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 2.3

Research needs in terms of statistical methodologies and new data

April 2020

Statistics Netherlands, Istat, Destatis, Southampton University, Pisa University, Trier University

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

Deliverable D2.3

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

Research needs in terms of statistical methodologies and new data

Authors

Statistics Netherlands:
J. van den Brakel, T de Jong

Southampton University:
P. Smith, N. Tzavidis

Istat:
F. Bacchini, L. Di Consiglio, A. Ferruzza, A. L. Palma, G. Tagliacozzo, T. Tuoto

Destatis:
M. Köhlmann, N. Rosenski, C. Schartner,

Trier University:
C. Caratiola, F. Ertz, L. Güdemann, R. Münnich

Pisa University :
C. Giusti, M. Pratesi

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. In Deliverables 2.1, 2.2 as well as 3.1, 4.1 and 4.3 it is discussed in detail how these new data sources can be used in the production of official statistics and measurement frameworks for well-being and sustainability indicators. This Deliverable extends on the experiences obtained in these preceding deliverables by pointing out the needs for new data sources and methods in this context.

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 and internet search behaviour from Google Trends.

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 deliverable 2.1 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 (van den Brakel et al., 2019). In Deliverable 2.2 the methodology required to use non-traditional data sources for measuring sustainable development goal indicators is described in more general terms (van den Brakel et al., 2019). The purpose of this deliverable is to identify future needs for new methods and data, based on the insights obtained in Deliverables 2.1 and 2.2.

This deliverable is organized as follows. Chapter 2 elaborates on the needs of a new, extended framework to measure wellbeing and SDG indicators in terms of data sources, methodology, and quality requirements and is based on contributions from F. Bacchini. Chapter 3 provides an extension of quality concepts and a quality frame work for non-probability data, and is a contribution by P. Smith. Chapter 4 describes needs for new methods to combine survey data with new data sources and methods to use new data sources as a primary data sources for SDGs and well-being indicators. This chapter is a contribution by A. Ferruzza, A. Laureti Palma, G. Tagliacozzo and J. van den Brakel. In Chapters 5 and 6 elaborates on the needs for new methods and data sources if remote sensing and deep learning is considered in the production of official statistics about SDGs and well-being indicators. This is a contribution by T de Jong (Ch. 5) and C. Caratiola, F. Ertz, L. Güdemann, and R. Münnich (Ch.

6). Chapter 7 describes two real life applications where non-traditional data are used to measure indicators related to SDGs. The first one is the use of mobile phone data to measure natural disasters in Italy and is based on a contribution by A. Ferruzza, A. Laureti Palma, G. Tagliacozzo. The second application is the use of remote sensors in Germany and is based on a contribution by M. Köhlmann, N. Rosenski and C. Schartner. The Deliverable concludes with a discussion in Chapter 8.

2. Extended framework to measure wellbeing and SDG indicators

In 2018, just few months after the project was launched, the consortium released a reflection paper on the *Future reasearch needs* (Rondinella and altri (2019). The paper addressed several issues for which the MAKSWELL project has tried to answer.

Starting form the description of the knowledge pyramid Eurostat (2017) where at the bottom there are *data* and at the top *knowledge*, we argue that *evidence-based policies have acquired great importance* pushing for the development of new framework, such as the Macroeconomic Imbalances Procedure, where a set of indicators are design and read together, sometimes across well defined domains, in a way to improve knowledge, A clear example is represented by the SDG'goals or, at national level in Italy, by the inclusion in the budget law of the 12 indicators of well-being against which the Government is required to measure the impact of the selected policies.

Increasing attention to well-being and sustainability and the impact of policies on those dimensions are now key drivers in the debate as documented by the work of Stiglitz's Commissione (Stiglitz et al. (2009) and its updating (Stiglitz et al. (2018)). This implies in turn a challenge both for the NSI, that are required to maintain new and updated set of indicators, and for researcher and academia that are require to release new methodology able to support the updating and the disaggregation of the indicators. Even new methods are required to provide a comprehensive but synthetic picture of well-being and SDG.

Against to these challenges, the starting point of the project has been to recognize how the development of national and internazional framework on well-being and SDG has spread out to European countries. The answers provided in del. 1.1 (Tinto et al. (2018)) and 1.2 (Tinto and Baldazzi (2018)) were extremely positive with most of the countries working on the measurement issues as well as on the relationship of the indicators with policy targets. It was interesting to observe that, althought the SDG framework is similar across the countries driven by the international regulation, even the well-being framework shared a common root in line with the Oecd's *Quality of Life*.

Starting from these evidences, the MAKSWELL project activities have been developed exploring new methodologies and data sources able to fill the gap on timeliness and disaggregation provided for well-being and SDG framework. Moreover, the project aims also to explore how macroeconometrics model could be extend to take into account for well-being and SDG dimensions.

Concerning the dimension of well-being we have concentrated our attention on the indicators related to poverty dimensione such as consumption, prices, income while for SDGs attention were related to energy, natural disasters, characteristics of the land. However these selected issues does not cover all the examples presented that, in specific cases, where suitable to the presentation of new sources of data and methodologies.

This deliverable illustrates the future research needs in terms of statistical methodologies and new data illustrating, among other topic, how the concept of the Total survey error is suitable to manage new sources of data such as big data as scanner data, electronic payment, google trends. At the same time the deliverable contains an application on monitoring natural disaster based on mobile phone tath, that is extremely actual along this coronavirus days.

This example provides more evidences on how policy targets, statistical measures, new data, new methodologies, well-being and SDG are extremely connected each other and we hope that MAKSWELL project has been able to make a step further toward a interlinkages system able to manage all the interactions amid these characteristics.

3. Quality frame work for non-probability data

There is a well-developed framework for assessing the quality of data arising from probability surveys, in particular benefitting from the theory of probability sampling which provides methods for the estimation of errors due to the sampling process, based on the observed data (Neyman, 1934). Various extensions to deal with complex sampling designs and estimators taking advantage of auxiliary information about the sample and/or population units are possible. The methodology for these more advanced processes also leads to estimates of accuracy due to sampling, in the more complex cases providing only asymptotically unbiased variance estimators. Despite this well-developed theory, it has long been known that there are many other sources of inaccuracy in surveys, known collectively as non-sampling errors, and the identification and quantification of these errors has been brought together using the concept of Total Survey Error (TSE). TSE seeks to measure each of the components of inaccuracy in terms of bias and variance (some error sources will have either bias or variance, some will have both), and to combine these in an overall estimate of quality through the mean squared error. For recent overviews of TSE see Groves and Lyberg (2010), Biemer et al. (2017).

The rapidly increasing use of administrative data, big data and automatically collected data - which we can wrap together under the heading 'alternative data sources' - poses a challenge for the assessment of quality. In these sources there is typically no sampling, so the support of a well-developed statistical theory is unavailable. The data do not naturally arise as a result of a probability mechanism, whether under the control of the researcher or not, and can be regarded as nonprobability data. To assess the quality of alternative data we need to investigate the different aspects of the sources, processes and analyses which affect the conclusions drawn from the data. In the next section we describe an analogue for TSE which is applicable to alternative data.

3.1. Total error framework

Amaya et al. (2020) present an adaptation of TSE to alternative data sources. In the same way as TSE, the different components of the total error must be assessed individually, and the authors denote this the Total Error Framework (TEF), by dropping the 'survey' from TSE. A detailed picture of the quality can be built up by examining the different components and how they interact with each other, and this gives a better assessment of the overall quality than merely comparing the outputs with some gold standard (if one such exists). The approach has some similarities to InfoQ (information quality, Kenett and Shmueli (2014)) which is designed to consider how well the whole process of deriving outputs from inputs answers the research question under consideration.

Amaya et al. (2020) describe eight components of the TEF which correspond with similar elements in the TSE framework:

- coverage error
- sampling error

- specification error
- nonresponse/missing data error
- measurement/content error
- processing error
- modelling/estimation error
- analytic error

and these can be applied to a wide range of alternative data sources with some careful consideration of what the different components mean for different types of data. For example, coverage error should refer to the units of interest for analysis, and this might be different depending on whether the analysis is of tweets by person (which would require some linkage of tweets and mean that the unit of the analysis was the person), or of the body of tweets (when the unit of analysis would be the tweet). Unangst et al. (2020) give an illustrative example of assessing components of the TEF in a situation with multiple panel surveys, including some with nonprobability selection methods.

Zhang (2012) presents a different framework for the identification of errors in alternative data sources, considering errors to arise from two processes, one of measurement and one of representation. His model contains many of the same elements as the TEF, but sometimes using different terminology - for example Zhang splits measurement error into different components depending on whether it arises from the (lack of) relevance (using an available measure that is a proxy for what the researcher actually wants to measure), or a mapping process (correcting an available measure to more closely match the researcher's concept). Both belong to the measurement side of Zhang's model. On the other hand, coverage error is included in both Zhang's and Amaya *et al.*'s models, and is part of the representation side of Zhang's framework.

Another approach was developed by Meng (2018) in which the data quantity and quality, and problem difficulty are brought together in a 'trio identity', which allows some comparison of the effective size of big data for answering a particular problem. This effective size is often much smaller than the real size of the data, and this is strongly affected by the non-probability nature of the data collection process for big data sources. This identity (Meng's equation (3)) gives an expression for the error in using a big data source rather than a designed, probability source. This has not yet been turned into a statistic for comparing errors in different problems, but could be developed in this way. There is a clear need to review the different approaches to obtaining an overview of the total error in using non-probability sources to answer substantive questions.

3.2. Quality in linked data

A further development derived from the wider availability of data and increasing mechanisms for researcher access to data is that datasets are regularly linked to derive information on associations

and outcomes which are not available from single sources. In some cases a unique identifier makes linkage straightforward, but often such an identifier does not exist, and then the process of linking datasets is also subject to error. A substantial amount of work has been undertaken on ways to account for this error in analysis, and we list some of the principal developments below.

1. Linkage error in population size estimation. One simple use of linked data is in population size estimation, through the dual-system (capture-recapture) estimator. Ding and Fienberg (1994) developed a model to allow for linkage errors in one direction only between the two sources, and this has been extended to multiple sources and linkage errors in any direction by Zult et al. (2019). These approaches produce estimated variances which can be used to assess the quality, and accounting for the linkage error should reduce the bias at the cost of a modest increase in variance for a suitably well-fitting model. Unfortunately however this does not lend itself to TEF, because the bias is very hard to assess - if we knew the population size to compare against, we would not need the estimation process in the first place.

2. Accounting for error in regression type analyses. Chambers (2009), Kim and Chambers (2012a,b) explored the use of a simple *exchangeable linkage error* (ELE) model to adjust the outcome of different types of regression analyses for the possibility of linkage error. The general finding was that accounting for the error tends to attenuate relationships. ELE is rather unrealistic in practice, and there is a need to extend it to account for more complex error structures, but (despite some progress on software for more complex exchangeability models, see Powell and Smith (2020)) little progress has been made.

The errors arising from combining data sources are included in the TEF framework. They include errors through failing to link, which can arise through missing data, which may include not having sufficient common variables in the datasets to make a link, as well as through missing values within records. False links can also be made, particularly where variable values are not unique. These errors in making links are included within processing error in TEF, but can also contribute to coverage error. It is in these situations that an appropriate population size estimation process may be combined with metadata about linkage (particularly estimates of linkage error) to give information about the coverage.

Zhang (2012) also points to the challenges with linkage when the the datasets to be linked belong to units at different levels. This can lead to different estimates when the data are analysed at different levels – a component of error which is relatively less considered, and known as the 'unit problem' (Delden et al., 2018).

Linkage also affects Meng's approach inasmuch as it induces correlation between the topic of interest and the process of identifying (linked) records.

3.3. Making an assessment of error components

It is a considerable task to calculate or produce indicators for the various error components in the TEF framework. As we have seen above, there is typically an approach to variance estimation associated

with classical statistical estimators. But many of the other components of the TEF framework are challenging to measure. Biases in particular are challenging, because they rely on a bias-free measure, and if we had such a measure we would use it directly.

Sometimes a bias-free or reduced bias measure can be made using a separate study; such studies generally use small samples and are done infrequently because they are often expensive, involving additional data collection and possibly recontacting previous data providers. Such an approach is generally not possible or not practical, and therefore an assessment will rely on whatever information is available and what indicators can be gleaned from them. For example, much can be learned about the potential weaknesses in administrative (and big) data from a detailed description of how it has been collected and processed. Such 'data biographies' have been suggested by Connelly et al. (2016) and Smith et al. (2019). Amaya et al. (2020) also suggest making speculative assessments of TEF quality components in place of data-driven ones; these can be replaced if and when suitable data sources or studies become available.

If computationally intensive processes such as machine learning are used, the process itself generally does not give an estimate of variance. But it may be possible, through the use of test data or cross-validation results, to get an estimate of the variability of a procedure. In general it is not possible to make an assessment of the bias from internal calculations within a dataset – some additional source is needed. One idea is to gather a library of bias studies with all their metadata, and then use a statistical learning algorithm to make predictions for other studies based on their metadata.

3.4. Displaying and communicating (total) error

Once estimates or indicators for the various error components in the TEF framework are available, there is a further question over how to present this information in a way which is both accessible and of practical value to users of the statistics.

The principle of TEF (and Total Survey Error) is to provide an overall MSE which encapsulates the accuracy of the statistic being used, and this can be used in statistical tests of hypotheses of interest. But because it is a single number it does not give information about the relative importance of the different components of the error. For this purpose a dashboard-type approach is more useful, enabling the most important error sources to be highlighted. This may affect the choice of data or error measures for particular questions. Amaya et al. (2020) effectively present a dashboard for their chosen case study.

It is also pertinent to ask whether the indicators for different error components are in fact measuring similar characteristics of the data being used. Some additional analysis of the error components can show this. Smith and Weir (2006) consider the use of principal components to identify a subset of quality measures which capture most of the quality information present. This approach could be extended to identify important measures within or across datasets in the TEF (or TSE) too.

4. Needs for new methods

A common problem with non-traditional data sources is that the process that generates the data is unknown and likely selective with respect to the intended target population. A challenging problem in this context is to use this data for the production of official statistics that are representative of the target population. There is no randomized sampling design that facilitates the generalization of conclusions and results obtained with the available data to an intended larger target population.

Broadly spoken, two approaches can be distinguished to use non-traditional data sources in the production of official statistics and measurement frameworks for well-being and sustainability. The first approach is to combine survey data with non-traditional data sources in model-based inference methods. In this case prediction models for the target variables are constructed where survey data serve as the dependent data and related non-traditional data sources are used as covariates. The additional value of the information in the non-traditional data sources is that it can improve the precision and timeliness of survey data. This can be achieved with multilevel models or time series models with the purpose to increase the effective sample size in small domains with sample information from other domains or preceding sampling editions. Another advantage of related time series derived from non-traditional data sources is that they are often more timely and observed at a higher frequency compared to sample surveys. This aspect can be utilized to make more precise first predictions if the auxiliary series become available but the survey information is still lacking. This is often referred to as nowcasting.

A second approach is to use the non-traditional data sources directly to construct official statistics or indicators for well-being and sustainability. Under this approach the problem that the data are selective has to be faced. This might require strong assumptions about the data generating process in order to correct for selection bias.

Both approaches are described in more detail in Deliverable 2.2. (van den Brakel et al. (2019)). In this Chapter the needs for new methods and data in this context is described.

4.1. Methods for dealing with selection bias, current issues and perspectives

Chapter 3 in Deliverable 2.2. (van den Brakel et al. (2019)) describes a list of methods to adjust the estimates for the bias due to a non random selection of the data, that is a likely setting when dealing with secondary sources and big data, e.g. Rosenbaum and Rubin (1984)), (Puza and ONeill (2006), Tam and Kim (2018), Elliott (2009),Elliott and Valliant (2017), Särndal et al. (1992), Rivers and Bailey (2009), Chen et al. (2018), Kott and Chang (2010), Heckman (1976), Heckman (1990), Manski (1989), Sverchkov and Pfeffermann (2004), Pfeffermann and Sverchkov (2009) Feder and Pfeffermann (2015).

The adjustment methods have been developed in the framework of sample data in the presence of non response and for non ignorable sample selection, that arises in web-surveys. However, the efficacy of those methods to effectively correct for the bias in practice is questioned.

First, to adjust the selectivity all methods rely on additional information on the records (units) of the big data source and an external auxiliary source that is unaffected by bias. These conditions for the adjustment can be very demanding for secondary data.

Sometimes a separate survey has to be implemented. In this case, being the survey devoted only to adjust the estimators obtained with secondary source, a large sample size is not likely needed. Pfeffermann (2017) highlights conditions and limitations of a sample survey aiming at adjusting for selection bias: the knowledge of the sample element membership to the big data source, the existence and availability of covariates explaining the membership to the big data sources.

Recent literature have compared the adjustment methods (see Matei (2018), Buelens et al. (2018)). In particular Buelens et al. (2018) have pointed out how in practice the demographic auxiliary variables-usually employed in the correction, do not sufficiently explain the data generating process of a non-probability sample to correct successfully for selection bias.

Strong auxiliary data that is capable of explaining the data-generating mechanism for an effective adjustment for the selection bias are needed to successfully apply the methods.

Even so, the availability of this very basic (e.g. gender and age) auxiliary information for the records in the big data source is not always straightforward. Daas et al. (2016) explored the possibility to *extract* those feature from the twitter accounts that could serve to correct selectivity.

However, feature extraction can produce a *measurement errors* on the covariates that are used for the selectivity adjustment.

Indeed when high measurement errors affect the covariates, estimates of correlations and effect sizes are attenuated by those measurement errors then reducing the possibility of bias correction.

The effect of a measurement error of the auxiliary variables that are used for the selectivity adjustment is not explored yet and the degree of effectiveness of the correction needs ad hoc analyses in real cases.

Moreover, the methods that require unit level information are not straightforwardly applicable when the links of the big data source with target units of the population cannot be carried out.

In fact, big data source often contains records that are different from the target statistical units. The process of unit's identification might produce errors, such as duplication of units or wrong association of records, that cause themselves coverage errors and might impact the suggested bias-corrections. In this way, other sources of errors, such as error in unit identification, might reduce the benefit of bias adjustment. Moreover, the bias corrections are based on a model that describes relationship between the target variable and some auxiliary variables. These methods are less effective when the auxiliary variables are affected by measurement error, for example due to wrong association of the observed and target units.

It should be worthwhile exploring models for the bias adjustment that take into account the identification errors in the covariates, see the classical literature on measurement errors in in the covariates, Stefanski and Carroll (1985), Schafer (1987), Fuller (2009) and the recent literature on how to deal with linkage errors in statistical analysis, Chambers (2009), Di Consiglio and Tuoto (2018).

Measurement error and unit identification do not only affect the selectivity adjustment but are causes of biases themselves. These aspects are explored in the next sections.

4.2. Causes of bias in big data

As mentioned above, big data source may be affected by various sources of errors besides selectivity that impact on the bias (and somehow, variability) of the estimates:

- The unit error, i.e wrong identification of the population units;
- Measurement error, i.e. the measure observed in the source is different from the target variable;
- Consistency over time.

For big data the measurement error can be caused by the technical instrument used to collect data but also it may occur if a biased algorithm is applied for extraction of the target variable itself during the processing phase.

Let us consider some examples of SDG indicators obtained with big data and let us analyse the errors that might arise. Starting with the use of smart meters for measuring the SDG 'Energy consumption in households'. Misalignment between the big data units (the meters) and the population units (the households) may introduce both selection bias and measurement error. Misalignment may introduce selection bias due to missed links between big data and population units: e.g. households are not provided with the meters and so not included in the observed population. Measurement error may arise due false links: e.g. the smart meter connect to an small economic entity is linked to a private household, resulting in an erroneous value for the consumption.

A very useful source for SDG evaluation is the remote sensors; some of the following SDG indicators can indeed be measured by its use:

- The statistical product - SDG indicator 11.7.1 - the 'average share of the build-up area of cities that is open space for public use for all, by sex, age and persons with disabilities', see WPH-Team (2019);
- the 'Annual change in forest area and land under cultivation'.

In these examples, the target variable Y is not directly observed but results from processing the original data. The algorithms might introduce bias, i.e. in the classification of the land cover.

The given examples also introduce another peculiar aspect of big data, i.e. its consistency over time. In fact, the collected data are often subject to changes due to technical reasons (the remote sensing), this makes the comparisons over time also affected by a bias derived from changes of the relationship between collected data and "true data".

A study on the use of remote sensing for measuring the 'average share of the build-up area of cities' WPH-Team (2019) analyses the sensitivity of the results to the input parameters of the sensors themselves. They observe that the results for city boundaries are not very sensitive to the radius on which neighbour pixels are considered, indeed they depend on the characteristics of the neighboring pixels (i.e. the share of neighbour pixels which are also built-up) .

A standard set of methods for dealing with these errors in big data is not yet well structured. We briefly consider these sources in the following.

4.2.1. Unit identification

As pointed out above, the big data are 'indirectly' related to the target population. This applies to data from the social network platforms, where the most atomic object (e.g. a tweet) is typically not the statistical unit of interest (e.g. a person); to mobile network data, where the basic object is a communication event produced by a mobile device while the statistical unit of interest for statistics is usually the person using the mobile device, if any. The alignment between the observed units and the statistical units of interest is often complex, even when the statistical units are the same of the Big data source. van Delden et al. (2019) investigate the linkage between business website addresses and a business register, highlighting difficulties in identifying units correctly. Probabilistic record linkage methods might help to this purpose. When the link between the observed data and the statistical unit of interest involves uncertainty and errors, this is commonly referred to as as unit identification error.

The definition of a linkage strategy might be a challenging task when dealing with big data, from the selection of the most discriminant common variables to use as linking variables to the identification of the most effective distance measures for extracting the proper information from unstructured strings. New techniques for data linkage, different with respect to the traditional ones (e.g. based on the theory of Fellegi and Sunter (1969) for record linkage), might be investigated to enforce the recognition of less structured data. Moreover, due to the theoretical continuous availability of big data sources, linkage activities need to define an incremental process able to analyze, integrate and validate each added data sources. An analysis of the complexity of linking big data sources in official statistics is provided in Tuoto et al. (2018).

Recently, an increasing interest has been devoted to the analysis of linked data that takes into consideration in a proper way the risk of linkage errors, namely false links and missing links. The methods proposed to adjust bias and variability introduced by linkage errors still rely on strong assumptions: the availability of linking probabilities/weights at record level and non informative linkage procedure, see Lahiri and Larsen (2005), Han and Lahiri (2019). A proposal by Chambers (2009), with interesting extensions in Chambers and Kim (2015), Samart and Chambers (2014), follows a secondary user approach, it does not require the knowledge of linking probabilities at record level at the price of relying on an Exchangeability Linkage Errors assumption.

It is worthwhile noting that generally linkage procedures do not release linkage probabilities and other summary measures, a so-called linkage metadata. Even if the probabilistic record linkage produces these measures, as a by-product of the procedure itself, however the estimates accuracy is often questionable, see for instance Chipperfield and Chambers (2015), Tuoto (2016). How to obtain accurate summary measures for linkage procedures is still an open research field.

4.2.2. Measurement Error

Measurement error is the difference between the true value of the measurement and the value obtained during the measurement process.

According the Eurostat Quality Guidelines for the Acquisition and Usage of Big Data WPK-Team (2019), the concept of measurement error when dealing with Big data should be enlarged to account also for errors arising from data acquisition and errors in measurement instruments (meters, satellites,...). Moreover, the measurement error arises due to the transformation process that relates the big data sources to the statistical variables. Indeed, whereas working in statistical processes with traditional data, it is usual distinguishing between measurement errors, model errors, and process errors, on the other side, when dealing with big data, the distinction between measurement error, model error and process error becomes much ambiguous.

Big data based estimates are mainly produced by models. Working with big data sources instead of survey data, the information about the target variable is not explicitly in the data source ; instead the information of interest has to be inferred from other variables in the data. Hence, modelling the information about the target variable plays a prominent role. The complexity of the algorithms and models needed to arrive at the required information about the target variable depends on how directly the information of interest and the available information from the big data source are connected.

A model in general is a simplification or an idealized form of the data-generating process (the truth), so model mis-specifications can occur for classical statistical models, e.g. linear regression, but also for advanced machine learning algorithms, like random forest or deep learning, e.g. simply by not including an important variable.

For instance, with telco data and AIS data one might encounter technical faults that affect the data acquisition as well as model mis-specification that impacts on the interpretation of the signals. Measurement errors in the geolocation coordinates of an AIS message results in ships appearing at impossible or illogical locations (i.e. in an inland area) Another example is a ship/phone suddenly bouncing a few kilometers back and forth during its journey, showing illogical paths. These measurement errors are particularly relevant when the geo-location of the measured event is one of the statistical objects. Errors corrupting the data acquisition process may be associated with both the attribute value and locations of the attribute values.

When using smart electricity meters, model errors arise when aiming to measure the consumption of self-produced energy, that cannot be recorded by smart meters.

In statistics based on social networks, model errors are given by the transformation of the natural language textual data into some variables of interest, e.g. the sentiment expressed in the communication, which is unobserved (i.e. not directly observed).

Dealing with measurement errors in the use of big data sources is still not fully explored. Measurement errors should firstly be evaluated.

For instance, when applying supervised algorithms, the bias of the model can be measured by the training data set. However, the availability of an unbiased training set in an open issue.

4.3. Small area estimation and time series methods

In Deliverable 2.2 an extensive account is given on the use of model-based inference methods for survey data using new data sources as auxiliary variables (van den Brakel et al., 2019). It was recognized that parallel on the literature of cross-sectional small area prediction models, literature arise where in particular satellite and arial images are used in developing countries to make low regional estimates for poverty and income. Some papers propose remote sensor information that correlates with survey data as a surrogate construct for variables like poverty. Other papers use machine learning algorithms to train survey data on remote sensor information and use these algorithms thus obtained to make low regional predictions using remote sensor information only. This appears to be suboptimal compared to the literature on small area estimation where small domain predictions are interpreted as a composite estimator of a direct estimator and a prediction under the assumed model. The literature where variables derived from big data sources or remote sensors in cross-sectional small area prediction models is limited, see Marchetti et al. (2015) for an example. More empirical research where the performance of formal cross-sectional small area estimation methods are compared with the literature on estimating low regional poverty with machine learning algorithms applied to remote sensing data is needed.

Most surveys conducted by national statistical institutes are conducted repeatedly over time. A natural approach for small area prediction as well as now-casting is to apply time series models to use related information from previous editions of the survey. This can be done as a form of small area estimation, but also as a form of now casting to obtain more precise provisional estimates of target variables in real time, i.e. already during the reference period when the data collection of the survey is not completed yet. In Delivarable 4.1 an application is described where time series based on a large set of search terms in google trends are used to nowcast the monthly unemployment figures in the Netherlands (van den Brakel et al., 2019). With this kind of big data sources a large set of auxiliary series is easily obtained. Including these series in a multivariate time series model, where each series has its own trend and seasonal component and correlations between trend disturbance terms are modelled to borrow information from the auxiliary series results in models with a large amount of parameters, which consequently reduces the predictive power of such models. To handle this so-called high-dimensionality problem a dynamic factor model is proposed, following the approach proposed by Doz et al. (2011).

The application described in Deliverable 4.1 shows that an auxiliary series derived from the registered

number of people receiving unemployment benefits contains information that is strongly correlated with the number of unemployed estimated with the Dutch LFS, while the information derived from Google trends is rather weak. This is, however, a first empirical result that requires further research. In this application the Google trends series were obtained by considering which search terms can more logically expected to be related to unemployment. It appears from this first attempt that deriving common factors from a large set of auxiliary series that contain to much series that are unrelated with the target series deteriorated the predictive power of the common factors. More research how to select the most relevant set of auxiliary series out of a large set of potential auxiliary series, without falling into the trap of data dredging, is required. Empirical applications that illustrate the benefits of auxiliary time series that can be derived easily, without additional costs, from big data sources is needed.

A strong assumption underlying the multivariate structural time series models described in Deliverables 2.2 and 4.1 is the assumption that the correlation between the disturbance terms of the trend in the auxiliary series and the trend of the target series are time invariant. The correlation between the number of people derived from a register on unemployment social benefits and he estimated unemployment from the LFS might gradually change over time, e.g. due to legislative changes with respect to people who are qualified to receive unemployment social benefits. Similarly it can be expected that comparability over time of auxiliary series derived from Google trends, social media platforms or other relatively volatile big data sources is low, which also violates the underlying assumption that correlations with target variables of series obtained with repeated surveys are time invariant. Further research how to account for time varying correlations in multivariate time series models is therefore needed, before this type of auxiliary series can be considered in model-based inference procedures for official statistics.

5. Deep Learning in Official Statistics

Deep Learning has outperformed by a large margin many of the more traditional machine learning techniques in many domains. While many traditional machine learning techniques saturate at a certain point, deep learning performs better the more data is added. As such, it lends itself for increasingly large datasets that are currently more and more common in official statistics. What is more, using machine learning techniques, in general, opens up possibilities for different types of data to be used. Traditionally, official statistics uses numerical and tabular data, mostly originating from survey and register data, for which the represented information is close to the concepts that are being measured. By using machine learning, also text, images, and a wide range of signal data can be used as a source of information. For these kinds of data, the information or patterns extracted need more processing to connect them to the concepts being measured.

Moreover, the techniques used in traditional statistics give detailed information about the uncertainty of the processed results, are transparent, and are often transferable from one problem domain to another. Especially in these areas, we can find the challenges of using deep learning. First of all, the *different types of uncertainty* are less well considered and less easily explained. Second, deep learning models need to be trained on a sample of the data (the target domain) they will be applied on. To train a model that generalizes well to the target domain, a lot of data is needed. Often then, models are trained with a process called transfer learning in which a model that was trained on a large number of images, for example the Imagenet ILSVRC dataset (Russakovsky et al., 2015), is taken as a starting point and retrained on the new problem domain. Training a model on a sub sample of the target population, or using transfer learning, raises questions of how well these models *generalize across domains*. Third, deep learning models have millions of weights, which makes it much *more difficult to explain* why certain model outputs are generated. Often, deep learning models are therefore referred to as black boxes. While not entirely true, we can look and analyse all the weights in the model, it is this sheer number of weights and the complexity stemming from them that makes model behaviour more difficult to explain.

In the following, we will shortly describe three areas of deep learning that can have a greater or lesser influence on making deep learning suitable to be applied in official statistics. In section 5.1, we will look at the causes of uncertainty and possible ways of dealing with them. After that, section 5.2, will lay out concerns about model generalizability and will point out how research in domain adaptation could be a possible direction for further research. Last, section 5.3 will raise some concerns about model interpretability and will give some pointers for further research.

5.1. Uncertainty in Deep Learning Models

Deep Learning models are amongst the most powerful machine learning techniques that exist. While powerful, even the best model will eventually not be able to predict a correct input for all the inputs given. While some of the errors can be prevented during training time by using a bigger training dataset or a sample better reflecting the target domain, other errors will be difficult to prevent; there is no such thing as a perfect deep learning model. It is therefore important to find ways to deal with

the uncertainty and errors in the predictions of a deep learning model. We should identify possible sources of uncertainty, prevent the uncertainty where possible, and model the uncertainty elsewhere; only then can we use the deep learning models as a source for official statistics.

Two types of uncertainty can be identified for deep learning models: (1) the aleatoric uncertainty or the uncertainty in the data, and (2) the epistemic uncertainty or the model uncertainty (Kendall and Gal, 2017, Loquercio et al., 2020). If we look at a supervised deep learning process, several steps can be identified, which can influence either of these uncertainties. Figure 5.1 illustrates the various steps in a supervised deep learning process. We will describe how each of these steps can influence uncertainty below.

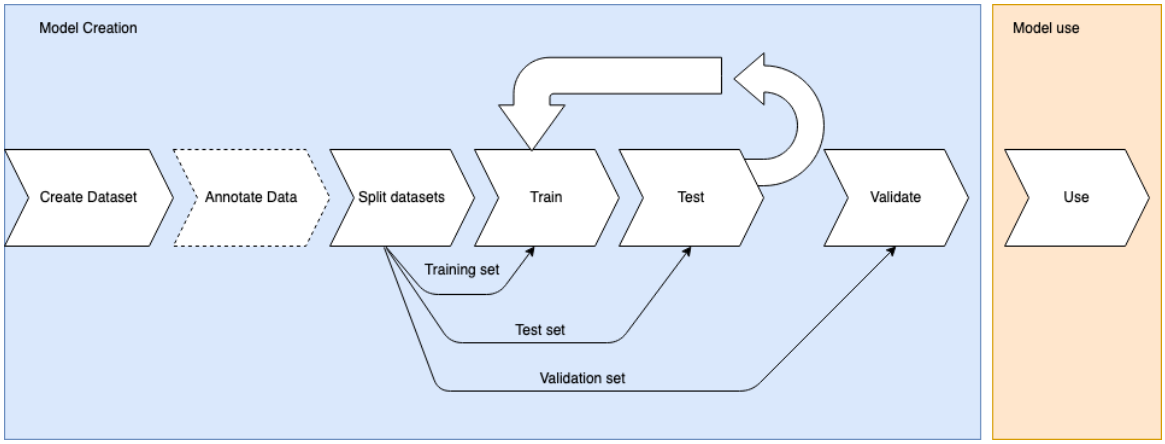


Figure 5.1: A supervised deep learning process visualized.

The process starts by *creating a dataset*. Sometimes, the data in the dataset is a random collection of found data on the Internet, like for example pictures or text from a social media platform. In other cases, as is often the case in official statistics, the data is collected from a source that describes a target population. The data collected can for example describe the population of a country, but also, in the case of aerial and satellite images, the dataset reflects the country as a whole. While in some cases the integral dataset is used in model creation, it is also common to create a sample out of the population. In the latter case, care should be taken that the sample is a good representation of the target population as a whole. An imbalance in the dataset, where certain groups of interest are over or unrepresented may introduce a model bias. But even if the integral dataset is used, a model bias towards certain years in the dataset may be introduced: a model that is trained using data of this year, may not work correctly for the previous or the next year. As such, the dataset taken to train the model can influence model uncertainty. Important issues for further research here are related to the issue of creating a good sample of the target population. In this respect, in some cases it is possible to use background information from registers as metadata to create a representative sample. In other cases, it may be necessary to look at the data itself to create a representative sample. In the context of image, text, or signal data, creating such a sample may not be straightforward.

Step two is *annotating the data*. In some cases, this step can be done automatically, as the annotation can be based on register data. For instance, aerial image data can be annotated with labels from

the housing or income register. In other cases, the annotation process is a laborious manual process, in which human annotators need to go through the whole dataset, specifying a label for each image. Labelling whether an aerial picture contains solar panels or not is an example of a manual annotation process. A manual annotation process can be used to get more information about the data uncertainty. By having more people annotate the same data points, the similarity or dissimilarity between annotations can be used to measure the uncertainty or noise in the data. Having such information about the dataset, it is possible to use a trusted subset to greatly improve final results Hendrycks et al. (2018), Li et al. (2017). Further research is needed to investigate how to shape the annotation process to measure the data uncertainty and to evaluate practical aspects of utilising trusted subsets.

Step three is *splitting the data* to create the training, test, and validation sets. The issues in this step are very similar to the ones described for the first step. The training, test, and validation sets should each form a representative sample of the target population. They should contain sufficient complexity and variety for the model to generalize adequately. What is more, the test and validation sets should be able to give a reliable impression of the model's performance.

Steps four and five *train and test* the model and directly influence model uncertainty. An issue that especially affects model performance is class imbalance, where samples of one class may largely outnumber samples of another. While related to the representative sampling methods mentioned in the previous paragraph, class imbalances are often part of the target population. The number of houses with solar panels in the Netherlands are, for instance, a much smaller subset than the houses without. A number of approaches have been suggested to deal with class imbalances which are systematically studied in (Buda et al., 2018). In this paper, a distinction is made between (1) *data level methods* and (2) *classifier level methods*. *Data level methods* consist of either oversampling the minority class(es) or undersampling the majority class(es). *Classifier level methods* modify the classifier itself and can consist of a variety of measures, among which adjusting the network output according to class distribution, adjusting the learning rate for the samples from different classes, or adjusting the the loss function. Since class imbalances can lead to model bias, it is important to consider ways to overcome the effects of these imbalances. Several suggestions for limiting the effects of class imbalances are given in (Buda et al., 2018). Similarly, the methods mentioned here can also be used to learn from wrongly classified samples. It has to be investigated if retraining a network with the wrongly classified samples leads to an improved model performance or that these samples need to be given more importance while retraining. A promising avenue of research related to this is so-called importance sampling that aims to focus model training on informative samples (Katharopoulos and Fleuret, 2018).

In step six, the trained model is *validated* on a previously unseen dataset. As such, the issues for this step are similar to those in the previous step. During validation, it is also measured how well a model generalizes across domains, more about this will be presented in section 5.2

Last, step seven *applies the model* and uses it in a production setting. In a production setting, the model will encounter data points not present in the training set. To be used in official statistics, the uncertainty in the model results caused by data and model uncertainty should both be taken

into account. Several studies have investigated various ways to deal with uncertainty. First, it can worthwhile to be able to predict when a model is likely to commit an error and to abstain from classifying in such cases. A model which is able to abstain is called an *selective predictor*. There is a lively branch of inquiry looking into how one might go about building such a model Geifman and El-Yaniv (2017), Hendrycks and Gimpel (2017). In some cases, just the magnitude of classifier output scores is enough to tell whether a sample is misclassified Hendrycks and Gimpel (2017). Second, another branch of research aims to model both the data and model uncertainty to the extent that this is possible (Gal, 2016, Kendall and Gal, 2017). Most of the studies use Bayesian methods to estimate the uncertainties. Some methods change the deep learning model to provide the Bayesian estimation. For example, in one case, a dropout layer is kept during prediction time to give an approximation of the uncertainty in model prediction (Gal and Ghahramani, 2016). A strong critique of the dropout method is however offered in Osband (2016). Another approach trains ensembles to be able to estimate model uncertainties (Lakshminarayanan et al., 2017). Conversely, a particularly promising study provides a way to estimate both data as model uncertainty without changing the model (Loquercio et al., 2020). A confounding factor for the use of uncertainty estimation are so-called adversarial examples. Adversarial samples are such that they are classified easily by a human, but neural networks have a tendency to assign a completely wrong class to them with high confidence Szegedy et al. (2014). Such samples can be crafted artificially given parameters of a model but, quite concerningly, they have also been found to occur in the absence of malicious tampering Hendrycks et al. (2019). Furthermore, research in adversarially-robust models implies such robustness may be at odds with usual metrics Tsipras et al. (2019). We feel that the directions given in this paragraph are especially interesting for further research, because they help make the causes of uncertainty in model predictions more explicit.

5.2. Model generalizability and Domain Adaptation

One important issue with deep learning models is how well the model generalizes to unseen data. To evaluate how well the model generalizes, the validation set should be sampled in such a way that it gives an accurate description of model performance on data not encountered before. Often there are multiple ways of creating a validation set from the target population. For instance, for geographical data we can distinguish between cross-region evaluation and cross-site evaluation (Wang et al., 2017). A cross-region evaluation samples both the training and validation set from several geographical dispersed regions, while a cross-site evaluation trains a model on one region and validates it on another. It was found in (Wang et al., 2017), that a model trained and evaluated in a cross-site evaluation performs substantially worse than a model trained in a cross-region evaluation. It has to be investigated what substitutes as a "good" validation set and a reliable validation of the model. An inadequate validation can give a too optimistic view of model performance, model uncertainty as well as data uncertainty.

Another issue that plays a larger role when validating a model, is that the evaluation metrics often used to benchmark a model, like the accuracy, precision, recall, and f1-score, can be largely influenced by imbalances in the dataset. As such, these evaluation metrics cannot be compared across datasets that have different imbalances and may give a too optimistic view of model performance. A scenario in which this issue comes forward in the geographical cross-site evaluation mentioned above. In a cross-site validation, the validation set is independently sampled from the training set, which can result in a different class imbalance than the training set. Therefore, further research should look into evaluation metrics independent of class imbalances, like has been presented in Luque et al. (2019).

An interesting point that should be considered that is related to model generalizability, is the issue of domain adaptation. Time and again, it is assumed that the domain a trained model is applied on, has the same feature space and distribution as the dataset the model was trained on, however this is not always the case. Several studies have already explored the problem of domain adaptation for more traditional machine learning models (Ben-David et al., 2006, Mansour et al., 2009, Ben-David et al., 2010, Pan and Yang, 2010). A same study applied to the domain of deep learning and computer vision would be a good direction for further research. Some specific aspects to consider in this respect are on the one hand how many data are needed to increase the performance of a model that was trained on one domain to be applicable to the other (Wang et al., 2017). Particularly, minimizing the amount of data to yield similar model performance on a different domain is worth further investigation. On the other hand, further research could look into whether the distributions of an image dataset during training time and one during application can be derived and compared on the basis of the image data alone. Finally, there is some work on the topic of estimating performance of a model trained on one domain and used on another, but this field of inquiry merits more work before it can be used (Wang and Deng, 2018, Elsahar and Galle, 2019).

5.3. Model Interpretability

An interesting aspect of deep learning for official statistics, but at the same time the most elusive one, is model interpretability. One of the most important issues in this respect identified by Lipton (2018), is that the term *model interpretability* has not been further specified. Subsequently, several parts of model interpretability are explored by Lipton (2018), split out in two major subcategories: (1) the aspects that cause a demand for interpretability and (2) the transparency notion of interpretability. The author furthermore identifies that even a linear model with complex input features may not adhere to all of the aspects of model interpretability. To use deep learning in the context of official statistics it has to be therefore identified which aspects of model interpretability are important. Is it important that a person can understand the model at once, should every part be intuitively explainable, or are post hoc interpretability aspects, like for example visualizations of intermediate layers (Selvaraju et al., 2016, Zeiler and Fergus, 2013, Montavon et al., 2017), while really local in nature, more important to look at?

6. Challenges and further research needs for remote sensing data

Satellite and other remote sensing data have become increasingly available at low or no cost. Therefore, and due to increasing computational capabilities and new methods, the interest in social sciences and statistics for large scale applications has increased in recent years. Similarly, the interest in interdisciplinary cooperation in remote sensing and geography to try out their information and knowledge in other fields increased (cf. Taubenböck et al., 2015, p. 1).

Main reason for the use of remote sensing data have been the non-availability or unreliability of official data under certain circumstances. Hence, this new data source is mainly used as a replacement or in addition to situations in which good quality data is only sparsely or not at all available. For example, to evaluate economic changes in North Korea, for which no data are officially published as done by Lee (2016). Other applications successfully use remote sensing data to track illegal deforestation when official reported information have been faulty, or for identification of mayor war crimes in remote areas, as demonstrated by Henderson et al. (2012). The global coverage of sun synchronous satellites made those applications possible, where no other data would be available or reliable (cf. Burgess et al., 2012, p. 3ff. and Henderson et al., 2012, p. 8).

When trying to improve the measurement of well-being development in the European Union member states, the situation is quite different than in the mentioned applications. Quality data are available in a timely manner, produced by established statistical institutions and agencies in each member state under common ideas. Although further improvement is always possible, remote sensing applications have to deliver excellent quality estimations and predictions or find different ways to assist in the statistical framework. Many problems still remain for the use of geographic and remote sensing data, which have to be tackled to effectively incorporate remotely sensed data into a framework of official statistics. For example, no comprehensive overview of methods and datasets are available, according to Dai et al. (2017), not even for the use night-time light data for GDP or population estimation (cf. Dai et al., 2017, p. 1).

With regard to the general topic of measurement, it has to be disccussed that well-being measurement is so divers that almost any satellite based study could qualify as relevant. While many opportunities have been discussed in WP 2.1, the integration of remotely sensed data into the framework of SDG and well-being measurement, at present state, are mostly experimental. Applications in forest and agriculture are rather established and started as early as 1930 as Monmonier (2002) describes, when cameras were fixed to balloons to map the total extent of the US American agriculture. For a modern use in forest inventories see for example Wagner et al. (2017). The interest in such data for the measurement of other SDG areas, however, is a newer development, on which this part in the MAKSWELL project is going to focus.

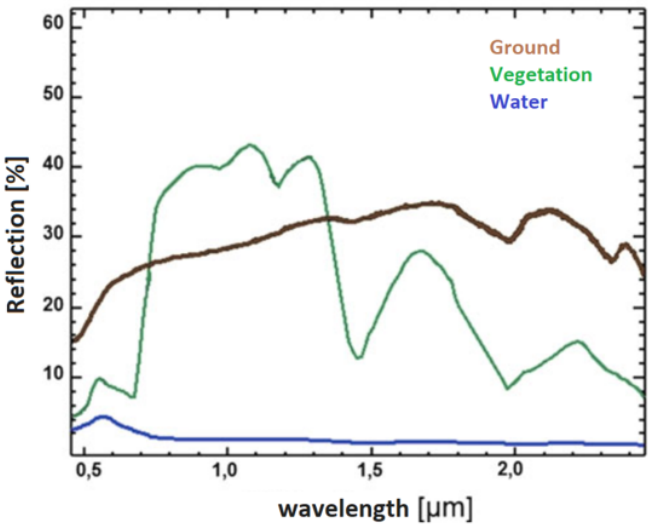
Remote sensing refers to drone, aircraft or satellite based data. To simplify the context, identified research and data needs will be illustrated for satellite-based approaches as a remote sensing data

sources, which is most widely available. Aircraft based images for example are usually much higher in their spatial details, but data are collected less regularly and only for small areas. To a great part both datasets share similar challenges.

Commonly additional geographic information, often called geographic information systems (GIS) data are required to combine remote sensing and traditional survey data. In the simplest version this can be a map of administrative areas. More complex are products such as the CORINE dataset which contains remote sensing information in combination with possibly many other data sources (Umwelt Bundesamt, 2019).

Taubenböck et al. (2015), page 50, warns to consider that: "Every city is always only so big, so divers, so heterogeneous or complex as the own perspective allows. This applies to any scientific approach. Scientific approaches document and analyse only a self-determined subsystem." This is especially true for remote sensing, where the potential possibilities seem endless in theory. Applications on the other hand show the limitations of these datasets to date and the creativity required to design new approaches.

The core challenges for measuring well-being and SDG using any remote sensing data is that satellites provide only measures of electromagnetic energy. Patterns in these measures represent physical and chemical properties of the earth's surface. An example is shown in figure 6.1. Well-being aspects are never measured directly, only structures and changes in the surrounding living environment of people for well-being aspects are described. Such, well-being aspects, which are reasonably related to the living environment of people might be subject to applications including remotely sensed data.



Spectral signature of three landcover types: water, vegetation and ground.

Figure 6.1: Translated from image 4.1 in Taubenböck et al. (2015) p. 25.

Two basic approaches are available when using satellite images. First a direct use of the image information, for example by creation of indices for urban or vegetation density. The information from

the images then might be used in a model or further developed. Problems related to this approach are discussed in section 6.1. The second approach are methods, which create new variables and features from the satellite image, but will not use the original image's data for estimation, discussed in section 6.2. Cluster and machine learning methods, which allow the identification and counting of object such as cars, planes or crop type and the quantities as new variables are commonly applied.

6.1. Model based approaches

Satellite based data might be valuable in SDG measurement due to several properties such as the global coverage, spatial resolution and high recover rates (see Deliverable 2.1). These properties also create many new problems, before unknown to official statistics, particularly in direct methods as described in this section.

The following example might illustrate some of the issues that are going to be discussed. NASAs Landsat 8 satellites collect images on 12 spectral bands with 100 up to 15 meter spatial resolution since 2013. To compile one image of Germany spatial mosaicking and temporal filtering is required to process a total of 483 Landsat 8 images. Each image covers the area of 185 square kilometres and is uncompressed 1.61 gigabytes in size. This means, to create one index image of Germany, for example the *Normalizes difference buildings index* (NDBI) image in figure 6.2, 777.63 GB of data would have to be downloaded to solve the composition on a local computer for each year. For multispectral data many combinations of spectral bands have been proposed to construct indices, which describe aspects such as vegetation density or surface concealment.

The NDBI is used as an example of satellite indices. By combining the Landsat near-infrared light (NIR) and mid-infrared (MIR) bandwidth in the following form:

$$\text{NDBI} = \frac{\text{MIR} - \text{NIR}}{\text{MID} + \text{NIR}} \tag{6.1}$$

(cf. Faisal et al., 2016, p. 16)
A NDBI indicator $\in \{-1, 1\}$ is created, which indicates dense artificial surface concealment, absence of natural ground or vegetation for high values and no ground concealment for low values. Because all spectral bands are equally coded from 0 to 255, the ratio of differences is at most 1 and minimally -1. Construction efforts in settlement areas might be a good indicator to changes in realized construction or economic development in industrial areas. Similar concepts exist for the vegetation density normalized difference vegetation index (NDVI) and other environment factors in more complex approaches (cf. Faisal et al., 2016, p. 16).

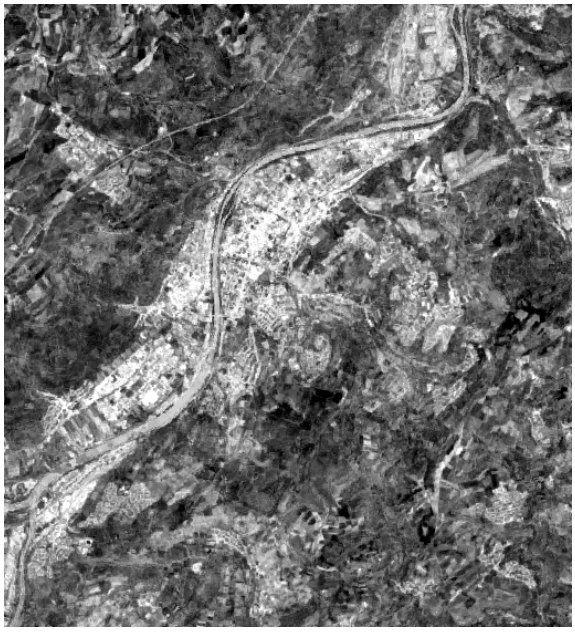


Figure 6.2: NDVI Index map of the Trier area composed of the 2017 Landsat 8 images

Figure 6.3: NDBI Index map of the Trier area composed of the 2017 Landsat 8 images

The images for this example were composed using Google Earth Engine (GEE). (cf. Gorelick et al., 2017, p. 19ff.). GEE is a satellite data catalogue and cluster computing interface containing several hundreds of petabytes of freely available satellite data from around the world. The cluster computing interface from Alphabet Inc. is free of charge for non-commercial applications and features as a data pool for freely available satellite data.

Direct approaches create a value for each pixel of the satellite image, which can be aggregated to any *area of interest* (AOI) and used as variable in models as long as the pixels are smaller than the AOI.

The following research areas and data needs were identified:

1. Technical Challenges

Data Volume and Computation:

Sun synchronous satellites produce information about the entire globe. Depending on the number of spectral bands and the spatial resolution of the produced image, a high volume of data is collected and published in form of single scenes, rectangle images of determined real world coverage. If the target area is not covered completely by one scene, multiple images must be combined. This is a common task in remote sensing and requires mosaicking, composition and filtering of multiple images. The required calculation and strorage capacities for larger scale applications are problematic. The example was only calculable due to third party resources from Google Earth Engine.

Often the data quantities can not be handled on personal computers for larger applications. Hence, many applications with complex data are reduced to small "areas of interest" (AOI) such as bigger cities instead of countries. (cf. Faisal et al., 2016, p. 1) Whether corresponding results can be generalized beyond these AOI is questionable.

To facilitate the wider development of satellite-based application, which include also remote areas specifically interesting for well-being and poverty analyses, an Europa-wide or even transnational computing infrastructure should be established. This could ensure the data safety when working with official statistics information as well as independence from commercial entities. The maintenance and creation of efficient remote sensing databases and cluster computing is not a simple task, which might not be solved effectively by individual institutes, while allowing access by outside users such as researchers at universities.

Image inconsistencies:

Although, the mosaicking of multiple satellite scenes to greater maps is a common task in remote sensing, the result is never perfect. Each scene is collected at different points in time, with different angles, seasons and illuminations possibly creating very inconsistent images. The following figure 6.4 depicts the recording path taken by the satellites for the GOME project.

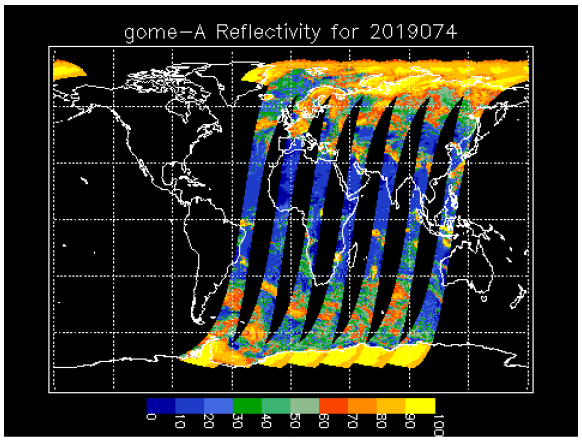


Figure 6.4: Image of the GOME mission by NOAA, National Centers for Environmental Information (2014)

Figure 6.5: Image near Frankfurt Hahn Airport from Google Maps static API

As most satellites, the GOME mission presented in figure 6.4 did not cover the entire surface in one rotation circumference. For a full surface record, neighboring areas might be recorded weeks or months apart.

Algorithmic composition of images often results in unrealistic compositions as often visible in Google Maps static API seen in figure 6.5.

How well the mosaicking works depends on many factors, but mostly the recording intervals of the corresponding satellites, the spatial resolution and the complexity of the collected data determine the

outcome. In contrast to information density in high resolution images, the combination of low spatial resolution and only one spectral band as in the NPP data, allows for very consistent compositions.

This makes the selection of appropriate data sources a difficult task. More complex data might result in worse predictions when the complexity is not cleanly solved technically.

The development of technical solutions to these effects is a pure remote sensing problem. However, the data will never be exact and perfect. To what degree such inconsistencies are problematic to predictions of well-being indices is not clear. Many publishing institutions of satellite images already provide data on different levels of homogeneity regarding adjacent images and previous images, for example Department of the Interior, U.S. Geological Survey (2018). Further studies to data quality and prediction sensitivities are required to allow evaluation of upcoming results and judgment of approaches and their re-usability to other areas.

A second type of image inconsistency are created by clouds, shadows and other atmospheric situations. While many images are pre-processed for temperature and moisture effects, clouds and shadows distort the image elements to a degree that underlying information can not be recovered from within the image. Commonly a type of imputation of effected pixels takes place by replacing covered parts of the image with information from previous records. To maintain spatial consistency, pixels are imputed from images of the same area, but earlier or later in time.

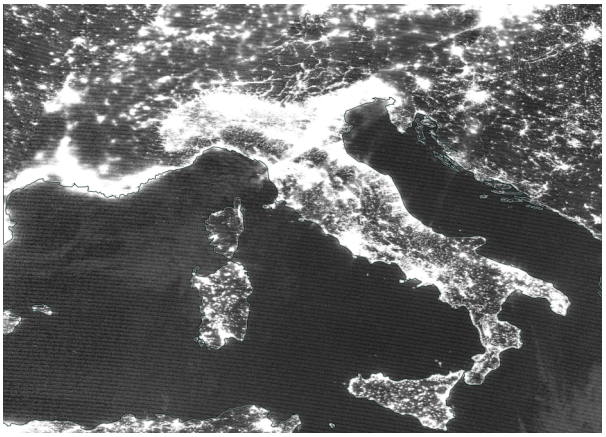


Figure 6.6: Raw NPP VIIRS night light image of Italy on march 13th 2019

Figure 6.7: Raw NPP VIIRS night light image of Italy on march 16th 2019

The figures 6.6 and 6.7 show raw images from the NPP data, which were not corrected and published as a scene. The differences in illumination over just 3 days is extreme due to cloud coverage on the 16th March 2019. But the final product is corrected and consistent only every month.

A simple way to achieve cloud free images is to apply a median filter to a stack of images of the same AOI. Clouds will have high values on any spectral band while shadows will have low digital values. The problem of this approach is that multiple recordings are collapsed to the median value of each pixel-location. It allows for consistent greater images, but the advantage of high frequency publications is lost in the process as in the NPP images from some days intervals to monthly published

scenes. (cf. Faisal et al., 2016, p. 1)

Approaches exist to combat images inconsistencies, but those come with great drawbacks. The target of remote sensing, the detailed description of the individuality of any area, is not the same as the target of social statistical analysis, which seeks to find a generalizable inferential relationship between observable variables and indicators of interest. To model the satellite information with administrative or survey data, data has to be scaled to match, a problem discussed hereafter. When satellite data is going to be aggregated at some point, the information must be consistent over the aggregate only and not over each pixel. Space for research remains on approaches for handling inter-temporal and spatial inconsistencies, which target the usability for statistical analysis regarding well-being rather than geographic precision. Sensitivity analysis for data quality and the influence on different applications will be required to make informed data and modelling decisions.

Scaling:

When satellite data are combined with other remote sensing information, or administration data, it is unlikely that each dataset is available at the same spatial level.

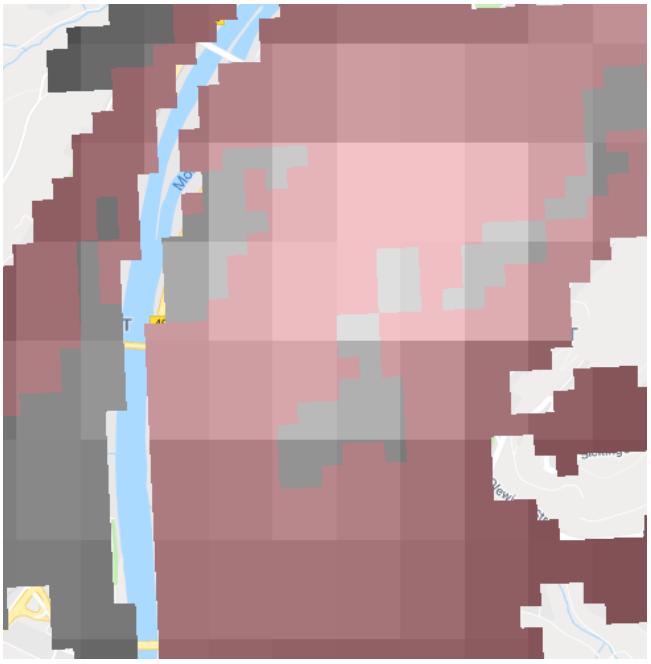


Figure 6.8: Image of own production created from 2017 NPP VIIRS-DNB data and CORINE 2012 urban and industrial land use classes

The image 6.8 demonstrates such an issue. Visible is the area of Trier. While the black and red areas show urban and industrial areas according to the CORINE dataset, the greater white to grey scale rectangles are NPP night-time light pixels. It could be interesting whether the night illumination stems from industrial areas of urban areas.

It is possible to calculate the amount of intersections between each dataset at each area, and simply reweigh the data. Further research regarding the development of better methods to topics in social statistics is still required. One approach has been tested by (Caratiola et al., 2019) under the MAKSWELL project. Using LiDAR data of the city of Trier in combination with Atkis data it was possible to determine the volume and position of all private household buildings. This information was used to reallocate the 100 times 100 meter micro census grid cells to the households in city districts rather then by surface. This allows the combination of several dataset via their location to create estimates for city districts for which no information are published elsewhere in a meaningful manner. At the same time it prevents that areas without inhabitants are assigned population counts or income.

2. Statistical Challenges

The nature of most applications, the inadequacy of official information and exploratory applications give reasons to why many statistical key questions are still unsolved. The information to identify the consequences of the approaches do not allow further research. What questions remain unanswered are discussed in the following.

Digitalization:

Satellite data are always digital information. This requires the transformation from continuous light measurements into numeric data. Similar to microphones, this transformation is automatic and depends of the equipment used. The relevant parameter is the bit value -*NPPs DNB at sensor radiance produces 16 bit images*-. sometime bit-grain of the recording (cf. Román, Wang, Shrestha, and Yao, Tian and Kalb, Virginia, Román et al., p. 12). It directly determines the integral increments in which differences of input information are transformed into digital values (cf. Taubenböck et al., 2015, p. 25).

The DMSP night-time lights image dataset has been rather extensively used for the prediction of population changes and economic development. Some example studies are Dai et al. (2017), Doll et al. (2006), or Gosh et al. (2010). The dataset is comparably small and easy to use, while a relationship between light production and population levels is feasibly explainable. A discussed drawback of the DMSP images is the conversion of total illumination into integer values from 1 to 63. (cf. Zhang and Seto, 2011, p. 2321 or Xu et al., 2014, p. 7711). This has the consequence that greater metropolitan areas are top coded. This prevents the differentiation within bigger cities and allows only the differentiation of 63 levels of light intensity.

The NPP system produces images with a resolution of 500 times 500 meters per pixel and differentiates the light intensity in a integer scale from 0 and 65534 integer values making top coding less likely and allowing the differentiation of finer differences.

Most simple methods are applied as if the satellite data is metric data. This is not the case, the smaller the range of digital values is, the more likely it is that the digital representation will cause biases in parameter and variance estimations. What are the consequences under different scenarios, and which

methods are more or less sensitive for non-metric data is a largely untouched field for research.

Evaluation:

Satellite data for the estimation of population parameters and well-being are rather new. As such, no generalized workflow has established itself in terms of best practices and methods. Direct satellite image applications often result in one of three outcomes. Either a new variable is estimated or predicted, existing information are downscaled using satellite data as auxiliary data, or parameters are nowcasted using new satellite images. Example studies are Zhang and Seto (2011), Caratiola et al. (2019), or Jean et al. (2016). In either of these cases the resulting values do not exist in any reference dataset for comparison. This was the reason for the application in the first place.

This makes the evaluation of models difficult. Even when applications developed on data of higher administrative levels are evaluated successfully, a corresponding application to other levels might not be valid.

Many statistical concepts are tested and developed in a simulation scenario. Simulations are difficult to apply to satellite image based applications as the creation of synthetic satellite images requires extensive knowledge about the interaction between parameters of interest, population structures and geographic properties, while at the same time creating images which are meaningful. This knowledge is not available, as ground research is just starting to provide such ideas.

An exception might be now-casting applications, either by waiting for official data, or by applying the methods on an older subset of the time series for which the official results are available.

To develop and test new methods, including remote sensing data it is necessary to develop methods, ideas or data environments which allow for a proper evaluation of results. Corresponding data might be available in some national statistical institutes. Several European countries have access to unit level data about several well-being relevant variables, such as the Dutch CBS with access to geo-located income information. Only in cooperation with such institutes it seems possible to properly evaluate model results and approaches on real data in the near future.

Target:

Presently there is no general concept, which allows to anticipate whether applications on one area are applicable to other areas. A pilot study might reveal that night-time light data are excellent for the prediction of economic performances of German municipalities, they might not be suitable for federal states of Germany, or municipalities in Poland. Xu et al. (2014) found that model relations are highly biased based on the ground types. The strong thermal reflection of sandy areas led to a strong bias in the NPP based model predictions for China, of up to 850% in certain areas. Further studies on determinants of biases in satellite data and model approaches are required specific for comparisons of European countries to answer the question whether there can be a unified approach to implementations of satellite data in the EU.

Outlier Detection and Statistical Quality:

A regression model based on night-time lights and municipality population data is simple formulated and was exercised exemplary here:

$$\hat{y}_{i,t} = \alpha + \beta_t TNL_{i,t} + \epsilon_{i,t} \tag{6.2}$$

where $\hat{y}_{i,t}$ is the estimated population total ($\hat{y}$) in municipality i of year t and is predicted as a linear function of the *total night lights* (TNL) in year t.

The night-time image of each year was intersected with the CORINE data to differentiate areas of agricultural use, forests, bodies of water, industrial areas and urban areas. For each of these sub areas the night light values were summed within each municipality of Germany. Each land use type was weighted by the amount of area within each use type covered of each NPP pixel.

$$\hat{y}_{i,t} = \alpha + \beta_{t,m} TNL_{i,t,m} + \epsilon_{i,t} \tag{6.3}$$

The subscript m is used for the 5 different land-use types.

The sum of night-time light intensity over each municipality area was used to model populations. Several models were tested and evaluated, but a persistent issue is outlier detection and handling.

Administrative areas are extremely different. The city of Berlin with millions of inhabitants is as much just a municipality as a small township in the Bavarian Alps is. This creates TNL datasets with high mass on low values with only some extreme high values from bigger cities.

From the point of satellite information, official statistics are aggregated arbitrarily. The way that boarders are drawn determines the outcome of estimates and resulting patterns. Using land-use types was one approach to account for uneven urban proportions in municipalities. This *modifiable area unit problem* (MAUP) was present multiple times throughout the application. This opens the area of p-value forging. When trying long enough a model or data combination will be found which indicates high p-values. While the populations predictions based on such a model shows mixed results, it became evident that given the data structure, all typical model and goodness of fit evaluations are misleading. Tests suggested an outlier problem, and robust methods weighted the models in a way that highly informative data points were taken as outliers and virtually eliminated from the fitting process. This resulted in extremely high R^2 values of up to 0.98 while also predicting highly biased population counts. Common figures of model quality seem unsuitable to most satellite generated datasets in combination with official statistics when a MAUP is apparent. This means that methods developed in survey environments might not result in expected outcomes and have to be reevaluated. To improve this groundwork, detailed geographically coded official statistics also on city and within city level are required. In Germany these are published incoherently by each city according to their

choice to do so.

Variance Estimation and Quality:

Precision and variance in the world of remote sensing has a different meaning than in social statistical applications. Even though information about the data generating process on the satellites are often listed in detail and minimum precision requirements for satellite scenes are defined for publication, these information find no respect in any well-being application found to date. The fact that satellite images are not perfectly accurate and come in different quality is not recognised as a potential variance component for modelling attempts so far. Henderson et al. (2012) proposed a weighted composition framework for quality improvement of official statistics using satellite data. Henderson et al. (2012) showed that in a combined indicator, the use of satellite images might improve bias and reduce variance in theory, no concept of compatibility of satellite parameters and survey reliability exists though. In this process only the quality of official statistics is tried to be evaluated and used for weighting the combined indices components, the quality of the satellite based results is not accounted. No variance estimation under the consideration of satellite data quality is currently conducted.

To allow respecting the quality requirements for official publication a concept of combined variance estimation has to be developed from a social statistics point of view. The creation of an international Register and the formulation of reporting conventions for satellite data will be of great assistance to such a development. Avoiding inconsistencies in reported information and reference systems would assist out-of-field users in the research of overarching variance concepts in their fields of studies.

Time Horizon:

Satellite programs have a limited lifetime. This is a natural cycle when satellites slowly drop out of their orbit and crash to earth or travel uncontrolled through space. In any case, after some time the technique behind the existing systems will become outdated. For time series models this is the absolute structural break. (compare Department of the Interior, U.S. Geological Survey (2019b) p.18 and Department of the Interior, U.S. Geological Survey (2019a))

This prohibits long term time series applications. Although satellite missions often have consecutive follow up missions, the data could become completely incomparable. Investigations of handling such breaks and possibly chaining some mission follow ups would increase the possible areas of applications.

As social statistical applications are third party users, it should not be expected that satellite missions are constructed in consideration of possible pilot applications although common bandwidth definitions across satellite missions would significantly improve the long term use.

Validity Analysis:

Using related indicators as measurement instruments to quantify constructs, which cannot be measured directly such as human behaviour, is a common approach in social science. In these cases it is crucial that the measurement instruments are valid and reliable. (Drost 2011, p. 105ff.) The term reliability describes the extent to which the measurement are repeatable and includes stability over time of a measurement, equivalence and internal consistency. Validity describes meaningfulness of research components and gives insights about whether the indicators used to measure a certain construct actually measure what is intended to be measured. (cf. Donaldson and Storygard, 2016, p. 106ff.) A special type of validity is the construct validity, which is important to test for in case indicators are used to measure a concept. Construct validity focuses on how well the concept was translated into indicators as mean for the operationalization of the measurement. (cf. Donaldson and Storygard, 2016, p. 116ff.)

To be in alignment with this research praxis, when using satellite data as indicator data for constructs of social science, reliability and especially validity analysis have to be constructed as well.

In order to test reliability test-retest can be used to test the temporal stability of the measurements. Satellite images are taken at different point in times and often information of the same area from several images are studied. In this case, a reliability test of the satellite images could be necessary and will give insights about the stability of collected information from satellite images. (cf. Donaldson and Storygard, 2016, p. 108)

An analysis of construct validity to confirm that the information on hand is actually reflecting on the construct of interest. Often this is done by performing several correlation analysis of the measure and a number of other measure which have been found to be in relation with the construct of interest. For example, in case the construct of poverty is to be measured with night-time light satellite data such as in Elvidge et al. (2009), the results on the night-time light data can be compared with well established measures for poverty such as the at-risk of poverty rate. (cf. Drew and Rosenthal, 2003, p. 609ff.)

Good practice of using satellite data to derive indicators of social science should include a discussion of reliability and validity of the indicators in order to ensure that the indicators are valuable and benefit in the measurement of the construct of interest.

6.2. Machine Learning approaches

Supervised and semi supervised learning algorithms are used for one of two purposes. Some studies use exploratory learning applications to search remote sensing data for features relevant to well-being of people. The second approach is to identify objects in the satellite image. This approach is for example used to count cars or airplanes. The problem with these approaches is that the research requires a clear idea of what objects are relevant and need to be found to contribute measurements of well-being.

The exploratory approach was used by Jean et al. (2016). Using geographically coded information of well-being, poverty and women participation from the DHS survey in an African country. The aim of Jean et al. (2016) was to demonstrate an accurate, inexpensive and scalable method for estimating consumption expenditure and asset wealth. Up-scaling household surveys to allow measurement of every Sustainable Development Goal target for every country on the world might be extremely expensive. Compared to this, the data might not be considered expensive, but the computation infrastructure behind the study would not be available to many other researchers. With access to GPS geo-referenced survey data from the Demographic Health Survey (DHS), information about health and household assets were available for measures of well-being and wealth.

The neuronal network discovered image features which seemed to be related to poverty in such a way that the best prediction performance is achieved when considering them. This way of implementing neuronal networks does not allow for greater inferential conclusions. While it was possible to mark seemingly important areas in the images it is not possible to interpret the 4096 features resulting from the network after principle component analysis. The only relevant measure to CNN is prediction precision.

A different perspective was taken by the work of Benjamin et al. (2017) in a working paper. Using high resolution images from GeoEye and DigitalGlobe, census household surceys data and local interviews, a model was trained to identify rusted roofs in a Kibera slums to identify degrees of poverty within the group of the poorest. It was found that rusted roofs allow for the identification of degrees of poverty when sufficient training data were generated.

1. Data Quality and Requirements

Both approaches require high resolution images. The following figures show the differences in spatial resolution between the Landsat 8 with 30 times 30 meters, Sentinel 2 with 10 times 10 meters and a airplane image with unspecified but highest resolution.

Figure 6.9: Resolution comparison at Frankfurt Hahn airport

Left: Aircraft image from Google Maps static API, center: real light composite using Sentinel 2 images, right: real light composite using Landsat 8 images

The Sentinel 2 images with 10 times 10 meter pixels are already in the area of high resolution public available data. Access to higher resolution earth observation data is currently only provided by commercial users.

Satellite systems such as GeoEye and DigitalGlobe do provide sub 1 meters resolution images. The drawback of commercial programs are the considerable costs. To investigate large scale applications of high resolution satellite images to the area of Germany would create initial costs into the millions, before even developing a method without any insurance that the results will be meaningful.

The ongoing development of satellite data will eventually result in high resolution images. The long timespan for development, construction, launch and data publishing creates unavoidable gaps in the development of data quality in public satellite data. To foster the development of methods for satellite images, finance options or cooperation programs for datasets on commercial satellite data would be helpful to develop methods in smaller applications in expectation of data coming in for large scale applications in the near future.

Apart from the satellite data quality, these methods require considerable computation power and storage. Most neuronal network application run much faster on Graphical Processing Units (GPU) to a degree that CPU based approaches are unreasonable to date. Statistical research centres would required new hardware and the required funding to work on these methods. (cf. Jean et al., 2016, p. 794)

2. Evaluation

The application by Jean et al. (2016) was only possible because Google Maps provided high resolution images, although the quality of these images is questionable, and because the DHS dataset provides geo-located unit level data on poverty and computation backup by NVIDIA. These were uncommon conditions. (cf. Jean et al., 2016, p. 19 f.)

Without unit level data with geolocation, the application would not have been possible. Such data are

almost never available in any European country. Georeferenced dataset are required for any satellite application. Opening more survey and administration data with precise geographic reference is the backbone of remote sensing and satellite application to well-being questions.

Many uses of remote sensing data in modern approaches, such as by Caratiola et al. (2019) would not be required if the information from agencies and institutes were available. Most information exist already, measurement and tracking of land use are core tasks in governance. In Germany, the possibility to charge money, combined with the federal structure does often not allow for a combined, coordinated distribution of valuable information. Hopefully access to further, local and geo coded data will become available with developing online data infrastructures in many agencies.

Only with high quality geographic located unit level data a true evaluation of results of estimation approaches will be possible while also giving a boost to the development of spatial applications.

3. Inference:

Exploratory applications of neuronal networks do not allow inferential analysis of the results. Object oriented methods do not provide estimations, they only create new data, which might be used in futher work. This might commonly lead to similar situations described in 6.1. Neither work will help to understand changes in poverty, but they might assist in the tracking of poverty development. In form of natural experiments, when laws change the effect on local poverty might analysed.

7. New data sources for SDG and well-being indicators

7.1. Monitoring natural disasters with mobile phone data

Monitoring natural disasters using mobile phone data is one of the possibilities offered by new data sources. These can support Disaster Risk Management (DRM) which is the application of policies and strategies to prevent new disaster risks, reduce existing disaster losses, and manage residual risks. In this context, the measures offered by official statistics can have an important role to play in the different phases of DRM; during and after disasters, in order to inform emergency response and recovery.

Since the beginning of the 1990s, the United Nations has been promoting efforts to change the paradigm of disasters, advocating for the incorporation of disaster risk reduction efforts worldwide as a way to reduce the effects of hazardous events and disasters on vulnerable communities. In 2015, United Nation facilitated the negotiations amongst Member States, experts and collaborating organizations; which led to the adoption of the Sendai Framework for Disaster Risk Reduction 2015–2030. National statistical systems are involved in these projects and could provide the basis to monitor and report on progress in achieving key goals and targets of the Sendai Frameworks of international policy for disaster risk reduction, and the Sustainable Development Goals (SDGs) of the 2030 Agenda. In 2015 UNECE promoted the constitution of the Task Force on Measuring Hazardous Events and Disasters that produced the The Task Force prepared the 'Recommendations on Measuring Hazardous Events and Disasters' which the Conference of European Statisticians endorsed in June 2019. The Recommendations were published in November 2019.

Beyond traditional competencies, official statistics are called to apply the newest techniques of social dynamic analysis based on new data sources to the field of DRM (Ferruzza et al., 2019). In this section we use mobile phone data, as new data sources, based on mobile network data to analyze the population behavior and dynamics during a flood event.

7.1.1. The case study: the flood in Livorno of September 10th, 2017

The analysis is based on anonymized CDRs of an Italian national network operator generated during a flood which append in Tuscany, Italy, in October, 2017. The flood was very intense and caused damage in the areas of Livorno and Pisa (the red circles in the picture below). In particular, especially in the city of Livorno there was 173 mm of rainfall and also 200 mm on nearby hill above Livorno. These combined rainfalls caused a great deal of damage in the city. In Pisa there was 178 mm of rainfall which caused alarm but not damage. Consider that the average rainfall of these areas is 9 mm per day. In the picture, the side graduated color bar indicate the intensity of rainfall in the region over the previous 24 hours.

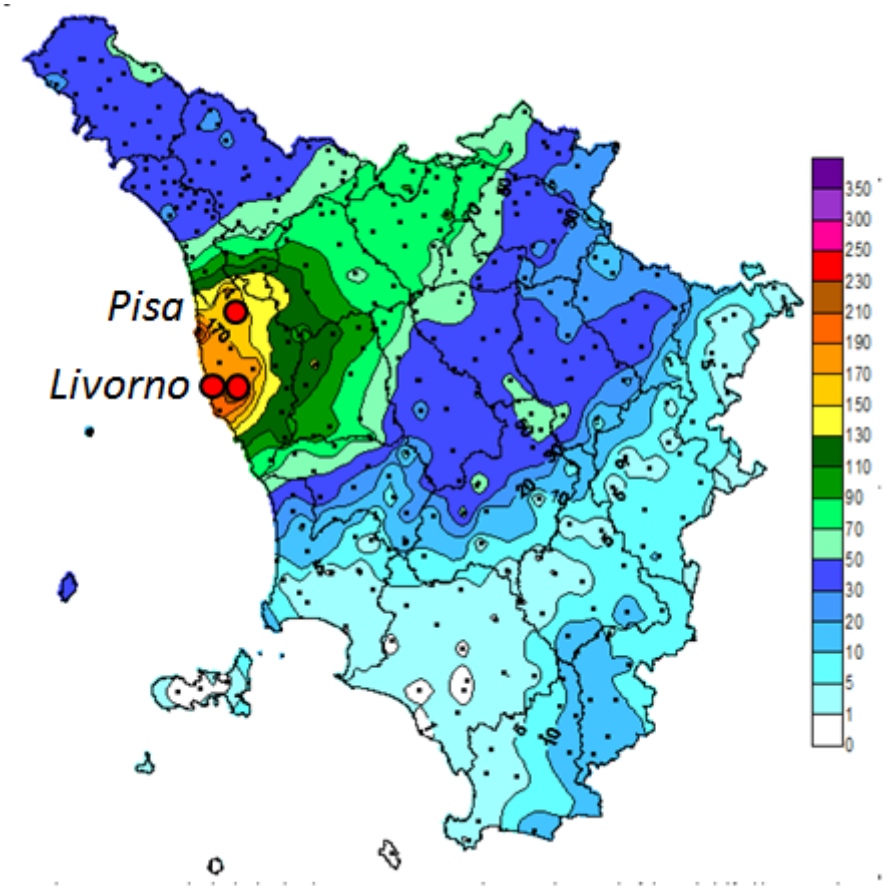


Figure 7.1: impact map of rainfall accumulated over the previous 24 hours â€“September 10th, 2017 at 8:00 am

The dataset used for this study contained only CDRs generated by the base transceiver stations (BTS) located in the provinces of Pisa and Livorno affected by the floods. Each BTS has a geographical location represented by its latitude and longitude. All the used data was not only anonymized but also aggregated. We analyzed aggregated CDRs for a period of time from 1st September 2017 to 8th October 2017, which included 140M phone calls and 1.4M IMSIs with more than one call. No personal data was collected, accessed or utilized for this study. No authors of this study participated in the extraction of the dataset.

7.1.2. Methodology: the timeline pattern changes of mobile phone calls

In general, the timeline activity of the aggregated number of calls during the same day of the week or weekend tend to have a similar pattern. Several studies have been carried out showing these regularities (Bagrow et al., 2011). These studies have also shown the differences between the weekend days, or public holydays, and working days. Of course, the similarities are between normalized patterns since there are possible differences in terms of the absolute number of calls. This depends on the location of the antennas: in the city center or in the countryside. For this reason standardized data were used and whenever necessary, in order to compare different areas analyzed, a daily scaling factor was applied.

Once a base timeline pattern of aggregated CDRs were found, we looked for pattern anomalies during

the day of the flood in order to assess the possible correlation with the event.

We initially analyzed the day timeline activity pattern in Livorno and then extended the analysis to each antenna of the provinces in the Region.

In the following map we identified a critical area in Livorno defined as the urban area included in a radius of five kilometers centered on the point where the flood was most intense (the red point on the map). In the map the BTS are identified as small green squares and the background is the census areas of the city. The circle includes all metropolitan areas in which we identified 19 BTS. In term of population, using the overlap with the census areas, we estimated a resident population of about 120.000 inside the circle.

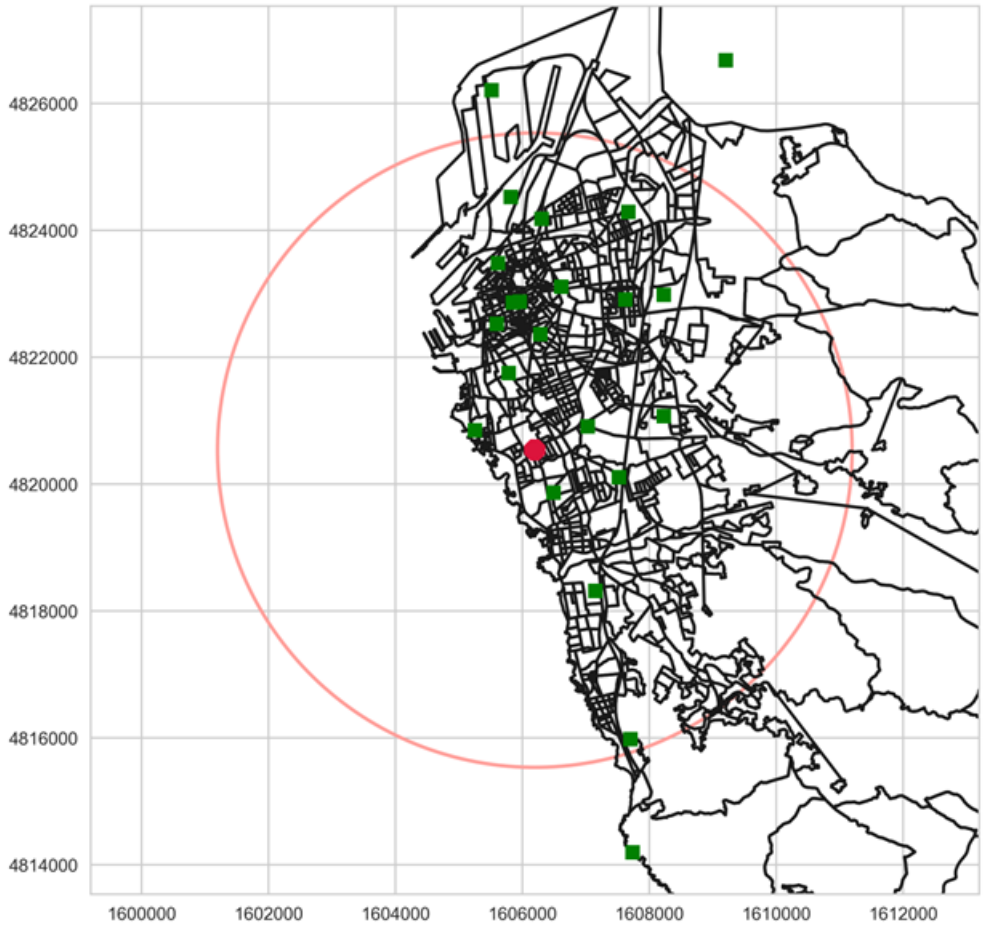


Figure 7.2: antenna identification in the Critical Area in Livorno

All CDRs elaborated by the antennas inside the critical area are grouped by hour for every weekends during the period analyzed, which includes the 10th of September when there was the flood. The sum of the CDRs frequencies are standardized in order to compare the different weekends. The plots of 48 hour weekend activities are shown in the following graph. The red relates to the weekend when the flood happened and clearly displays an anomalous behavior. The other three lines, blue, green and

violet, have the typical daily timeline pattern, without significant inhomogeneity.

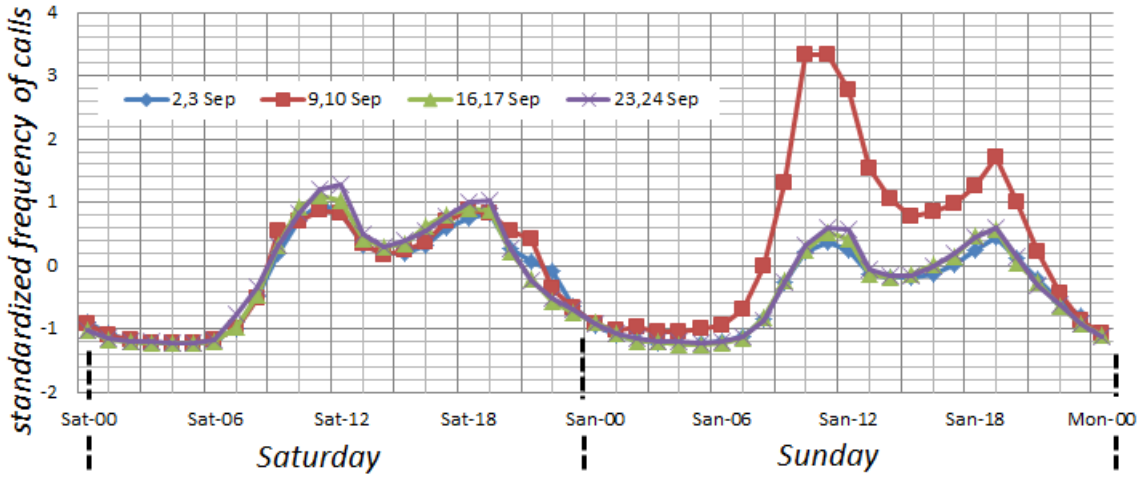


Figure 7.3: standardized frequency of grouped outgoing calls made inside the Critical Area in Livorno during different weekends–four weekends compared

It is evident that the red line starts to show a different path from the other paths already during the night of Saturday. In particular, the red line starts to vary from the regular timeline pattern at about the 3:00 am on Sunday. This can be seen as the initial signal of anxiety among the population. The differences became very relevant during the day on Sunday with a peak at midday.

If these anomalies are correlated with social protection warnings they could provide civil protection organizations with information of the location of a possible disaster, the starting time as well as an estimation of the number of the people involved. These analysis show that the network data, or better the sensor data, have the potential for detecting anomalous behavior which could provide a warning. Of course, to interpret any anomalous behaviors correctly there must be other environmental indicators such as weather warning, earthquake events or any other possible catastrophic events related to the same geographic area.

In order to extend the analysis and to identify the anomalies, we defined the residual function, or the noise, as difference between the standardized frequency value and the averaged standardized frequency value at the same hour of the day. This allowed us to estimate the average residual value and its variance for different time windows. These values were used to check anomalous behavior in the activity timeline pattern (outlier values). In this way, outlier checks could be realized in real time, with a lag of a few hours, or as post analysis as in what follows.

We also extended the analysis spatially at the provincial level for Livorno and Pisa. On the map below we show the administrative territorial outlines of the two provinces (black lines), the three red circles indicate the area where the flood was particularly intense, small green squares indicate the position of the BTSs and the red point on the BTSs indicate anomalous behavior during the flood event.

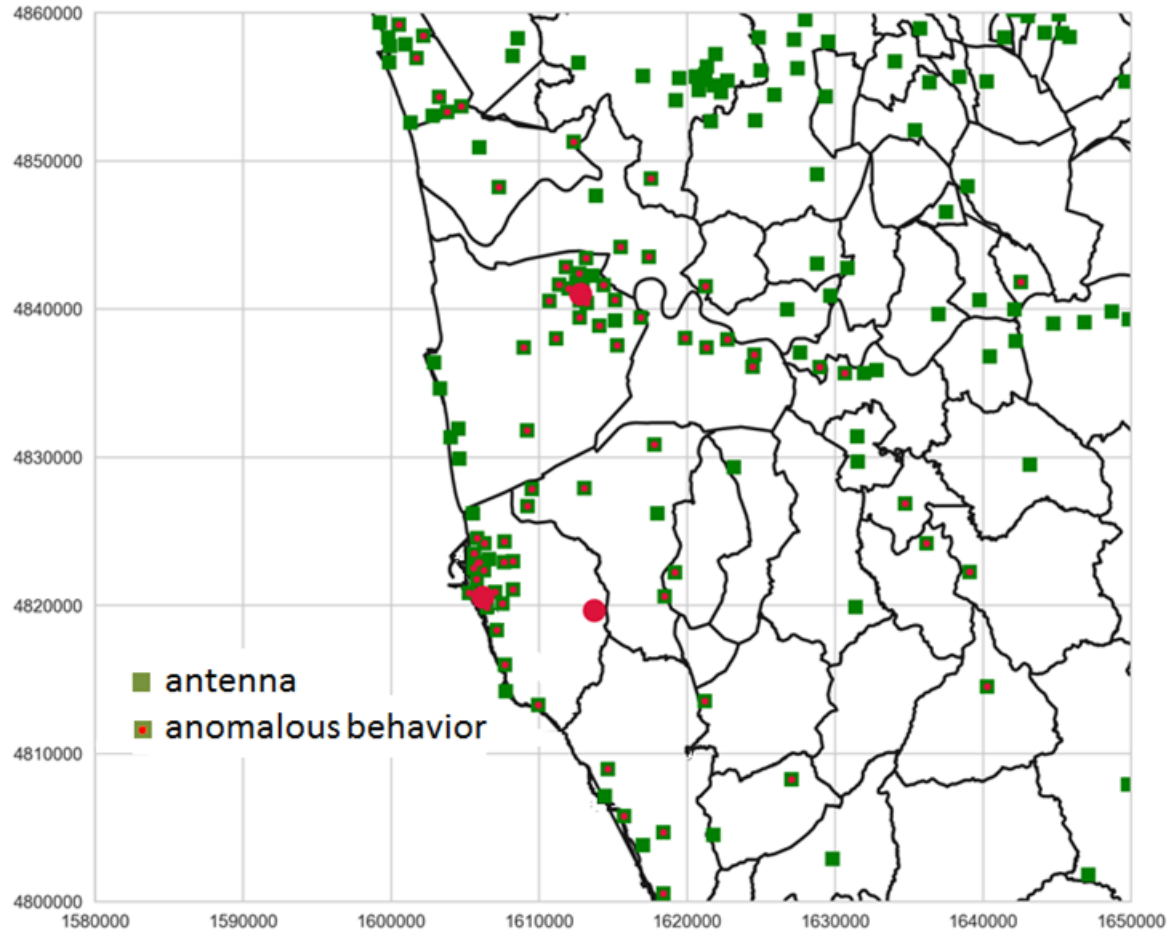


Figure 7.4: anomalous behavior of frequency patterns in each cell in the area during the event

The anomalous behavior was detected as residual outliers of the CDRs grouped in the same BTS. The results on the map show that the outliers are localized around the red circles and the critical areas nearby. This map representation make it possible to distinguish between the areas able to absorb the impact of heavy rain from the potentially critical areas. For example, near the coast at the latitude of Pisa the rainfall was very intense but no anomalous behavior from the BTSs was detected; on the contrary, in the valley to the east of Pisa there were many outliers indicating a worried population. Using this method it is possible to efficiently focus on the critical areas, estimate the population involved, assess needs and therefore allocate resources (for example, sending supplies to affected areas). This information can be used also as early warning signal in the worst affected areas so as to improve and direct public communications and safety alerts, as well as help measure the effectiveness of such early warning announcements. (Pastor-Escuredo and Morales-GuzmÃ¡n, 2014)

7.1.3. Population dynamics in the Critical area (metropolitan area of Livorno)

In order to further emphasize the benefit of using mobile phone data, the daytime population and mobility, was analysed focusing on the critical area of Livorno, before and after the flood. With this analysis we want to show the possibility of assessing the percentage of the population involved, their movements and the relaxation time necessary to return to normality, as a potential indicator of the

rate of recovery for a resilience measurement. It could also add analysis dimensions to decision makers' understanding of vulnerability and behavior helping them to combine analysis with crowdsourced data from disaster–affected communities (for example, by conducting phone surveys via SMS).

The population dynamics was then analysed, evaluating the incident population and the commuters inside and outside the critical area before and after the flood. We applied the proposed Bayesian method to estimate the probability that a mobile phone is present, or not, in the critical area with a prior probability based on the BTS densities function. This approach was chosen because we were comparing two large geographical areas: the metropolitan area, with a high density of antennas, and a much larger area in the countryside with a lower density of antennas. Moreover, to reduce the error impact estimation and to simplify the analysis we modeled the daily human mobility to a two–node model: place of leaving (home–place) and place of working or studying (work–place).

The two–nodes model is based on the assumption that people are mostly at home during the nighttime and are at work, or at school, during daytime. With this aim, we used CDRs aggregated by a spatial–temporal array for each IMSI. The spatial dimension identifies two areas: inside or outside the critical area. The temporal dimension is articulated in four levels of values: before and after the flood, both of these articulated in nighttime (20:00 pm to 8:00 am) and working time (8:00 am to 20:00 pm). With this two–node model it is assumed that the home–place can be identified through the most frequent event during the nighttime and the work–place as the most frequent event during the working time.

The population dynamics results are shown in the following graph: the blue bars represents the percentage of population living inside the critical area before the flood and the red bars are the percentage of population living outside the critical area before the flood. It is evident that the blue bars show a higher percentage of home or work changes compared with the red bars.

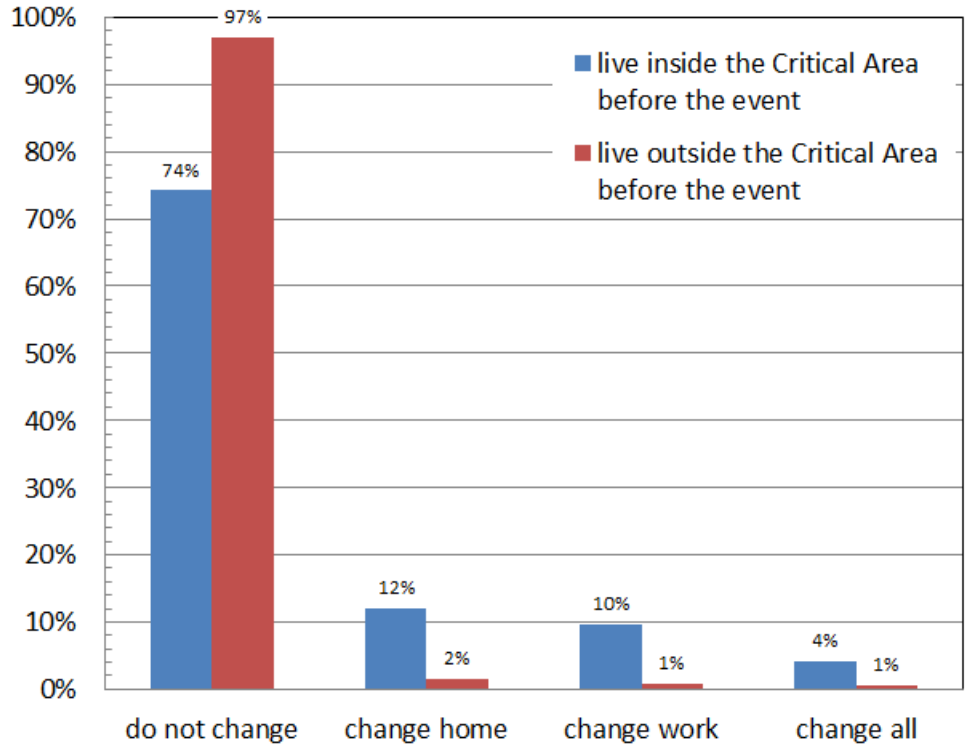


Figure 7.5: percentage of estimated population flow, inside and outside the Critical Area (metropolitan area of Livorno), before and after the event.

Using a longer observation time, this analysis could also be applied to study the time necessary to return to normality, which would correspond to similar change rates between the two observed areas.

While a more extensive analysis is required, these results suggest the high potentiality in using mobile phone activity information to improve early warning signals and emergency management. These results also underline the value of a public–private partnership in using new data sources to indicate flooding impacts accurately.

7.2. Remote sensing

The United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) implemented a global framework that addresses stakeholders in the individual countries and promotes the integration of remote sensing data into national reporting. One good example to integrate remote sensing data is the reporting on the SDGs which are located at global level, since the monitoring of the indicators requires a database that has full coverage and is easily accessible. In addition to statistical data, which are already systematically and regularly collected in many countries, new data sources and methods of data evaluation are gaining in importance. Satellite data have the advantage that they are available on a global scale; the ground-based monitoring of various parameters would be neither technically nor financially practicable. Data sets from earth observation can provide reliable information on a great range of different topics. The advantages of Earth observation are already being used to monitor the atmosphere, oceans, snow and ice, natural resources, infrastructure, forests and water bodies. Besides the comprehensive coverage, remote sensing data have the advantage

of a high temporal resolution. This allows changes over time to be captured in even shorter intervals. ZKI-DE (ZKI-DE)

New data sources for SDG and well-being indicators:

7.2.1. SDG 11.7.1

The indicator 11.7.1 describes the "average share of the built-up area of cities that is open space for public use for all, by sex, age and persons with disabilities". Based on the tier classification of the Inter-agency and Expert Group (IAEG) on SGD indicators 11.7.1 is classified as an tier III indicator meaning that "no internationally established methodology or standards are yet available for the indicator". In the following, different handlings about the data to report on SDG 11.7.1 are described.

Definitions

Regarding the UN SDG indicator, some definitions are agreen upon: cities are defined by their urban extent. This relies on the analysis of satellite imagery to define the city boundaries. Only cities with more than 100 000 inhabitants are included, as this indicator is defined for urban areas.

Public space is definced as all places that are publicly owned or of public use, accesible and enjoyable by all, free and without a profit motive. This definition also includes streets. In Germany the Authoritative Real Estate Cadastre Information System (ALKIS) is used to categorize the land use which is then used for the calculation of the indicator.

EU SDG

Since UN Indicators are selected for reporting on a global level they are not always relevant for the EU. The EU SDG indicator set is the basis for Eurostat,s annual monitoring report, which evaluates the SDGs in an EU context. The EU SDG indicator set is aligned as far as appropriate with the UN list of global indicators. The goal 11 defines a sub goal 'Share of urban population without green urban areas in their neighbourhood' which is currently kept on hold for further consideration in future reviews. This definition additionally adds the constraint of green space for which remote sensing can be used.

7.2.2. Analysis with New Digital Data

Research Question

In this analysis, Destatis investigates the combination of different sources of New Digital Data, that is mobile network data and remote sensing imagery. The starting point for the analysis is the sustainability indicator 11.7.1, which describes access to public places. The corresponding EU SDG indicator differentiates further and only considers green space. Based on this the goal of the analysis at hand is to investigate the access to public green space in urban areas. However, these indicators only consider the place of residence and not the whereabouts that inhabitants have throughout the day. Mobile network data allows to analyse the sorroundings of population thoughout the day - that is the access to green spaces while the residents are at work, running errands etc. Our research question is thus, does access to public green spaces differ by time of day or week? The basic idea of the study is to

use mobile phone signal data to illustrate the differences in access to urban green spaces in the course of the day and the week. Mobile network data can be used as a proxy for the population density at different times of the day. Geodata can provide information about the percentage of green space in the corresponding cell.

The feasibility study is limited to the cities with more than 100 000 inhabitants in the federal state of North Rhine-Westphalia (NRW). Due to its high city density NRW is particularly suitable for investigations in urban areas.

Data and Methods

The mobile network data from T-Systems was used as a proxy for population density at different times of day and week. This data is aggregated to grid cells between 0.5 km and 8 km. The number of mobile phone signals staying longer than 30 minutes within a cell is counted. To ensure privacy and generalizability, the data only contains averaged values for one month for periods of one hour. The size of the grid cells is based on the population density to ensure that at least 30 signals are located within a cell, and thus the anonymity of the users is secured. As a result, grids in rural areas have a larger size than in inner city areas. Due to data protection regulations, the mobile phone activities are only available as an aggregated data set, averaged by month. The mobile activities include the average activity for each hour, by day of week (Tuesday to Thursday is aggregated).

ALKIS which is used for the calculation of the national indicator does not allow to determine a difference between public areas and green public areas. For the determination of the urban green area the Urban Atlas from 2012 was used, which is based on remote sensing data. The Urban Atlas provides comparable land use and land cover data for urban areas. In this analysis the urban green areas were used. This includes the vegetation areas that are planted by humans and maintained regularly. These areas are mainly for recreational purposes (e.g. gardens, parks and zoos). Forests or green spaces are mapped as urban green spaces if at least two sides of urban areas and structures are limited and traces of recreational use are discernible. The data is published by the European Environment Agency (EEA) and is subject to the principle of full, open and free access. The share of green space for each grid cell was determined based on the Urban Altas. This was weighted with the number of signals within the cell for all available intervals. This led to a distribution of 'greenness' for every interval, and changing for every interval because of the changing mobile phone signals.

Results

Figure 7.6 illustrates the distribution of access to green spaces by weekday. Each graph illustrates the distribution for different hours of the day. There are only small discrepancies in the distribution of access to public green spaces. This would allow the assumption that the access to urban green does not differ by time of day or time of week.

Discussion

The grid cells are of unequal size, to ensure privacy. The sizes depend on the population density and are thus smaller in the center and larger at the outskirts of a city. For each share the proportion is calculated. This means that the same share can have different implications on the quality of life that

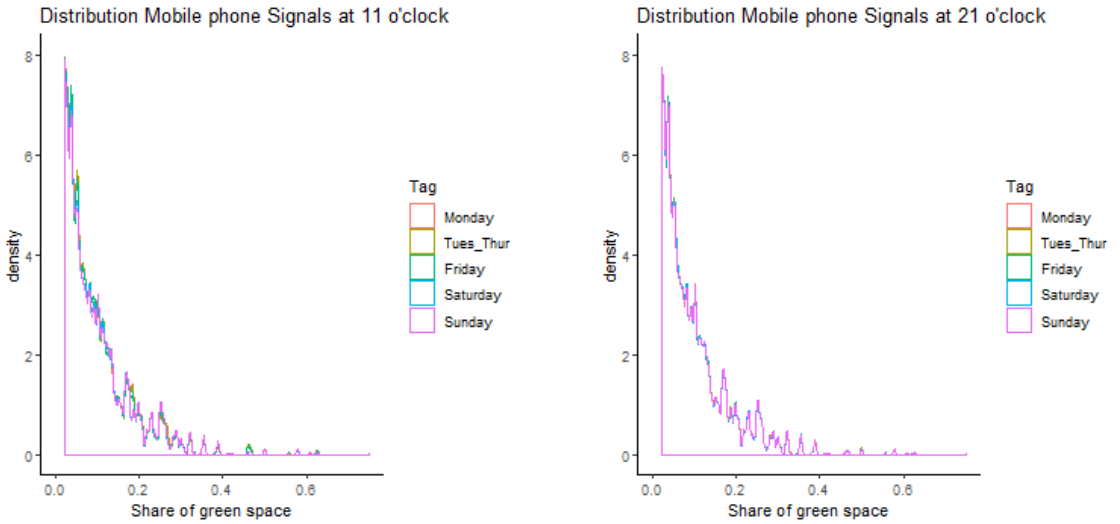


Figure 7.6: Distribution of access to urban green by hour

people experience with regards to accessibility of urban green. A potential difference in distribution might be disguised because of the large grid cells in less densly populated areas. With a finer grid it might be possible to capture a difference, however privacy protection laws do not allow for it at the moment. The mobile phone data used is only from one provider and thus only represents a sample of the population. This subpopulation is not representative due to different consumer choices.

The cities were defined by their administrative boarders, which can lead to some inconsistencies between cities. Furthermore only the mobile network signals in urban areas were considered. This means that people living outside the cities were considered for the periods they spent in cities and inhabitants were not considered for the time they spent outside of the cities. Thus the population within the analysis varies between periods.

For the calculation of the indicator 11.7.1. ALKIS data is used. However, the goal of this analysis is to extend the indicator by adding other meaningful data sources. The urban atlas provides a suitable data source to differentiate between the areas of open public access. This way not only areas that are open to public use but the additional constraint of green areas is added. Using this new digital data source allows for analysis which deviates from the indicator but potentially adds value to discussions about environmental justice. However, the results are not comparable to the indicator.

8. Discussion

For probability samples there is a well-established frame work for quantifying sampling and non-sampling errors. Sampling errors can be quantified in a relative straightforward manner through variance estimation. For non-sampling errors there is an extended frame work to quantify the total survey error. Currently there is no well-developed theory to assess the quality of statistical information derived from non-probability data sources. Therefore the total survey error frame work needs extensions to situations where non-probability data, big data or non-traditional data sources are used to compile statistical information. First proposals are available in the literature under the name Total Error Framework to emphasize the use of alternative data not obtained through survey sampling. This frame work must account for the different aspects of these new data sources. In particular error components through processing fussy volatile big data sources (identifying units of interest, linkage errors, errors due to extracting information using machine learning and AI algorithms) require additional research.

Two approaches to use new data sources in the production of official statistics about SDG and well-being indicators are distinguished. The first approach is to use new data sources as the primary or direct data source to construct official statistics. In this case methods are required that account and correct for selectivity in these data sources. Different methods that correct for selection bias in non-probability samples can be found in the literature. These methods, however, assume that the non-probability data source is structured in a sense that there is a link between the records in the data source and the units of the target population. In many situations this is not the case. On top of that, correction methods, assume the availability of sufficient good auxiliary variables that explain the data generating process of the big data source. In many practical situations this is not the case. Therefore there is much need for methods that handles these issues in a proper way. One can think of feature extraction to derive auxiliary information from big data sources, unit identification methods to identify the correct units in big data sources and matching techniques, to match these records with other sources. These kind of techniques introduce of all kind of errors in the auxiliary variables. Research into which extend this reduces the effectivity of these methods to correct for selection bias is also required.

A second approach is to combine survey data with new, non-traditional data sources by using the latter as auxiliary variables in model-based inference methods like small area estimation methods and time series models for now casting. It is recognized that more empirical research is required to compare to which extend formal small area estimation methods are superior to machine learning algorithms that are applied in the literature as an alternative to extract information concerning poverty from satellite and aerial images. Concerning time series modelling, it is recognized that more insight in methods that select relevant auxiliary series from a large amount of potential series without falling into the trap of data dredging is required. Finally there is a strong need for methods that allow for time varying correlations between target series and auxiliary series.

Another point of further research is to obtain more empirical evidence of the usefulness of using non-traditional data sources. Statistical institutes generally have, for good reasons, a low risk appetite. For this reason national statistical institutes prefer to base their official publications on data obtained with probability samples in combination with design-based inference methods. Moving towards the use of non-traditional data sources in the production of official statistics requires the acceptance of moving towards the use model-based inference methods, either to correct for selection bias if new data sources are used as primary data, or for small area estimation methods and now-casting methods where non-traditional data sources are used as covariates to obtain more detailed and timely estimates from survey samples. More empirical research on the use of these methods in the context of official statistics is required to illustrate the benefits of these data sources and inference methods for national statistical institutes.

Remote sensing data are successfully used in situations where no reliable official data are available, like developing countries, combat areas or countries with unstable political systems. Improving measurement of well-being in the European states is a different situation, since high quality official data are already available. To further improve regional detail, precision and timeliness of these data, remote sensing data must meet higher quality requirements. It is observed the integration of remotely sensed data in the frame-work of SDGs and well-being measurement is predominantly experimental. There is need for a comprehensive overview of datasets and methods to facilitate their use in official statistics. Two approaches for the use remote sensed data to measure SDGs and well-being are identified. The first one is to derive information from images and relate that in a model with target indicators. The second approach is to extract covariates from satellite and aerial images that are expected to be correlated with the target variables of interest.

Several issues that require further research for methods are identified. For statistical purposes, interest is focused on large areas. This implies that multiple images have to be combined. Methods to avoid or reduce image inconsistency are needed as well as insight into the effects of image inconsistency on predictions of well-being and SDGs. In a similar way, inconsistencies between images of the same area over time due to different atmospheric conditions might obscure or distort temporal analysis and predictions for period-to-period change. Further research to handle inter-temporal and inter-spatial inconsistencies on predictions for poverty, well-being and SDGs and methods to reduce these error sources is needed. There is also need for research how these sources of image inconsistencies effect variance and precision of predictions for well-being, poverty and SDGs. Another type of research is to validate the reliability of constructs for poverty and well-being that are derived from remote sensed data.

Methods for processing satellite and aerial images for large areas for longer periods require considerable computational power and data storage capacity. The availability of the required hardware as well as the required knowledge to handle these data on appropriate AI machines is not standard available at national statistical institutes.

Another issue is that satellite images of sufficient quality or resolution are only commercially available. To evaluate methods, geo-coded unit level data on income and poverty are required but are currently

hardly available, due to confidentiality restrictions. To facilitate wider development of satellite-based applications an Europe wide computing infrastructure should be established. This could ensure data safety and handle problems with computational power and data storage capacity and make applications independent form commercial entities.

Deep learning is an important modern AI tool to extract information from satellite and aerial images. In Deliverable 2.1 an example of deep learning to count solar panels from aerial images is described. For successful application of these techniques in the production of official statistics, several direction of further research are identified. Better understanding of different error sources is required to evaluate the uncertainty of results obtained with deep learning and to optimize the precision of predictions obtained with deep learning. This concerns e.g. further research how to create training sets, test sets and validation sets to maximize model generalizability to intended target populations and avoid model bias. Another points for further research are methods that minimize the effect errors in annotating the data, the impact of class imbalance on uncertainty measures and methods to quantify estimation uncertainty. Finally model interpretability and transparency are aspects that are hardly explored and requires further research. Deep learning algorithms are complex black boxes. To improve model generalizability, some understanding and insights which feature are important for classification and prediction is necessary.

Finally, it is important to address the question on how the new methodologies and sources of data presented along this deliverable could be suitable for policy dimension.

A first answer to this question is related to the expected gains in timeliness of the indicators. The examples provided are important, among the others, concerning labour market, prices and regional disparities, energy and natural disaster. This implies that the results of the project could improve the quality and timeliness of the indicators included in the Macro Imbalance Procedure, that is the reference framework for policy evaluation used by European Commission (see also Bacchini et al. (2020) and deliverable 4.2).

A second important results stemming from the project is on how macro models could shed lights on well-being and sustainability. From this prespective deliverable 2.2 (van den Brakel et al. (2019) has presented an application to Italy of the so called I-S-O framework (Input-State-Output) that provides an integrated picture for Sustainable development that can be viewed as a process of 'interaction among three elements: the biological and resource system, the economic system, and the social system' (Barbier (1987)). Meanwhile deliverable 5.2 is facing the issue on how inequality and energy could be addressed in a macro-econometric model interacting directly with the other traditional macroeconomic aggregates (Bacchini et al. (2015)). A case study related to Italy and Hungary complete the analysis of the interaction amid well-being and SDG indicators and policy targets (Deliverable 5.3).

Bibliography

Amaya, A., P. Biemer, and D. Kinyon (2020). Total error in a big data world: adapting the tse framework to big data. *Journal of Survey Statistics and Methodology 8*, 89–119.

Bacchini, F., R. Golinelli, and C. Jona-Lasinio (2015). An Energy-Environment-Macro model for the Italian economy:2E- MeMo-It. *Working papers* (4).

Bacchini, F., R. Ruggeri-Cannata, and E. Doná (2020). Evaluating economic and social convergences across european countries: Could macroeconomic imbalance procedure indicators shed some light? *Statistical Journal of the IAOS* (Preprint), 1–11.

Bagrow, J. P., D. Wang, and A. L. Barabasi (2011). Collective response of human populations to large-scale emergencies. *PloS One 6(3)*.

Barbier, E. B. (1987). Stiglitz development. *Environmental conservation 14*(2), 101–110.

Ben-David, S., J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan (2010, May). A theory of learning from different domains. *Machine Learning 79*(1-2), 151–175.

Ben-David, S., J. Blitzer, K. Crammer, and F. Pereira (2006). Analysis of Representations for Domain Adaptation. *Advances in neural information processing systems.*, 8.

Benjamin, M., S. Thomas, and T. Suri (2017). There is No Free House: Ethnic Patrinage in a Kenyan Slum. *Working Paper MIT*.

Biemer, P., E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. Tucker, and B. West (Eds.) (2017). *Total Survey Error in Practice*, Hoboken. John Wiley & Sons.

Buda, M., A. Maki, and M. A. Mazurowski (2018, October). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks 106*, 249–259. arXiv: 1710.05381.

Buelens, B., J. Burger, and J. A. van den Brakel (2018). Comparing inference methods for non-probability samples. *International Statistical Review 86*(2), 322–343.

Burgess, R., F. Consta, and B. Olken (2012). The Political Economy of Deforestation in the Tropics. *Quarterly Journal of Economics 127*(4), 1707–1754.

Caratiola, C., H. Dieckmann, R. Münnich, M. Gerhards, and T. Udelhoven (2019). Measuring well-being and poverty at local level using remote sensing data. Presented at the ITACOSM conference; Florence.

Chambers, R. (2009). Regression analysis of probability-linked data. *Official Statistics Research Series 4*.

Chambers, R. and G. Kim (2015). *Secondary analysis of linked data*. Wiley Online Library.

Chen, J., R. Valliant, and M. Elliott (2018, 12). Calibrating non-probability surveys to estimated control totals using lasso, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 68*.

Chipperfield, J. O. and R. L. Chambers (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics 31*(3), 397–414.

Connelly, R., C. Playford, V. Gayle, and C. Dibben (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research 59*, 1–12.

Daas, P., J. Burger, Q. Le, O. Bosch, and P. M. (2016). *Profiling of Twitter users: a big data selectivity study*. Statistics Netherlands Discussion Paper.

Dai, Z., Y. Hu, and G. Zhao (2017). The Sustainability of Different Nighttime Light Data for GDP Estimation at Different Spatial Scales and Regional Levels. *Sustainability 9*.

Delden, A. v., B. Lorenc, P. Struijs, and L.-C. Zhang (2018). Letter to the editor: On statistical unit errors in business statistics. *Journal of Official Statistics 34*, 573–580.

Department of the Interior, U.S. Geological Survey (2018). Landsat 8 (l8) Data Users Handbook.

Department of the Interior, U.S. Geological Survey (2019a). Landsat 7 (L7) Data Users Handbook.

Department of the Interior, U.S. Geological Survey (2019b). Landsat 8 (L8) Data Users Handbook Version 5.

Di Consiglio, L. and T. Tuoto (2018). When adjusting for the bias due to linkage errors: a sensitivity analysis. *Statistical Journal of the IAOS 34*(4), 589–597.

Ding, Y. and S. Fienberg (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology 20*, 149–158.

Doll, C., J. P. Muller, and J. G. Morley (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecol Econ 57*, 75–92.

Donaldson, D. and A. Storygard (2016). The View from Above: Applications of Satllite Data in Economics. *Journal of Economic Perspectives 30*(4), 171–198.

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics 164*, 188–205.

Drew, W. and R. Rosenthal (2003). Quantifying Construct Validity: Two Simple Measures. *Journal of Personality and Social Psychology 84*(3), 608–618.

Elliott, M. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* (2(6)), 1–6.

Elliott, M. R. and R. Valliant (2017). Inference for nonprobability samples. *Statistical Science 32*(2), 249–264.

Elsahar, H. and M. Galle (2019). To Annotate or Not? Predicting Performance Drop under Domain Shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 2163–2173. Association for Computational Linguistics.

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.

Eurostat (2017). Toward an harmonised methodology for statistical indicators - part 3 relevance of indicators for policy making. Technical report, Eurostat, Manuals and guidelines.

Faisal, K., A. Shaker, and S. Habbani (2016). Modelling the Relationship between the Gross Domestic Product and Built-Up Area Using Remote Sensing and GIS Data: A Case Study of Seven Major Cities in Canada. *International Journal of Geo-Information 5*(23).

Feder, M. and D. Pfeffermann (2015). Statistical inference under non-ignorable sampling and non-response. an empirical likelihood approach. *Technical Report, University of Southampton.*.

Fellegi, I. P. and A. B. Sunter (1969). A theory for record linkage. *Journal of the American Statistical Association 64*(328), 1183–1210.

Ferruzza, A., A. L. Palma, and G. Tagliacozzo (2019). Mobile phone data to support disaster risk management. *A.I.S.Re XL ANNUAL SCIENTIFIC CONFERENCE, September 16-18, 2019 L'Aquila*.

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Gal, Y. (2016). *Uncertainty in Deep Learning*. Ph. D. thesis.

Gal, Y. and Z. Ghahramani (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. pp. 10.

Geifman, Y. and R. El-Yaniv (2017). *Selective Classification for Deep Neural Networks*, pp. 4878–4887. Curran Associates, Inc.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Gosh, Tilottama, Sutton, Paul, Elvidge, Christopher, Powell, Rebecca, Baughm, and Kimberly (2010). Shedding Light on the Global Distribution of Economic Activity. *The Open Geograpohy Journal 3*(1).

Groves, R. and L. Lyberg (2010). Total survey error: past, present, and future. *Public Opinion Quarterly 74*, 849–879.

Han, Y. and P. Lahiri (2019). Statistical analysis with linked data. *International Statistical Review 87*, 139–157.

Heckman, J. (1990). Varieties of selection bias. *The American Economic Review 80*(2), 313.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pp. 475–492. NBER.

Henderson, V., A. Storeygard, and D. Weil (2012). Measuring Economic Growth from Outer Space. *American Economic Review 102*(2), 994–1028.

Hendrycks, D. and K. Gimpel (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv: 1610.02136.

Hendrycks, D., M. Mazeika, D. Wilson, and K. Gimpel (2018). *Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise*, pp. 10456–10465. Curran Associates, Inc.

Hendrycks, D., K. Zhao, S. Basart, J. Steinhardt, and D. Song (2019). Natural adversarial examples. *arXiv preprint arXiv:1907.07174*.

Jean, N., M. Burke, M. Xie, M. Davis, D. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *SCIENCE 353*(6301), 790–794.

Katharopoulos, A. and F. Fleuret (2018, 03). Not all samples are created equal: Deep learning with importance sampling. In *ICML*.

Kendall, A. and Y. Gal (2017). *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?*, pp. 5574–5584. Curran Associates, Inc.

Kenett, R. and G. Shmueli (2014). On information quality. *Journal of the Royal Statistical Society: Series A 177*, 3–38.

Kim, G. and R. Chambers (2012a). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis 56*, 2756–2770.

Kim, G. and R. Chambers (2012b). Regression analysis under probabilistic multi linkage. *Statistica Neerlandica 66*, 64–79.

Kott, P. S. and T. Chang (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association 105*(491), 1265–1275.

Lahiri, P. and M. D. Larsen (2005). Regression analysis with linked data. *Journal of the American statistical association 100*(469), 222–230.

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. pp. 12.

Lee, S. (2016). International Isolation and Regionial Inequality: Evidence from Santions on North Korea. *Stanford Working Papters No. 575*.

Li, Y., J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li (2017, Oct). Learning from noisy labels with distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1928–1936. IEEE.

Lipton, Z. C. (2018, June). The Mythos of Model Interpretability. *Queue 16*(3), 30:31–30:57.

Loquercio, A., M. Segu, and D. Scaramuzza (2020). A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*.

Luque, A., A. Carrasco, A. MartÃn, and A. de las Heras (2019, July). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition 91*, 216–231.

Manski, C. (1989). Anatomy of the selection problem. *The Journal of Human Resources 96*(24-3), 343–360.

Mansour, Y., M. Mohri, and A. Rostamizadeh (2009). Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory*.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimators using Big Data sources. *Journal of Official Statistics 31*, 263–281.

Matei, A. (2018). On some reweighting schemes for nonignorable unit nonresponse. *The Survey Statistician 77*, 21–33.

Meng, X. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics 12*, 685–726.

Monmonier, M. (2002). Area Photography at the Agricultural Adjustment Administration: Acreage Controls, Conservation Benefits, and Overhead Surveilance in the 1930s. *Photogrammetric Engineering and Remote Sensing 76*(11), 1257–1261.

Montavon, G., S. Bach, A. Binder, W. Samek, and K.-R. Müller (2017, May). Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition 65*, 211–222. arXiv: 1512.02479.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society 97*, 558–625.

NOAA, National Centers for Environmental Information (2014). Global Ozone Monitoring Experiment 2 (GOME-2) Products from METOP-A.

Osband, I. (2016). Risk versus Uncertainty in Deep Learning: Bayes, Bootstrap and the Dangers of Dropout.

Pan, S. J. and Q. Yang (2010, October). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering 22*(10), 1345–1359.

Pastor-Escuredo, D. and A. Morales-GuzmÃ¡n (2014). Flooding through the lens of mobile phone activity. *IEEE Global Humanitarian Technology Conference*.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: A unified approach based on s. n. roy memorial lecture in the symposium. *Calcutta Statistical Association Bulletin 69*(1), 35–63.

Pfeffermann, D. and M. Sverchkov (2009). Inference under informative sampling. *Handbook of Statistics 29 Part B*, 455–487.

Powell, B. and P. Smith (2020). Computing expectations and marginal likelihoods for permutations. *Computational Statistics (in press)*.

Puza, B. and T. ONeill (2006). Selection bias in binary data from voluntary surveys. *Mathematical Scientist 31*, 85–94.

Rivers, D. and D. Bailey (2009). Inference from matched samples in the 2008 us national elections. *Proceedings of the joint statistical meetings 1*, 627–39.

Román, M. O., Z. Wang, R. Shrestha, and Yao, Tian and Kalb, Virginia. Black Marbel User Guide Version 1.0.

Rondinella, T. and altri (2019). Future research needs in terms of statistical methodologies and new data. Technical report, MAKSWELL project deliverable 5.1.

Rosenbaum, P. and D. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*(387), 516–524.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV) 115*(3), 211–252.

Samart, K. and R. Chambers (2014). Linear regression with nested errors using probability-linked data. *Australian & New Zealand Journal of Statistics 56*(1), 27–46.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York, NY: Springer.

Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika 74*(2), 385–391.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra (2016, October). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv:1610.02391 [cs]*. arXiv: 1610.02391.

Smith, P., A. Berrington, and P. Smith (2019). Administrative vs survey data for longitudinal analyses. Technical report, ESRC Commissioned Report, University of Southampton.

Smith, P. and P. Weir (2006). Characterisation of quality in sample surveys using principal components analysis. pp. 86–93.

Stefanski, L. A. and R. J. Carroll (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, 1335–1351.

Stiglitz, J., A. Sen, and J.-P. Fitoussi (2009). Report by the Commission on the Measurement of Economix Performance and Social Progress.

Stiglitz, J. E., J.-P. Fitoussi, and M. Durand (2018). Beyond gdp: Measuring what counts for economic and social performance.

Sverchkov, M. and D. Pfeffermann (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology 30*, 79–82.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus (2014). Intriguing properties of neural networks. In Y. Bengio and Y. LeCun (Eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tam, S. and J. Kim (2018). Big data ethics and selection-bias: An official statistician's perspective. *Journal of IAOS 34*, 577–588.

Taubenböck, H., T. Wurm, T. and Esch, and S. Dech (2015). *Globale Urbanisierung*. Berlin Heidelberg: Springer-Verlag.

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, Deliverable 1.1. Deliverable `https://www.makswell.eu/attached_documents/output_deliverables/deliverable_1.1.pdf`, Eurostat.

Tinto, A. and B. Baldazzi (2018). Definition of the existing database on Beyond GDP initiatives within official statistics, MAKSWELL, WP1, Delivarable 1.2. Deliverable `https://www.makswell.eu/attached_documents/output_deliverables/deliverable_1.2.pdf`, Eurostat.

Tsipras, D., S. Santurkar, L. Engstrom, A. Turner, and A. Madry (2019). Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Tuoto, T. (2016). New proposal for linkage error estimation. *Statistical Journal of the IAOS 32*(3), 413–420.

Tuoto, T., D. Fusco, and L. Di Consiglio (2018). Exploring solutions for linking big data in official statistics. *Studies in Theoretical and Applied Statistics*.

Umwelt Bundesamt (2019). CORINE Land Cover - CLC.

Unangst, J., A. Amaya, H. Sanders, J. Howard, A. Ferrell, S. Karon, and J. Dever (2020). A process for decomposing total survey error in probability and nonprobability surveys: A case study comparing health statistics in us internet panels. *Journal of Survey Statistics and Methodology 8*, 62–88.

van Delden, A., D. Windmeijer, and O. ten Bosch (2019). Searching for business websites.

van den Brakel, J., B. Buelens, R. Curier, P. Daas, Y. G. Gootzen, T. de Jong, M. Puts, M. Tennekes, R. Willems, A. Brunetti, S. Fatello, F. Polidoro, A. Simone, A. Ferruza, A. Palma, G. Tagliacozzo, N. Rosinski, K. Wichmann, T. Zimmermann, F. Ertz, R. Münnich, and L. Güdemann (2019). Aspects of existing databases, traditional and non-traditional data sources and collection of good practices, MAKSWELL, WP2, Delivarable 2.1. Deliverable `https://www.makswell.eu/attached_documents/output_deliverables/deliverable_2.1.pdf`, Eurostat.

van den Brakel, J., C. Schiavoni, N. Tzavidis, R. Iannaccone, D. Zurlo, F. Bacchini, I. Benedetti, and T. Laureti (2019). Report on the use of time series models for sdgs and well-being indicatros, MAKSWELL, WP4, Delivarable 4.1. Deliverable `https://www.makswell.eu/attached_documents/output_deliverables/deliverable_4.1.pdf`, Eurostat.

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. Deliverable `https://www.makswell.eu/attached_documents/output_deliverables/deliverable_2.2.pdf`, Eurostat.

Wagner, J., R. Münnich, J. Hill, and Stoffels, Johannes and Udelhoven, Thomas (2017). Nonparametric Small Area Models using Shape-Constrained Penalized B-Splines. *Journal of the Royal Statistical Sciety A 2017*.

Wang, M. and W. Deng (2018, October). Deep visual domain adaptation: A survey. *Neurocomputing 312*, 135–153.

Wang, R., J. Camilo, L. M. Collins, K. Bradbury, and J. M. Malof (2017, Oct). The poor generalization of deep convolutional networks to aerial imagery from new geographic locations: an empirical study with solar array detection. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–8.

WPH-Team (2019). Work package h earth observation, deliverable.

WPK-Team (2019). Work package k methodology and quality, deliverable.

Xu, T., T. Ma, C. Zhou, and Y. Zhou (2014). Characterizing Spatio-Temporal Dynamics of Urbanization in China Using Time Series of DMSP/OLS Night Light Data. *Remote Sens 6*, 7708–7731.

Zeiler, M. D. and R. Fergus (2013, November). Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*. arXiv: 1311.2901.

Zhang, L.-C. (2012). Topics of statistical theory for register based statistics and data integration. *Statistica Neerlandica 66*, 41–63.

Zhang, Q. and K. Seto (2011). Mapping urbanization dynamics at regional and global scales susing multitemporal DMSP/OLS nighttime light data. *Remote Sensing environment 115*, 2320–2329.

ZKI-DE. Der einsatz von fernerkundungs- und geodaten zum monitoring der sgds: Stand der forschung und methodenüberblick für den sdg 11.7.1. september. *unpublished report*.

Zult, D., P. de Wolf, B. Bakker, and P. van der Heijden (2019). A general framework for multiple-recapture estimation that incorporates linkage error correction.