while it is 0.24 in the North and 0.15 in the Centre). The main factors explaining this divergence seem to be the relatively higher value of the standard deviation of the differences of sub-indices and the relatively high value of the linear correlation coefficient $R_{\pi_{nj},\delta_{nj}}$ (higher differences in the level of sub-indices tend to have higher weights).

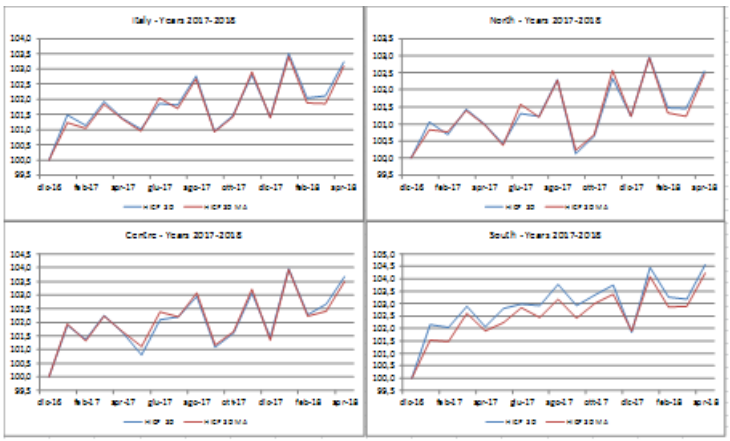


Figure 3.3: Comparison between HICP SD and HICP SD MA - Years 2017-2018

Regarding the comparison between HICP SD and HICP SD 80P, major divergences tend to be concentrated in the South of Italy (3.5) as well. This result reflects the fact that the share of provinces participating to the survey before the introduction of scanner data, in this part of the country, was relatively low as compare to the North and the Centre.

Generally speaking, for what concerns modern retail trade distribution and comparing indices compiled from scanner data source in a time span of 16 months, the improvement in terms of accuracy coming from the coverage of the entire provincial territories are limited to three months at national level

	Italy	North	Centre	South
K	8.368	3.713	1.738	2.917
$s_{\pi_{n,j}}$	0,0002	0,0004	0,0012	0,0005
$s_{\delta_{n,j}}$	3,3349	2,9171	2,8525	4,0235
$R_{\pi_{n,j}\delta_{n,j}}$	0,0139	-0,0088	0,0082	0,0460
$\bar{\delta}_{n,j}$	0,1522	0,2732	0,1046	0,0267
HICP SD	100,6830	100,2363	101,2078	101,2760
HICP SD MA	100,4517	100,0011	101,0555	100,9741
HICP SD - HICP SD MA	0,2314	0,2352	0,1523	0,3019

Figure 3.4: Decomposition of the difference between HICP SD and HICP SD MA. March 2018

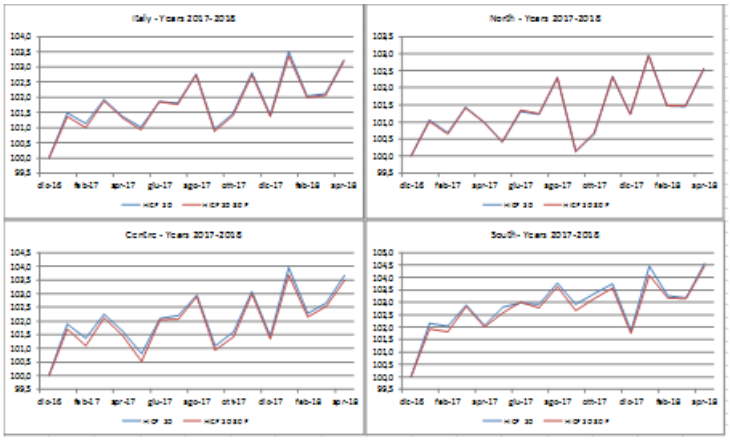


Figure 3.5: Comparison between HICP SD and HICP SD 80P - Years 2017-2018

with a maximum of three decimal points of differences between grocery index calculated inside the municipal borders and that one compiled with the outlets of the entire municipal areas. For the South the differences between the two indices are more frequent and wider along the time span considered. In all the cases, the level of the indices referred to the entire provincial areas are slightly higher than those ones compiled just within the municipal borders.

If we consider the grocery index compiled taking into consideration the outlets of the 80 provinces that were involved in 2017 in the territorial data collection, the comparison with the grocery index calculated from the data of all the 107 Italian provinces, shows just some local differences (in the South and in the Centre of Italy) but without important consequences in the estimation of the grocery index at national level.

3.4. Next steps

Scanner data project (brought forward by ISTAT) is still on the way and it is possible to sketch some further steps. The first one is the adoption of the so called dynamic approach (abandoning the static one) to the selection of the elementary items (GTINs) to be considered for the compilation of indices. It should be implemented in the near future and it means the use of all the elementary price quotes of all the GTINs sold monthly in the sample of outlets selected. This requires that the main

crucial issues need to be solved in sight of this next step, starting from that regarding relaunches and IT environment and procedures. Dynamic approach is the choice towards other National Statistical issues at European level are converging and could represent a further improvement in the accuracy of Italian CPI/HICP. The following steps concern the extension of the use of scanner data to other retail trade channels (discount, outlets with surface between 100 and 400 square meters) and other goods such as fresh products with variable weight and no grocery products. Toward these aims the role of the collaboration with the modern distribution and the representative association (Association of Modern Distribution, ADM) keeps its crucial importance.

4. Mobile phone network data

Usage of mobile phones generates massive amounts of data. Two types of data can be distinguished based on their sources: sensory data and network data. Sensory data are generated by the sensors in a mobile device such as the gyroscope, GPS sensor, camera and microphone. These sensor data can be used by the mobile operating system or a specific app, but explicit individual consent of the user is required to use and share these data.

Network data is data generated by the network of antennas owned by a mobile network operator (MNO). The MNO facilitates mobile communication and charges the corresponding costs to its customers. Most countries have more than one MNO, who each own and maintain their own network of antennas. When using a mobile phone abroad, the foreign MNO that provides mobile communication forwards the usage costs to the primary MNO. Data collected by the antenna network can be analysed with consent from the MNO, while conforming to strict privacy measures. This project focuses on network data, leaving sensory data out of the scope.

Each time a device interacts with an antenna in the network, a record is generated. Two types of records can be distinguished: call detail records (CDR) and network events. CDR data contain information about calls, SMS messages and mobile data usage and are used to bill customers. This information contains the date, time, location, and duration (only for calls) of events, but not the content of the communication. Network events are passively generated when a mobile phone moves from one antenna to another, even if it is not actively being used.

Signalling data includes both CDR data and network event data. The primary use of signalling data is network analysis and optimisation. The analyses in this report are described for signalling data, but can be reinterpreted for just CDR data without difficulties. Signalling data have an international standard, which is convenient when using these data for other purposes as described below.

Signalling data has a tabular format. Each row corresponds to an event, which can be the start or end of a call (both initiating and receiving), sending or receiving an SMS, information about mobile data use, or a network event (e.g. antenna handovers). The latter is not included in CDR data. The main variables that are included in signalling data are: International Mobile Subscriber Identity (IMSI) which corresponds to a telephone number, date, time, antenna id, country code, network code, and event type. For call and SMS events, it also contains the telephone number of the other party. The size of these data depends on many factors such as how much a mobile phone is used, what kind of mobile phone and operating system are used, the used network technology, and what kind of events are registered by the MNO. Signalling data may contain hundreds of events per device per day. Note that the number of events per hour may be a lot smaller during nighttime.

The MNO also owns the *cell plan* data, which contains the physical settings (including geographical

coordinates) of the antennas. This data is used to approximate the geolocation of the events, which is described below.

This project aims to process mobile phone network data for applications in official statistics and sustainable development goals (SDG). Section 4.1 describes how the data can be used for several domains within official statistics and for SGDs. An origin-destination-time cube contains the estimated number of people by home region, present region, and one hour time period. Section 4.2 discusses the methods to produce these aggregates from signalling data, and proposes a concrete indicator to measure the SDG poverty.

4.1. Applications for official statistics

Within official statistics, mobile phone network data can be used in the following domains.

Daytime population The daytime population is the number of people in a certain region at a certain time. This is useful for many applications, including visitor counts during events, and emergency management (Deville et al., 2014, Ahas et al., 2015, De Meersman et al., 2016, Xu et al., 2018, Kondor et al., 2017). In countries which do not have integral population registers, such information can be used as an auxiliary data source to improve the quality of population estimates (Deville et al., 2014, Salgado et al., 2018).

Mobility Obviously, people are not always located at the same place, but move for many reasons, for example commuting, school, shopping, recreation, and social events. Mobile phone network data can be used to estimate mobility flows between regions (Alexander et al., 2015, Diao et al., 2016, Iqbal et al., 2014, Jiang et al., 2016, Jonge et al., 2012, Pucci et al., 2015, Widhalm et al., 2015, Zagatti et al., 2018). Moreover, it is possible to estimate the mode of transport, which is useful for infrastructure planning.

Tourism Statistics on tourism cover the places that tourists visit, where they stay overnight, and where they come from. Since signalling/CDR data contains country code, it is possible to estimate the number of tourists by home country (Tennekes et al., 2017).

Migration Large migration flows can be observed from mobile phone network data. Such flows are expected after wars or natural distastes (Lu et al., 2016, Wilson et al., 2016).

Social networking With CDR data, it is possible to reconstruct a social network by using information on who has mobile communication with whom. Such network may provide insights in social cohesion within and between regions (Toomet et al., 2015, Xu et al., 2017).

Although mobile phone network data cannot directly be used for SDGs, the following indicators can be derived from it:

Employment It is possible to derive commuting flows from mobile phone network data. A device can be classified as owned by a commuter if it travels between two regions at rush hours during

normal working days (i.e. weekdays except holiday weeks). This is just an approximation, since people can also go to cities for other reasons and people can also travel to work at non-standard times. However, from the numbers of classified commuting devices over time, a valuable indicator for employment can be derived.

Poverty From CDR data, metrics can be derived that are used to predict poverty (Steele et al., 2017). Examples of such metrics are basic phone usage, daily activity patterns and social network activity. In Section 4.2.5, we provide an example of an indicator for poverty using signalling data.

Economic growth The population size in touristic city centres, large shopping centres, and other large hotspots for recreation and tourism can be interpreted as an indicator for economic growth. People are expected to spend money in these areas, which indicates a certain degree of welfare, and therefore also economic growth.

4.2. Methods

The origin of a device is not seen as the location of the device during the previous time period, but rather as the device's place of residence. A Bayesian model allows for combining knowledge of the signal strength model and network properties into a location estimate, as described in Section 4.2.1. The place of residence is determined by antenna usage over a period of time, as explained in Section 4.2.2. The previous steps result in an origin-destination matrix of number of devices, which is explained in Section 4.2.3. The process of transforming this to a matrix of expected number of persons is described in Section 4.2.4.

4.2.1. Geographic location

The exact geographic location is often neither measured nor stored in signalling or CDR data. In these cases, the geographic location of mobile phones can be estimated based on the location of the connected antennas. The vast majority of studies on mobile network data use Voronoi tessellation to approximate the geographic location of mobile phones. This method assumes that antennas have a 360 degree coverage and that the coverage areas do not overlap. However, both assumptions do not hold in reality.

We propose a Bayesian method to estimate the probability that a mobile phone is present at a specific location given the antenna it is communicating with. This method takes overlapping coverage areas into account. Furthermore, we model the signal strength for each antenna using not only its location, but also other properties, such as direction and tilt.

The central question is what the location of a mobile device is, given the connected antenna. Our approach is the following. We place a grid of 100 by 100 meter cells over the area of interest. For each grid cell g, we calculate the probability that a mobile phone is located inside g given that it is

connected to antenna a. This probability, $\mathbb{P}(g|a)$, can be calculated using the formula of Bayes.

$$\mathbb{P}(g|a) = \frac{\mathbb{P}(g)Pr(a|g)}{Pr(a)}. \tag{4.1}$$

The probability $\mathbb{P}(g)$ is the prior, i.e. the expected probability of presence in g. This prior can be set as uniform, so the probability that a device is on a certain grid cell does not depend on the grid cell itself. One might also argue to use a prior based on land use. The philosophy behind this is that the probability that a mobile phone is located on water areas is far less likely than on land areas. Moreover, the probability of presence in residential areas is more likely than on farmland. Another approach is to use antenna densities for this prior probability, because it is to be expected that the number of mobile phones is higher in grid cells with coverage from a high number of antennas.

Let the probability $\mathbb{P}(a|g)$ that a mobile phone is connected to antenna a given that it is located in grid cell g be defined as:

$$\mathbb{P}(a|g) = \begin{cases} 0 & \text{if grid cell } g \text{ is not covered by antenna } a, \\ \dfrac{s(g,a)}{\sum_{a'} s(g,a')} & \text{if raster cell } g \text{ is covered by } a, \end{cases} \tag{4.2}$$

where $s(g, a)$ is the relative signal strength of antenna a in grid cell g. Signal strength can be modelled using physical data of the antenna, such as direction, height, and tilt (Salgado et al., 2018).

Finally, the probability $\mathbb{P}(a)$ can be seen as a normalisation constant since the connection to antenna a is given. Therefore, Equation (4.1) can be formulated as

$$\mathbb{P}(g|a) \propto \mathbb{P}(g)\mathbb{P}(a|g). \tag{4.3}$$

4.2.2. Place of residence

Raw signalling data does not contain any demographic data, it merely contains encrypted phone numbers (IMSI). An MNO has some demographic data of its customers, at least the name and postal address. However, an MNO is only allowed to join these with signalling data in order to bill its customers. Therefore, it is not possible to join customer data with signalling data for statistical purposes.

Although it is not an option to join postal address with IMSI, the place of residence can be approximated from the data itself. A logical approach is to take the position of the most frequently connected antenna during nighttime during a longer period of time, for instance one month.

For the majority of people, it is to be expected that their place of residence can be estimated by taking

the position of the most frequently connected antenna during nighttime over a longer period of time, such as one month. The approach is based on the assumption that people who work or go to school, do this during daytime. This assumption is violated for people with a nighttime job or people who are on holiday for the majority of the observed period. An approach with slightly fewer drawbacks is to take the location of the most frequently connected antenna, regardless of the time of day. People who spend more time in their place of work than at home, perhaps due to long working days or social activities, may receive an incorrect approximation of their place of residence.

Both of the above approaches come with challenges regarding location and time. The location is estimated by the grid cell of the physical antenna location, whereas the antenna's coverage may reach many grid cells that should be considered for home location. Time-wise, the number of events of a device during a time period is influenced by the activity of its user during this time period. The number of events generated by a device during a period of time is correlated with the hour of day, since device often generates more events when moving or messaging. The importance of a signalling event followed by a small period of inactivity might be different from the importance of a signalling event followed by a larger period of inactivity.

The proposed method accounts for these two problems by using not just the most frequently used antenna, but the top x (say 5) of most frequently used antennas. Each event is assigned a weight proportional to the time between that event and the consecutive event. The weight of each antenna is defined as the sum of weights over all events of the device corresponding to the antenna. For each antenna a, the probability for all grid cells g in coverage of the antenna $\mathbb{P}(g|a)$ (derived from the method described in the previous section) are multiplied by the weight of a. Hence, the outcome per device is a vector of normalised weights over all grid cells. These probabilities can be aggregated to administrative regions. The region with highest probability can be viewed as place of residence. Alternatively, the vector of probabilities can be used as such in the aggregation, which is described in the next section.

4.2.3. Aggregation of devices

The signalling data can be aggregated to an origin-destination-time cube D, in which an element $d_{h,g,t}$ corresponds to the number of unique devices with home location h and present location g during time period t. The home and present location can be specified in administrative regions such as neighbourhoods or municipalities, or squares (grid cells). Henceforth, we specify both on the level of municipalities. The time period can also be chosen as desired. We use time periods of one hour.

We create the origin-destination-time cube as follows. For every device i, a matrix M^i is derived in which each element $m^i_{g,t}$ is the fraction of presence of this device in region g during time period t. For each each period t the corresponding column total should add up to 1. As described in Section 4.2.2, the concept of a single place of residence is replaced by the concept of weights over multiple municipalities. Each weight represents the belief of a municipality being the true place of residence. Let v^i be the vector of these weights for device i, with element v^i_h being the belief that municipality h is the municipality of residence. Rather than estimating one exact origin, the origin is thought of as a fuzzy concept and consist of a set of municipalities, each with a corresponding weight. Let N be the total number of devices. The elements $d_{h,g,t}$ of the origin-destination-time cube D can be calculated

by combining all $m^i_{g,t}$ and v^i_h over all devices:

$$d_{h,g,t} := \sum_{i=1}^{N} m^i_{g,t} \cdot v^i_h. \tag{4.4}$$

Note that $d_{h,g,t}$ can be a non-integer value, although it is still an estimate of the total number of devices with municipality of residence h that are present in municipality g during time period t.

For incoming roaming data, i.e. signalling data from foreign devices in the Netherlands, a similar approach can be used. Here, the home regions are not municipalities of residence, but home countries. Furthermore, v^i remains a vector of weights for device i, but in this case, v^i_h is defined as 1 if h is the home country of device i and 0 otherwise.

4.2.4. Daytime population

The elements of the origin-destination-time cube D represent expected numbers of devices. In order to estimate the number of people, D has to be calibrated for several reasons. For instance, not every person owns a mobile phone and some persons might carry multiple devices. Furthermore, D represents the estimated number of devices of the MNO(s) from which the signalling data is used, excluding the portion of the population that communicates via a different MNO. We propose the following method to transform the estimated numbers of devices to estimated numbers of persons.

Let X denote the origin-destination-time cube of persons, such that $x_{h,g,t}$ is the expected number of persons with place of residence h that is present in municipality g during time period t. Let $w_{h,t}$ be the calibration factor between $d_{h,g,t}$ and $x_{h,g,t}$, such that:

$$x_{h,g,t} := w_{h,t} \cdot d_{h,g,t}. \tag{4.5}$$

Since signalling data includes roaming data of foreigners and the countries they originate from, we can not simply think of h as a native municipality. Two types of regions can be distinguished: municipalities and foreign countries. For each municipality h, let the number of residents, derived from population registers, be denoted as z_h. When no populations registers of foreign countries are available, $w_{h,t}$ should still be estimated somehow.

MNOs have bilateral agreements between countries on which mobile network is preferred abroad. For instance, a German device from T-Mobile Germany will by default use the mobile network from T-Mobile Netherlands. When we use roaming data from the Dutch T-Mobile network, then we can therefore assume that all German devices are from T-Mobile Germany. Therefore, we use the market share from the foreign MNO, in this case T-Mobile Germany. With this method for roaming devices, we correct for market share. However, we do not take into account that not all foreign people use mobile phones. Further research is needed for this issue.

	A	B	C
A	950	100	100
B	10	200	10
C	40	50	140

	A	B	C	dtp
A	4750	300	400	5450
B	50	600	40	690
C	200	150	560	910
pop	5000	750	1000	

Table 4.1: Example of calibration. Left the estimated number of devices, right the estimated number of people. The columns represent the home municipality and the rows the present municipality.

Let the estimate of the MNO market share be denoted by $\pi_{h,t}$ for each country h. This market share is often seen as a constant over time, especially when shorter time periods are considered. Let the weight $w_{h,t}$ of region h and time period t be defined as follows:

$$w_{h,t} := \begin{cases} \dfrac{z_h}{\sum_g d_{h,g,t}} & \text{if } h \text{ is municipality,} \\ \dfrac{1}{\pi_{h,t}} & \text{if } h \text{ is foreign country.} \end{cases} \tag{4.6}$$

The following example illustrates this scaling method. Suppose there are only three municipalities A, B, and C, in which the numbers of residents are 5000, 750, and 1000 respectively. Suppose for a fixed time period, the estimated number of devices is given by Table 4.1(left).

The total estimated number of devices for per home municipality are the are column totals of Table 4.1(left): 1000, 250, and 250 for A, B, and C respectively. Since the population numbers of the three municipalities are 5000, 750, and 1000, the calibration weights are 5, 3, and 4. Therefore, the number of devices with home municipality A, B, and C are multiplied by 5, 3, and 4 respectively. Hence, the Table 4.1(right) is derived. The row sums correspond to the estimated daytime population numbers.

This simple calibration method is based on the assumption that the device to person ratio is homogeneous for all persons from the same region of residence. Obviously, this assumption is violated for a lot of people, in particular very young children and elderly people who often have fewer devices per person and tend to move less between municipalities. People with two mobile phones, one for business and one for private use, are expected to travel more between municipalities. Note that a Dual-SIM phone also counts as two different devices, since the IMSI number represents the SIM card and not the mobile phone itself. Another assumption that is used, is that the mobility pattern of person does not depend on the subscribed MNO. In reality, this assumption is also violated, since MNOs often target specific markets, for instance young entrepreneurs or elderly people. Violations of these assumptions result in biased estimates. Further research is needed to quantify this bias, and if possible, to correct for it.

Another issue is outgoing roaming. In the case of the Netherlands, it is important to know how many Dutch people are abroad. Therefore, it is necessary to use population numbers that represent the number of people who are not abroad, preferable per home region. These numbers can be estimated

from tourism or holiday statistics.

4.2.5. SDG indicators

SDG indicators can be derived from the origin destination cubes. These indicators are especially useful in countries for which no administrative sources exist. It is also possible to use these indicators as auxiliary variables for model-based estimation methods.

As an example, we propose the following indicator for poverty:

$$\text{pov}(h,t) = \text{ts}\Big(\frac{x_{h,h,t}}{\sum_g x_{h,g,t}}, t\Big), \tag{4.7}$$

where $\text{ts}(y,t)$ represents a time series y over time t where the cyclic weekly and seasonal patterns are removed. The fraction in Equation 4.7 is the number of residents who are in their home municipality divided by the total number of residents. The assumption behind this indicator is that poor people are more likely to be in their home municipality than other people. A couple of reasons can be stated to support this assumption. Obviously, travelling to other municipalities costs money, which is therefore easier for rich people. Also, the purpose of travelling is often related to wealth; for instance, people travel to go to work, to go shopping, and to do (paid) recreational activities. For poor people, there are less reasons to travel to other municipalities.

Note that this indicator should not be used for comparisons between municipalities, since each municipality has its own profile regarding employment, shopping centres, etc. For instance, a person from the municipality of Amsterdam that is observed in Amsterdam, a city with many employment and recreational opportunities, has a smaller probability of being poor than a person from a small town with relatively low employment and recreational opportunities. Another reason to not compare between municipalities is that certain age groups might be less mobile, even though they are not poor. If municipalities have different ratios of an elderly population, this will impact their score on the proposed indicator. This indicator is solely meant to monitor the poverty over time within a municipality.

4.3. Applications at Destatis

As a result of the generally increasing use of digital technology, statistical offices are facing the challenge to explore and employ new data sources and to organise their processes and procedures accordingly. For that reason, the Federal Statistical Office of Germany (Destatis) conducts several feasibility studies, together with national and international partners, to determine the usefulness of new digital data, such as mobile phone data, for official statistics. The use of such data is considered to have a great potential for a quicker, more precise and more cost-effective production of official statistics and to help reducing the burden on respondents.

Regarding mobile phone data, Destatis has established a cooperation with T-Systems International GmbH and Motionlogic GmbH (both wholly-owned subsidiaries of Deutsche Telekom AG) in September 2017. The conceptual design of the planned feasibility studies was developed in coordination with

the Federal Network Agency, the Federal Commissioner for Data Protection and Freedom of Information and T-Systems. The medium and long-term objective is to use mobile phone data to provide a valid picture and estimation of the daytime and resident population, of commuting flows and of tourist distributions in the whole of Germany.

First results show that, to some extent, the available mobile phone data could provide a good picture of the population. The figures were used to determine the correlation between mobile phone activities and census values by type of day and time. Overall, the values reveal a high correlation of 0.8 between mobile phone activities and census values throughout Saturday and Sunday. On weekdays, the correlation declines to less than 0.7 between 5 a.m. and 4 p.m., which indicates significant differences in the resident population according to the 2011 census and according to the location of mobile phone activities within the given period. The differences observed between the population figures based on mobile phone data and those based on census values may partly be explained by the time difference between the mobile phone data from 2017 and the census data from 2011, but they may also result from the extrapolation method used by the mobile network operator (Hadam, 2018). The issue of biases and selectivities will be discussed in future papers.

Furthermore, the studied data allow to distinguish between daytime and nighttime population. Using the data currently available, it is however not yet possible to describe the commuting patterns as such, i.e. the movement between areas. Nevertheless, the results allow to deduce commuter regions. The forthcoming 'Pendler Mobil' project, which will be carried out in cooperation with the statistical office of North Rhine-Westphalia (IT.NRW), will have the objective of identifying the domains where mobile phone data may contribute to complementing the commuter accounts. Using origin-destination matrices, it is possible to employ mobile phone data for mapping commuting flows during the day. If the SIM card nationality is taken into account, it is even possible to record cross-border commuters.

At European level, Destatis uses an equivalent data record for the ESSnet project 'City data from LFS and big data'. This project examines whether and to what extent indicators of the Labour Force Survey (LFS) can be estimated at the level of Functional Urban Areas by using mobile phone data as auxiliary information. According to the model described by Schmid et al. (2017) and as part of collaboration between Destatis and Freie Universität Berlin, mobile phone data are linked to the unemployment rate and are estimated for smaller areas by using a small area estimation method. The basic question is where and to what extent small area estimation could be used in combinations of mobile phone data and official statistical data. The results show that it is possible to estimate unemployment rate by using aggregated and anonymised mobile phone data at spatially disaggregated level and obtain reliable results for FUAs. Furthermore, reliable estimators for areas without observations can be estimated by this method. The results show also a gain in accuracy compared to the direct estimators, since the uncertainty in the estimates will be reduced due to smaller confidence intervals.

With the start of the feasibility studies, the first steps have been taken on the use of mobile phone data in official statistics. The representativity of the data across Germany will continue to be of essential relevance. To ensure this, further steps have already been taken to obtain data from another mobile operator in Germany. In the next step, the data from both mobile phone providers will be compared

with each other and checked for distortion and representativeness. In addition, it will be investigated whether the population can be extrapolated with the available data.

4.4. Mobile phone data for Disaster Management in Italy

Natural disasters affect hundreds of millions of people worldwide every year. During the events, the timely, accurate information about movement and communications of affected populations can strongly help humanitarian response and, consequentially, reduce the impact of the event on population.

In this context, the mobile phones can offer relevant opportunities for accessing such information and, in the meantime, can provide valuable insights about the behavior of affected populations. Demographic measurements extracted from mobile phone records combined with the indicators provided by an official statistics institute could improve the quality, timeliness and spatio-temporal granularity of statistical information.

This kind of statistical information can also be used building statistical measures correlated with SDGs indicators that are requested for Goals 1, 11 and 13.

During and after severe disasters, mobile phone data can be used to understand how people communicate and the patterns of mobile phone activity. This understanding helps constructing indicators of impact on infrastructure and population and public awareness of the disaster (UN, 2014). Moreover, mobile phone data sources could also allow building indicators on how long it takes to stabilize and return to normal levels after a disaster, as a potential indicator of the rate of recovery for resilience measurement.

In order to obtain the best level of spatiotemporal information, the mobile phone data should be combined with remote sensing data (satellite images), rainfall data, census and civil protection data. In this context real time mobile phones data allows possible immediate distribution and communications of alerts, interaction among all the stakeholders; and most importantly can have a substantial effect in diminishing the negative impacts of the event (UN, 2014, Serrano-Santoyo and Rojas-Mendizabal, 2017)

Although in this work we will focus only CDR record, we want to emphasize that mobile phone, and in particular smart phones, could be useful also in more other ways. In particular has been identified four potential level of information (Poblet et al., 2014):

- data traffic due to apps: these data are collected by platforms (Big Data);
- mobile phone sensors: mobile phones are continuously generating data from their internal sensors, including GPS, accelerometers, gyroscopes and magnetometers;
- mobile social network connections: users generate data by using social media and offer their own information on events (e.g. taking a photo of damage, tweeting about weather conditions, etc.).

- microtaskers by mobile phone: users create content such as adding roads or buildings to satellite images, in this way users became active participants.

5. Webscraping

Many businesses nowadays have a website on which they display information about their company. As a result more and more NSI's study these websites and their content to be used as a source of information. Because much of the information is available as text, text analysis is an important part of these studies. Such studies usually are tailored to a particular use, which may affect the precise sequence of steps used. In this document, we will describe the approach followed by Statistics Netherlands during the study of small innovative companies performed at as an example for the text analysis of websites (van der Doef et al., 2018). The aim of this study was to develop a classifier capable of determining if a company is innovative based on the text on their website. Its approach can fairly easy be generalized to other topics.

Getting an overview of the innovative companies in a country is a challenging task. One of the ways of doing this is setting up a survey to contact a sample of companies; for instance, by phone or via a questionnaire. The response can be used to derive how many innovative companies there are in a country or area. This approach, however, puts a burden on companies and may result in a considerable non-response. Another downside is the fact that usually the focus of such a survey is on large companies and less on smaller companies. In this way, a lot of information on small innovative companies, such as start-ups, is missing. Therefor an alternative approach was developed focussing on using the text on the main web page of a company to determine if it is innovative. The following steps were applied, namely:

1. Selecting a set of known innovative and non-innovative companies;

2. Making sure that for each company the correct URL of their web site was available;

3. Scraping the main page of each web site and preprocessing the text displayed;

4. Developing a model to determine if a company is innovative or not based on the pre-processed texts.

We started with a sample of 3000 innovative and 3000 non-innovative companies as determined by the Community Innovation Survey of Statistics Netherlands. The companies were randomly selected from the sample of 10.000 companies surveyed in 2016. This provided a training and test set.

5.1. Getting web texts

The first thing observed in the sample of 6000 companies was that two-thirds of their URL's were absent from the business register. These URL's were added via the fully automated URL finding approach developed in WP2 of the ESSnet Big Data: this approach uses the Google API and the name and location of the company as search terms to determine its URL (Stateva et al., 2017). Next, the results were manually checked and -at the same time- any ambiguous results were dealt with. For

around 300 companies, URLs needed to be additionally searched by hand. Subsequently, for each URL, the text displayed on the page referred to was scraped. This was done by scripts written in Python or R.

5.2. Processing texts

All subsequent processing steps were performed in Python 3.6. First the language of the text was determined with the langdetect library. The text was usually Dutch or English. Language detection was needed because i) many of the subsequent processing steps are language dependent and ii) it was used an additional feature in the classification step. Next, the texts were converted to lower case, punctuation marks and stop words were removed and the remaining words were stemmed. The resulting processed text was used as input for the classification step.

5.3. Classification

A model was trained to optimally classify companies as either innovative or non-innovative. Accuracy was used as the performance measure. Various classifiers included in the scikit-learn library were compared. With a 70%-30% training and test set, Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machines and Neural Networks were tested. For all classifiers the document-term matrix used Term Frequency-Inversed Document Frequency (TFIDF) as weights and 10-fold cross validation was used. All classifiers performed reasonably well; all above 80%. Logistic Regression with the L1-norm performed best. Here an accuracy of 91% was obtained. Despite the addition of other features, only language was included in the final model; An English web site increased the change that a company was innovative. All other features in the model were text based. However, upon careful study of the other terms included in the model, it was found that special attention needed to be paid to words with a length of two characters; the default setting in Python includes these words. Web pages that displayed a lot of email-addresses and URL's were found to contain large amounts of two character length words after processing; such as nl and eu. As a result, including words of two characters would make a model very sensitive to such features. However, excluding them resulted in a decrease of the model's accuracy; 63%. Here again the best case was logistic regression with L1 norm. Despite the reduction in accuracy, it was decided to focus on an approach solely using words of three character lengths or more. After various attempts, the combination of unigrams and word embeddings was found to perform best, resulting in an accuracy of 93% for the logistic regression model. The word2vec library was used for the latter purpose. Advantage of word embeddings are that the contexts of the words, their surroundings, also become included in the model which usually enhances its performances.

5.4. General remarks

Our text analysis experiences obtained so far suggest that the steps followed to detect innovation from web pages can be generalized to many other topics. As long as text based documents with considerable amounts of words are available. There are however cases in which this is much more troublesome. The first exceptions to this rule are social media messages and other 'documents' with a limited number of words or characters. Here, the processing steps followed are much more sensitive to the research question at hand. For instance, removing stop words from social media messages will very likely negatively affect the overall performance of a model. In short messages any word may provide valuable information. In some cases more advanced modelling approaches may be needed, such as

Deep Learning, or the addition of some carefully constructed extra feature is required. An example of the latter is adding information of the occurrence of distinctive grammatical structures in a text. Our current studies into the annual reports of sustainable companies suggest the need for such more advanced approaches.

6. Social media studies

Social media messages are created in large amounts on various platforms. Routinely collecting all of them is a tremendous effort. For many of our studies huge amounts of social media messages were needed from preferably the entire Dutch population active on as many platforms as possible. We therefore purchased access to the collection of public social media messages gathered by the Dutch company Coosto (2014). This company routinely collects public social media messages written in the Dutch language on the most popular social media platforms in the country, such as Twitter, Facebook and Instagram. Their data collection additionally includes Dutch messages and reactions posted on public blogs and forums and on many publicly available web pages, such as those of newspapers and news sites. A total of 400,000 sources are continuously monitored. This has resulted in a collection composed of more than 4 billion messages covering the period of 2009 until the present. Around 3 million new publically available messages are added per day. Through a secure online interface, the messages can be queried in a convenient fashion. A free accessible alternative for academic Twitter studies is the collection of Dutch Twitter messages collected by Twinl. The web site of this organization (https://twinl.surfsara.nl/) provides access to academic researches and associated organizations without costs.

In all our studies we attempt to look at messages of Dutch people only. By using specific selection criteria, such as all locations within the Netherlands or by using specific Dutch selection words, this population is tried to be selected as good as possible. However, it was found that occasionally a considerable amount of messages from people living in the upper part of Belgium, i.e. Flanders, were included. These people also speak and write Dutch, but there are considerable differences in its use. To assure that only messages from the target population of Statistics Netherlands are studied, a specific 'Netherlands Dutch' and 'Flemish Dutch' classifier was developed. This enabled the exclusion of Flemish texts and is used were needed.

6.1. Social tension indicator

At Statistics Netherlands research has been done on the development of a high frequent indicator for feelings of unsafety; it is called the social tension indicator (CBS, 2018a). Because of the rapid availability of social media and the behavior of Twitter users in particular, this data source was investigated for this purpose.

The work started with input from the safety monitor survey of Statistics Netherlands. The questions and interview instructions were used during in-depth interviews to determine which words Dutch people associate with feelings of safety and unsafety. To the list of words obtained synonyms and antonyms were added, resulting in a final list of 350 words. This list was checked by Statistics Netherlands experts. Subsequently, the use of each word associated with feelings of unsafety was checked on Twitter. Only words that were used more than 10.000 times in Twitter messages over a period of 6 years were kept. This resulted in a final set of nearly 150 words used to select messages indicative for expressing a feeling of unsafety. It was found that messages on sport events and politics

often included one or more of these words and severally disturbed the indicator. These messages do not only occur in large amounts but many of them also expressed a negative sentiment. It was therefore decided to exclude these groups of messages. To validate the findings of the indicator, the peaks observed were checked by i) looking at events that took place at and just before the date of the peak in the Netherlands and by ii) looking at the words included in the messages send. Both should match. It was also checked to what kind of events the peaks referred. It was found that these were nearly all national or international events and indicated a collective feeling of unrest and tension in the Dutch society active on Twitter. Hardly any local event generated enough amounts of messages to produce peaks. This revealed two important findings. The first one is that the indicator developed is not specific for feelings of unsafety but is more focussed on feelings of unrest and tension. The second one is that the indicator focusses on the tension in the Dutch society as whole. The results of the social tension indicator developed are shown in Figure 6.1.

Figure 6.1 reveals that the social tension indicator displays peaks on top of a more or less fairly stable baseline; at least until the end of 2017. On the day of a peak an event took place that resulted in the creation of Twitter messages in the Netherlands expressing feelings of unrest and tension. The large peak in 2010 for instance, is caused by the disturbance of World War II commemoration day on May 4th in Amsterdam. A considerable number of peaks are caused by terrorist attacks, the MH17 disaster (July 17th 2014), the attacks in Paris (November 13th, 2015) and in Brussels (March 22nd, 2016) are examples of this. At the end of 2017, from December onwards, the baseline starts to increases steadily. This indicates a clear change in social tension on Twitter in the Netherlands from that period onwards. In the messages included the words 'police', 'live', and 'problems' were increasingly used in combination with the words indicative for social tension. The combination of the two most occurring words in all 2018 messages included, made more clear what is going on: these two words are: ous society. This strongly suggests that the messages included indicate worries on societal changes in the Netherlands. During 2018 a number of events occurred that are indicative for these changes. The first one is a peak on September 8th in which people react on the migration plans of the Dutch government and in particular to the plans to expel two Armenian teenagers. The second is a peak on November 17th composed by messages on the riots during the entry of Saint Nicolas in the Netherlands (the Dutch father Christmas) including the discussion on the race of its helper 'Black Pete'. The third peak concerns messages on the possible violation of the privacy law by Dutch intelligence services. Messages pointing to the 'yellow vest' movement start to occur in November 2018 but do not lead to peaks in the social tension indicator.

Future plans focus on developing a real time indicator, comparing the findings with those provided by others on similar phenomena and a more in depth validation of the results obtained.

6.2. Sentiment and consumer confidence

The Consumer Confidence Index (CCI) is based on a monthly survey, called the Consumer Confidence Survey (CCS), and measures the opinion of households residing in the Netherlands about the economic climate in general and their own financial situation. In an attempt to reduce administration costs and response burden, Daas and Puts (2014) developed a sentiment index from social media sources that could be used as an alternative indicator for the CCI. They used messages posted on the most popular social media platforms in the Netherlands, written in the Dutch language.

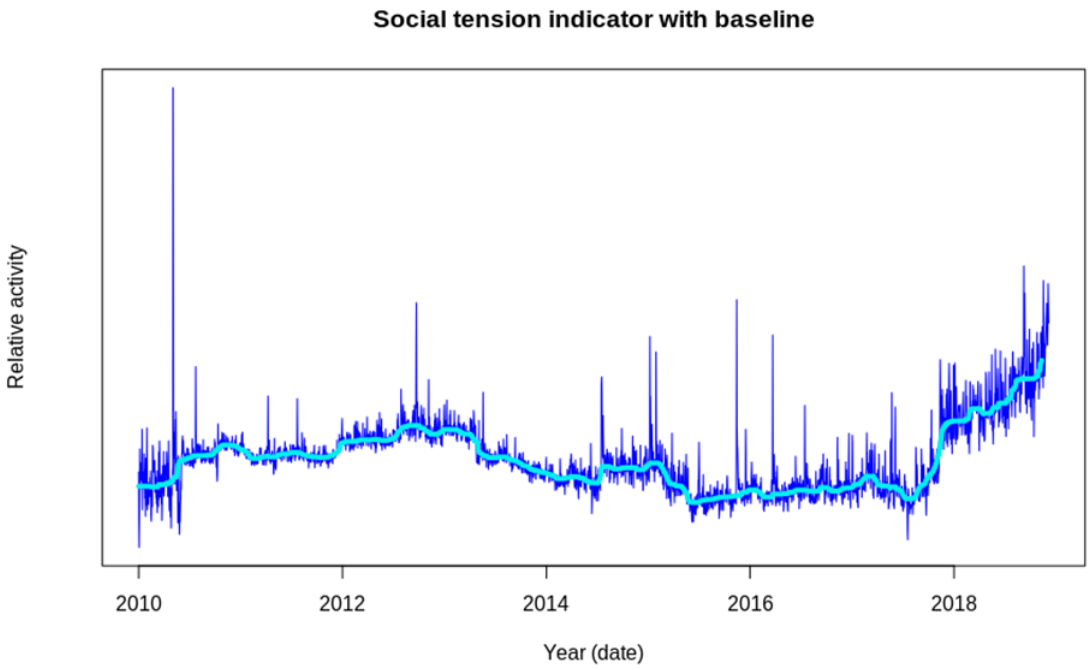


Figure 6.1: Social tension indicator

6.2.1. Sentiment determination

Apart from the message's content and some basic information of the user, the sentiment of the messages collected is automatically determined by Coosto. This is done by checking whether a message expresses a negative or positive opinion. For this purpose a proprietary variant of a sentence-level based classification approach is used (for an overview see Pang and Lee (2008). The approach strictly determines the overall sentiment of the combination of words included in each message. The sentiment classification of the words in the Dutch lexicon is used, in a fashion similar as described by van Assem et al. (2013), to which the sentiment of the informal words and emoticons used on social media are added (Velikovich et al., 2010). The overall sentiment of a message is assigned essentially as described by Esuli and Sebastiani (2006). This results in messages to which either a positive, negative or neutral label is assigned. Neutral messages exhibit no apparent sentiment, e.g. objective sentences, and account for around 60% of all messages. At the level of individual messages sentiment classification will obviously contain errors. However, since we are only interested in the aggregated sentiment of messages created during specific intervals (e.g. days, weeks, months), such errors will generally cancel out because of the enormous amounts of messages produced, see O-Connor et al. (2010) for more details. They may however still be potentially biased. Our studies usually included aggregates of 2 to 75 million messages per time interval studied.

6.2.2. Data selection and analysis

Via a secure web interface the database of collected public Dutch social media messages of Coosto was accessed. In the interface keywords, a time period and the various social media platforms to include were specified. Query results, such as the total number of messages and the number of positive and negative sentiment assigned messages included in the period studied, were exported at an aggregated

level. Routinely, results were exported as daily aggregates in CSV-format for more rigorous analysis. For this the open source statistical software environment R was used. In R, the CSV-files were loaded and the total number of positive and negative assigned sentiment messages were aggregated at selected time intervals, e.g. 7, 14, 21 or 28 days. The average sentiment for each interval was calculated by subtracting the percentage of negative classified messages from the percentage of positive classified messages included. Next, the social media sentiment findings were aligned with monthly consumer confidence data covering the same period and Pearson correlation coefficients were determined.

In Figure 6.2 the Social Media Index (SMI) is compared with the CCI for the period June 2010 until March 2015. Both series are clearly on a different level but show a more or less similar evolution. During the presented period, the CCI is always negative, while the SMI is always positive. The size or amplitude of the movements of the CCI are also considerably larger compared to the SMI. Many factors are responsible for this difference since the CCI is based on a survey where data collection is conducted by telephone and the SMI is based on classifying messages on Twitter and Facebook. The interesting question is to which extent the evolution of both series is similar.

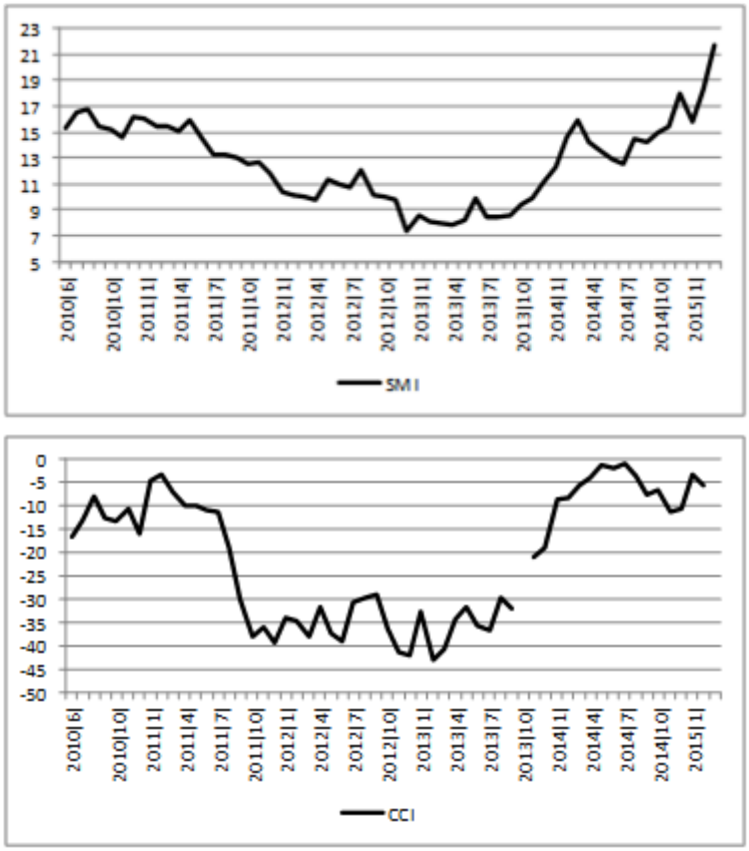


Figure 6.2: comparison of the Social media index (SMI, upper panel) with the Consumer confidence index (CCI, lower panel)

6.2.3. Structural time series model for CCI ands SMI

van den Brakel et al. (2017) adressed the question how this additional information can be used in the context of official statistics. The SMI can ofcourse be published as an indicator on its own. It might, however, also contain valuable information to improve the CCI estimates. van den Brakel et al.

(2017) explored how the CCI and SMI time series can be combined in a bivariate estructural time series model with the purpose to improve the accuracy as well as the timeliness of the CCI. In this context the SMI serves as an auxiliary series in a bivariate structural time series model that allows to model the correlation between the unobserved components of the structural time series models, e.g. trend and seasonal components. If such a model detects strong positive correlations between these components, then this might further increase the precision of the time series estimates for the sample survey. Indicators derived from social media are generally available at a higher frequency than related series obtained with periodic surveys. This allows to use this time series modelling approach to make early predictions for the survey outcomes in real time at the moment that the outcomes for the social media are available, but the survey data not yet. In this case the social media are used as a form of nowcasting.

In the bivariate model developed for the CCI and sMI, both times series are modelled with a smooth trend model (Durbin and Koopman, 2012). The CCI also contains a trigonometric seasonal component (Durbin and Koopman, 2012) and a level intervention for the break in September 2011. To investigate the additional value of the SMI as an auxiliary time series, the trend estimates for the CCI obtained with the bivariate model are compared with trend estimates based on a univariate model for the CCI that contains a smooth trend, a trigonometric seasonal component and the same level intervention for September 2011.

The model detects a strong positive correlation of about 0.92 between the slope disturbances of the CCI and the SMI. In Figure 6.3 compares the direct estimates for the CCI series with the smoothed trend plus intervention under the univariate model and the bivariate model. The level and evolution of the smoothed estimates for the CCI series are almost identical under the univariate and bivariate model.

Figure 6.4 compares the standard errors of the direct estimates for the CCI series with the smoothed trend plus intervention under the univariate model and the bivariate model. As follows from Figure 11, the standard error under the bivariate model is slightly smaller compared to the standard error under the univariate model, as expected given the strong and significant positive correlation between the trend disturbance terms of both series.

6.2.4. Nowcasting excersise

A drawback of sample surveys, however, is that they generally are less timely compared to social media sources. The additional value of the SMI becomes more clear when the higher frequency of this series is used to produce early predictions or nowcasts for the CCI with the bivariate state space model. If during month t or directly at the end of month t a first early prediction for the CCI is required, the univariate model can only produce a one-step-ahead prediction. As soon as during month t or at the end of month t results for the SMI series become available, the bivariate model exploits the strong correlation between the series to make a more precise prediction for the CCI, already before the direct estimate for month t becomes available.

To illustrate the additional value of the SMI in a nowcast procedure for the CCI, we compare in the upper panel of Figure 6.5, the one-step-ahead predictions for the trend plus intervention of the

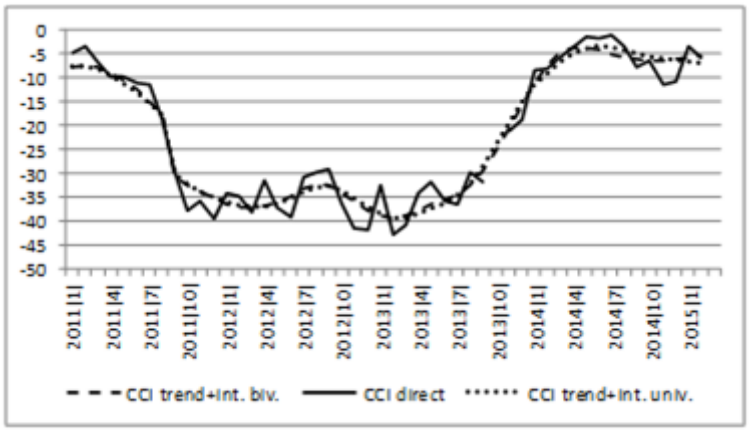


Figure 6.3: CCI comparison of the direct estimates and smoothed trend plus intervention under the bivariate and univariate model for CCI

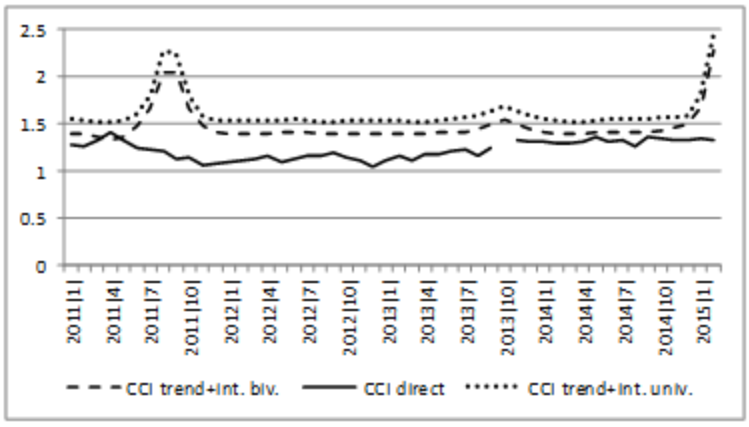


Figure 6.4: CCI comparison of standard errors direct estimates and smoothed trend plus intervention under the bivariate and univariate model for CCI

CCI series obtained with the univariate model with the estimate obtained with the bivariate model if the SMI for month t is available but the direct estimate of the CCI is still missing. The smoothed estimates for the trend plus intervention of the CCI obtained with the univariate model are included as a benchmark. In the lower panel the standard errors of these three estimates are compared. The Figure illustrates that the SMI improves the stability and precision of nowcasts for the CCI

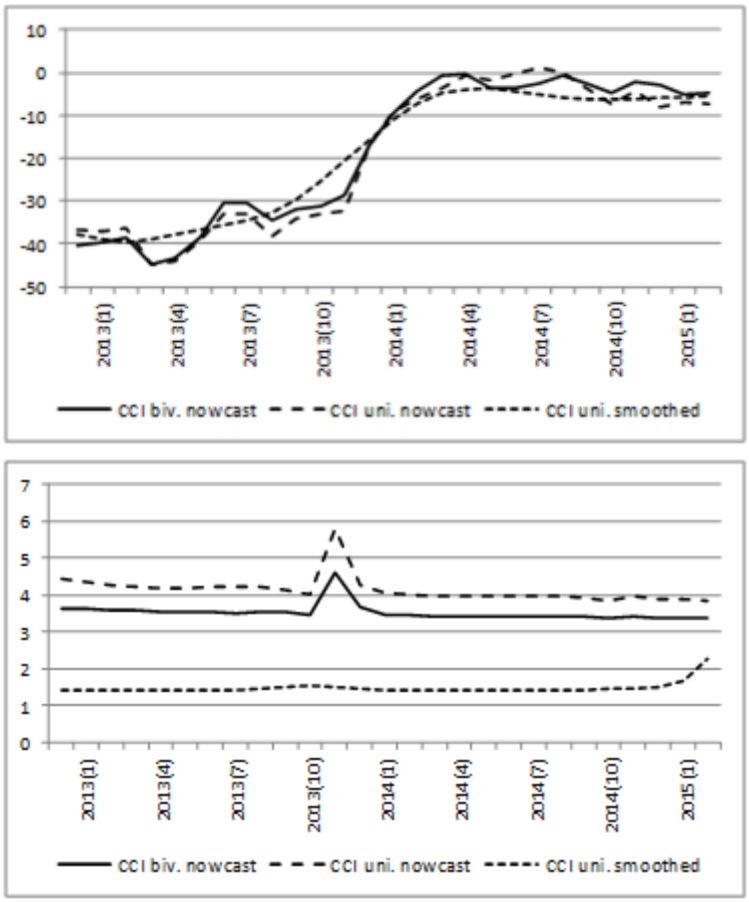


Figure 6.5: comparison estimates for trend plus intervention CCI series; one-step-ahead prediction univariate model (CCI uni. nowcast), bivariate model if the SMI for month t is available but the direct estimate of the CCI is missing (CCI biv. nowcast) and smoothed estimates with the univariate model (CCI uni. smoothed). Upper panel compares point estimates. Lower panel compares standard errors

7. Data found on the Web

7.1. Estimating unmetered photovoltaic power consumption

This chapter describes an example where time series data on electricity exchange on the high voltage grid and meteorological time series that can be downloaded from the internet are used to estimate the amount unmetered photovoltaic power consumption.

Energy accounting encompasses the compilation of coherent statistics on energy related issues in countries, including the production and consumption of electricity. A complete picture of demand and supply of electricity must include data on electricity production outside the energy industries, such as electricity produced by domestic photovoltaic (PV) installations. These PV installations are rarely metered by distribution net operators, hence, their production remains invisible to statistical agencies responsible for the energy accounts. Consequently, the renewable electricity production is difficult to estimate while monitoring it is crucially important for climate policy evaluation.

In the Netherlands—the country studied in this article—an incomplete register of PV installations is available. Such registers can be used to estimate power produced by PV installations using a modelling approach relating installed capacity to produced electricity. In the present article we propose inferring solar power production from causal relations between solar irradiance and consumption of grid power. Since the production of solar power by domestic PV installations results in a reduced consumption of electricity from the high-voltage grid the combination of time series of electricity exchange on the high-power grid and series of solar irradiance contain a hidden signal of unmetered solar power produced by domestic PV installations. In this paper a causal model for these time series to estimate unmetered solar power production is developed.

7.2. Methods

When PV installations produce a lot of electricity, consumers need less electricity from the high voltage grid. The total electricity use in the country is catered for partly by small PV installations, and complemented with grid power. For a given—but unknown—installed PV capacity, the solar power is proportional to the solar irradiance. In this paper the amount of produced domestic solar power is estimated by combining time series of electricity exchange from the high power grid in combination with time series on meteorological data, in particular solar irradiance. Data on electricity exchange on the high power grid in MWh at a daily frequency have been used covering the period from Jan 1st, 2004 through Dec 31st, 2017, which are freely available from the website of the Dutch Transmission System Operator (Tennet). Data on solar irradiance (in J/cm^2) and the temperature (in 0.1^{o}C) at a daily level as well as day length were obtained from the Royal Netherlands Meteorological Institute for the same period.

For solar energy, the most important exogenous variable is the solar irradiance, $I_{\odot t}$. The subscript t indicates that the accumulated total for time period t in days and $\odot$ is the symbol for the Sun. Interest is in the effect of solar irradiance on public grid demand Y_t. The effect of $I_{\odot t}$ depends on how much

solar power ($P_{\odot t}$) is generated. Public grid demand is not only determined by solar power, it mainly dependent on the total electricity demand, D_t and factors like solar irradiance $I_{\odot t}$ weather conditions like the average temperature T_t, length of day L_t and calendar effects C_t. Figure 7.1 shows the causal relationships among these variables by means of a directed acyclic graph (DAG), see (Pearl, 1995). Interest is in estimation of $P_{\odot t}$ based on observations of $I_{\odot t}$ and Y_t. From the DAG (Figure 7.1) it is clearly seen that there are two causal paths between $I_{\odot t}$ and Y_t,

$$I_{\odot t} \rightarrow P_{\odot t} \rightarrow Y_t \tag{7.1}$$

$$I_{\odot t} \rightarrow D_t \rightarrow Y_t. \tag{7.2}$$

The role played by $I_{\odot t}$ renders $P_{\odot t}$ and D_t only conditionally independent; $P_{\odot t} \perp D_t | I_{\odot t}$, which complicates estimation of the direct effects of $I_{\odot t}$ on Y_t, in particular the effect of $P_{\odot t}$ on Y_t. In order to obtain this desired estimate, the causal path (7.2) must be closed. This is achieved by conditioning on—or adjusting for—D_t. This can be conducted using data where $P_{\odot t} = 0$.

In the Netherlands there was no significant presence of domestic PV installations prior to 2010. PV installations were introduced gradually in 2011 and 2012 and started to become more widespread from 2013 onwards. We use the aforementioned observational time series data and construct subset A containing the data from the period 2004—2010 to estimate relation (7.2), where it can be assumed that $P_{\odot t} = 0$. Subsequently we use subset B for the period 2013—2017 to estimate relation (7.1) where we control for the effect $I_{\odot t} \rightarrow D_t \rightarrow Y_t$ from the previous analysis.

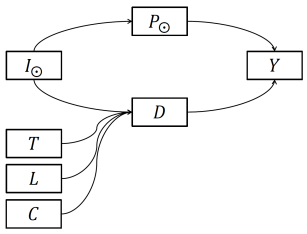


Figure 7.1: Directed acyclic graph (DAG) for the solar power causal model, with $I_{\odot t}$ solar irradiance, $P_{\odot t}$ solar power, Y grid power, D total demand, T temperature, L length of day and C calender effects.

The time series on electricity exchange Y_t for data set A and B are modelled with autoregressive integrated moving average (ARIMA) models (Box et al., 2015). Relations with $I_{\odot t}$, $P_{\odot t}$, T_t, L_t and C_t are included by adding additional covariates terms to the ARIMA model and are known as ARIMAX models. First, an ARIMAX model is fitted to data set A, in which no solar panels are present, hence it is known that in this case $P_{\odot t} = 0$. From this model we establish the effect of $I_{\odot t}$ on D through the regression of Y_t on $I_{\odot t}$. The corresponding regression coefficient $\hat{\beta}_I^{[A]}$ can be interpreted as the effect of solar irradiance on the total electricity demand Y_t. This relation is used to correct Y_t in data set B for the effect of $I_{\odot t}$ on D, i.e. $\tilde{Y}_t = Y_t - \hat{\beta}_I^{[A]} I_{\odot t}$. Then an ARIMAX model is fitted again to the corrected data $\tilde{Y}_t$ in set B and the regression of $\tilde{Y}_t$ on $I_{\odot t}$ is used to derive the production of solar power. Let $\hat{\beta}_{I,y}^{[B]}$ denote the regression coefficients from the regression of $\tilde{Y}_t$ on $I_{\odot t}$ in data set B. The subscript y stand for year and denote that separate regression coefficients $\hat{\beta}_{I,y}^{[B]}$ are assumed for the years $y = 2013$, ..., 2017. Solar power estimates on a daily basis are obtained by $\hat{P}_{\odot t} = \hat{\beta}_{I,y}^{[B]} I_{\odot t}$. Annual estimates are

finally obtained by aggregating over the days within a year.

7.3. Results

The order for differencing the series d, the required number of AR lags p and MA lags q are chosen by minimising the Akaike Information Criterion. It turns out that first order differencing ($d = 1$) is required to render the Y series stationary. The AR order was determined to be $p = 6$ and the MA order $q = 1$. As covariates $I_{\odot t}$, T_t, L_t and C_t are included in the ARIMAX model. We found seasonal components to be insignificant and conjecture the reason for this to be that calendar effects are present in the model, essentially fulfilling a role similar to seasonal components.

Daily solar power is estimated following the method described in Section 2 and is shown in Figure 7.2. There is a clearly increase over time due to the increasing number of PV installations in the country, collectively producing more and more power. The estimated regression coefficients and their standard errors are given in Table 7.1. In addition, the table contains the estimated annual solar power $P_{\odot t}$, the estimated total electricity demand $\hat{D}$ which is the sum of total grid demand and solar power, and in the last column the solar power as a percentage of the total. In the years 2013 and 2014 the regression coefficient is not significantly different from zero. However, considering the five years together a gradual uptake of solar power is clearly visible. In 2017, unmetered solar power is estimated to account for just under 2 percent of total electricity use.

To evaluate how well the ARIMAX model fit the series several diagnostic checks are applied to the standardized residuals to test the assumptions that these residuals are identically and normally distributed. In addition the estimate of solar power production based on the ARIMAX model are compared with the official figures of Statistics Netherlands. Figure 7.3 shows estimated unmetered solar power consumption (our model, solid line) and estimates by CBS of total solar power consumption (dashed line) and of the consumption of power from domestic PV installations (dashed line). The latter is derived from an incomplete register of PV installations and assumptions about the power production of these installations. The trends of the CBS total and our model are remarkably well aligned showing predominantly a level difference. Since the estimated total in the official statistics of CBS includes metered solar power as well—from large solar power farms—it is expected that the unmetered component is lower than the total. The unmetered solar power is often thought to be supplied mainly by PV installations at private homes. In the years 2013, 2014 and 2015 the official estimate of domestic PV consumption is close to our model estimates, but the two diverge in 2016 and 2017. No data are publicly available on the share of metered versus unmetered solar power in the total CBS estimates, nor in the domestic versus business consumption. If all estimates shown in Figure 7.3 are correct, one may conclude that businesses must have installed smaller—hence unmetered—PV installations in the most recent years, 2016 and 2017, before which it was domestic solar power consumption that accounted for most unmetered solar power.

7.4. Conclusions

Reliable statistical information on the use of renewable energy is relevant in order to monitor the implementation of sustainable development. To this end, a time series model is proposed to estimate the unmetered solar power production as a hidden signal in time series of exchange of electricity on the high voltage grid and meteorological time series on solar irradiance, temperature and day length.

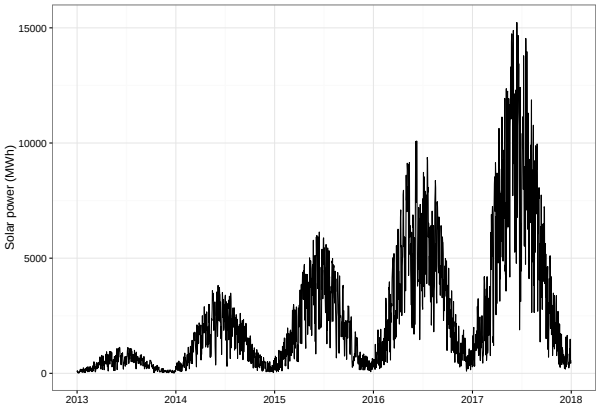


Figure 7.2: Estimated solar power for the years 2013—2017 in MWh.

Table 7.1: Results of the ARIMAX model fit on data set B.

Year	$\hat{\beta}_{I,y}^{[B]}$	SE	$\hat{P}_{\odot t}$ (MWh)	$\hat{D}$ (MWh)	Percentage solar
2013	-0.390	0.787	140,877	101,554,484	0.14%
2014	-1.296	0.797	485,381	99,549,220	0.49%
2015	-2.004	0.755	774,212	100,436,422	0.77%
2016	-3.409	0.828	1,275,643	102,065,655	1.25%
2017	-5.086	0.807	1,867,628	103,223,204	1.81%

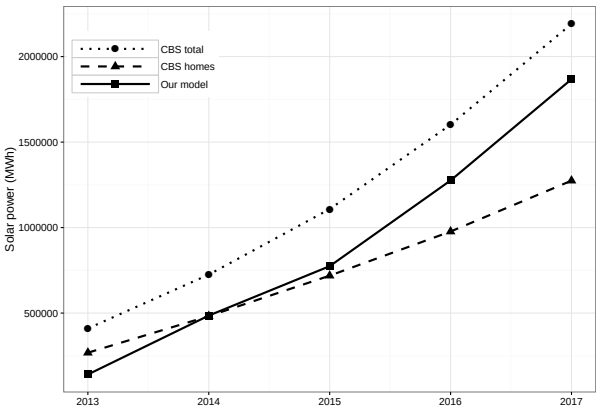


Figure 7.3: Comparison of our model results (solid line) with official statistics published by CBS on total solar energy consumption (dotted line) and the amount consumed by households (dashed line).

All time series data used in this analysis are freely available from the internet.

The model estimates are not in disagreement with official statistics on solar power consumption. Estimates have been produced on unmetered solar power production and consumption. While official statistics are at annual level, our modelling approach produced daily estimates. In contrast with the regular official statistics, no administrative or survey data on PV installations in the country was required. Hence, the proposed model can be applied easily, quickly and widely, and could be particularly useful in countries where no good estimates of unmetered PV electricity are available yet.

8. Use of Satellite Data to Measure Indicators for the Sustainable Development Goals

8.1. Introduction

In the natural and environmental science remote sensing applications and the use of satellite imagery are well-established as a source of information about vegetation composition, housing structures, weather forecasting, monitoring forestry and agricultural activities. In the social and economic science however, the application of this type of data source is less developed due to potential pitfalls of this data type and integration problems with traditional data sources such as survey data (cf. Hall, 2010, p. 1 and Donaldson and Storeyard, 2016, p. 190ff.). A popular sensor for earth observation data used in social and economic science is the LANDSAT sensor due to the wide range of collected data by the sensor and the long lifespan of this program. LANDSAT 1 has been launched by the National Aeronautics and Space Administration (NASA) in 1972 and was the first satellite designed with the purpose to monitor the surface of the earth. Since then a series of LANDSAT sensors has been launched with improved spatial, spectral and radiometric resolution (cf. Hall, 2010, p. 3 and Donaldson and Storeyard, 2016, p. 182). Other satellite sensors often used as data sources for social and economic science studies are the SPOT series of satellites launched by CNES (Centre National d' Études Spatiales), IKONOS and Quickbird (cf. Hall, 2010, p. 4 and Donaldson and Storeyard, 2016, p. 182f.).

This chapter aims to provide a summary about possible uses of satellite data for constructs which are traditionally of concern in the social and economic science and connect these ideas to the possibility of using satellite data as an information source to measure and monitor the Sustainable Development Goals (SDGs) of the Agenda 2030. In Section 8.2 the results of the Copernicus project are summarized. In Section 8.3 the benefits of data derived from satellite imageries for the social and economic science with reference to traditional surveyed data are explained. This Section highlights potential areas for which satellite data can help to overcome common pitfalls of survey data such as costs and time lags. Two examples of constructs of interests in the social and economic science and examples for the integration of satellite data with traditionally surveyed data are given in Section 8.4. The constructs *poverty* and *quality of life* are analysed. In this context, satellite data is used to detect slums with classification techniques or as proxies for the environmental welfare dimension of quality of life. The question for which SDGs satellite data might be appropriate is summarized in Section 8.5. The overview table in this section represents a summary of results from existing literature which discusses the use of satellite data for all SDGs. Not all SDGs can be measured with satellite data and several indicator variables proposed are assumed to be applicable for more than one SDG. Problems arising from this approach and potential pitfalls of satellite data are explained in Section 8.6. In Section 8.7 a summary and conclusion of this chapter is given.

Satellite	Pixel resolution	Number of satellite coverages
Landsat-8	30m	2
Sentinel-2	10m	3
Sentinel-1	30m	4
RapidEye	5m	2-7

Table 8.1: Summary information on the satellite characteristics

8.2. Evaluation of the Copernicus project in Germany

Cop4Stat_2015plus was a joint project of the German Federal Statistical Office (Destatis) and the German Federal Agency for Cartography and Geodesy (BKG). The scope of the project was to explore the use of remote sensing data for providing statistical information on land cover and land use. The content and the results of the remote sensing part of the project, which were produced by the BKG, will be summarized in this section based on the unpublished project report from (Stephan Arnold [STBA], Sylvia Seissiger [BKG], Sarah Kleine [STBA], Michael Hovenbitzer[BKG], Angela Schaff [STBA], 2019). The project was designed to act as a feasibility study for the use of the Copernicus data to produce statistics related to land use and land cover. While the term 'land use' indicates socio-economic consumption of land, for instance, agriculture, recreation or residential use, the term 'land cover' expresses information on the presence or absence of physical materials of the Earth's surface (e.g. different types of vegetation like crops, grass, woodland, bare rocks or artificial surfaces).

In this project, information about land cover had to be derived from the processing and analysis of satellite data in compliance with the LUCAS (Land Use and Coverage Area frame Survey) [1] nomenclature, which is used by Eurostat. The methodology was tested in the administrative area "Regierungsbezirk Darmstadt", located in the southern part of the federal state Hesse, covering an area of approximately 7500 m^2 and a population size of about 4 million people.

To answer the research questions, whether satellite data can be used to generate statistics on land cover according to the LUCAS nomenclature and whether they are comparable with respect to official German area statistics, a variety of different data from remote sensing sources have been explored. The use of data from optical, multispectral scanner (LANDSAT 8, SENTINEL-2 and RapidEye) as well as SAR (Synthtic Aperture Radar) data (SENTINEL-1) was investigated. Furthermore, raster data from the Copernicus high resolution layers [2] regarding imperviousness and forests as well as digital elevation data were assessed. Finally, vector data comprising topographic data and the German land cover model (LBM-DE 2015) were used and administrative boundaries in vector format served as spatial statistical reference units. Summary information on the pixel resolution and the number of satellite images available for the region is given in Table 8.1. It can be seen that the highest resolution is provided by images obtained from the RapidEye satellites. While the images from the other satellite systems (Landsat 8, Senitel-1, Senitel-2) cover the whole study area, the RapidEye data were processed

[1] Eurostat (2018): Statistics explained - LUCAS - Land use and land cover survey `https://ec.europa.eu/eurostat/statistics-explained/index.php/LUCAS_-_Land_use_and_land_cover_survey` (accessed on 05 February 2019).

[2] `https://land.copernicus.eu/pan-european/high-resolution-layers` (accessed on 05 February 2019).

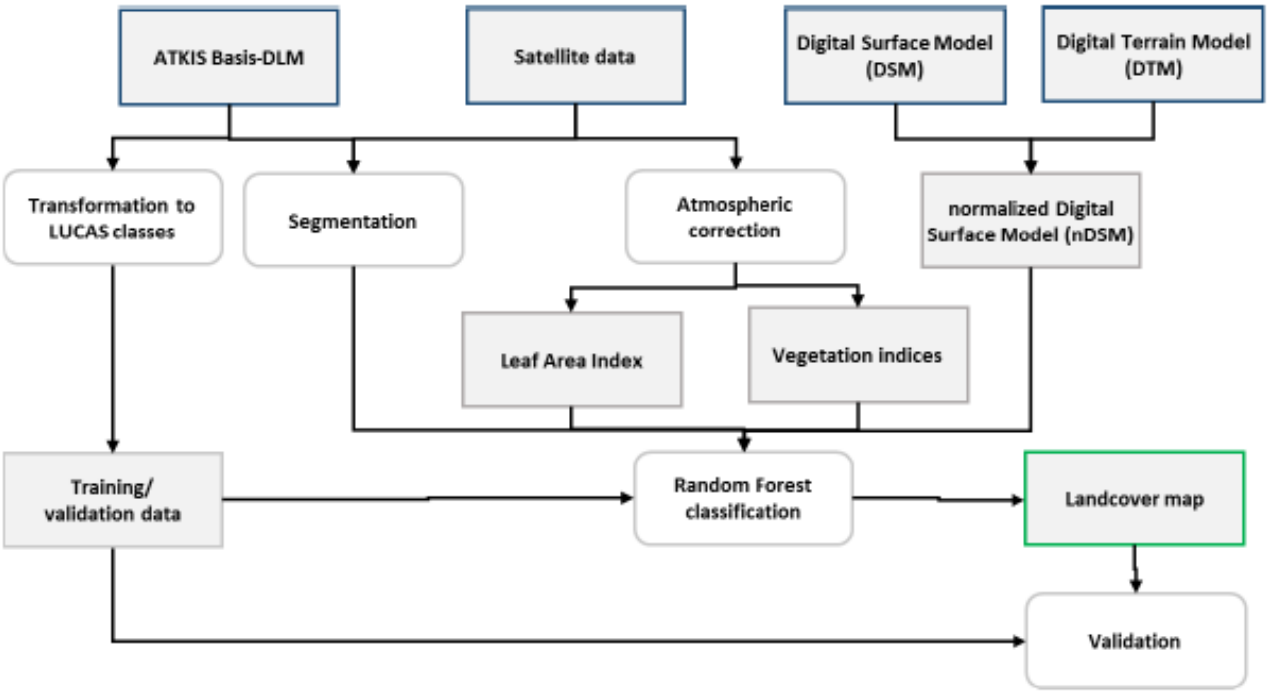


Figure 8.1: Workflow used to derive land cover information from satellite and height data (Stephan Arnold [STBA], Sylvia Seissiger [BKG], Sarah Kleine [STBA], Michael Hovenbitzer[BKG], Angela Schaff [STBA], 2019)

for a subset of four municipalities only.

The combination of data from sensors with different acquisition times, spatial resolution and spectral characteristics enhanced the temporal and spectral coverage of the designated test area. After the data had been acquired, an atmospheric correction was applied to the multispectral data using ESA's "Sen2Cor" tool. This pre-processing was necessary to obtain precise information about the Earth's surface and to be able to compare optical remote sensing data sets spatially and temporally (DLR, 2019). Senitel-1 data were pre-processed by correcting for orbit displacements and speckle, calibrated for radiometric distortions and georeferenced. Thereafter, various vegetation indices were calculated using remote sensing data. In order to train the classification algorithms, topographic reference data from German land surveying authorities (ATKIS Basis_DLM) were used and semantically translated to the targeted LUCAS classes where possible. To overcome anticipated problems of distinguishing between similar types of land cover, for instance shrubs and trees, data from a normalized digital surface model (nDSM) compromising absolute height differences were used. The workflow of the classification process is illustrated in Figure 8.1.

The results from the work conducted in this project showed that Copernicus data, and images from

the optical sensor Sentinel-2 in particular, are useful to derive land cover information. Special emphasis was given to the main categories of land cover, in particular, the level-1 classes according to the Eurostat's LUCAS nomenclature. The use of additional data sources such as digital height data and Sentinel-1 radar backscatter further increased the overall accuracy of the classification. The final algorithm developed by the BKG within the project successfully discriminated between the classes *Artificial Land*, *Cropland*, *Woodland* and *Water Surfaces* within an overall accuracy of above 90 percent (according to the AKTIS Basis_DLM). Classification of smaller classes such as *Wetlands*, *Bare soils* and *Shrubland* were more challenging to classify correctly.

Future work will aim at improving the classification process by optimizing the training procedure of the input data and parameters. Moreover, it is planned to apply the methodologies over the whole of Germany.

8.3. Benefits of Using Satellite Data in the Context of Official Statistics

Many national statistical institutes and international institutions have acknowledged that satellite data has the potential to improve the analysis of concepts of interest which are traditionally measured using survey data. For example, the United Task Team on Satellite Imagery and Geospatial Data has written the handbook *Earth Observations for Official Statistics*, which aims to provide a guide for National Statistical Offices considering the use of satellite data to produce official statistics (cf. UN GWG for Big Data, 2017, p. 2). In the light of the data-driven framework of the Agenda 2030 the use of satellite imagery is considered to be of importance for measuring, monitoring and reporting the indicators of the 17 SDGs (cf. Paganini et al., 2018, p. 11f.). The aim of using satellite data for measuring concepts of the SDGs can be based on three beneficial aspects of this data type when compared to traditional survey data. These benefits are the possibility to generate data from satellite imagery in areas or about concepts of interest which are difficult to cover with survey data, the potential of more timely data and cost efficient panel data and the availability of detailed spatial information (cf. Paganini et al., 2018, p. 11, Anderson et al., 2017, p. 81 and Donaldson and Storeyard, 2016, p. 171). In the following section, the three beneficial aspects of satellite data are explained in more detail.

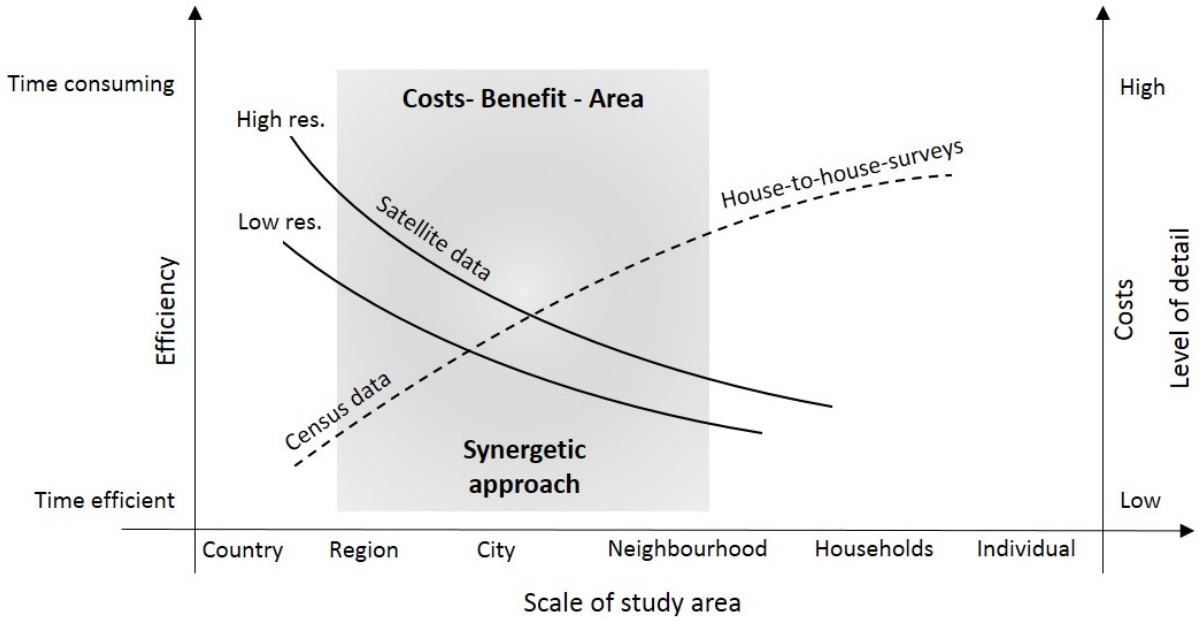
Enrichment of Data Availability for Constructs Usually Measured Using Survey Data:

Using satellite data as a complement to survey data opens up the possibility to increase the data volume for example for surveys from data-poor countries which suffer from inadequate coverage of the target population (cf. Hall, 2010, p. 9ff.) . Furthermore, it can be possible to gain information about hard-to-measure statistics with satellite imagery which are not easily obtained by survey data. Concepts of interest like deforestation or pollution can be hard or impossible to measure with traditional surveys and sensitive topics could also be affected by manipulation or misreporting in traditional surveys. In such cases, satellite imagery offers the possibility to collect information which would not have been available otherwise (cf. Donaldson and Storeyard, 2016, p. 171 and Kit et al., 2011, p. 661). Another interesting application of satellite data to advance the use of survey data is the identification of initial sampling units to be sampled in a ground survey. This has been applied by demographers for the analysis of population counts and may be especially helpful in hard-to-reach or unstructured spatial areas (cf. Hall, 2010, p. 10). In the study by Henderson et al. (2012) for example satellite data

is used to increase data on GDP because the available survey data is lacking in consistent sub-national levels and also suffers from measurement errors. Satellite data on night-time lights and information on light growth are utilized as proxies to augment the data from countries with sparse data on the national income accounts. As a result, they conclude that variables from satellite night-time light data are adequate proxies of economic activity. Especially the model which maps light growth onto a proxy for economic growth proves to be valuable for counties with low quality of national accounts data (cf. Henderson et al., 2012, p. 24f.). In Section 8.4, example studies are mentioned for which satellite night-time light data are used to derive proxies of poverty.

Collection of Timely and Cost Efficient Data:

Depending on the construct of interest, satellite data can be collected more cost and time efficient than household data which has been collected via traditional survey methods. In Ebert et al. (2009) the cost-benefit concept which applies to the use of satellite data for assessing the construct of social vulnerability is explained. The authors conclude that the integration of satellite data and census data is beneficial because satellite data has the ability to overcome the problem that detailed house-to-house survey data is often expensive and time-consuming to collect and census data is in turn less detailed but more efficient. With the integration of satellite data they can overcome this problem because this data type can be collected more frequently at lower costs. The cost-benefit concept is shown in Figure 8.2. Here the dashed line represents the characteristics of census and house-to-house ground surveys in the dimensions time, costs and the scale of the study area. Two types of satellite data with high and low resolution are marked with solid lines. The main statement made with this graph is the identification of the costs-benefit area following from a synergetic approach by integrating satellite and ground survey data (cf. Ebert et al., 2009, p. 290f.).



Source: Ebert et al., 2009, p. 291

Figure 8.2: Cost-Benefit Concept for the Assessment of Social Vulnerability with a Synergetic Approach

Another argument for the use of satellite data is that compared to survey data the time gap between data collection and publication of the estimates can be reduced which can lead to more up to date information and reduce the risk of making decisions based on outdated information (cf. Ebert et al., 2009, p. 277). Often, official statistics are collected with a focus on the observation of changes in the respective construct of interest, like changes in poverty. This is done using labour and cost intensive panel surveys with which data is collected for the same sample units in several reporting periods. Panel data can be for example collected as rotating panel in which the sample units will be part of the survey for a fixed number of reporting periods and will drop out of the survey afterwards. In order to evaluate change adequately, the sample weights have to be adapted to account for the rotation. Panel surveys are also prone to suffer from panel attrition. On the other hand, it is possible to take satellite images in a very frequent manner from the same position and therefore panel data can be collected via satellites cost efficient and with low marginal costs of the data generation method (cf. Donaldson and Storeyard, 2016, p. 172). For example in the study by Lunetta et al. (2006) multi-temporal satellite imagery data MODIS NDVI 16-day composite grid (MOD13Q1) on vegetation indices is used which was acquired between February 2000 and December 2005 from the NASA Earth Observing System (cf. Lunetta et al., 2006, p. 8). This dataset provides an average of the normalized difference vegetation index (NDVI) for a 16 day period and therefore made it possible to develop a detailed model for monitoring locations and distributions for land cover changes in the study area (cf. Lunetta et al., 2006, p. 2ff.).

Provision of High Spatial Resolution and Wide Geographic Coverage:

The third aspect about satellite data often mentioned in the literature is that in comparison to traditional data sources, satellite data has the advantage of being available at higher spatial resolution. This makes it possible to implement satellite data in high-spatial-resolution research designs with low spatial levels and take spatial heterogeneity of the construct of interest into account (cf. Baud et al., 2010, p. 36f., Donaldson and Storeyard, 2016, p. 174 , UN GWG for Big Data, 2017, p. 3 and Anderson et al., 2017, p. 79). An example of an economic science study for which high resolution satellite data was used is Xie et al. (2016). The authors utilized satellite data to develop proxies for the variables income, wealth and poverty rate for the analysis of economic well-being. These variables are typically measured by data collected via traditional surveys, but Xie et al. (2016) have chosen to use high resolution satellite data because reliable survey data is often scarce, not available with sufficient coverage or labour-intensive to collect in developing countries (cf. Xie et al., 2016, p. 3929). Donaldson and Storeyard (2016) state that in this field satellite data has the possibility to gain estimates on economic well-being at spatial and temporal frequency higher than from the data usually available (cf. Donaldson and Storeyard, 2016, p. 190). A great benefit of satellite data also is the wide geographic coverage of this data type. This benefit comes along with the favourable characteristics that satellite data can be collected in a consistent manner and with uniform spatial sampling of the region of interest. Images are taken with substantial temporal coverage from the same location and also at constant frequencies which can be as prompt as daily images (cf. Hofmann et al., 2008, p. 539 and Donaldson and Storeyard, 2016, p. 175f.). The coverage possibilities for satellite data are global and data can be collected without regarding local events like political strife or natural disasters and across borders (cf. Donaldson and Storeyard, 2016, p. 175f.). This could be of great benefit to gain additional information in cases where the collection of survey data is limited due to regional borders. In such cases, the analysis of constructs of interest may be incomplete because influencing cross-border effects cannot be analysed.

8.4. Using Satellite Data to Analyse Constructs of Interest for Official Statistics: Two Examples

The benefits of satellite data described in Section 8.3 have already been acknowledge by many researchers and used as reasoning to integrate satellite data and survey data to analyse constructs for which data is traditionally collected by ground surveys. In the following section some examples for studies analysing the constructs *quality of life* and *poverty* will be explained in more detail with a focus on how satellite data is used to enrich the available data on these constructs as well as the approaches to integrate the two data types. Poverty for example is an important construct for the Agenda 2030 because it is implemented in the SDG 1: *End poverty in all of its forms everywhere.* Quality of life is a multidimensional construct which is measured in the described studies with versatile dimensions of economic, social and environmental welfare (cf. Berrada et al., 2013, p. 407). Socio-economic welfare variables are traditionally measured with ground survey data and satellite imageries are able to gather information about the well-being of the environment. Therefore, quality of life is an interesting example for a construct for which methods of data integration of the two data sources has to be applied. In the Agenda 2030 the environmental welfare dimension of quality of life is reflected by several SDGs for example by SDG 11: *Make cities and human settlements inclusive, safe, resilient and sustainable* (cf. UN General Assembly, 2015, p. 14ff.).

Measuring Poverty with Day-Time Satellite Data:

Numerous studies use object-based classification algorithms on satellite imagery to detect slums in a given area of interest. The underlying assumption of this approach in order to analysis poverty is that residents of slums are deprived of healthy sanitary conditions and do not have enough income to meet the basic needs of daily life and therefore live in poverty (cf. Rhinane et al., 2011, p. 217). In an object-based classification approach areas of slums are detected using panchromatic satellite data of urban areas with sufficient spatial resolution. Different pattern recognition algorithms are available to process the satellite data with the purpose to group identified objects into categories. The classification is based on spatial pattern recognition and the choice of algorithm depends on the available data and the purpose of the study (cf. Rhinane et al., 2011, p. 219). Groups of pixels from the satellite imagery which are uniform or nearly similar are considered to be spectral classes and the identified spectral classes are matched to the information classes of interest such as for example slums and organized living areas. Methods from supervised or unsupervised classification are available. For the supervised methodologies pixels are statistically grouped based on their numerical similarity and the groups are assigned to different information categories. The supervised methodologies rely on a pre-identification of homogeneous representation samples for each information class from the analyst. This pre-identification is used as training areas and the characteristics of each pre-defined area is used to classify the study area into the information classes (cf. Hall, 2010, p. 7 and Taubenböck and Kraff, 2015, p. 110ff.). Interesting applications of object-based classification for detecting slums can be found for example in Rhinane et al. (2011) for the identification of slum areas in Casablanca Morocco with SPOT-5 2.5m resolution georeferenced satellite data from 2004 or in Kohli et al. (2012) who use information from Kisumu, Kenya and Ahmedabad, India and detect slums via object-based classification. In Taubenböck and Kraff (2015) a more detailed explanation on how structural information from satellite imageries is taken to identify slums is given. Kit et al. (2011) use a different approach to identify informal settlements in Hyderabad, India with the concept of lacunarity which is computed with line detection algorithms and principle component analysis. Both methods apply the original multispectral satellite image and produce binary matrices of filled pixels based on the spectral information of the images. The lacunarity is calculated for different sampling windows in the image and informal settlements are identified by this lacunarity information (cf. Kit et al., 2011, p. 663f.). Problems of these methods are that for urban structures it is difficult to classify information classes due to a lack of unique and easy to identify spectral structures, different from agricultural land for which the classification methods can be applied more easily. This could result in identification errors from satellite data (cf. Kit et al., 2011, p. 661).

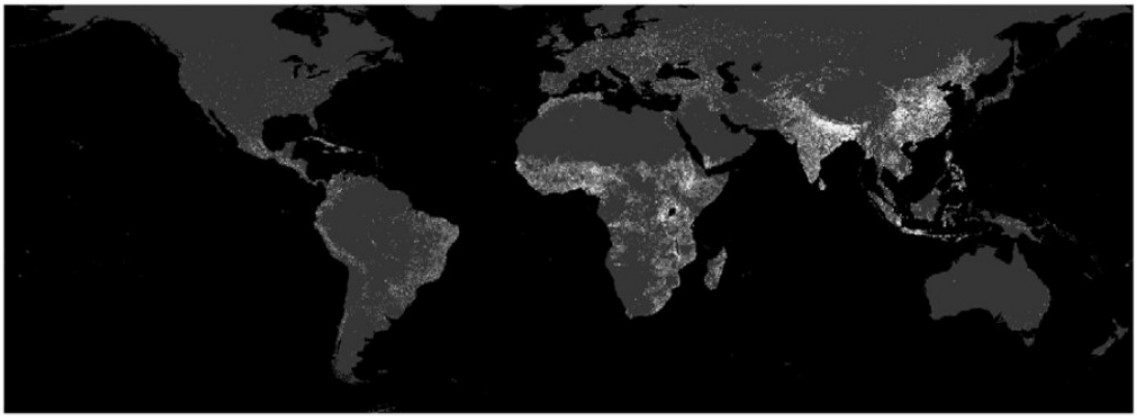
Other studies use a classification-based approach to detect land cover or land use and compare these information with texture information to identify slums in certain areas. Stoler et al. (2012) derive information about vegetation, impervious surface and soil as components of urban environment from multispectral satellite data from QuickBird, ASTER and LANDSAT TM (cf. Stoler et al., 2012, p. 36f.). The aim of the study is to analyse which image-derived information provide a good indicator for slum characteristics which can then be used to generate supplementary data for other cities for which reliable demographic data is not available from traditional surveys. The indicators from

satellite imagery are analysed with a regression analysis together with a socio-economic slum index obtained from census data (cf. Stoler et al., 2012, p. 38ff.). As a conclusion from this analysis the authors identify the strength of correlation with which the three variables derived from satellite data are correlated to the slum index. Vegetation index showed a moderate but the strongest correlation with the slum index out of the three variables, indicating that more impoverished slum like areas can be associated with sparse vegetation. The variable impervious surface showed also a modest positive correlation with the slum index. Soil on the other hand, only showed a week correlation with the slum index and the authors conclude that this variable might not be a suitable indicator to identify slum like areas (cf. Stoler et al., 2012, p. 45ff.). Another study in which the slum index generated by census data is used to verify the application of satellite data for the analysis of poverty is applied is Stow et al. (2007). They use multispectral and panchromatic QuickBird data for image classification and the Universal Transverse Mercator map projection by DigitalGlobe for georeferencing the images. A land variable with a specification for vegetation, impervious surface and soil (V-I-S urban land cover by Ridd (1995)) was derived from the satellite imageries and mapped for the study area. This map is compared with a map derived for the census data based slum index. For the study area it is concluded that areas with high socio-economic indicator values are connected with more patches of vegetation. Areas with low socio-economic indicator values on the other hand are characterized with small patches of soil and impervious land (cf. Stow et al., 2007, p. 5168ff.). Weeks et al. (2007) also use the V-I-C concept and additionally derive a variable from satellite data called texture which indicates the degree of variability in land cover. Slum like areas are associated with little variability because the buildings are made of similar material and other areas as lawns, parks, roadways and so on are not planned. The authors concluded from their regression analysis that the vegetation index is a much better predictor for the slum index compared to the texture variable (cf. Weeks et al., 2007, p. 6f.).

Measuring Poverty with Night-Time Satellite Data:

Night-time satellite data is used to observe the relative brightness and spatial extent of lights and some studies use this information to analyse poverty or economic activity on a global level and compare countries, but also for finer geographical units such as cities (cf. Donaldson and Storeyard, 2016, p. 183f. and Doll, 2008, p. 25). Elvidge et al. (2009) compile a global poverty map constructing a normalized poverty index with night-time light satellite data of the US Air Force Defence Meteorological Satellite Program. These satellite imageries provide information about the average visible band digital number from the lights as extent of brightness and additionally information about population counts are utilized from LandScan 2004 (cf. Elvidge et al., 2009, p. 1654). A normalized global poverty index is calculated by dividing the population count by the brightness of night-time light and multiplying it by 100. In order to verify the connection between brightness and poverty, the normalized poverty index is regressed on the percentage of population which is living with less than 2 dollars per day as reported by the World Development Indicators 2006. This regression showed a good explanatory power with an R^2 of 0.7217. Therefore, the authors conclude that satellite data of this kind can be valuable to improve the knowledge about socio-economic living conditions (cf. Elvidge et al., 2009, p. 1659). Figure 8.3 shows the normalized poverty index in a grey-scale image where lighter areas indicate areas with higher poverty indices. In accordance to the calculation of the poverty index, lighter areas are marked in regions with high population values and no or dim night-time lights (cf.

Elvidge et al., 2009, p. 1655ff.).



Source: Elvidge et al., 2009, p. 1654

Figure 8.3: Normalized Poverty Index

Poverty and economic activity are just two examples of constructs which have been studied using night-time light satellite data. Further descriptions for examples such as urban extent or population can be found in Doll (2008).

Measuring poverty on a global, country and regional level is especially important for the monitoring and measuring of SDG 1: *End poverty in all its forms everywhere and all targets of this goal.* Goal 1.b demands the creation of sound policy frameworks at national, international or regional level to develop appropriate poverty reducing strategies. Proper poverty indicators, which reflect on different aspects of poverty and data of wide geographical span as well as with the ability to capture the spatial heterogeneity will be crucial to achieve this goal (cf. UN General Assembly, 2015, p. 15). As argued above, the benefits of satellite data could be used to foster this development.

Measuring Quality of Life with Satellite Data:

The construct quality of life is often analysed as a composite indicator consisting of several dimensions like social-, environmental- and economic welfare, to reflect on its multidimensionality. Satellite data is often used to measure indicator variables for the environmental dimension which is a vital dimension of quality of life. Indicator variables of other dimensions are typically measured using survey data. The construct itself is not measured with the same indicator variables in each dimension in every study, but some variables with which the socio-economic welfare dimensions are measured are household income, expenditure variables, and variables on the housing situation, education level, unemployment rates, or demographic variables such as population counts (cf. Afsar et al., 2013, p. 374, Berrada et al., 2013, p. 408f., Ogneva-Himmelberger et al., 2013, p. 189 and Stathopoulou et al., 2012, p. 26ff.).

Berrada et al. (2013) for example construct a composite indicator for an urban quality index with

which they want to measure quality of live for the city Casablanca, Morocco. Three dimensions are used to represent the urban quality index: social welfare, economic welfare and environmental welfare. All three dimensions are measured with various indicator variables, the first two dimensions from census data and the environmental indicators are obtained by satellite data (cf. Berrada et al., 2013, p. 407f.). Interestingly, the authors use two types of satellite imagery to derive different sets of variables. The LANDSAT 5 Thematic Mapper image consists of different spectral bands and spatial resolution and can collect thermal infrared data. With the information about reflection of infrared light for example from plants, data can be obtained about green spaces, presence of vegetation, heat islands or cool areas in areas of interest. With the panchromatic SPOT-5 images the variables density of impervious areas and slums are obtained using classification techniques as explained above (cf. Berrada et al., 2013, p. 408f.). Using z-scores to standardize the derived variables, the three dimensions are aggregated with a weighting scheme to a single indicator of urban quality. Finally, the distribution of the urban quality index is discussed by mapping the results for the area of interest in an urban quality index map (cf. Berrada et al., 2013, p. 413). Another common indicator which is used to measure the environmental welfare dimension of quality of life is the normalized difference vegetation index (NDVI). This index can be measured from satellite imageries with different bands and uses the information of reflection contrast of near infrared (NIR) and visible red (VIS) wavelength from vegetation. Therefore, the index reflects the greenness of an area due to vegetation and is often used to evaluate the environmental health and vegetation abundance of areas. It is calculated using the following formula (cf. Elmore et al., 2000, p. 93f.):

$$NVDI = \frac{NIR - VIS}{NIR + VIS} \quad .$$

Lafary et al. (2008) use this indicator calculated from Spaceborne Thermal Emission and Reflection Radiometer (ASTER) data and regress it on socio-economic variables like income for different areas in Vanderburgh County, Indiana. Estimates for the regression parameters were obtained using global regression and geographically weighted regression which does not assume spatial stationarity of the observed values. The regression results showed significant relationships between the NDVI and the socio-economic variables and in general it is concluded that more affluent blocks in the observed area have more access to greater abundance of green areas (cf. Lafary et al., 2008, p. 55ff.). The NDVI is also used in other studies of quality of life such as Stathopoulou et al. (2012), Ogneva-Himmelberger et al. (2013) or Li and Weng (2007). Even though the NDVI is applied in many studies, Weng et al. (2004) state that there is a need for further indicators of greenness because the NDVI is not measured without errors from satellite data. Rather the values can be affected by soil reflection or plant species and therefore need adaptation (cf. Weng et al., 2004, p. 469f.). Other variables for the construct of quality of life which are collected using satellite data are for example land surface temperature, impervious surface, land use or land cover such as for example buildings, roads, parks or forests as local recreation area (cf. Stathopoulou et al., 2012, p. 26, Ogneva-Himmelberger et al., 2009, p. 480, Ogneva-Himmelberger et al., 2013, p. 188, Lo and Faber, 1997, p. 147 and Münnich et al., 2016, p. 96ff.).

In the Agenda 2030 these indicator variables could be of interest for example for SDG 11, in more

detail goal 11a which describes the aim to foster the links between positive economic, social and environmental aspects of urban, peri-urban and rural areas due to more developed planning. The target 11.7 which brings attention to the provision, universal access to safe, inclusive and accessible green public spaces also seems to be measurable with indicators of greenness from satellite imagery (cf. UN General Assembly, 2015, p. 21f.). Table 8.2 of Section 8.5 provides an overview of the SDGs which can be supported using satellite data. This overview shows that the variables which are mentioned in this section to represent the environmental welfare dimension of quality of life are considered to be useful measuring and monitoring other SDGs as well, such as for example the variables land cover or land use.

8.5. Satellite Data to Measure SDG Indicators

8.5.1. Two Principles of the Agenda 2030 Framework: Cooperation Between Partners and Free Data Access

In the light of the Agenda 2030 it is understood and also specifically addressed by SDG 17 that partnerships and cooperation on international level are of great importance for the realization of the goals and targets. Furthermore, it is stated that the success of the Agenda depends strongly on the availability of high quality data with reference to timely production of necessary indicators and accessible data for transparency. This will make cooperation necessary between national statistical institutions, geospatial institutions and so-called custodian agencies like the UN Committee of Experts on Global Geospatial Information Management aiming to support countries with methodologies to monitor the SDG indicators. For different expertise fields such as for example using earth observation (EO) and satellite data for the SDGs, institutions and initiatives have been launched and authorised. The EO4SDG initiative by the Group on Earth Observation (GEO) consists of 104 governments and 106 participating organizations and aims to increase the potential of EO, including satellite data, used to monitor SDGs to foster the realization of the Agenda 2030 and create social achievement through the realization of the SDGs. One part of this mission is the implementation of a Global Earth Observation System of Systems (GEOSS) which provides a broad range of EO data. For EO4SDG GEO works together with national statistical institutes as well as with custodian agencies to develop methodologies to apply EO data for SDG estimation and to build national pilot studies which integrate EO with national statistical information. GEO also works closely together with the Committee on Earth Observation Satellites (CEOS) and this cooperation has published the booklet *Earth Observation in support of the 2030 Agenda for Sustainable Development* which highlights the potential of EO data for the SDG framework (cf. Paganini et al., 2018, p. 21ff. and Anderson et al., 2017, p. 78f.).

The principle *leave no on behind* of the Agenda 2030 also leads to the necessity of an open data policy which encourages mutual accountability as well as transparency of results and helps to improve the decision making processes also in countries which lack data for certain SDGs (cf. Anderson et al., 2017, p. 80). As an example, Copernicus, the EO and monitoring program of the European Union in cooperation with the European Space Agency (ESA), provides full, free and open data access on satellite data from the Sentinel series of satellites of optical, radar, ocean and atmosphere sensors which collect continuous data and provide a range of different measurements (cf. Paganini et al., 2018, p. 26). The program provides huge amount of satellite data and information collected from on-site or local measurement systems. The user can benefit from the integrated service of transforming the

satellite data into value-added information, analysing the data, integrating it with other data sources and also validating the results. The topics covered with Copernicus satellite data concern atmosphere monitoring, marine environment monitoring, land monitoring, climate change, emergency management and security. Sentinel-1 collects data on land and ocean monitoring, Sentinel-2 data about land monitoring and emergency management, Sentinel-3 provides data for marine and land monitoring, Sentinel-4 and Sentinel-5 both are collecting data on atmospheric composition and finally Sentinel-6 provides valuable data on global sea-surface height. Therefore, the data is viable for measuring some of the SDGs concerned with land, ocean and atmospheric monitoring (cf. United Nation, 2018, p. 26ff.).

8.5.2. Measuring SDGs with Satellite Data

For the framework of the Agenda 2030 satellite data plays a role in the measurement and monitoring of almost all of the 17 goals and around a quarter of the corresponding targets (cf. Paganini et al., 2018, p. 9). It has been acknowledged that EO can improve national statistics by being spatially-explicit and therefore contribute to inform about the SDGs with providing direct indicators but also by providing a possibility to validate national statistics (cf. Anderson et al., 2017, p. 81 and Holloway et al., 2018, p. 5). But it has to be understood that satellite data is not able to measure the SDGs on its own and can only contribute in achieving the SDGs. The data has to be integrated in a wider framework to validate the results (cf. United Nation, 2018, p. 44).
In Table 8.2 an overview about SDGs and variables derived from satellite data to measure and monitor the corresponding SDG is given. In the second column of the table contribution possibilities from satellite data for the SDGs are given, whereas in the third column information about specific variables derived from satellite imagery for the specific SDG is summarized. The information is gathered from Anderson et al. (2017), Paganini et al. (2018) , United Nation (2018), Holloway and Mengersen (2018) as well as from the literature of Section 8.4 for analysing the constructs *poverty* and *quality of life* with satellite data.

It is often difficult to fully distinguish to which SDG the indicator form satellite data contributes because several indicators can serve the purpose of measuring and monitoring aspects of more than one SDG. A good example is the monitoring of crops via satellite data. Anderson et al. (2017) assign the classification of crop conditions via satellite data to the SDG 2: *End hunger, achieve food security and improved nutrition and promote sustainable agriculture*, whereas United Nation (2018) describes that this indicator has value in monitoring SDG 1: *End poverty in all its forms everywhere*, because it appropriately reflects on the access to food important for the construct of poverty (cf. Anderson et al., 2017, p. 85 and United Nation, 2018, p. 48). Another example which shows that the SDGs and their targets are often times not distinct and therefore the same satellite indicator can be used for different SDGs, is the identification of slums. Clearly, the identification of slums is important to monitor poverty for SDG 1 as explained in Section 8.4. But also for SDG 11, which is concerned with the planning of sustainable cities, it is important to monitor the existence of informal living areas and their expansion (cf. United Nation, 2018, p. 71). It is therefore difficult to organize a summary of the satellite data used for the different SDGs without mentioning variables repeatedly for different SDGs.

Table 8.2: Support of Satellite Data for SDGs

SDG	Support of Satellite Data	Indicator from Satellite Data	Sources
1	- forecasting natural disasters and improving the coordination of aid - maximization of the exploitation of natural resources - improving the efficiency of support for vulnerable people	- crop conditions - weather information - lacunarity - vegetation - impervious surface - soil - night-light brightness - normalized poverty index - slum identification	Elvidge et al., 2009, p. 1654 Kit et al., 2011, p. 663f. Rhinane et al., 2011, p. 217 Stow et al., 2007, p. 5168 United Nation, 2018, p. 47f.
2	- providing information for monitoring the vegetation and water cycle - optimization of the agricultural productivity management - life stock management	- crop conditions - crop stage - crop area - crop yield - soil moisture - structure of the agricultural land - grazing identification - condition of cattle and pastures - land cover - proportion of agricultural area	Anderson et al., 2017, p. 85 Holloway and Mengersen, 2018, p. 1368 Paganini et al., 2018, p. 96 United Nation, 2018, p. 51ff.
3	- studying disease epidemiology - identification of ecological and environmental factors contributing to disease spread - monitoring the over-occurrence of mosquitos	- greenness - brightness - temperature - tracking of gases like ozone, nitrogen dioxide, methane aerosols	United Nation, 2018, p. 55

Table 8.2: Support of Satellite Data for SDGs

SDG	Support of Satellite Data	Indicator from Satellite Data	Sources
6	- monitoring water availability - provision of precipitation forecasts, soil moisture contents and evapotranspiration data - data provision for the better understanding of the world water cycle - hydrological modelling - short- and medium-term meteorological forecasting	- soil moisture - vegetation - surface water - water quality - water level - snow cover - water storage dynamics - wetland extent - wetland categorization - chlorophyll-a - phycocyanin - cyanobacteria scums - soil water index - proportion of bodies of water with good ambient water quality - cloud detection - precipitation (rain and snow) - water cycle information - groundwater storage	Anderson et al., 2017, p. 86ff. Holloway and Mengersen, 2018, p. 1368 Paganini et al., 2018, p. 98 United Nation, 2018, p. 59f.
7	- infrastructure monitoring - power grid synchronization - seismic surveying - solar and wind energy production forecasting	- spread and location of electric lighting by night	Anderson et al., 2017, p. 83

Table 8.2: Support of Satellite Data for SDGs

SDG	Support of Satellite Data	Indicator from Satellite Data	Sources
8	- help generating efficiencies in transportation - creation of new jobs in the field of EO - improving the establishment of safe and secure working environments for lone workers	- device tracking for localization of lone workers	United Nation, 2018, p. 66f.
9	- monitoring road structure - achievement of optimal network capacity of radio frequencies - monitoring disaster areas - fleet management for the transportation sector and reduction of CO2 emissions - smart waste management - improving transport efficiency - topographic surveys for city infrastructure planning - monitoring the world's cultural heritage	- distance of communities to nearest roads - changes in impervious surface cover - land cover - land use - ground motion information - air quality - information about thermal patterns - urban housing density	Anderson et al., 2017, p. 90 United Nation, 2018, p. 68ff.
11	- detection of structural risks - mapping of informal settlements - distribution of basic services	- air pollution - air quality - slum identification - land cover classifications - land use - lacunarity - normalized difference vegetation index - land surface temperature - impervious surface	Anderson et al., 2017, p. 83f. Berrada et al., 2013, p. 413 Kit et al., 2011, p. 663 Ogneva-Himmelberger et al., 2013, p. 188ff. Paganini et al., 2018, p. 101 Rhinane et al., 2011, p. 217 Stathopoulou et al., 2012, p. 26

Table 8.2: Support of Satellite Data for SDGs

SDG	Support of Satellite Data	Indicator from Satellite Data	Sources
12	- monitoring the use of rural resources in a frequent manner - promoting sustainable resource management - monitoring the food supply chain - improving the efficiency, security and safety of the supply chain	- land cover - land cover change	United Nation, 2018, p. 76
13	- create better understanding on the driving forces of climate change - monitoring changes in climatic conditions - planning tool to reduce CO2 emissions - reduction of the response time to natural disasters	- precipitation - temperature - drought events - number of extreme weather events - level of greenhouse gases - glacier outline, area and calving front - glacier surface type and snow line - ice velocity - glacier crevasses and surge - glacier lakes	United Nation, 2018, p. 79ff.

Table 8.2: Support of Satellite Data for SDGs

SDG	Support of Satellite Data	Indicator from Satellite Data	Sources
14	- monitoring and evaluation of marine resources - mapping, monitoring and managing natural and protected areas - improvement of the productivity of fishing activities - detection of illegal fishing activities	- coverage of protection areas in relation to marine areas - level of phytoplankton in the ocean - sea-surface temperature - ocean colours - chlorophylla levels - suspended sediments - dissolved organic matter - chlorophyll-a on the ocean surface - information on fishing activity and vessel compliance	Anderson et al., 2017, p. 81f. Paganini et al., 2018, p. 102f. United Nation, 2018, p. 82

Table 8.2: Support of Satellite Data for SDGs

SDG	Support of Satellite Data	Indicator from Satellite Data	Sources
15	- monitoring the status and evaluation of the land surface - description of the state and disturbances of the vegetation - mapping of zones affected by fire events - improving the understanding of animal tracing for protecting animal species	- structural attributes of tree cover and height - forest cover extent - proportion of forest area - proportion of degraded land - loss and gain of forest cover - afforestation and reforestation - net primary productivity (photosynthetic activity) - nutrient retention (leaf nitrogen retention, leaf phosphorus limitations) - habitant structure (cover, height, clumping) - land cover and land cover change - land use - desertification - land degradation loss of biodiversity - fraction of radiation absorbed for photosynthesis - mountain green cover index - soil moisture - soil erosion and salinity - carbon stocks above and below ground - water resources	Anderson et al., 2017, p. 82ff. Holloway and Mengersen, 2018, p. 1368 Paganini et al., 2018, p. 104 United Nation, 2018, p. 85

Holloway et al. (2018) define three minimum requirements which must hold in order to be able to measure SDGs with satellite data. Firstly, the indicator must be measurable from the satellite imagery which means that it is possible to see and extract the information of interest. One problem that can arise using satellite data could be for example that bad weather conditions are present by the time the picture is taken. This can make it difficult or even impossible to observe the information of interest from the satellite image. The second requirement states that the satellite imagery needs to be available for the region of interest and the images must be ready for statistical analysis. The pre-processing of satellite imagery to prepare them for statistical analysis for example with cloud shadow or topographic adjustment makes special knowledge from earth science necessary. Lastly, validation of the statistical model output is needed with for example field measurements, survey data or weather data. This auxiliary data for validation is also called *ground truth* (cf. Holloway et al., 2018, p. 10f.).

8.6. Possible Challenges with Using Satellite Data

Despite the advantages of using satellite data this data type just as any other information tool has flaws and limitations which have to be known and understood in order to be able to adequately use this data and obtain sound information to analyse the constructs of interest (cf. Hall, 2010, p. 13f.). In order to benefit from the advantages of satellite data the possible pitfalls have to be understood and in each situation its applicability has to be assessed based on in-depth knowledge about the characteristics of the data. In that follows, some of the limitations of satellite data are explained in more detail.

Like traditional survey data, information derived from satellite imageries can suffer from errors which can be based for example in the data acquisition or the data conversion. Some of these errors from data acquisition due to atmospheric conditions such as clouds and bad weather conditions or the natural variability of the landscape cannot be controlled for, whereas for other types like geometric or radiometric errors it is possible to control for or adapt the collected information (cf. Lunetta et al., 1991, p. 678). Identifying land cover types and differentiating between certain objects in the images can be very difficult and prone to errors. Stoler et al. (2012) for example describe the problem in their application of classification algorithms that often confusion in the identification exists between bare rocks, vehicle tracks, land under construction or other complex surface material composition (cf. Stoler et al., 2012, p. 34).

Using satellite data for further statistical analysis can be difficult due to its characteristics. The variables derived from satellite imagery are often discrete or truncated which makes linear analysis in a statistical model inappropriate. Also, satellite data often contains spatial dependencies which have to be considered in every subsequent analysis. For example in regression analysis, spatial dependencies would lead to error terms which are not independent and identically distributed and therefore a core assumption of this method is violated. Using satellite data without taking this violation into account will lead to biased parameter estimates and wrong conclusions about the relationship in the regression equation (cf. Donaldson and Storeyard, 2016, p. 190ff.).

Further issues need to be taken into consideration when satellite data is integrated with traditionally collected survey data and analysed together. In many studies which have been discussed in the previous sections satellite data and survey data is integrated to obtain more in-depth knowledge about

the construct of interest such as quality of life. One problem which occurred in many studies is that the two integrated data sources provided data from different reporting periods which was integrated without any adaptation with the implicit assumption of no change in between the periods for the variables of interest. It is rather questionable how close to reality this assumption is. In Elvidge et al. (2009) for example the satellite data was derived in 2003 whereas the data on poverty indicators originated from survey data reported in 2006. Another example is the study of Weeks et al. (2007) in which this problem has also been pointed out by the authors (cf. Weeks et al., 2007, p. 6). In order to be able to integrate survey data and satellite data it is necessary to geocode the survey data. Often, the problem arises that geocoding of survey data is only available on higher aggregation levels and therefore, it is difficult to connect satellite data and survey data (cf. Rindfuss and Stern, 1998, p. 20f.). Even if georeferenced survey data is available, the question must be raised to which pixel of the satellite data (as smallest unit of satellite imagery) the reporting unit of the survey data should be linked. This can be a difficult question to answer because people move but pixels do not. If for example the effect of behaviour on the environment is analysed, it is necessary to take into consideration the travel or commuting behaviour of individuals and not only consider the residence areas. This is because humans do influence their environment not only at their usual place of living but also where they work and travel (cf. Rindfuss and Stern, 1998, p. 16). Further problems are based on different aggregation levels of the data sources. Often, socio-economic data is not available on the same fine spatial level as satellite data which will lead to a situation in which interpretation about the results may be not appropriate or sound. The problem is known as the *Modifiable Area Unit Problem* which describes the situation in which the disaggregated satellite data has to be aggregated on a higher level to be able to analyse it on a common spatial scale with the survey data. Such an aggregation though can lead to differing results on the aggregated and the disaggregated level and therefore change the interpretation (cf. Donaldson and Storeyard, 2016, p. 190ff. and Stow et al., 2007, p. 5167). Also, privacy issues may arise from the use of satellite data in connection with survey data in the case of high spatial resolution. In case of high spatial resolution and the integration of survey data it might be easier to identify single observation units from the data and therefore it is important to discuss which satellite data should be freely available (cf. Donaldson and Storeyard, 2016, p. 190ff. and Rindfuss and Stern, 1998, p. 11f.).

For all its problems, Hall (2010) concludes that the use of satellite data for socio-economic issues is not uncommon any more but still immature (cf. Hall, 2010, p. 13). Accuracy assessment, validation using ground truth data and advanced methods are needed to integrate satellite data with survey data for socio-economic research. Therefore, Hall (2010) states that the use of satellite data is most valuable when connected with traditional survey data. He warns that the usefulness of satellite data might be overestimated and the methods used to connect satellite data with survey data might be oversimplified due to the complexity of both information sources. In general, satellite data cannot measure the construct of interest from social science directly but the information which are obtainable from satellites like roads and buildings are rather artefacts of social development. These driving variables are more abstract variables, but nevertheless most of the times they are of interest for social scientists and in official statistic, rather than the observable artefacts (cf. Hall, 2010, p. 13 and Rindfuss and Stern, 1998, p. 6f.). In the case of measuring SDGs with satellite data and the

suggestions from the literature about which indicator from satellite imagery can be used to measure or monitor which SDG it can be noticed that similar indicators such as land cover and land use are suggested to be used for SDGs with very different targets. This is a good example of why satellite data should not be used to measure constructs of interest for example from social or economic research on its own. This information should rather be integrated in an analysis which highlights different aspects of a construct of interest with different information sources. Lastly, the identification of some of the pitfalls of satellite data points out the need for special knowledge to be able to make this data source a useful information tool for different disciplines. Therefore, experts from different research fields of earth science advanced in satellite data and researchers from official statistics with advanced knowledge about ground surveys need to effectively collaborate (cf. Hofmann et al., 2008, p. 539, Donaldson and Storeyard, 2016, p. 190ff. and Rindfuss and Stern, 1998, p. 2f.).

8.7. Conclusion

The opportunity to collect data in areas or about aspects of constructs difficult to reach and measure with traditional survey data, as well as the ability to gain further insights about spatial relationships and heterogeneity are great advantages which make the use of satellite data for social or economic research attractive. Poverty and quality of life are just two examples of constructs which can be analysed with satellite data by integrating this data type with traditional survey data as for example in Elvidge et al. (2009) or Berrada et al. (2013). Other examples for constructs are mentioned in Section 8.4 like social vulnerability and economic well-being. The variables for these studies derived from satellite images like land use, land cover, the normalized difference vegetation index or slum identification are also mentioned in United Nation (2018), Paganini et al. (2018) or Anderson et al. (2017) as valuable proxies to measure and monitor some of the SDGs. As suggested in the literature and summarized in Table 8.4, Section 8.5 especially SDG 6: *Ensure availability and sustainable management of water and sanitation for all* and SDG 15: *Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss*, could benefit from satellite data due to the high amount of indicator variables which are suggested to be applicable and can be estimated from satellite data. The SDGs with a focus on sustainable development of the environment such as SDG 6, 15 and also 14 can benefit greatly from the experience of the natural and environmental sciences with regard to the use of satellite data. As in these fields the data type is well-established and its use widely discussed with advanced methods to extract necessary information (cf. Hall, 2010, p. 1). Other SDGs more related to questions of social and economic science such as SDG 1: *End poverty in all of its forms everywhere* or SDG 11: *Make cities and human settlements inclusive, safe, resilient and sustainable*, could also benefit from new data sources such as satellite data. In Section 8.4 studies for the constructs of poverty and quality of life have been introduced which showed how satellite data could be used to detect slum areas via classification algorithms and the environmental welfare dimension could be measured with satellite information such as impervious surface. The studies by Berrada et al. (2013) or Ogneva-Himmelberger et al. (2013) are examples of how data from satellite images is integrated with survey data to measure all dimensions of quality of life. It is noticeable that in most of the studies satellite data is not used on its own but rather validated with available survey data, the ground truth (cf. Rindfuss and Stern, 1998, p. 2f.). Satellite data as a source of information has some pitfalls which have to be addressed explicitly, such as specific errors. It also has to be taken into account that whenever satellite data is used as proxies for constructs in social and economic science most of the information taken from

images are showing artefacts of human behaviour or economic activity and not specifically the driving forces of these artefacts. This is also important when measuring and monitoring the corresponding SDGs and when comparisons over time or between units of observations, such as countries, are made.

9. Remote sensing data

9.1. Measuring urban extension with satellite images

In 2015, a set of goals to end poverty, protect the planet and ensure prosperity for all as part of a new sustainable development agenda was adopted by the UN (UN, 2019). Each goal has specific targets to be achieved over the next 15 years and required commitment from all parties governments, private sector, civil society but also the citizens. Earth Observation (EO) datasets offers many opportunities to improve the monitoring of these SDGs for both reaching the SDG targets and reporting on progress. In recent decades, the added-value of remote sensing datasets as an information source in support of many sectors of government and industry has been proven, and was reconfirmed by world leaders, whilst adopting the 2030 Agenda for Sustainable Development. Earth observation will enable the tracking of global change at high resolution and in real time. The geospatial information provided by EO data will allow for implementation at local to national levels while still allowing for a monitoring and reporting based on the global indicator framework. However, National Statistics Institutes should also be aware that Earth Observations data come with various limitations. Though, Earth observation does not provides statistical indicators as default output, it provides some spatial, spectral, and temporal information which can be related.

This study aims to evaluate to what extent can National Statistic Institutes benefit from Earth observation to monitor and report on the SGDs at local to national level. This case study will focus on the characterization of urban sprawl (SDG 11.7.1) across urban areas in Europe. Urban sprawl was only recently officially acknowledge as an issue in Europe (Hennig et al., 2016) and numerous attempts at characterizing urban sprawl have been made in recent years however a consensus remains to be reached. A first step in urban sprawl characterization is to correctly classify the land cover and land cover change, below a first example test of land cover classification for the characterization of urban sprawl by means of data-driven machine learning methods is presented.

9.1.1. Context

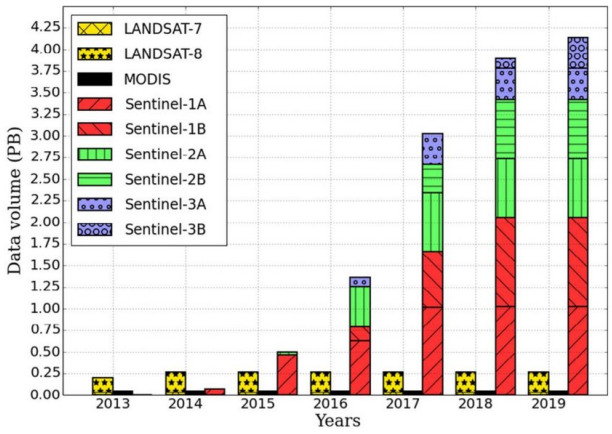


Figure 9.1: Yearly data volume (from Soille et al. (2018) without changes).

At European and International scale various action have been taken to evaluate the added value of Earth observation for official statistics. Current NSO's, lack the expertise, infrastructure and internet bandwidth to efficiently and effectively access, prepare, process, and utilize the growing volume of raw space-based data for local, regional, and national decision-making. In practice, EO data needs to be corrected for many effects due to the atmosphere or the land, the solar angle, the viewing angle from the sensor to the target etc. Such expertise when available is limited to to a small number of specialists which hinder the successful utilization of space-based data for statistics production. On-going effort at ESA will ensure that the wealth of data, cf. Figure 9.1, being generated by the COPERNICUS program will be Analysis Ready Data/Services, therefore allowing a shift of the workload from the end users to the data providers. ARD data are more accessible and useful and allow for real-time analysis and data mining to derive better indicators. In practice, ESA together with EC is deploying five DIAS platforms which will allow users to discover, manipulate, process and download the Copernicus data and information. Hence, various datacubes/services which could be of interest for official statistics are being developed by instance,

1. The Monitoring Agricultural Resources (MARS)[1] bulletins report the latest predictions on crop vegetation growth (cereal, oil seed crops, protein crops, sugar beet, potatoes, pastures, rice) including the short-term effects of meteorological events. These reports are supported since 1992 by near real-time crop growth monitoring and yield forecasting information called the MARS Crop Yield Forecasting System (MCYFS).

2. The Global Human Settlement Layer (GHSL) produces new global spatial information, evidence-based analytics and knowledge describing the human presence on the planet, (Corbane et al., 2017). GHSL aims to provide scientifics methods and system for reliable ad automatic mapping of built-up areas from remote sensing data. GHSL operates in an open and free data and methods access policy (open input, open method, open output)

3. Mapping the surface deformation at national scale through the Amazon Web Services (AWS) Cloud Computing platform (Luca et al., 2017). In this service an automatic pipeline was implemented within the Amazon Web Services (AWS) Cloud Computing platform for the interferometric processing of large Sentinel-1 multi-temporal SAR datasets, aimed at analyzing Earth surface deformation phenomena at wide spatial scale. The current experimental results are focused on a national scale DInSAR analysis performed over the Italian territory. Validated results for the whole Europe are expected to be available by end Q2 2019.

4. The Copernicus Atmosphere Monitoring Service (CAMS)[2] provides continuous data and information on atmospheric composition. The service describes the current situation, forecasts the situation a few days ahead, and analyses consistently retrospective data records for recent years. CAMS supports many applications in a variety of domains including health, environmental monitoring, renewable energies, meteorology and climatology. The service focuses on five main areas: Air quality and atmospheric composition; Ozone layer and ultra-violet radiation; Emissions and

[1] https://ec.europa.eu/jrc/en/mars
[2] https://atmosphere.copernicus.eu/

surface fluxes; Solar radiation; Climate forcing.

5. The Copernicus Land Monitoring Service (CLMS)[3] provides geographical information on land cover and its changes, land use, vegetation state, water cycle and earth surface energy variables to a broad range of users in Europe and worldwide in the field of environmental terrestrial applications. It supports applications in a variety of domains such as spatial and urban planning, forest management, water management, agriculture and food security, nature conservation and restoration, rural development, ecosystem accounting and mitigation/adaptation to climate change. CLMS is jointly implemented by the European Environment Agency and the European Commission DG Joint Research Centre (JRC) and has been operational since 2012. CLMS consists of five main components:systematic monitoring of biophysical parameters; land cover and land use mapping; Thematic hot-spot mapping; Imagery and reference data.

9.1.2. Case Study: Measuring urban sprawl with satellite images

At global scale, urban expansion is one of the primary factors for habitats loss and species extinction as it results in land cover changes. Locally, urban areas and urbanization have great, irreversible impacts on their surrounding environments, further affecting local climate and hydrological systems through the modification of albedo and evapotranspiration. In recent years, a general consensus has been reached that remote sensing is a viable means to provide measurement-based characterization on a local to global scale. Previous studies, at high ($< 10m$) (e.g., SPOT, IKONOS, QuickBird) and medium (10-100 m) (e.g., Landsat TM/ETM+, ASTER) spatial resolution remote sensing imagery have been applied worldwide in mapping urban areas or built-up areas for individual cities or city-regions [add ref here]. Here, we propose to deploy and compare a set of classification algorithms to evaluate the potential of remote sensing data in mapping urban areas (i.e. Artificial surfaces) at regional to local scale for the Netherlands.

datasets

1. **MODIS NDVI:**

   In this study, will primarily exploit the spectral and temporal information present in the MODIS[4] 16-day NDVI data composite from MODIS/Terra and Aqua(MOD13Q1: MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid V006). This dataset runs from 2004 to present with a fortnightly temporal and 250 to 500 m spatial resolution. The datasets is provided with 16 days pixel reliability which allow for cleaning of the data as well as masking sea pixels. The MOD13Q1 Version 6 product provides a Vegetation Index (VI) value at a per pixel basis for two primary vegetation layers,the first is the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), which has improved sensitivity over high biomass regions. The algorithm chooses the best available pixel value from all the acquisitions from the 16 day period. The criteria used is low clouds, low view angle and the highest NDVI/EVI value. Along

[3] https://www.copernicus.eu/en/services/land

[4] The MODerate resolution Imaging Spectrometer (MODIS) onboard of Aqua is part of the A-train constellation. MODIS also flies on Terra which is orbiting around the Earth in a descending mode passing across the equator in the morning (10:30 local sun time), while Aqua has a ascending orbit and passes south to north over the equator in the afternoon (13:30 local sun time). MODIS has 36 channels between 0.44 μm and 15 μm with spatial resolution ranging from 250 m to 1 km.

with the Vegetation layers and the two QA layers, the HDF file will have MODIS Reflectance bands 1 (Red), 2 (NIR), 3 (Blue), and 7 (MIR), as well as four observation layers. Detailed description on the MODIS NDVI retrieval algorithm can be found in Didan (2015).

For the purpose of this study, the available 16-day composite were re-projected into a latitude/-longitude coordinate system using the Application for Extracting and Exploring Analysis Ready Samples (AppEEARS) to produce data for our domain of interest. The 250m 16 days pixel reliability quality flags are used to screen the data. Cloud cover is the major limitation of using MODIS data for various applications, hence cloud removal to obtain cloud-free images of prior importance. This flag contains simplified ranking of the data that describes overall pixel quality.

2. **CORINE Land cover:** For validation and training purposes the CORINE Land Cover (CLC) for the year 2006 will be used. The CLC inventory was initiated in 1985 (reference year 1990) and updated version have been produced in 2000, 2006, and 2012. It consists of an inventory of land cover in 44 classes. CLC uses a Minimum Mapping Unit (MMU) of 25 hectares (ha) for areal phenomena and a minimum width of 100 m for linear phenomena. The time series are complemented by change layers, which highlight changes in land cover with an MMU of 5 ha. The Eionet network National Reference Centres Land Cover (NRC/LC) is producing the national CLC databases, which are coordinated and integrated by EEA. CLC is produced by the majority of countries by visual interpretation of high resolution satellite imagery. In a few countries semi-automatic solutions are applied, using national in-situ data, satellite image processing, GIS integration and generalization. The 2012 version of CLC is the first one embedding the CLC time series in the Copernicus program, thus ensuring sustainable funding for the future.

Classification Methods

In this study, for comparison purposes various machine learning approaches are under consideration for implementations:

- support vector machine The SVM classifier has been widely used and reported as an outstanding classifier (Cortes and Vapnik, 1995). The basic idea of SVM is to classify the input vectors into two classes using a hyperplane with maximal margin.

- random forest Random Forest (Breiman, 2001) is an ensemble method, which constructs many decision trees to be used for classifying a new instance by the majority vote. Each decision tree node uses a subset of attributes randomly selected from the original set of attributes. Additionally, each tree uses a different bootstrap sample data.

- k-nearest-neighbors The k-NN method is a non-parametric method used for classification and regression (Altman, 1992). For k-NN classification, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors; Typically, k (number of neighbors) is a positive integer less than 20 [29], we set k equal to 13 in this study .k optimal was determined by using a tenfold cross-validation study on a subset of the dataset.

Here, the python module scikit-learn is used to implement all the classifiers. This module integrates a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems (Pedregosa et al., 2011). Table 9.1 provides with an overview of the current parameter sets for each algorithm, this is subject to evolution as the study progresses.

First Findings
In order to evaluate urban expansion, it is of prior importance to be able to accurately classify the land cover. Hence, this study will focus on training a robust classifier to create land cover classification based on MODIS NDVI pixels. The period of interest for the training of the classifier spans from January 2004 to December 2012, this choice is motivated by the recurring cycle of the CORINE databases for which **validated** updates have been produced in 2000, 2006, 2012[5]

First Experiment In this experiment, a SVM, RF and k-NN classifiers are deployed. Table 9.1 provides with an overview of the parameter set-up for each classifier. For computational reasons a subset of 500000 pixels was randomly created. For each pixel statistical parameters the standard deviation,the extrema ,the 25^{th},50^{th} and 75^{th} quartile,the kurtosis, the skewness were computed and land cover information is retrieved from the 2006 and 2012 CORINE database [REF]. The pixel are classified into five classes: Agricultural areas, Artificial surfaces, Forest and semi natural areas, Water bodies, Wetlands). Further as per usage, 2/3 of the subset is randomly allocated for training purposes while the remaining third is kept for validation purposes. Features such as the mean, the standard deviation,the extrema ,the 25^{th},50^{th} and 75^{th} quartile,the kurtosis and the skewness representing the MODIS NDVI timeseries for each pixels are under investigation.

Table 9.1: Parameters for classifiers

Classifier	Parameters
k-NN	number of neighbor is set to 13[6]
SVM	Kernel :RBF Cost :1 Gamma:0.1 Decision function shape='ovr' Probability=False
RF	number of trees: 500

Figure 9.2 provides with an overview of the importance of each features for the classification. It appears that the most important features in this experiment stems for the higher value of NDVI.

Figure 9.3 shows the confusion matrix[7] of the best performing classifier for this experiment. It can be concluded that, the classifiers solely classifies agricultural areas confidently, while it confuses the other classes, this correlate with the result of importance feature test.

[5] the authors are aware that a new update was released in Q1 2019. However, these data have not been yet validated.
[7] A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

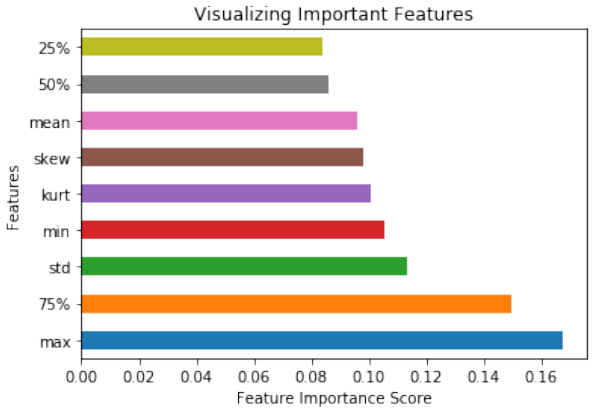


Figure 9.2: Random Forest features importance

In this experiment, it was observed that all tested classifiers reached an accuracy higher then 0.75, with an accuracy of 0.807 Random Forest approach is the best performing on the subset (Results shown only for Random Forest). The confusion matrices of all classifiers shows that agricultural areas could be confidently identify as 97% of the agricultural areas pixels were correctly classified. However, in all the classifiers misclassified at least 35% of the Artificial surfaces as Agricultural areas which is a problem when trying to identify urban sprawl. It should be however noted that Random Forest is a ML method which does not consider an order of importance in the feature, hence when timeseries are used as features each point in time is considered as an independent feature and the order of measurement is ignored. Therefore, this method can not reproduce serial correlation or detect trends in data, while we do know that at our latitude the NDVI of vegetated area present a strong seasonal behavior.

Second Experiment The second experiment was solely carried out on the Random forest classifier. This choice was motivated by the fact that it was the best performing classifier in the first experiment on one hand. On the other hand, it is well-known that they have a good predictive performance, low overfitting, and easy interpretability as it is straightforward to derive the importance of each variable on the tree decision. Setup-wise this experiment is similar to the first experiment, however, the histogram of the NDVI timeseries for each pixels are provided as additional features.

In this experiment,the overall accuracy and the F-1 score are computed in order to evaluate the accuracy of the produce land cover classification. The F-1 score is a number between 0 and 1 which correspond to the harmonic mean of the precision and recall, with 1 representing perfect precision and recall. In this experiment, the RF classifier prove its ability to tackle the classification of land cover using MODIS NDVI by reaching an overall accuracy of 87%. Figure 9.5 is a snapshot of the interactive map which was developed in order to visualize the land cover classification derived from this experiment. Here, the python module folium was used to implement this map. In this visualization, each MODIS NDVI pixel is represented by a circle with a 250m radius, the prediction from the trained RF classifier are represented by the color in the inner circle while the outer circle represent the corresponding land cover classification from 2012 CORINE land cover database. Agricultural areas, Artificial surfaces, Forest and semi natural areas Water bodies and Wetlands are marked in white,

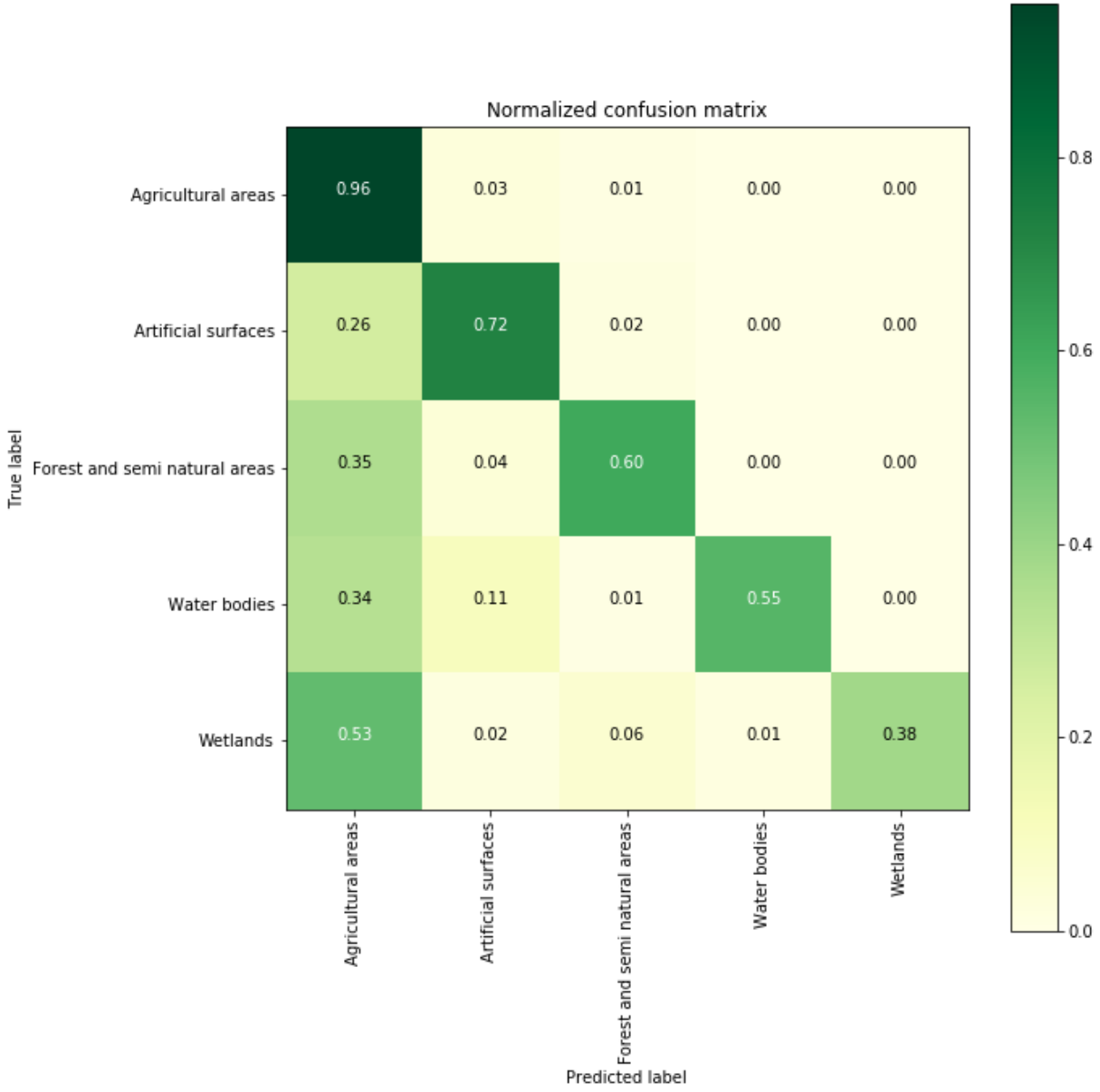


Figure 9.3: Confusion matrix of the Random Forest experiment on the proposed NDVI datasets. Best performing experiment

red, green, blue and purple respectively.

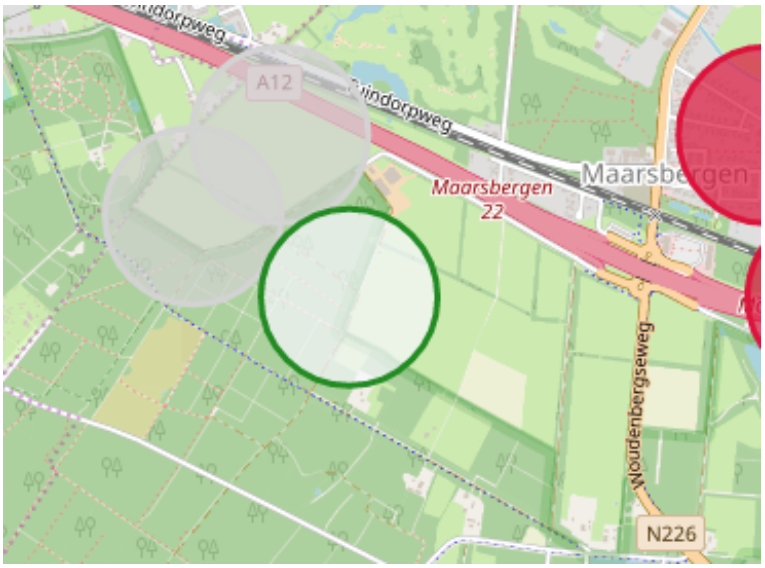


Figure 9.4: Misclassification Example

The statistical performance of the classifier are presented Table 9.2. It is observed that the variations per land cover are quite large, this is mainly due to unbalanced datasets which affects the F-scores, by instance the unbalance samples between Agricultural areas and wetlands. Further, it can be noted that in the case Forest and semi natural areas, Water bodies and Wetlands classes the model is only able to identify a small subset of these categories (low recall), this poor performance is attributed to the size of the pixels, i.e. 250m which as illustrated in Figure 9.4 can be a combination of land cover.

Table 9.2: Experiment 2: Overview precision recall f1-score

	precision	recall	f1-score
Agricultural areas	0.88	0.96	0.92
Artificial surfaces	0.83	0.72	0.77
Forest and semi natural areas	0.86	0.61	0.71
Water bodies	0.82	0.55	0.66
Wetlands	0.92	0.40	0.56
weighted avg	0.87	0.87	0.86

9.1.3. Outlook

In this study, we addressed the challenge of classifying land use cover by means of earth observation datasets and machine learning methods. To this end, a dataset was created using MODIS NDVI and CORINE land cover meshed to our domain of interest, i.e. ARD. First findings show that a robust random forest classifier for land cover mapping by means of MODIS NDVI was trained for the Netherlands. This trained algorithm will allow for land cover change detection and therefore enable us to estimate the urban sprawl which occured in recent decades. Further tasks will be carried out with Sentinel-2 images (10m) with a more refined classification which will allow for the discrimination

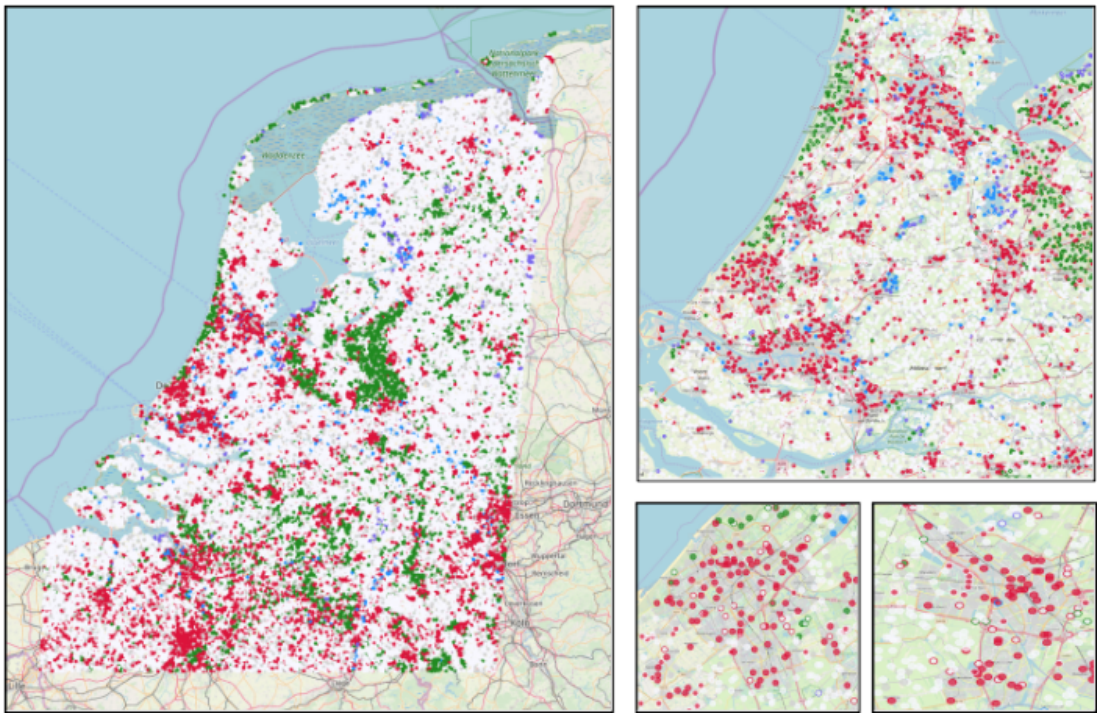


Figure 9.5: Experiment 2: Mapping RF land cover classification. Each MODIS NDVI pixel is represented by a circle. Inner circle represent the prediction from the trained RF classifier, outer circle represent the land cover classification from 2012 CORINE land cover database. Agricultural areas, Artificial surfaces, Forest and semi natural areas Water bodies and Wetlands are marked in white, red, green, blue and purple respectively.

between various type of artificial surfaces. The overall availability of free earth observation datasets offer a broad range of possible application. Nevertheless, National Statistics Institutes should be aware that EO data have limitations. EO will not provide any statistical indicators by default; it provides some spatial, spectral, and temporal information which can be related to indicators.

9.2. Detecting Photovoltaic Solar Panels in Aerial Images with Convolutional Neural Networks

Climate change and international agreements, like for example the Paris Agreement, force governments to come up with policies to keep the temperature rise within acceptable bounds and emphasize green alternatives to fossil fuels. As such, governments more and more need reliable statistics about renewable energy sources. Statistic Netherlands provides several official statistics about renewable energy, among others an estimate of the photovoltaic installations and the electric energy produced (CBS, 2018b). At this moment, the official statistics on solar energy are based on sources that are either incomplete or have a large uncertainty. In an attempt to get a more complete, more accurate, and more timely estimation of the energy production by photovoltaic installations, Statistic Netherlands is looking at additional data sources such as smart meter data, energy consumption data, citizen-generated open data, and aerial images.

Aerial images provide a complete and high resolution snapshot of the entire Netherlands at two mo-

Table 9.3: Overview of the datasets being used

Year	RGB	CIR	Resolution	Number of tiles
2013	√	√	DOP25	175034
2014	√	√	DOP25	175034
2015	√	×	DOP25	87517
2016	√	√	DOP25	175034
2017	√	√	DOP25	175034

ments each year. While manual inspection would be in principle a reliable method of retrieving the locations and sizes of photovoltaic installations, it is also a labor intensive process that may take several months or even years to complete. To successfully use aerial images as an additional source of statistical information, methods to (semi-) automatically extract relevant information have to be investigated and developed. In this section, the ongoing investigation into several computer vision and machine learning algorithms for the (semi-) automatic extraction of statistical information about photovoltaic installations will be presented. The work presented here aims to provide information for the UN sustainable development goals (SDG) 7 "Affordable and clean energy" and 11 Sustainable cities and communities. In the following sections, the available data sources will be described first in section 9.2.1. After that, the methods used to prepare and analyze the aerial image data will be described in 9.2.2.

9.2.1. Data sources

Two times a year, aerial images are taken of the Netherlands in commission of the Dutch Government [8]. A high-resolution image (10 cm per pixel, DOP10) is taken during the winter, and a low-resolution image (25 cm per pixel, DOP25) is taken during spring and summer. The aerial images are then made available to several governmental agencies. Statistics Netherlands owns various low-resolution aerial images taken since 1995. These aerial images are available in RGB as well as near-infrared (CIR). As the main aim of the project is the validation of a register, the aerial images since the start of the register (2013) will be used. Table 9.3 gives an overview of the data used. Since 2016, the aerial images are additionally made available as open data via pdok.nl[9].

For the detection of photovoltaic installations the aerial images have to pre-processed. At first, to limit the amount of data to be processed, a certain region of interest has been selected: the city of Heerlen in the South of the Netherlands. This region is furthermore divided into square tiles of 75 ×75 pixels (18.75m ×18.75m); a convenient size that can be processed with computer vision and machine learning algorithms without using too much processing power (see figure 9.6). For the city of Heerlen, dividing the region into tiles this size results in 87517 tiles per year per image type (RGB vs. CIR). In total, this results in 175034 tiles per year and 787653 tiles across all years (the CIR for 2015 is missing).

[8] http://www.beelmateriaal.nl
[9] http://www.pdok.nl

Figure 9.6: Examples of image tiles, left-to-right: 2016 RGB, 2016 CIR, 2017 RGB, 2017 CIR.

9.2.2. Methods

To automatically analyze aerial images on the presence or absence of solar panels, two types of machine learning algorithms are considered. On the one hand, a state-of-the-art deep learning classifier is re-trained to the domain of aerial pictures. The algorithm is trained to carry out a classification of the whole picture and predicts whether a solar panel appears in the picture or not (picture-based classifier). On the other hand, a Random Forest classifier from the literature is reproduced that uses manual specified features (Malof et al., 2016). The Random Forest classifier is trained to predict if one pixel in the picture belongs to a solar panel or not (pixel-based classifier). A classifier like this can be used for the segmentation of image in solar panel versus background regions.

- **Preparation of the data:** Training a machine learning classifier involves annotating the data with the categories that the classifier should be able to distinguish. The two types of machine learning classifier considered also need two different forms of annotation. For the picture-based classifier, the images in the dataset need to be classified in the two categories: the positive pictures that contain solar panels and the negative pictures that do not contain solar panels, see figure 9.7 for an example. The dataset will be split up into three subsets for training (70%), testing (20 %), and validation (10 %).

Figure 9.7: Example of a picture-based classification: a tile with a negative classification (left, without solar panels) and a tile with a positive classification (right, with solar panels).

  For the pixel-based classifier, the pixels that belong to the solar panel regions should be additionally annotated. By drawing polygons around the solar panel regions, pixels within the polygons can be given a positive classification and pixels outside the polygons a negative one. See figure 9.8 for an example of a pixel-based classification. The dataset will be split up in three subsets in the same way as the dataset for the picture-based classifier.

- **Training a deep learning picture-based classifier:** For the picture-based classifier, a standard VGG16 convolutional neural network from the literature is taken as inspiration (Simonyan

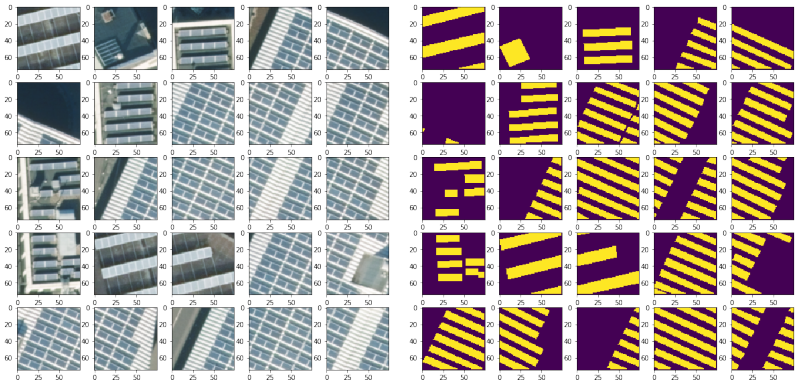


Figure 9.8: Example of a pixel-based classification: the original pictures with the solar panels on the left vs. the classification of the pixels into background (purple) and solar panel (yellow) on the right.

and Zisserman, 2014). On top of the VGG16 convolutional layers, with the weights pre-trained on ImageNet, custom layers specific for the classification at hand are added. The custom layers consist of a global average pooling layer directly after the last convolutional layer, a fully-connected layer of size 512, and a sigmoid layer of size 1 as the output. Supervised learning is used to train the network; each picture in the training set is presented as input to the neural network along with its classification: 0 for a picture without solar panels, 1 for a picture with solar panels. The network is trained with a RMSprop optimizer (Tieleman and Hinton, 2012) for a maximum of 100 epochs. During the training process early-stopping is used; if the validation accuracy does not change during 10 epochs, training is stopped. After training, the classifier can identify whether an image contains solar panels or not. The classifier however cannot find the location and size of the photovoltaic installations. As such, the picture-based classifier is more suitable as pre-processing step for semi-automatic classification. It can be used to find pictures containing solar panels in a much larger dataset consisting of thousands of pictures.

- **Training a Random Forest pixel-based classifier:** A first version of the pixel-based classifier was based on the work by (Malof et al., 2016). A feature vector is created for each pixel by extracting the mean color values and standard deviations in several areas around the pixel. A Random Forest classifier is then trained to perform a pixel-based classification on the basis of the feature vector. By combining a classification of all pixels in an image and some post-processing steps, solar panel regions in a picture can be identified. The pixel-based classifier can therefore be used to count the number of solar panel regions in a picture, as well as the area of these regions, allowing it to be used to create detailed statistics and solar energy estimates.

9.2.3. Future needs

Areal images provide a wealth of detailed visual information about large geographical areas. By using novel machine learning techniques integral and detailed geographical information can be acquired about a county and its regions. This geographical information can in turn be used to provide detailed statistics and indicators (1) about the availability of solar panel and solar boiler installations, wind-

mills, and hydro-electric power stations (SDG 7: Affordable and clean energy), (2) the status of the infrastructure and the industrialization of a country (SDG 9 Industry, Innovation and Infrastructure), (3) the growth and living conditions within cities, i.e. ratio of parks/nature to buildings (SDG 11 Sustainable Cities and Communities), and (4) availability of forests, deforestation, and the ratio of urbanization (SDG 15 Life on Land). The use case presented in this report focuses on the automatic detection of solar panel installations on rooftops (SDG 7). In principle, detailed statistics about the number of solar panel installations but also their area (and thus their estimated output) can be made using aerial images. As such, aerial images are an interesting additional source of data next to already existing (incomplete) registers and surveys.

The use of aerial images as a data source for official statistics comes with its own set of challenges however. Machine learning algorithms have to be trained and evaluated to automatically process visual information. To be able to deal with the large variety of visual information in vast geographical areas, specific care should be taken to make machine learning algorithms generalize well. In this respect, especially the composition of the data sets used for training, testing, and validation are important. All of these data sets should reflect the composition of the target population correctly. If not, standard evaluation measures for machine learning (accuracy, recall, precision, f1, etc.) may give a biased and maybe even too optimistic view of algorithm performance. This is particularly true for unbalanced data sets, like the solar panel dataset presented before; only a small subset of the houses in the data set actually have solar panels. When training the algorithm, this bias should be taken into account. In addition, the data sets should contain sufficient variability in training examples. Building styles, environmental cues, and solar panel types vary from region to region. When training on a small subset of the data, some examples may go unrecognized. The geographical spread of data therefore should be minded if algorithms need to work on the whole region of interest.

The quality and amount of detail may furthermore not be sufficient to detect the phenomenon of interest. Low resolution images, 25cm per pixel,not provide sufficient detail to detect some types of solar panels or other phenomena like stonification of gardens. What is more, in low resolution images, the difference between a solar panel and a roof top window or sunroof is often not apparent. In this case, higher resolution images, at 10cm per pixels, will have to be used (if available). Lighting conditions in the image may further influence algorithm performance: shadows can obscure the phenomenon of interest but also weather conditions can result in underexposed or overexposed images. Weather conditions and time of day moreover cause a variability in lighting conditions from year to year. This in turn causes different color distributions throughout the years and can greatly influence algorithm performance if the algorithm was only trained on the data of a certain year. Last, aerial images are not made that frequently, in the Netherlands twice a year, in other countries even less often. As a consequence, aerial images are thus less suitable for rapidly changing phenomena, nor generating daily, weekly, or monthly statistics. For these more frequent statistics, satellite imagery is a lot more suitable, provided that the resolution is sufficient for the phenomena of interest.

To be able to use aerial images in the production of official statistics more research has to be carried out. First of all, computer vision techniques for classification and segmentation have to be evaluated for their suitability and accuracy when applied to aerial images. Second, specific attention should be given

to the composition of training, test, and validation sets; they should reflect the target population/area as close as possible. Methods that measure, evaluate and guarantee a certain variety in these data sets should be investigated. Third, a related point of further research is evaluating how well the algorithm generalizes to unseen situations across both space and time; the external validity of the algorithm. Fourth, (partly) explaining the internal workings of the algorithm is another important point that should be considered. Last, performing machine learning algorithms on aerial images on large geographical areas requires a scalable and reliable IT infrastructure. It should be investigated which IT infrastructures are suitable to handle large amounts of image data.

9.3. Road sensor data

Road sensors measure the number of passing vehicles every minute. The most common road sensor type is the induction loop, which is installed in a road. The counts from these sensors provide a very detailed image of traffic over time. Intentionally, these data are used for congestion prediction and traffic flow optimization. However, these data can also be used for traffic intensity statistics. These statistics can be a source for SDGs, as we will show in this section.

In the Netherlands, minute based vehicle counts are gathered at 24,000 sites by approximately 60,000 road sensors. We focus on the data collected by 20,000 sensors on the Dutch highways. For the period 2010 until 2014 a total of 115 billion records were collected, resulting in files comprising a total volume of 80TB. Although the data is very structured in a technical sense, the content of the data is not that well-structured. For instance, there were many missing values among the vehicle counts, and the metadata of the sensors was often inconsistent.

Traffic is related to the SDG regarding air quality (SDG 11.6.2), since obviously, traffic will cause air pollution. Therefore, the total amount of traffic can be seen as an indicator for air pollution. However, not all vehicles will produce the same amount of air pollution; obviously an electric car is much better for the environment than an old truck. Air pollution can be measured with satellites and measurement devices (Bechle et al., 2013). Although traffic intensities cannot be translated directly to air pollution, they could be used as additional data source to increase the spatial and temporal detail of the air pollution maps.

We will describe the method to produce traffic intensity statistics (Puts et al., 2019) in Section 9.3.1 and 9.3.2. In section 9.3.3 we describe the relation between traffic intensities and air pollution. Finally, we provide some concluding remarks in Section 9.3.4.

9.3.1. Processing measurement data

Figure 9.9(left) shows the raw, unprocessed, data from one road sensor. The chart displays the counts per minute for one specific day. Missing values are coded as -1. Note that there are many missing values during the night and morning. Furthermore, these data are very noisy. Apart from the question whether the counts have been measured correctly, we need to derive a smooth signal from these data that represents the traffic intensity in a more natural way. Apparently, the traffic intensity for this sensor increases during the afternoon, with a peak around 17:00, and decreases in the afternoon.

For deriving such a smooth signal, we use a Bayesian Recursive Estimator (Puts et al., 2019). Fig-

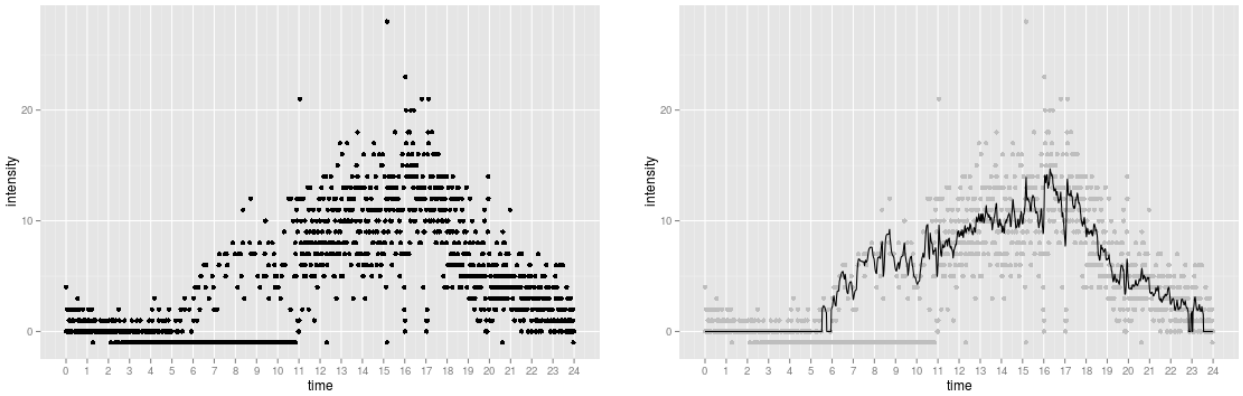


Figure 9.9: Counts per minute from one road sensor (left) and the processed signal (right)

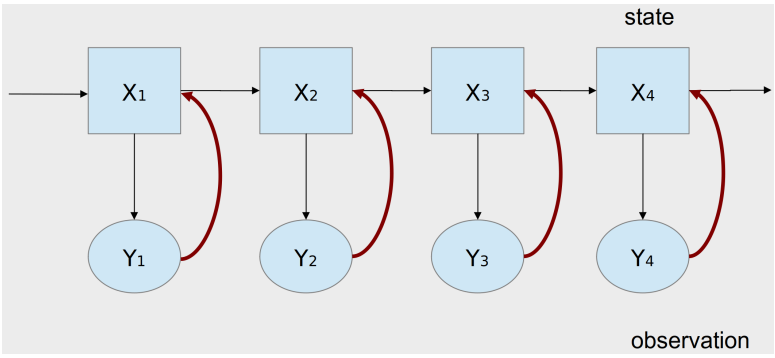


Figure 9.10: Bayesian Recursive Estimator

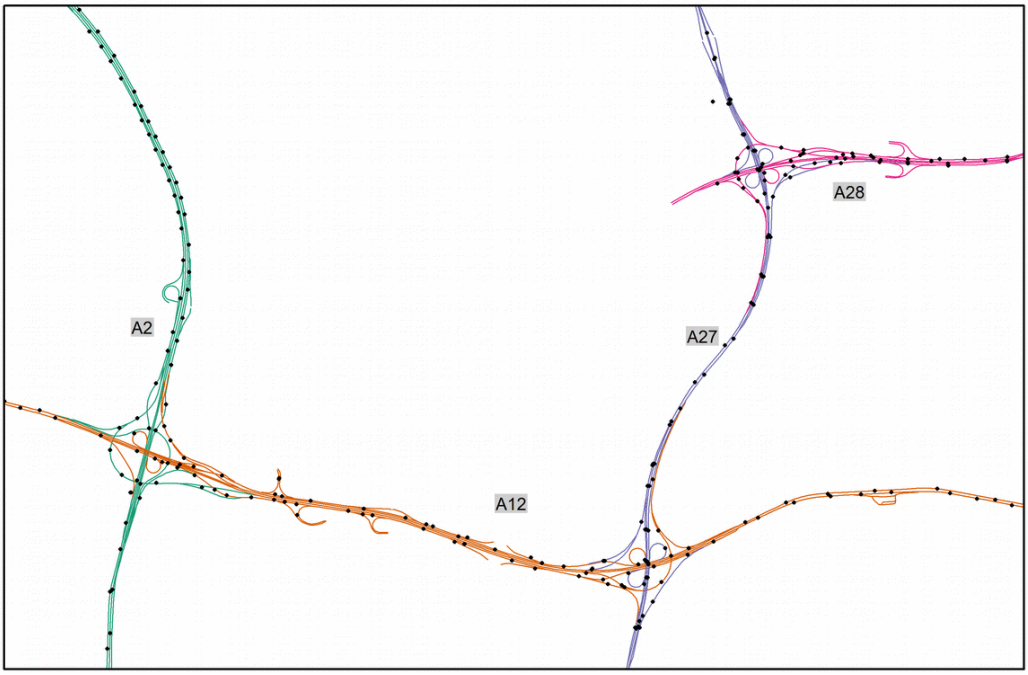


Figure 9.11: Dutch highways and road sensors

ure 9.10 depicts this model, where y_k is the observed value and x_k the hidden state for time k (in our case, in minutes). Here, we assume that the observed values have been generated from a Poisson distribution. The results of this signal are shown in Figure 9.9(right).

Indicators have been developed to access the quality of the processed signals (Puts et al., 2019). Road sensors for which the quality is below a certain threshold will be left out in the further processing. Note that in many cases, the traffic flow can still be estimated using nearby sensors, which will be covered in the next section.

9.3.2. Calibrating

The sensors only measure the traffic at specific locations. However, since there is traffic intensity across the whole network, it is important that the road sensor data are calibrated to the whole network. A single road sensor does not only represent the traffic intensity on its specific location, but it also represents a segment of the road. In this section, we will describe how these road segments, which we will use for calibration, are derived.

Figure 9.11 shows the intersections of four Dutch highways and the location of the road sensors. For the traffic intensity statistics, we only consider the main routes, so the traffic intensities from one end point of a highway to the other, and in the opposite direction. Note that some parts of the highway are on- and off-ramps and interchanges between highways. We will require the coordinates of these parts for the calibration, as we will describe next.

A schematic view of the calibration process is depicted in Figure 9.12. It shows a road that is directed eastwards. On the road are five sensors. Observe that the traffic flow along this road only depends

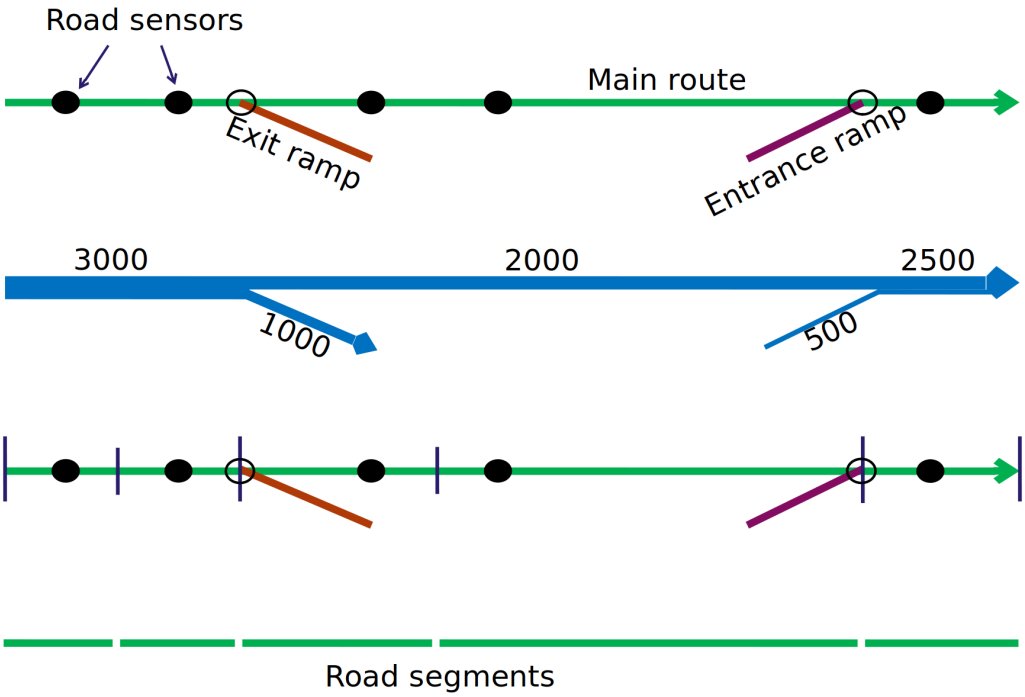


Figure 9.12: Calibration process

on the exit and entrance ramps. This is illustrated in the second row of the figure, where the traffic flow has three levels. Accordingly, the road is split into three segments. Each segment that contains more than one road sensor is subdivided in such a way that the middles are taken as split points. In our example, the first and the second road segment are split since each contains two road sensors (see the third row of the figure). It also may occur that a part does not contain any road sensors. In that case, the statistical outcomes for that part are imputed. Hence, the five resulting road segments are depicted in the bottom row of the figure. The lengths of these road segments correspond to the calibration weights. Suppose that the fourth (and longest) segment takes 2/5 of the total road length, that the weight of the fourth road sensor will be 2/5, since this road sensor represents that part of the road.

9.3.3. Air pollution

Urban growth is driving land-use change in Europe, with periurban areas developing at four times the rate of towns and cities (EEA, EEA) and an efficient road traffic system is of high importance for society. On the other hand, air pollution currently represent the single largest environmental health risk in Europe (EEA, EEA) despite successful the European mitigation policies which have resulted in significant decreases in emissions of air pollutants and noticeable improvements in air quality.

Most people living in European urban area are exposed to poor air quality and the latest EEAâ€™s estimates of the health impacts attributable to exposure to air pollution indicate that $PM_{2.5}$ concentrations in 2014 were responsible for about 428,000 premature deaths originating from longâ€‘term exposure in Europe, of which around 399,000 were in the EU-28. The estimated impacts on the population in European countries of exposure to NO_2 and O_3 concentrations in 2014 were around 78,000 and 14,400 premature deaths per year, respectively, and in the EU-28 around 75,000 and 13,600

premature deaths per year, respectively (EEA, EEA). Air pollution also has considerable economic impacts, cutting lives short, increasing medical costs and reducing productivity through working days lost across the economy. Europe's most serious pollutants in terms of harm to human health are PM, NO_2 and ground-level O_3.

The anthropogenic NO_x emissions in Europe are dominated by combustion processes in road transport with a 40% share, followed by power plants, industry, off-road transport and the residential sector(Pouliot et al., 2012). Traffic is a major source of NO_2 and nitrogen oxide (NO), which reacts with ozone (O_3) form NO_2 (Zariņš et al., 2014). This percentage can be much higher for the areas close to major roads (Henschel et al., 2015).

Road sensor data can potentially be used as an auxiliary source to estimate the air quality on street level. The primary source to estimate air quality still comes from satellites and measurement stations, but road sensor data provide valuable information about one of the main sources for air pollution, namely traffic. Future research is needed to develop estimation methods.

The estimations would improve when not only the vehicle counts are known, but also the type of vehicles. Some road sensors are able to measure counts per vehicle length class. Counts per length class are very useful, since compact cars will produce less air pollution than trucks.

Another source that can be used are license plate registers, which contain data of all registered cars, including the type, full consumption, and yearly mileage. Although the road sensor counts can not be directly joined to these registers (unless the road sensors are cameras), they can be used to estimate the average air pollution per car for a specific urban area. For instance, it is to be expected that the average car in the USA will produce more air pollution than the average car in Norway, where the percentage of electric cars is higher.

9.3.4. Conclusion

Road sensors measure the number of vehicles passing by. We described a method to use these data to produce traffic intensity statistics. Besides traffic intensities, road sensor data provide a potentially useful data source to estimate air pollution on street level. However, future research is needed to develop methods to use road sensor data in combination with air quality satellites and measurement stations.

10. Discussion

This report illustrates the potential benefits of using non-traditional data sources for compiling official statistics and measuring sustainable development goal indicators. An inventory among three European countries (Italy, Germany and the Netherlands) illustrates that traditional data sources like sample surveys and registers are predominantly used to produce SDG indicators. The inventory also indicates that there are many alternative data sources that can be used to either construct SDG's directly or combine them with survey data to produce more accurate regional figures using small area estimation methods as well as more timely indicators using nowcasting methods.

Scanner data are increasingly used by national statistical institutes for the production of price indices. Scanner data have a clear advantage over traditional price collection by interviewers, since such data sets offer a better coverage of items being sold. There are nevertheless a lot of methodological issues to be solved related to the use of scanner data. Sudden prices changes due to replacing products by new similar versions are hard to detect automatically from barcodes only and might result in bias or drift in price indices. An important issue is how to control the homogeneity of the products used to construct price indices and requires further research in multilateral methods and index theory (Chessa et al., 2017) as well as methods to correctly match barcodes from new products that replace outdated products.

Information from websites can be obtained by webscraping and using text mining methods. It can be used for classification purposes, e.g. estimating the number of companies that have sustainable goals in their missions. Generally classification methods such as logistic regression, random forests and support vector machines are successfully applied after language specific removal of stopwords and stemming. These methods are particularly successful if the documents contain a considerable amount of text. For small texts, like social media messages, the performance of these methods is less successful. In such cases more specific preprocessing methods and feature extraction techniques need to be considered.

Information from social platforms can be obtained by classifying messages based on lists of specific words that are related to the topic of interest, e.g. sentiment and other emotions. In this way time series can be constructed to monitor the evolution of indicators like sentiment and social tension. These data typically come at high aggregation levels and are timely because they can be observed at a very high frequency. Issues with these data is that it is unknown to which extend they represent a target population and to which extend period-to-period changes are affected by increasing or decreasing use of particular social media platforms.

Mobile phone network data, in particular signalling data, have potential for official statistics: either for new statistics or to enhance current statistics. A crucial difference with administrative data sources is that mobile phone network data can be used to estimate where people actually are rather than where they formally live or work.

Mobile phone network data have the potential to be used for measuring SDGs, not only because mobile phone network data reflect the actual mobility rather than the administrative expectations, but also since the quality of administrative data may vary across countries. Moreover, administrative data are not always available, especially in development countries. In contrast, mobile phone technology meets an international standard, and the format and quality of signalling data is homogeneous across countries. An alternative is to use them as covariates in small area prediction models as done by e.g. Schmid et al. (2017).

Although mobile phone network data cannot be directly translated to SDGs, there are many SDGs that are related to human mobility. For these SDGs, indicators can be defined using mobile phone data, either as a single source, or in combination with other data sources. As an example of a single source, we have proposed an indicator for measuring poverty. Further research is needed to analyse the correlation between this indicator and existing validated poverty indicators. Future research is also needed to define, analyse, and validate other SDG indicators.

Satellite and aerial images provide a great oppertunity to collect data about geographical and spatial aspects that might be difficult to measure with traditional surveys. Various SDG indicators related to land use, land cover, sustainable water management, sustainable use of terrestrial ecosystems, sustainable forest management, desertification, and protecting biodiversity can be measured directly using satellite images. Satellite and aerial images also contain information that can be used to make better constructs for measuring poverty and quality of life. In most application where satellite data are used for social econonmic constructs, they are combined with survey data.

Aerial images also provide a wealth of detailed visual information about large geographical areas. They can be used to provide detailed statistics and indicators (1) about the availability of solar panel and solar boiler installations, windmills, and hydro-electric power stations (SDG 7: Affordable and clean energy), (2) the status of the infrastructure and the industrialization of a country (SDG 9 Industry, Innovation and Infrastructure), (3) the growth and living conditions within cities, i.e. ratio of parks/nature to buildings (SDG 11 Sustainable Cities and Communities), and (4) availability of forests, deforestation, and the ratio of urbanization (SDG 15 Life on Land). The use case presented in this report focuses on the automatic detection of solar panel installations on rooftops (SDG 7). In principle, detailed statistics about the number of solar panel installations but also their area (and thus their estimated output) can be made using aerial images. As such, aerial images are an interesting additional source of data next to already existing (incomplete) registers and surveys.

The use of satellite and aerial images as a data source for official statistics comes with its own set of challenges however. Machine learning algorithms have to be trained and evaluated to automatically process visual information. To be able to deal with the large variety of visual information in vast geographical areas, specific care should be taken to make machine learning algorithms generalize well.

What is needed in order to make satellite and aerial data a sound information source for measuring and monitoring the SDGs, are strong cooperation between scientists and users of different fields and open data access, so countries for which only scarce data is available can benefit in the light of the

principle of *leave no one behind*. Also, further methods and applications have to be developed to integrate satellite data and survey data. If satellite data is integrated with survey data, both data sources should be available in the same time period and georeferenced information of survey data is needed to connect it to satellite data information on lower spatial levels (cf. Hall, 2010, p. 9ff., United Nation, 2018, p. 44 and Donaldson and Storeyard, 2016, p. 190ff.). Computer vision techniques for classification and segmentation have to be evaluated for their suitability and accuracy when applied to aerial images. Specific attention should be given to the composition of training, test, and validation sets; they should reflect the target population/area as close as possible. Methods that measure, evaluate and guarantee a certain variety in these data sets should be investigated. A point of further research is evaluating how well the algorithm generalizes to unseen situations across both space and time; the external validity of the algorithm. Last, performing machine learning algorithms on aerial images on large geographical areas requires a scalable and reliable IT infrastructure.

The results of this report have the following relations with the other Work Packages of the MAKSWELL project. In the second deliverable of WP2, methodology for using non-traditional data sources are treated in a more general way. Insights obtained from these two delivarales will be used in the third deliverable to identify future research needs in terms of statistical methodologies and new data. The experiences and insights obtained with analysing satellite images and aerial images will be picked up in Work Package 3 to make poverty indicators. In Work Package 4 time series derived from non-traditional data sources like google trends will be used as auxiliary series to nowcast survey estimates. Finally the methodology and experiences with collecting and using non-traditional data sources will be used in the pilot of Work Package 5, where an indicator will be constructed from a non-traditional data source.

Bibliography

Afsar, S., S. S. Ali, and S. J. H. Kazmi (2013). Assessment the quality of life in karachi city through the integration of space and spatial technologies. *Journal of Basic and Applied Sciences 9*, 373–388.

Ahas, R., A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru, and M. Zook (2015). Everyday space-time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science 29*(11), 2017–2039.

Alexander, L., S. Jiang, M. Murga, and M. C. González (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies 58*, 240 – 250.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician 46*(3), 175–185.

Anderson, K., B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl (2017). Earth observation in service of the 2030 agenda for sustainable development. *Geo-spatial Information Science 20*(2), 77–96.

Baud, I., M. Kuffer, K. Pfeffer, R. Sliuzas, and S. Karuppannan (2010). Understanding heterogeneity in metropolitan india: The added value of remote sensing data for analyzing sub-standard residential areas. *International Journal of Applied Earth Observation and Geoinformation 12*(5), 359–374.

Bechle, M. J., D. B. Millet, and J. D. Marshall (2013). Remote sensing of exposure to no2: Satellite versus ground-based measurement in a large urban area. *Atmospheric Environment 69*, 345 – 353.

Berrada, A., H. Rhinan, A. Hilali, and Y. Bedraoui (2013). Application of remote sensing and geographic information system to elaborate uqi (urban quality index): A case of casablanca, morocco. *Journal of Environmental Science and Engineering 2*(7B), 406–415.

Biggeri, L., A. Brunetti, and T. Laureti (2008). The interpretation of the divergences between cpis at territorial level: Evidence from italy. In *Joint UNECE/ILO meeting on Consumer Price Indices, May*, pp. 8–9.

Biggeri, L. and A. Giommi (1987). On the accuracy and precision of the consumer price indices. methods and applications to evaluate the influence of the sampling of households. *Bulletin of the International Statistical Institute 52*, 137–154.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breiman, L. (2001, Oct). Random forests. *Machine Learning 45*(1), 5–32.

CBS (2018a). Social tension indicator based on social media. https://www.cbs.nl/en-gb/our-services/innovation/project/social-tension-indicator-based-on-social-media.

CBS (2018b). Zonnestroom; vermogen, bedrijven en woningen, regio (indeling 2017).

Chessa, A., J. Verbug, and L. Willenborg (2017). A comparison of price index methods for scanner data.

Coosto (2014). Social media management software website. https://www.coosto.com/.

Corbane, C., M. Pesaresi, P. Politis, V. Syrris, A. J. Florczyk, P. Soille, L. Maffenini, A. Burger, V. Vasilev, D. Rodriguez, F. Sabo, L. Dijkstra, and T. Kemper (2017). Big earth data analytics on sentinel-1 and landsat imagery in support to global human settlements mapping. *Big Earth Data 1*(1-2), 118–144.

Cortes, C. and V. Vapnik (1995, Sep). Support-vector networks. *Machine Learning 20*(3), 273–297, doi: 10.1007/BF00994018.

Daas, P. and M. Puts (2014). Social media sentiment and consumer confidence. European central bank statistics paper series no. 5, frankfurt germany.

De Meersman, F., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, and H. Reuter (2016). Assessing the quality of mobile phone data as a source of statistics. In *European Conference on Quality in Official Statistics*. Eurostat.

Deville, P., C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111*(45), 15888–15893.

Diao, M., Y. Zhu, J. Joseph Ferreira, and C. Ratti (2016). Inferring individual daily activities from mobile phone traces: A boston example. *Environment and Planning B: Planning and Design 43*(5), 920–940.

Didan, K. (2015). Mod13q1 modis/terra vegetation indices 16-day l3 global 250m sin grid v006 [data set]. nasa eosdis lp daac. Technical report.

DLR (2019). Atmospheric correction.

Doll, C. N. H. (2008). *CIESIN Thematic Guide to Night-time Light Remote Sensing and its Applications*. Center for International Earth Science Information Network of Columbia University, Palisades, NY.

Donaldson, D. and A. Storeyard (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives 30*(4), 171–198.

Durbin, J. and S. Koopman (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Ebert, A., N. Kerle, and A. Stein (2009). Urban social vulnerability assessment with physical proxies and spatial metrics derived from air- and spaceborne imagery and gis data. *Natural Hazards 48*(2), 275–294.

EEA. *Air quality in Europe 2018*.

Elmore, A. J., J. F. Mustard, S. J. Manning, and D. B. Lobell (2000). Quantifying vegetation change in semiarid environments: precision and accuracy of spectral mixture analysis and the normalized difference vegetation index. *Remote sensing of environment 73*(1), 87–102.

Elvidge, C. D., P. C. Sutton, T. Ghosh, B. T. Tuttle, K. E. Baugh, B. Bhaduri, and E. Bright (2009). A global poverty map derived from satellite data. *Computers & Geosciences 35*(8), 1652–1660.

Esuli, A. and F. Sebastiani (2006). Senti wordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of language resources and evaluation (lrec), pp. 2200-2004.

Eurostat (2017). Pratical guide for processing supermarket scanner data. Technical Report `https://circabc.europa.eu`, Statistics Netherlands.

Hadam, S. (2018). Use of mobile phone data for official statistics. methods - approaches - developments.

Hall, O. (2010). Remote sensing in social science research. *The Open Remote Sensing Journal 3*, 1–16.

Henderson, J. V., A. Storeyard, and D. N. Weil (2012). Measuring economic growth from outer space. *Economic Review 102*(2), 994–1028.

Hennig, E. I., T. Soukup, E. Orlitova, C. Schwick, F. Kienast, and J. A. Jaeger (2016, June). Urban sprawl in europe. joint eea-foen report. no 11/2016. Technical report, European Environment Agency and the Swiss Federal Office for the Environment, Luxembourg. The Annexes are in a separate document titled "Annexes 1-5: Urban Sprawl in Europe. Joint EEA-FOEN report" (141 pp).

Henschel, S., A. L. Tertre, R. W. Atkinson, X. Querol, M. Pandolfi, A. Zeka, D. Haluza, A. Analitis, K. Katsouyanni, C. Bouland, M. Pascal, S. Medina, and P. G. Goodman (2015). Trends of nitrogen oxides in ambient air in nine european cities between 1999 and 2010. *Atmospheric Environment 117*, 234 – 241.

Hofmann, P., J. Strobl, T. Blaschke, and H. Kux (2008). *Detecting informal settlements from QuickBird data in Rio de Janeiro using and object-based approach*, Chapter 6.1, pp. 531–560. Blaschke, T.; Lang, S. and Hay, G.

Holloway, J. and K. Mengersen (2018). Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sensing 10*(9), 1365–1386.

Holloway, J., K. Mengersen, and K. Helmstedt (2018). Spatial and machine learning methods of satellite imagery analysis for sustainable development goals. Paper prepared for the 16th Conference of IAOS OECD Headquarters, Paris, France, 19-21 September 2018.

Iqbal, M. S., C. Choudhury, P. Wang, and M. C. Gonzalez (2014, 03). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies 40*, 63–74.

Istat (2017). Indici dei prezzi al consumo: aspetti generali e metodologia di rilevazione. Technical Report `https://www.istat.it/it/files/2013/04/Indice-dei-prezzi-al-consumo.pdf`, Statistics Netherlands.

Jiang, S., Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González (2016). The timegeo modeling framework for urban motility without travel surveys.

Jonge, E. d., M. Pelt, and M. Roos (2012). Time patterns, geospatial clustering and mobility statistics based on mobile phone network data.

Kit, O., M. Lüdeke, and D. Reckien (2011). Texture-based identification of urban slums in hyderabad, india using remote sensing data. *Applied Geography 32*(2), 660–667.

Kohli, D., N. Kehle, and R. Sliuzas (2012). Local ontologies for object-based slum identification and classification. *Environs 3*, 201–205.

Kondor, D., S. Grauwin, Z. Kallus, I. Gódor, S. Sobolevsky, and C. Ratti (2017). Prediction limits of mobile phone activity modelling. *Royal Society open science 4*(2).

Lafary, E. W., J. D. Gatrell, and R. R. Jensen (2008). People, pixels and weights in vanderburgh county, indiana: toward a new urban geography of human-environment interactions. *Geocarto International 23*(1), 53–66.

Li, G. and Q. Weng (2007). Measuring the quality of life in city of indianapolis by integration of remote sensing and census data. *International Journal of Remote Sensing 28*(2), 249–267.

Lo, C. P. and B. J. Faber (1997). Integration of landsat thematic mapper and census data for quality of life assessment. *Remote sensing of environment 62*(2), 143–157.

Lu, X., D. J. Wrathall, P. R. Sundsøy, M. Nadiruzzaman, E. Wetter, A. Iqbal, T. Qureshi, A. Tatem, G. Canright, K. Engø-Monsen, and L. Bengtsson (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in bangladesh. *Global Environmental Change 38*, 1 – 7.

Luca, C. D., I. Zinno, M. Manunta, R. Lanari, and F. Casu (2017). Large areas surface deformation analysis through a cloud computing p-sbas approach for massive processing of dinsar time series. *Remote Sensing of Environment 202*, 3 – 17.

Lunetta, R. L., G. R. Congalton, L. K. Fenstermaker, J. R. Jenson, K. C. McGwire, and L. R. Tinney (1991). Remote sensing and geographic information system data integration: Error sources and research issues. *Photogrammetric Engineering & Remote Sensing 52*(6), 677 – 687.

Lunetta, R. L., F. K. Knight, J. Ediriwickrema, J. G. Lyon, and L. Worthy (2006). Landcover change detection using multi-temporal modis ndvi data. *Remote Sensing of Environments 105*, 142–154.

Malof, J. M., K. Bradbury, L. M. Collins, and R. G. Newell (2016). Automatic Detection of Solar Photovoltaic Arrays in High Resolution Aerial Imagery. *arXiv preprint arXiv:1607.06029v1 183*, 229–240.

Münnich, R., J. Wagner, J. Hill, J. Stoffels, H. Buddenbaum, and T. Udelhoven (2016). Schätzung von holzvorräten unter verwendung von fernerkundungsdaten. *AStA Wirtschafts- und Sozialstatistisches Archiv 10*(2-3), 95–112.

O-Connor, B., R. Balasubramanyan, B. Routledge, and N. N.A. Smith (2010). From tweets to polls: Linking text sentiment to public opinion time series. Proceedings of the fourth international aaai conference on weblogs and social media, may 23-26, washington dc, usa.

Ogneva-Himmelberger, Y., H. Pearsall, and R. Rakshit (2009). Concrete evidence & geographically weighted regression: A regional analysis of wealth and the land cover in massachusetts. *Applied Geography 29*(4), 478–487.

Ogneva-Himmelberger, Y., R. Rakshit, and H. Pearsall (2013). Examining the impact of environmental factors on quality of life across massachusetts. *The Professional Geographer 65*(2), 187–204.

Paganini, M., I. Petiteville, S. Ward, G. Dyke, M. Steventon, J. Harry, and F. Kerblat (2018). *SATELLITE EARTH OBSERVATIONS IN SUPPORT OF THE SUSTAINABLE DEVELOPMENT GOALS*. Euorpean Space Agency.

Pang and Lee (2008). Opinion and sentiment mining. Foundations and trends in information retrieval 2 (1-2), 1-135.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–688.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Poblet, M., E. Garcá-Cuesta, , and P. Casanovas (2014). *Crowdsourcing Tools for Disaster Management: A Review of Platforms and Methods. In AI Approaches to the Complexity of Legal Systems.*

Pouliot, G., H. Simon, P. Bhave, D. Tong, D. Mobley, T. Pace, and T. Pierce (2012, 01). *Assessing the Anthropogenic Fugitive Dust Emission Inventory and Temporal Allocation Using an Updated Speciation of Particulate Matter*, Volume 4, pp. 585–589.

Pucci, P., F. Manfredini, and P. Tagliolato (2015, 02). *Pucci P., Manfredini F., Tagliolato P. (2015), Mapping urban practices through mobile phone data, PoliMI SpringerBriefs Series.*

Puts, M. J. H., P. J. H. Daas, M. Tennekes, and C. d. Blois (2019). Using huge amounts of road sensor data for official statistics. *AIMS Mathematics 4*(1), 12–25.

Rhinane, H., A. Hilali, A. Berrada, and M. Hakdaoui (2011). Detecting slums from spot data in casablanca morocco using an object based approach. *Journal of Geographic Information System 3*(03), 217–224.

Ridd, M. K. (1995). Exploring a v-i-s (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: comparative anatomy for cities. *International Journal of Remote Sensing 16*(12), 2165–2185.

Rindfuss, R. R. and P. C. Stern (1998). Linking remote sensing and social science: The need and the challenges. *People and pixels: Linking remote sensing and social science*, 1–27.

Salgado, D., M. Debusschere, O. Nurmi, P. Piela, E. Coudin, B. Sakarovitch, S. Hadam, M. Zwick, R. Radini, T. Tuoto, M. Tennekes, C. Alexandru, B. Oancea, E. Esteban, S. Saldaña, L. Sanguiao, and W. Williams (2018). Proposed elements for a methodological framework for the production of official statistics with mobile phone data., essnet big data, wp5, deliverable 5.3.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in senegal. *Journal of the Royal Statistical Society, Series A 178*, 239–257.

Serrano-Santoyo, A. and V. Rojas-Mendizabal (2017). Emergency telecommunications for managing disasters: A complexity science perspective. *European Scientific Journal*.

Simonyan, K. and A. Zisserman (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. pp. 1–14.

Soille, P., A. Burger, D. D. Marchi, P. Kempeneers, D. Rodriguez, V. Syrris, and V. Vasilev (2018). A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems 81*, 30 – 40.

Stateva, G., O. ten Bosch, J. Maslankowski, G. Barcaroli, M. Sannapieco, M. Greenaway, I. Jansson, and D. Wu (2017). Web scraping enterprise characteristics, deliverable 2.2: Methodological and it-issues and solutions. Essnet big data work package 2, version 31-07-2017.

Stathopoulou, M., S. Iacovides, and C. Cartalis (2012). Quality of life in metropolitan athens, using satellite and census data: Comparison between 1991 and 2001. *Journal of Heat Island Institute International 7*(2), 25–32.

Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. de Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem, and L. Bengtsson (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface 14*(127).

Stephan Arnold [STBA], Sylvia Seissiger [BKG], Sarah Kleine [STBA], Michael Hovenbitzer[BKG], Angela Schaff [STBA] (2019). Cop4Stat_2015plus: Studie zur Verwendung von Copernicus-Daten für Zwecke der Flächenstatistik im Bereich Landbedeckung/Landnutzung.

Stoler, J., D. Daniels, J. R. Weeks, D. A. Stow, L. L. Coulter, and B. K. Finch (2012). Assessing the utility of satellite imagery with differing spatial resolutions for deriving proxy measures of slum presence in accra, ghana. *GIScience & Remote Sensing 49*(1), 31–52.

Stow, D., A. Lopez, C. Lippitt, S. Hinton, and J. Weeks (2007). Object-based classification of residential land use within accra, ghana based on quickbird satellite data. *International Journal of Remote Sensing 28*(22), 5167–5173.

Taubenböck, H. and N. J. Kraff (2015). Das globale gesicht urbaner armut? siedlungsstrukturen in slums. In *Globale Urbanisierung*, Chapter 9, pp. 107–119. Taubenböck, H. and Wurm, M. and Esch, T. and Dech, S. Springer-Verlag Berlin Heidelberg.

Tennekes, M., M. Offermans, and N. Heerschap (2017). Determining an optimal time window for roaming data for tourism statistics. In *Proceedings of the NetMob 2017 Conference*.

Tieleman, T. and G. Hinton (2012). Lecture 6.5-RMSProp, COURSERA. *Neural Networks for Machine Learning Technical report*.

Tinto, A., F. Bacchini, B. Baldazzi, A. Ferruzza, J. van den Brakel, R. Willems, N. Rosinski, T. Zimmermann, Z. Andrasi, M. Farkas, and Z. Fabian (2018). Report on international and national experiences and main insight for policy use of well-being and sustainability framework, makswell, wp1, delivarable 1.1.

Tinto, A. and B. Baldazzi (2018). Definition of the existing database on beyond gdp initiatives within official statistics, makswell, wp1, delivarable 1.2.

Toomet, O., S. Silm, E. Saluveer, R. Ahas, and T. Tammaru (2015, 05). Where do ethno-linguistic groups meet? how copresence during free-time is related to copresence at home and at work. *PLOS ONE 10*(5), 1–16.

UN (2014). Using mobile phone activity for disaster management during floods, global pulse project series no.2, 2014.

UN (2019). About the sustainable development goals - united nations sustainable development.

UN General Assembly (2015). *Resolution adopted by the General Assembly on 25 September 2015 - Transforming our world: the 2030 Agenda for Sustainable Development*. United Nations.

UN GWG for Big Data (2017). *Earth Observations for Official Statistics: Satellite Imagery and Geospatial Data Task Team report*.

United Nation (2018). *European Global Navigation Satellite System and Copernicus: Supporting the Sustainable Development Goals - BUILDING BLOCKS TOWARDS THE 2030 AGENDA*.

van Assem, M., A. Isaac, and J. van Ossenbruggen (2013). Wordnet 3.0. http://datahub.io/nl/dataset/vu-wordnet.

van den Brakel, J., P. Smith, N. Tzavidis, R. Iannaccone, D. Zurlo, F. Bacchini, L. Di Consiglio, T. Tuoto, , M. Pratesi, C. Giusti, S. Marchetti, S. Bastianoni, G. Betti, A. Lemmi, F. Pulselli, and L. Neri (2019). Methodological aspects of using big-data, makswell, wp2, delivarable 2.2.

van den Brakel, J., E. Söhler, P. Daas, and B. Buelens (2017). Social media as a data source for official statistics; the dutch consumer confidence index. *Survey Methodology 43*(2), 183–210.

van der Doef, S., P. Daas, and D. Windmeijer (2018). Identifying innovative companies from their website. Bigsurv18 conference, october 27, barcelona, spain.

Velikovich, L., S. Blair-Goldensohn, K. Hannan, and R. Mc-Donald (2010). The viability of web-dervied polarity lexicons. The 2010 annual conference of the north american chapter of the association for computational linguistics, pp. 777-785.

Weeks, J. R., A. Hill, D. Stow, A. Getis, and D. Fugate (2007). Can we spot a neighborhood from the air? defining neighborhood structure in accra, ghana. *GeoJournal 69*(1-2), 9–22.

Weng, Q., D. Lu, and J. Schubring (2004). Estimation of land surface temperature-vegetation abundance relationship for urban heat island studies. *Remote sensing of Environment 89*(4), 467–483.

Widhalm, P., Y. Yang, M. Ulm, S. Athavale, and M. C. González (2015). Discovering urban activity patterns in cell phone data. *Transportation 42*(4), 597–623.

Wilson, R., E. zu Erbach-Schoenberg, M. Albert, D. Power, S. Tudge, M. Gonzalez, S. Guthrie, H. Chamberlain, C. Brooks, C. Hughes, L. Pitonakova, C. Buckee, X. Lu, E. Wetter, A. Tatem, and L. Bengtsson (2016, Feb). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 nepal earthquake. *PLOS Currents Disasters*.

Xie, M., N. Jean, D. Burke, M. Lobell, and S. Ermon (2016). Transfer learning from deep features for remote sensing and poverty mapping.

Xu, Y., A. Belyi, I. Bojic, and C. Ratti (2017). How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data. *Trans. GIS 21*(3), 468–487.

Xu, Y., A. Belyi, I. Bojic, and C. Ratti (2018). Human mobility and socioeconomic status: Analysis of singapore and boston. *Computers, Environment and Urban Systems*.

Zagatti, G. A., M. Gonzalez, P. Avner, N. Lozano-Gracia, C. J. Brooks, M. Albert, J. Gray, S. E. Antos, P. Burci, E. zu Erbach-Schoenberg, A. J. Tatem, E. Wetter, and L. Bengtsson (2018). A trip to work: estimation of origin and destination of commuting patterns in the main metropolitan regions of haiti using cdr. *Development Engineering 3*, 133–165.

Zariņš, A., J. Smirnovs, and R. Lācis (2014, 11). Evaluation of air pollution measurements in urban environment considering traffic intensity. *Construction Science 15*.