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 2.2

Methodological aspects of using Big data

April 2019

Leading partner: Statistics Netherlands (CBS)

Authors:

J.A. van den Brakel (CBS), P.A. Smith, N. Tzavidis (Soton), R. Iannaccone, D. Zurlo, F. Bacchini, L. Di Consiglio, T. Tuoto (Istat), M. Pratesi, C. Giusti, S. Marchetti (UP), S. Bastianoni, G. Betti, A. Lemmi, F.M. Pulselli and L. Neri (US)

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

Deliverable D2.2

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

Methodological aspects of using Big data

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. The report presents an overview of methods to use big data sources in official statistics and measurement frameworks for well-being and sustainability indicators. One approach is to combine survey data with big data in prediction models, where the big data sources are used as covariates to improve the precision for low regional estimates or the timeliness of the predictions of the first releases of official statistics. Another approach is to construct official statistics or indicators directly from big data sources and apply corrections to account for selection bias. An Input-State-Output framework is proposed to describe and compare the level of sustainability of national and regional economies as an alternative for just presenting a large number of juxtaposed indicators from a measurement framework.

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).

The purpose of this deliverable is to describe in more general terms the methodology required to use non-traditional data sources for measuring sustainable development goal indicators and related official statistics. 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. A major drawback of design-based inference methods from classical sampling theory is that standard errors of sample estimates become large if the sample sizes is relatively small. This problem typically occurs if statistics are required for sub-populations or domains. In such cases multilevel models or time series models can be applied to increase the effective sample size in a particular domain with sample information from other domains or preceding sampling editions. This is known in the literature as small area estimation. These methods work better as strongly correlated covariates are available. 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.

Another aspect of these measurement frameworks is its multidimensionality. Once a large set of indicators is constructed, they can be presented as a large number of juxtaposed indicators to monitor developments of countries or regions. As an alternative the relations between indicators can be visualized and explained with Input-State-Output framework. This framework can be used to describe and compare the level of sustainability of national and regional economies.

This deliverable is organized as follows. Chapter 2 describes general methods to use non-traditional data sources as auxiliary information in model-based estimation procedures and is based on contributions from P. Smith, N. Tzavidis, M. Pratesi, C. Giusti, and S. Marchetti (Subsection 2.1), J.A. van den Brakel (Subsection 2.2), F. Bacchini, R. Iannaccone and D. Zurlo (Subsection 2.3). Chapter 3 describes various methods to correct for selection in non-traditional data sources and is based on a contribution of L. Di Consiglio and T. Tuoto (Istat). In Chapter 4, an Input-State-Output framework is introduced and applied to classify regions in Italy based on their level of sustainability. This is a contribution of S. Bastianoni, G. Betti, A. Lemmi, F.M. Pulselli and L. Neri (University of Siena). The report concludes with a discussion in Chapter 5.

2. Combining survey data with non-traditional data sources

In this chapter we review literature that uses new forms of data for estimating SDG related indicators. Section 2.1 describes the use of new data sources in cross-sectional prediction models. The terms new forms of data and alternative sources of data will be used interchangeably. Although the focus here is on the use of remotely sensed data and data from mobile phone networks, we further present two recent studies that use web-scraped data. Moreover, emphasis is on the production of estimates at fine spatial level and hence the use of small area methods becomes relevant. The structure of this section of the report is as follows. We first briefly review small area estimation methods including direct and model-based methods. Emphasis is placed upon area-level models since the literature that utilises new forms of data does so by aggregating the data at some level of spatial scale. We then review a body of literature that uses alternative sources of data for estimating SDG related indicators and other official statistics. Different typologies of methods are being described and linked to the mainstream small area estimation literature. We conclude by describing some of the challenges with using new forms of data for producing official statistics and some directions for future research.

The focus of Section 2.2 and 2.3 is on time series models. Since many sample surveys conducted by national statistical institutes are conducted repeatedly in time, it make sense to apply time series models that use sample information observed in preceding periods to improve the precision of sample estimates. If new data sources can produce related time series, they can be combined with time series observed with repeated sample surveys in multivariate time series models to further improve precision and timeliness of the sample estimates.

2.1. Cross-sectional Small Area Estimation

We briefly review small area estimation (SAE) methods before focusing on examples of the use of alternative forms of data (also referred to as big data) in SAE applications. To formulate and implement policies and allocate funds at local, geographically disaggregated, levels there is a need for timely and reliable estimates of economic, social and other indicators. Examples of such indicators include the average household income, the unemployment rate, the proportion of individuals below a poverty line, inequality indicators – for example, the Gini coefficient or the Quintile Share Ratio – and SDG indicators. Here, we must emphasise that areas do not have to be defined geographically. For example, interest might be in estimating the poverty rate of ethnic minority groups living in geographical areas of interest. Groups defined by combining geographical and other variables are sometimes referred to as domains. The terms areas and domains can be used interchangeably. Although it should be obvious why having access to local estimates is useful, estimation can be difficult when insufficient data or perhaps no data at all are available from the domains of interest. The term small area is used to describe domains whose sample sizes are not large enough to allow for precise direct estimation.

Direct estimation uses only domain-specific survey data for producing small area estimates. For example, denoting by y_{ij} the outcome of interest for a unit j in area i, the Hajek-Brewer estimator of the population average is defined as follows,

$$\hat{\theta}_i^{Direct} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}}.$$

When direct estimation leads to unreliable estimates due to high sampling variability, one has to rely upon alternative model-based methods for producing small area estimates. National Statistical Institutes and other organisations producing official statistics are cautious about the use of models for producing official statistics. The reason for this is that models depend on assumptions that are hard to verify, which raises concerns about the validity of the estimates especially when these are used as national statistics for designing and implementing policies. Despite the scepticism, SAE is one of the areas in the production of official statistics where the use of models is now widely accepted as being necessary, but to mitigate the risks of model use considerable emphasis is placed on model-testing and validation (Tzavidis et al., 2018).

Model-based methods assume a model for the population and sample data and construct optimal predictors of the target parameters under the model. The term predictor instead of estimator is conventionally used as, under the model, the target parameters are assumed to be random. Generally speaking, SAE models can be classified in two broad classes, namely unit-level models (Battese et al., 1988) and area-level models (Fay and Herriot, 1979). Over the last fifteen years the literature on model-based small area estimation methods has rapidly grown with many contributions both in methodological and applied work. A comprehensive presentation of small area estimation methods can be found in Rao and Molina (2015). In addition, the uptake of methods by producers of official statistics, for example National Statistical Institutes (NSIs) has improved thanks to improved communication between academic researchers and researchers at NSIs. In this report we focus on the use of area-level models. The data requirements for fitting area-level models are less detailed, and therefore pose fewer confidentiality challenges in giving researchers access than the detailed micorodata necessary for unit-level modelling. Moreover, many of the applications that explore the use of new forms of data consider the use of area-level covariates. Without loss of generality the presentation in the following sections focuses on producing estimates for linear statistics such as means and proportions. Area-level models for non-linear statistics (Fabrizi and Trivisano, 2016) have been proposed in the literature and the use of alternative forms of data can be extended to cover these models.

2.1.1. Area-level Models

Let n_i denote the sample size for area i. The direct estimator of the population mean θ_i can be estimated using the Hajek-Brewer estimator. Provided that we have access to information about the design of the survey, the variance of the direct estimate can be computed by using standard survey estimation techniques. Assuming that the variance of the direct estimate (sampling variance) is known, the area-level model – also known as the Fay-Herriot (FH) model – can now be defined, and it has two stages. The first stage models the sampling variation, with the sampling errors ϵ_i assumed to be independent and normally distributed $\epsilon_i \sim N(0, \sigma^2_{\epsilon_i})$.

$$\hat{\theta}_i^{Direct} = \theta_i + \epsilon_i.$$

The second stage of the FH model is to fit a linear model for θ_i,

$$\theta_i = x_i^T \boldsymbol{\beta} + u_i.$$

where x_i^T denotes the area-level covariates, $\boldsymbol{\beta}$ denotes the regression parameter vector and u_i represents the random effects which are assumed also to be normally distributed, $u_i \sim N(0, \sigma_{u_i}^2)$. The combination of the two stages of modelling leads to the FH model

$$\hat{\theta}_i^{Direct} = x_i^T \boldsymbol{\beta} + u_i + \epsilon_i.$$

Assuming $\sigma_{\epsilon_i}^2$ is known, estimates of $\boldsymbol{\beta}$, u_i and σ_u^2 are obtained by using maximum likelihood, residual maximum likelihood or Bayesian methods that are available via *R* packages such as the *SAE* package. Using the estimates of the regression parameters, the estimated variance component and the predicted random effect, small area predictors of the target population parameter can then be derived. Estimates of the uncertainty are also required for assessing the quality of the small area estimates. Uncertainty in this case is quantified by the estimated Mean Squared Error that can be obtained both analytically using for example a Prasad-Rao estimator (Prasad and Rao, 1990) or by using parametric bootstrap (see for example, Gonzalez-Manteiga et al. (2008)). A number of important extensions for the FH model have been proposed. Here we refer to two that may be relevant in the context of using new forms of data. The FH model has been extended to account for measurement error in the covariates (Ybarra and Lohr, 2008). Assuming that the sampling variance of the covariates (quantifying the uncertainty in the covariates) can be estimated, the EBLUP estimator is adjusted to account for this additional uncertainty. This extension can prove useful when working with alternative forms of data that can be measured with error. The challenge, however, is how to quantify this error in new forms of data, such as satellite imaging or mobile phone data. In addition, extensions of the FH model that use a transformed outcome, e.g. by the arcsin transformation when working with proportions, have been proposed (Casas-Cordero et al., 2016, Schmid et al., 2017, Slud and Maiti, 2006).

2.1.2. Uses of new forms of data with emphasis on estimation for fine spatial levels

Identifying what data are needed for small area estimation is important as this determines what methods can be used. SAE is a prediction problem and typically relies on the use of survey data and data from the census, administrative or register sources. The survey data contain information on the target variable and auxiliary variables that are potentially correlated with the target one. The target variable is not available in the census but the census contains auxiliary data on the same variables as the survey. Access to census and administrative data sources is usually challenging because of confidentiality constraints. More commonly, access to census aggregate (area or domain) level data is possible. Users of small area statistics have raised concerns about the use of Census data as covariates. The main objection is the lack of frequent updating of the Census data. In the mainstream small area estimation literature this issue can be tackled either by using register or administrative data that are more frequently updated or auxiliary data from large continuous surveys that have relevant auxiliary variables. In the latter case of course one should account for the survey error in the covariates.

In the last decade there has been a growing body of literature that proposes using alternative forms of data (big data) for producing estimates at very fine spatial scales. In particular, applications have used mobile data, remotely sensed data or a combination of the two. Let us first understand what is the basis of using these alternative forms of data. Most of the applications that involve the use of new forms of data are in a developing country context for which Census data are either unavailable or out-of-date. For many countries survey data are available but there are examples where due to the high costs of conducting surveys, survey data are only collected infrequently. Hence, the reason for employing alternative data sources e.g. satellite images, night-time light data and mobile data is a pragmatic one as there are are limited sources of auxiliary data. Exploring the use of new forms of data may also be beneficial for developed countries that have access to regularly updated survey data. This is because alternative sources of publicly available data can be less costly to collect, hence offering benefits to survey organisations. In addition, the dynamic aspect of new forms of data can allow for more frequently updates of official statistics (see later work by Powell et al. (2017)). However, before deciding the feasibility of using alternative forms of data for producing official statistics estimates, it is important to review the currently available literature and critically assess the challenges associated with the use of such data. In what follows we review literature that uses alternative forms of data for estimating SDG related indicators mainly in a developing country context. Some additional non-SDG related examples are also reviewed.

The literature on the use of non-traditional sources of data for estimating geographically disaggregated parameters of interest developed independently from the mainstream small area estimation literature. In the last decade the focus of the so-called mainstream small area literature was on advancing model-based (unit and area level) methodologies for improving point estimation with more complex e.g. spatial, outlier robust and non-parametric models and improving the accuracy of uncertainty (MSE) estimators via higher-order asymptotic approximations, the bootstrap and other computer intensive methods. On the other hand, researchers working in developing countries and conflict areas were also interested in estimation at very fine geography but faced a lack of access to Census and survey data. This gave rise to the use of alternative forms of data that are publicly available. More recently, researchers representing these two strands of work have started discussing how to cooperate to bridge the different strands of work. The uses of new forms of data vary from purely algorithmic methods, for example machine learning methods and network analysis, to the use of area-level models with auxiliary variables coming from alternative data sources. As previously mentioned, due to the lack of access to sufficient data, the use of area-level models is more relevant in these applications.

Marchetti et al. (2015) identified three possible approaches to the use of new forms of data in the small area estimation framework, with a specific reference to the computation of poverty and living condition indicators in developed countries. The first opportunity is to use the new data sources to create local indicators and compare them to those obtained with small area estimation methods. The idea is to reconcile data from the two independent sources - big data and sample surveys - to use available local measures extrapolated from big data to compare and benchmark measures on related aspects of the phenomenon under study (e.g. poverty and social exclusion) obtained from survey data and vice versa. Measures from big data sources are usually obtained very quickly; however, they can be affected by a serious self-selection bias. Conversely, small area estimates are methodologically sound,

but they require timely survey and population data that can be difficult to obtain. Comparing the two alternative sets of measures referring to the same areas can provide useful insights on the potential of big data to benchmark small area estimates. If there is accordance between big data and survey data in a given small domain/area with respect to the recorded level of deprivation and poverty, then analysts and policy makers may rely on a strong evidence. Otherwise, if there is a discrepancy between the results obtained from the two sources of data, then there is a need for further investigation of those domains/areas.

The second possibility is to use alternative data sources to generate new covariates for small area models. However, as already underlined, the extension of the covariates to include variables such as social media search loads or remote-sensing images (e.g. in crop-yield surveys, and also in social surveys) or tracking of human mobility opens up difficulties and challenges. Due to technical problems and legal restrictions, it is unfeasible at this stage to have unit-level data that can be linked with administrative archives, census or survey data. To overcome this problem we can use the so-called area-level models, such as the Fay-Herriot model. However, attention should be paid to the fact that under the Fay-Herriot model it is assumed that the auxiliary variables are measured without error, that is, that they are available for all the areas and they come from census or archives covering the entire population of interest. When auxiliary variables come from surveys, they suffer from sampling errors and may also suffer from nonsampling errors, and thus thry can be considered as measured with error. Generally, auxiliary variables coming from alternative sources of data are not measured on all (or on a big proportion) of the units of the target population, nor are they collected using a random sample. For these reasons, in their application Marchetti et al. (2015) consider that big data are subject to measurement error and use the Ybarra and Lohr (2008) model.

The last opportunity suggested byMarchetti et al. (2015) for the use of big data with small area estimation is to use survey data to check and remove the self-selection bias of the values of the indicators obtained using big data. The idea is that big data could be used directly to measure poverty and social exclusion, appropriately taking into account the self-selection problem. The author envision that survey data could be used to check and remove this bias, provided that unit-level information from big data sources will be available.

Noor et al. (2008) present work on the use of remotely sensed night-time light as a proxy for poverty in Africa. In this paper nationally representative survey data are used to construct asset-based poverty indices at aggregate level (administrative I level) for 37 countries in Africa. Geographical information systems are used to compute average brightness and distance to night time lights, also at aggregate level. Correlation analysis is then used to explore the relationship between poverty and remotely sensed data. Frias-Martinez and Virseda (2012) present a study analysing the relationship between socio-economic factors and the use of mobile-phone data. In this paper the focus is on comparing new forms of data with national level Census data. In particular, the latitude and longitude of mobile phone towers are used for mapping data on mobile usage to geographical units for which information from Census variables is available. Census and mobile phone data are compared using correlations. Smith-Clarke et al. (2014) also present work on estimating poverty maps using aggregated mobile communications networks. Their methodology relies on the use of two datasets. The first one is

a representative dataset of a country's population that is used to automatically extract indicators at fine spatial scales that are assumed to be associated with poverty. For this dataset the authors use mobile data provided from local network providers and linked to mobile towers. The second dataset that is required is one used for validation purposes and provides a picture about disaggregated poverty. For this dataset the authors use socio-economic datasets provided by surveys and Censuses. Similarly to other papers, spatial aggregation of indicators constructed from the network data are defined for Voronoi polygons. Finally, correlation coefficients are being computed between the call network extracted characteristics and poverty rates. Jean et al. (2016) propose a machine learning approach for extracting socioeconomic data from high resolution daytime satellite imaging. Validation of their machine learning methodology is then done using recent data on economic outcomes in five African countries.

A common feature of the methods described above is that they rely only on the use of alternative forms of data that are publicly available. Using these new forms of data, indicators are constructed that are assumed to be correlated with poverty characteristics. Other data sources (e.g. from surveys) are used for validation purposes. This research offers evidence for the usefulness of alternative sources of data. Indeed, these methods can provide a useful first step when working in a context with limited access to official data. However, more work is needed before the use of such data can enter the production of reliable official statistics. Steps in this direction are offered by the literature summarised below.

Blumenstock et al. (2015) explore the prediction of poverty and wealth using mobile phone data combined with a small survey in what they call resource-constrained environments where survey and Census data are limited. Anonymised call data are supplemented by a geographically stratified random sample of subscribers. Using informed consent, survey data on wealth and mobile data are being merged and used for prediction purposes. Watmough et al. (2016) explore the relationship between poverty and remotely sensed satellite data in Assam, India. The authors used highly disaggregated (community) socio-economic data from the Census on education, land ownership, caste, and access to water to compute an indicator of relative wealth. Data obtained from geographical information systems were then used to extract remotely sensed environmental condition data for each community. Census and remotely sensed data were then modelled using classification trees and random forest methods. The research proposed by these papers makes positive steps towards the use of alternative forms of data, however, in our view additional significant work is needed if interest is in the use of these methods in the production of official statistics.

In their application, Marchetti et al. (2015) use two different Fay-Herriot models accounting for the measurement error in the covariates to produce estimates of the poverty incidence, measured with the At-risk-of-Poverty-Rate, and of the mean household equivalised income for the Local Labour Systems of the Tuscany Region, Italy. The authors use area-level data from the EU-SILC survey 2011 to compite the target (response) estimates and, due to the unavailability of updated Census information, the EU-SILC survey is also as source for covariate information together with remote-sensing data on private vehicles' mobility. The hypothesis of the authors is that mobility data can be predictive of well-being measures, representing a valuable course of information for out-of-sample areas in the EU-SILC survey. The authors also provide some discussion on the methodological hypothesis that are necessary

to use survey data and nre sources of data as covariates in the Ybarra and Lohr (2008) model. More specifically, the source of big data on mobility in the application presented by Marchetti et al. (2015) is a dataset of private vehicles in central Italy, tracked with a GPS device. The dataset is comprised of information on approximately ten million different car journeys made by 150,000 vehicles tracked during May 2011. Focusing on Tuscany, the dataset refers to 37,326 vehicles, which correspond to 1.5 percent of the total vehicles registered in Tuscany in 2011. Thus, the data cannot be considered as covering all the population of vehicles in the Tuscany Region. The GPS traces are collected by OCTO Telematics S.p.a., a company that provides a data collection service for insurance companies. The GPS device is automatically turned on when the car is started, and the global trajectory of a vehicle is formed by the sequence of GPS points that the device transmits every 30 seconds to the server. When the vehicle stops no points are logged or sent. The authors exploited these stops to split the global trajectory into several sub-trajectories, which corresponded to the single journeys undertaken by a vehicle. Vehicle traces were then mapped on the road network and their position during the stops was associated with the census sectors, provided by the Italian National Institute of Statistics (ISTAT). The covariates used in the small area estimation model are a measure of mobility and a measure of entropy computed using the GPS data for each LLS.

Marchetti et al. (2015) estimate the small area income means and ARPRs applying the Ybarra and Lohr (2008) model with covariates coming form the EU-SILC survey (as the response variables) and also covariates computed using the remote sensing data on vehicles' mobility. The author underline that estimates obtained using data taken from the EU-SILC survey are design unbiased. Thus, in the variance-covariance matrix they use the estimated variances of the auxiliary variables? mean estimates, setting the covariances equal to zero. As regards mobility data, the authors argue that, as an alternative source of data, they can be considered as coming about according to a survey design or not. In the second case, there is no need to make any inference about unobserved population units. The first case, the one chosen by the authors, follows instead a design perspective, and then there is uncertainty in the GPS data. From this perspective, Marchetti et al. (2015) consider the data on mobility as collected on a self-selected sample of car journeys. However, referring to Bethlehem (2002), the authors argue that the bias due to the self-selection process is related to the correlation between the target variable (mobility index) and the response behaviour (having or not having a GPS). Using the results shown in Pappalardo et al. (2013), the author argue that this correlation coefficient can be considered very small in their application, and hence the bias due to the self-selection process could be negligible. In fact, Pappalardo et al. (2013) showed that the mobility index measured using the sample of cars with GPS was coherent with the mobility registered for all the vehicles in the municipality of Pisa (data derived from traffic sensors spread around the city). Given this evidence, it seemed reasonable to use the hypothesis of independence between the mobility indexes and "having a GPS", so that the authors could handle the mobility data as if they were a simple random sample from the population of vehicles. The variances in the variance-covariance matrix of the model were then computed using a simple random sample design variance formula (considering negligible the correction term for finite populations).

Irrespective of whether the design perspective is chosen or not, the use of alternative sources of data as auxiliary variables in small area models is motivated by their predictive power, which results in

improved efficiency of the small area estimates for sampled and out-of-sample areas. The discussion and the perspective presented by Marchetti et al. (2015) could be useful in many other applications using new sources of data with small area estimation models. Another application of the same ideas was presented in Marchetti et al. (2016), where the authors used an indicator computed using Twitter data as covariate in an area-level model for estimating a poverty-like indicator, households' share of food consumption expenditure.

Similar in spirit but employing statistical modelling closer to the mainstream SAE literature and more established survey sources is the work by Steele et al. (2017). The authors present methodology for poverty mapping in Bangladesh that utilises mobile and satellite auxiliary data when Census data are out-of-date or unavailable. The paper uses call detail records (CDRs) and remote sensing and geographical information system (RS) data. CDR data are used for extracting metrics such as phone usage, top up amounts and network information related to mobile phone usage. RS data include metrics likely to be associated with welfare indicators and include night-time lights, vegetation indices, climatic conditions and distance from roads and major urban areas. Turning now to the outcome of interest, poverty is measured by three measures, namely a wealth index, an indicator of household consumption and reported household monetary income. The three measures were obtained from various established survey data sources in Bangladesh for example the Demographic and Health Survey (DHS). At what spatial scale we decide to operate is also a point of interest. In mainstream SAE literature the level of spatial scale is usually determined by policy considerations and involves discussions between users and producers of official statistics. The feasibility of producing small area estimates at certain spatial scales also depends on data availability and striking a balance between the over-reliance on models and the ability to validate the estimates (see Tzavidis et al. (2018) for a discussion). In the literature that uses new forms of data, the spatial scale of the analyses appears to be driven by the data. In particular, as with other papers, in Steele et al. (2017) the spatial scale of the analysis was based on approximating the mobile tower coverage by Voronoi polygons of varying sizes (from 60m to 5km) depending on whether a rural or urban area was under consideration. Each polygon was then assigned aggregate values of the CDR and RS data. Survey data were also matched to Voronoi polygons via latitude and longitude information provided either by GPS data, the centroids of survey clusters or the home cell tower of survey respondents.

The authors then model poverty indicators via a hierarchical Bayes geo-statistical model that includes a spatial random effect with the neighbourhood structure specified by the Voronoi polygons. Poverty is predicted for each Voronoi polygon and uncertainty is estimated using the posterior distribution. The model results are assessed using cross-validation and by comparison with estimates from other studies. This is an example of a model that operates at the area-level (Voronoi polygons). Although it is not clear why fitting models at the level of the Voronoi polygon might be of policy interest, estimates can be aggregated to other meaningful spatial scales and benchmarking to reliable official statistics can be employed. It is also not clear why the particular model is preferred over possibly simple models such as the FH model. Additional research is needed to explore the use of tools used extensively in small area estimation, for example benchmarking, and to try to bridge the methodological gaps.

Schmid et al. (2017) use mobile phone data to estimate literacy rates in Senegal. This work is similar

in spirit to the paper by Steele et al. (2017) but closer to the mainstream SAE literature in the sense that it employs a FH model that uses DHS survey data, covariate information from anonymized call detail records in the form of tower to tower traffic and benchmarking methods. Hence, this paper offers some answers to the questions producers of official statistics may have about the use of new forms of data. In particular, in this paper mobile phone data were aggregated over time and commune by matching mobile phone towers to communes. Around 70 covariates are constructed using CDR data. Among other variables, covariates include the volume of calls, the distance of calls, the number of calls to the capital city (Dakar) and behavioural indicators. The authors obtain direct and model-based estimates of the proportion of literate people by gender at commune level. Since a proportion is being model, a transformed FH model that benchmarks commune-level model-based estimates to higher levels of geography is also considered. As the authors mention, one of the main advantages of using such covariate information is in the ability to update small area estimates more frequently. Nevertheless, uncertainty in the mobile phone data arises from the fact that the coverage of the mobile phone towers differs and is unknown. Hence, certain types of users may be excluded, which presents one of the main drawbacks of using mobile data. The authors also raise the important point of data processing and cleaning and the need to develop online platforms that will enable easier use OF data and models.

Up to this point the review focused on the use of mobile phone and remotely sensed data. Another form of data that has gained popularity in the production of official-type statistics in recent years is the use of web-scraping. Here we provide two examples of papers that work with this type of data. Powell et al. (2017) research the possibility of using web-scraping of online prices for producing the consumer price index more frequently than existing monthly estimates. The authors focus on product-specific disaggregated CPI and claim that as the completeness of web-scraped data increases, combining this with survey data will possibly allow for dynamic inter-month prediction of the CPI. The key here is in the use of alternative forms of data and existing representative survey data to achieve this goal. Steorts et al. (2018) present methodological work on simultaneous spatial smoothing and benchmarking for estimating average rental prices for small areas using web-scraped rental price data from Berlin. Although the focus of this paper is on developing new methodology, benchmarking to official small area rental prices is used to try and account for the possible lack of coverage of the web-scraped database used by the authors.

2.1.3. Discussion

In the last decade the uses of new forms of data such as remotely sensed and mobile data have begun to be explored as potential sources for producing official statistics. The first step in the literature is to establish the association between indicators constructed from the new forms of data and the targets of estimation, for example the incidence of poverty and SDG-related indicators. Although not directly recognised by this literature, this is a necessary step if new forms of data are to be used in producing official statistics with model-based or model-assisted methods. The case for using such data is a compelling one when working in countries that lack survey and Census data. Additional attractive features of new forms of data include public access and dynamic updating, which can offer cost benefits to survey organisations. Although the literature has made some important steps in establishing the potential value of these new forms of data, it is our view that significant effort is required before we

will be able to bring these ideas into the mainstream production of official statistics.

An important aspect when producing official statistics is assessing their quality. Hence, being able to evaluate the quality of the estimates produced with the use of new forms of data is of paramount importance. Although the literature uses correlational analysis and cross-validation, more needs to be done to bring in ideas about the quality of estimates from the survey literature. New forms of data are treated as fixed. Is this a reasonable assumption? And if not, how can we quantify the error associated with these data sources and account for this in estimation? The issue of coverage of new forms of data is also an aspect that concerns producers of official statistics. It is our view that some access to representative data from surveys is needed alongside methodological tools that will allow for the combined use of survey and new forms of data. Defining appropriate spatial scales at which estimates are produced is another aspect that requires additional work. The fact that you can use a model to extrapolate estimates at very fine spatial scales does not mean that the estimates are of acceptable quality. Over-reliance on a model can be viewed as a risky strategy when the aim is to produce official statistics. Related to this, there is a need for research on how benchmarking techniques can be used when working with new forms of data. Finally, if new forms of data are to be used for producing official statistics, survey organisations must intensify the investment in gathering, cleaning and processing such data.

2.2. Time series methods

In the previous section extensions of the Fay-Herriot model (Fay and Herriot, 1979) are developed to use new data sources in cross-sectional area level models for small domain predictions. Most surveys conducted by national statistical institutes are conducted repeatedly over time. A natural approach for small area prediction as well as nowcasting is to extend the Fay-Herriot model with related information from previous editions of the survey. Rao and Yu (1994) extended the area level model by modelling random domain effects with an AR(1) model. Other accounts of regional small area estimation of unemployment, where strength is borrowed over both time and space, include Datta et al. (1999), You et al. (2003), You (2008), Pfeffermann and Burck (1990), Pfeffermann and Tiller (2006), Krieg and van den Brakel (2012).

In this section multivariate structural time series models are developed to combine series obtained with repeated samples with related auxiliary series. This serves two purposes. Extending the time series model with an auxiliary series allows modelling the correlation between the unobserved components of the structural time series models, e.g. trend and seasonal components. If the model detects a strong correlation, then the accuracy of domain predictions will be further increased. Harvey and Chung (2000) propose a time series model for the Labour Force Survey in the UK extended with a series of claimant counts.

Information derived from non-traditional data sources like Google trends or social media platforms are generally available at a higher frequency than series obtained with repeated surveys. This allows to use this time series modelling approach to make predictions for the survey outcomes in real time at the moment that the outcomes for the social media are available, but the survey data not yet. In this case the auxiliary series are used as a form of nowcasting (van den Brakel et al., 2017). Google Trends in particular has been used already in the economic forecasting literature for this purpose, see

e.g. Vosen and Schmidt (2011) and Choi and Varian (2012) and the references therein.

With a structural time series model a series is decomposed in a trend component, seasonal component, other cyclic components, regression component and an irregular component. For each component a stochastic model is assumed. This allows the trend, seasonal, and cyclic component but also the regression coefficients to be time dependent. If necessary ARMA components can be added to capture the autocorrelation in the series beyond these structural components. See Harvey (1989) or Durbin and Koopman (2012) for details about structural time series modelling.

We first introduce a bivariate structural time series model to illustrate the concept how a time series obtained with a repeated survey can be modelled with an auxiliary series. In the case of big data sources like Google trends it is easy to derive many related series. For example in the case of combining a series of unemployment many related series based on search activities for finding jobs can be derived easily. To handle the high-dimensionality problem, dynamic factor models are introduced as a next step.

There are competing time series modelling approaches available in the literature. A well-known alternative approach for structural time series models are the Auto Regressive Integrated Moving Average (ARIMA) models developed by Box and Jenkins (1989). Contrary to structural time series models, stationary of the time series models play a crucial role in this class of time series models. Non-stationary time series are first made stationary by taking one or more differences with lagged observations. Once a stationary series is obtained, it is modelled with lagged observations (AR components) or lagged residuals (MA components). Combining multiple time series in this class of models proceeds by including them as a regression component in an ARIMA model, resulting in so-called (ARIMAX) models or stacking them in a vector that is modelled with Vector ARIMA (VARIMA) models. For an introduction in this class of models we refer to Lütkepohl (2005).

2.2.1. Structural time series models

Let $\hat{\theta}_{it}$ denote a direct estimate for area i and period t based on data observed in period t and $v(\hat{\theta}_{it})$ an estimate for the variance of $\hat{\theta}_{it}$. Widely applied estimators are the Horvitz-Thompson estimator or the general regression (GREG) estimator (Särndal et al., 1992). Developing a time series model for survey estimates observed with a periodic survey starts with a model, which states that the survey estimate can be decomposed in the value of the population variable and a sampling error:

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \tag{2.1}$$

with θ_{it} denote the real finite population value under a complete enumeration of the target population and e_{it} the sampling error. Under a structural time series modelling approach, the series of the finite population parameter can be decomposed in a stochastic trend, say L_{it}, a seasonal component, say S_{it}, to model systematic deviations from the trend within a year, and a white noise component, say ε_{it}, for the remaining unexplained variation. These considerations lead to the following model for the series of the finite population parameter:

$$\theta_{it} = L_{it} + S_{it} + \varepsilon_{it}. \tag{2.2}$$

The trend and seasonal component are time dependent by modelling them with stochastic processes. Popular models for the trend are the local level model, the local linear trend model and the smooth trend model (Durbin and Koopman (2012), Ch. 3). For illustrative purposes we specify the smooth trend model, which is widely applied in econometric time series modelling:

$$\begin{aligned} L_{it} &= L_{it-1} + R_{it-1}, \\ R_{it} &= R_{it-1} + \eta_{it}. \end{aligned} \tag{2.3}$$

This model specifies that the trend contains a level L_{it} that is equal to the level of the level of the previous period and an adjustment R_{it}, which is often interpreted a the slope. The slope R_{it} on his turn is time dependent by modelling it as a random walk. the level can change gradually over time through the slope parameter. The slope disturbances η_{it} are assumed to be normally and independently distributed; $\eta_{it} \simeq \mathcal{N}(0, \sigma^2_{i\eta})$. In the context of structural time series models, L_{it} and R_{it} are called the state variables. This are the parameters of the unobserved components like the trend and the seasonal component.

The local linear trend model is defined as (2.3) but also contains a disturbance term in the equation for L_{it}, which result in a more flexible trend. The local level model assumes a random walk for L_{it} and does not have a separate slope parameter. Since the trend models are time dependent, they have the flexibility to capture cyclic movements. Models of the form (2.3) are sometimes called the trend-cycle component.

The seasonal component S_{it} in (2.2) is also based on a stochastic process, which gives the model the flexibility to change the seasonal pattern gradually over time, depending on the dynamic behaviour of the observed series. Popular models are the dummy seasonal model and the trigonometric seasonal model. We refer to Harvey (1989) or Durbin and Koopman (2012) for expressions. Similar to the trend models, the seasonal components contain state variables, which describe the seasonal pattern, and disturbance terms that allow the state variables to change gradually over time. The disturbance terms are assumed to be normally and independently distributed.

The unexplained variation in the time series of the population parameter is modelled as white noise, i.e. $\varepsilon_{it} \simeq \mathcal{N}(0, \sigma^2_{i\varepsilon})$ and are uncorrelated over time.

Inserting (2.2) into measurement model (2.1) gives:

$$\hat{\theta}_{it} = L_{it} + S_{it} + \varepsilon_{it} + e_{it}, \tag{2.4}$$

In a cross-sectional survey it might be convenient to combine the population white noise and the sampling error in one disturbance term, say $\nu_{it} = \varepsilon_{it} + e_{it}$ and assume $\nu_{it} \simeq \mathcal{N}(0, \sigma^2_{i\nu})$. To allow for nonhomogeneous variance in the sampling errors, Binder and Dick (1990) proposed a measurement error where the disturbance terms νit are proportional to the sampling errors of $\hat{\theta}_{it}$, i.e.

$$\nu_{it} = \sqrt{v(\hat{\theta}_{it})}\tilde{\nu}_{it}$$

with $\tilde{\nu_{it}} \simeq \mathcal{N}(0, \sigma^2_{i\tilde{\nu}})$ Such a model would be useful if the sampling error dominates the white noise in the population parameter. The question how to account for sampling variance is also an issue in seasonal adjustment variances (Pfeffermann and Sverchkov, 2014, Bell, 2005) studied the contribution of the sampling variance in the variance of the estimation error of seasonally adjusted series and in the nonseasonal component. In Boonstra and van den Brakel (2016) it is shown how structural time series models can also be expressed as time series multi-level models in an hierarchical Bayesian framework as an extension of the cross-sectional Fay-Herriot model. This approach provides another way to separate sampling error from population white noise in time series observed with repeated sample surveys. In the case of (rotating) panels, the sampling error can be separated from the population white noise (van den Brakel and Krieg, 2016).

2.2.2. Multivariate structural time series models

The univariate structural time series discussed in the previous subsection can be seen as a form of small area estimation to borrow strength over time. In this subsection these models are extended to multivariate structural time series models. This serves two purposes. First, if time series of repeated surveys of different domains are combined in one multivariate model, more precise domain predictions are obtained by using both temporal and cross-sectional correlations, i.e. borrowing strength over time and space. In this case the series obtained in the different areas are stacked in one vector $\hat{\boldsymbol{\theta}}_t = (\hat{\theta}_{1t}, ..., \hat{\theta}_{Mt})^t$. The separate time series for the M domains have their own stochastic trend and seasonal component. By modelling the correlation between the trend disturbance terms or the seasonal disturbance terms between the domains, the effective sample size for each domain i in period t is increased with sample information from previous sampling occasions but also from other domains (Pfeffermann and Burck, 1990, Pfeffermann and Bleuer, 1993, Pfeffermann and Tiller, 2006, Krieg and van den Brakel, 2012, Boonstra and van den Brakel, 2016). Secondly, if a related auxiliary series derived from a register or another non-traditional data source is available, then the target series observed with a repeated survey can be combined in a multivariate structural time series model with one or more auxiliary time series. Correlations between the unobserved components can be modelled in a similar way to improve the estimates of the sample survey with information from related auxiliary series. This approach is applied by Harvey and Chung (2000) for the Labour Force Survey in the UK extended with a series of claimant counts.

Let x_{it} denote an auxilairy series that is considerd to be combined with $\hat{\theta}_{it}$. One approach to use the auxiiary series to improve the predictions for θ_{it} is to extend model (2.4) with a regression component; $\hat{\theta}_{it} = L_{it} + S_{it} + \varepsilon_{it} + \beta x_{it} e_{it}$. The major drawback of this approach is that the auxiliary series will partially explain the trend and seasonal effect in θ_{it}, leaving only a residual trend and seasonal effect for L_{it} and S_{it}. This hampers the estimation of a trend for the target variable. This problem is circumvented by modelling both series in a bivariate structural time series model:

$$\begin{pmatrix} \hat{\theta}_{it} \\ x_{it} \end{pmatrix} = \begin{pmatrix} L_{it}^{[\theta]} \\ L_{it}^{[x]} \end{pmatrix} + \begin{pmatrix} S_{it}^{[\theta]} \\ S_{it}^{[x]} \end{pmatrix} + \begin{pmatrix} \nu_{it}^{[\theta]} \\ \epsilon_{it}^{[x]} \end{pmatrix}. \tag{2.5}$$

In (2.5) both series has their own trend, seasonal and distrurbance term. In the case of a smooth trend model, the relation between both series can be modelled via the covariance structure of the

slope disturbances, i.e.

$$
\begin{aligned}
&L_{it}^{[z]} = L_{it-1}^{[z]} + R_{it-1}^{[z]}, \\
&R_{it}^{[z]} = R_{it-1}^{[z]} + \eta_{it}^{[z]}, z \in (\theta, x) \\
&\begin{pmatrix} \eta_{it}^{[\theta]} \\ \eta_{it}^{[x]} \end{pmatrix} \simeq \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} {\sigma^{[\theta]}}_{i\eta}^{2} & \rho\sigma_{i\eta}^{[\theta]}\sigma_{i\eta}^{[x]} \\ \rho\sigma_{i\eta}^{[\theta]}\sigma_{i\eta}^{[x]} & {\sigma^{[x]}}_{i\eta}^{2} \end{pmatrix} \right).
\end{aligned}
$$

If the model detects a strong correlation between the trends of the $\hat{\theta}_{it}$ and x_{it}, then the trends of both series will develop into the same direction more or less simultaneously. In this case the additional information from the auxiliary series will result in an increased precision of the estimates of the target series $\hat{\theta_{it}}$. In the case of strong correlation between the disturbances of the trends, i.e. if $\rho \to 1$, the trends are said to be cointegrated. This implies that the slope disturbances of both series simultaneously move up or down and that the slope disturbances of the auxiliary series can be perfectly predicted from slope disturbances of the target series. In that case there is one underlying common trend that drives the evolution of the trends of the two observed series. Cointegration increases the precision of the estimated trend and signal of the target series and allows for formulating more parsimonious models, which increases estimation efficiency. For a more detailed discussion about cointegration in the context of state space modelling, see Koopman et al. (2009), sections 6.4 and 9.1. The correlation between seasonal disturbance terms of both series can be modelled in a similar way.

The bivariate structural time series model applied to the consumer confidence index and the social media indicator in Subsection 6.2 of Makswell deliverable 2.1. This is an example where an auxiliary series derived from social media platforms is used to improve the precision and timeliness of estimates obtained with a repeated survey.

The extension of model (2.5) to multivariate models for more than one auxiliary series or models for time series of M domain series is relative straightforward. Examples can be found in the above cited literature.

2.2.3. Estimation of structural time series models

A widely applied approach to fit structural time series models is to write them in state-space form and analyse them with the Kalman filter. The Kalman filter is a recursive procedure that runs from period $t = 1$ to T and gives, for each time period, an optimal estimate for the state variables based on the information available up to and including period t. These estimates are referred to as the filtered estimates. The filtered estimates of past state vectors can be updated, if new data after period t become available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the complete time series. The Kalman filter assumes that the hyperparameters, i.e. the variance components of the stochastic processes for the state variables are known. This is generally not the case. In practise maximum likelihood estimates for the hyperparameters are obtained using a numerical optimization procedure (BFGS algorithm, Doornik (2009). Expressions for the state space representation of structural time series models and details of the Kalman filter can be found in Durbin and Koopman (2012).

Several software packages are available to fit structural time series models. Most standard structural time series models can be fitted with STAMP (Koopman et al., 2009). For more advanced models, more advanced software is required. One option is to implement these models in OxMetrics in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman et al. (2009, 2008). Another possibility is to implement these models in R (Team, 2017) using packages like KFAS (Helske, 2017) or DLM (Petris, 2010).

If the dimensions of a multivariate state-space model become large, estimation can become inefficient. One alternative is to express the structural time series models as time series multilevel models in an hierarchical Bayesian framework and fit these models using MCMC simulations, in particular the Gibbs sampler. Connections between structural time-series models and multilevel models have been explored before from several points of view in Knorr-Held and Rue (2002), Chan and Jeliazkov (2009), McCausland et al. (2011), Ruiz-Cárdenas et al. (2012), Piepho and Ogutu (2014). A comparison between multilevel time-series models and state-space models applied to time series of the Dutch National Travel Survey is carried out in Bollineni-Balabay et al. (2017). In Boonstra and van den Brakel (2016) it was found that multilevel time series models in a hierarchical Bayesian formulation have advantages in terms of flexibility and computational efficiency if the number of time series becomes large.

2.2.4. Dynamic factor models

Over the last decade, the number of non-traditional data sources that can be considered in the production of official statistics, is rapidly increasing. Particularly information derived from social media messages from Twitter and internet search behaviour from Google Trends easily result in a large number potential auxiliary series. Combining them in a full multivariate structural time series model as outlined in the previous subsection limits the degrees of freedom for model fitting. Due to the so-called "curse of dimensionality" prediction power of such models will be low. From this perspective, factor models are developed to formulate parsimonious models, despite a large number of auxiliary series are considered. Factor models are developed and widely applied by central banks to nowcast GDP on quarterly frequency using a large amount of related series observed on a monthly frequency (Boivin and Ng, 2005, Stock and Watson, 2002a,b, Marcellino et al., 2003). More recently, Giannone et al. (2008), Doz et al. (2011) proposed a state-space dynamic factor model. They propose a two-step estimator. In a first step a small amount of common factors are extracted from a large set of series using principal component analysis. In a second step, the common factors are combined with the target series in a state space model and are fitted using the Kalman filter.

Let $\mathbf{x}_t$ denote the vector with n auxiliary series, where n is large. In a first step a dynamic factor model is assumed for the auxiliary series

$$\mathbf{x}_t = \boldsymbol{\Delta}\mathbf{f}_t + \boldsymbol{\epsilon}_t \tag{2.6}$$

$$\mathbf{f}_t = \mathbf{f}_{t-1} + \boldsymbol{\omega}_t \tag{2.7}$$

with $\mathbf{f}_t$ a p dimensional vector containing a small set of common factors that capture the major part of co-movements from the set of auxiliary series $\mathbf{x}_t$, where $p << n$. Furthermore $\boldsymbol{\Delta}$ denotes a $n \times p$ dimensional matrix with factor loadings and $\boldsymbol{\epsilon}_t$ an n-vector containing variable specific shocks

(idiosyncratic components). If the series in $\mathbf{x}_t$ are stationary, then $\mathbf{f}_t$ can be estimated with the principal components on $\mathbf{x}_t$ and $\boldsymbol{\Delta}$ through OLS. Generally potential auxiliary series will be non-stationary. If it is assumed that the series in $\mathbf{x}_t$ are I(1), then $\mathbf{f}_t$ and $\boldsymbol{\Delta}$ can be estimated with principal components on the differenced data $\mathbf{x}_t - \mathbf{x}_{t-1}$ (Bai, 2004).

In a second step the target series is combined with the auxiliary series in the following structural time series model

$$\begin{pmatrix} \hat{\theta}_{it} \\ \mathbf{x}_t \end{pmatrix} = \begin{pmatrix} L_{it}^{[\theta]} \\ \hat{\boldsymbol{\Delta}}\mathbf{f}_t \end{pmatrix} + \begin{pmatrix} S_{it}^{[\theta]} \\ 0 \end{pmatrix} + \begin{pmatrix} \nu_{it}^{[\theta]} \\ \boldsymbol{\epsilon}_t \end{pmatrix}, \tag{2.8}$$

where $\hat{\boldsymbol{\Delta}}$ are the estimates for the factor loadings obtained in the first step, which are plugged in the design matrix of the measurement equation if the state-space representation of (2.8). The common factors $\mathbf{f}_t$ are re-estimated with the Kalman filter recursion. For simplicity it is assumed that the common factors do not contain a seasonal component. This can be added to (2.8), if necessary.

In the state space representation of (2.8), a random walk is assumed for the common factors: $\mathbf{f}_t = \mathbf{f}_{t-1} + \boldsymbol{\omega}_t$ with $\boldsymbol{\omega}_t$ a p-vector containing the disturbance terms of the common factors. The model can account for non-zero correlation between the slope disturbances of the trend in $\hat{\theta}_{it}$ and the disturbance terms of the common factors. In this way the target parameter estimates benefit from the common factors extracted from the auxiliary series using principal components in the first step.

To exploit the timeliness of the auxiliary series obtained with big data sources model (2.8) can be expressed the frequency of the target series observed with the survey. If $\hat{\theta}_{it}$ is observed on a monthly frequency, then initial results for the auxiliary series for the last month can be inserted in (2.8) to obtain more precise nowcasts for the target variables $\hat{\theta}_{it}$. This approach was followed in Subsection 6.2 of Makswell deliverable 2.1. It might be anticipated that this procedure works well if the auxiliary series are more or less final and not subject to large revisions. To produce reliable early nowcasts it might be necessary to express model (2.8) at a higher frequency, e.g. weeks instead of months. This requires a disaggregation of the unobserved time series components of the target series to this higher frequency. After fitting the model, estimates for the survey parameters are obtained by aggregating the underlying components to a monthly frequency. Details of mixed frequency state-space models are described in Harvey (1989), Ch. 6.3, Durbin and Quenneville (1997), Moauro and Savio (2005).

2.3. The use of electronic payment to nowcasting consumption

[1] Following previous studies related to the use of big data for nowcasting (Baldacci et al. (2016) for a survey), in this section we analyse whether electronic payments data can be used to improve the accuracy of nowcasts (current quarter forecasts), of Italian private household consumption for durable and non durable goods. Starting from the Istat VAR model (Bovi et al. (2000)), that provides internal quarterly forecast of GDP and its main aggregates, we explore if the components of payments data can improve the accuracy of the model (for an application of payment system data to forecast of Italian economy see (Aprigliano et al. (2017)). The series of the electronic payments system (hereafter

[1] Roberto Iannaccone, Davide Zurlo, Istat, Division for data analysis and economic, social and environmental research, Iannacco@istat.it, zurlo@istat.it

BI-COMP series) contain information on retail payment transactions, ie transactions with low priority and/or low amount (usually equal to or less than 500 thousand of Euros) between customers of banks and financial institutions (households, businesses, public administrations - PA), whereby the transfer of money from one person to another does not take place in cash, but in the form of electronic transactions. Another source that has been explored is that one related to the aggregated anti-money laundering reports (hereafter SARA series) born with the purpose of "... allowing the carrying out of analyzes aimed at bringing out any phenomena of money laundering", the data are sent only by financial intermediaries and derives from the aggregation of transactions of an amount equal to or greater than 15.000 of Euros. Data draws from this source are mainly related to the economic sector and the institutional characteristics (households) The SARA and BI-COMP series are at monthly frequency with different time span as shown in Table 2.1.

Finally we consider both the TARGET series, that are the aggregate data collected on payments using the infrastructure called Trans-European Automated Real-Time Gross Settlement Express Transfer System setup by Bank of Italy, and the POS series are data on payments using cards and point of sale payments (POS).

Table 2.1: Electronic series

Domain	**Name**	**Periodd**
BI-COMP	Transfers	January 2000 - November 2017
	Credit Cards	January 2000 - November 2017
	Receipts	January 2000 - November 2017
	Cheques	January 2000 - November 2017
	Total	January 2000 - November 2017
Target 2	Target	May 2008 - November 2017
Pos	Pos	June 2006 - November 2017
SARA	Wholesale trade	January 2001 - November 2017
	Retail Trade	January 2001 - November 2017
	Households	January 2001 - November 2017
	Services	January 2001 - November 2017
	Total	January 2001 - November 2017

Moreover they are characterized by a strong seasonal component and affected by different regulatory changes specific to each domain which could be reflected in deterministic effects (different types of outliers) in the series. The series were therefore analyzed using the standard Istat procedure for the seasonal adjustment (TRAMO-SEATS) implemented in JDemetra+ (Grudkowska (2016)). The preliminary analysis for the series has been organized in 5 steps:

- backcasting of the Target and Pos series to obtain the same time spans as for the other series in BI-COMP;
- identification and estimation of the deterministic effects for each series;
- linearization of the series removing the deterministic component;

- seasonal adjustment of the monthly linearized series;

- aggregation of monthly series and calculation of seasonally adjusted quarterly series.

The backcasting was carried out for the Target series by the Bank of Italy, while the POS series has been backcasted for the period from January 2000 to May 2006 using a simple rescaling method. Given the series of total (ATM + POS) available for the whole period January 2000 to November 2017 the ratio POS/(ATM + POS) was calculated for the period from June 2006 to November 2017. The series is strongly characterized by a monthly seasonal component and, therefore, the averages for each month of the POS/(ATM + POS) series were used for the backcasting. These seasonal factors were subsequently multiplied by the values of the ATM + POS series for the period January 2000-May 2006 to obtain an estimated series of POS to be used in following analysis.

Given the strong influence of regulatory changes on the series, in the second step the deterministic effects were estimated using the TRAMO software in JDemetra+ . Once these effects were detected, they were removed through the linearization of the series. The linearized series therefore corresponds to:

$$y_{lin,t} = y_t - X_t\hat{\beta} \tag{2.9}$$

where $y_{lin,t}$ is the linearized series, y_t is the original one, X_t is the (n, k) matrix of k deterministic effects which include the dummies variables related to regulatory events or working days regressors (the number of Monday, Tuesday, etc. in the month), estimated through an Arima model, $\hat{\beta}$ is the $(k, 1)$ coefficients vector. For the BI-COMP series, total and targets, the presence of a ramp effect has been estimated in the period December 2013-March 2014 as effect of a regulatory change that significantly lowered the values of the series in that time span. In general, a ramp effect in a period from t_0 to t_1 is estimated using the following variable:

$$RP_t = \begin{cases} -1 & t < t_0 \\ \frac{t-t_1}{t_1-t_0} & t_0 < t < t_1 \\ 0 & t > t_1 \end{cases}$$

The regulatory changes have also affected the SARA series; it has been possible to eliminate these effects through an intensive use of additive or level shifts outliers. The graphs in figure 2.1 and figure 2.2 show the original BI-COMP series in blue and the respective series linearized in red. As can be seen, the ramp effect has led to a translation of the BI-COMP series (bank transfers, receipts, cheques, total) and targets. In such way it has been possible to adjust the values before March 2014 to the values of the following period (April 2014 -February 2017). The series of card payments (credit cards) was corrected taking into account a level shift at the beginning of the series and the presence of an outlier in June 2015; the same outlier in June 2015 is also present in the POS series.

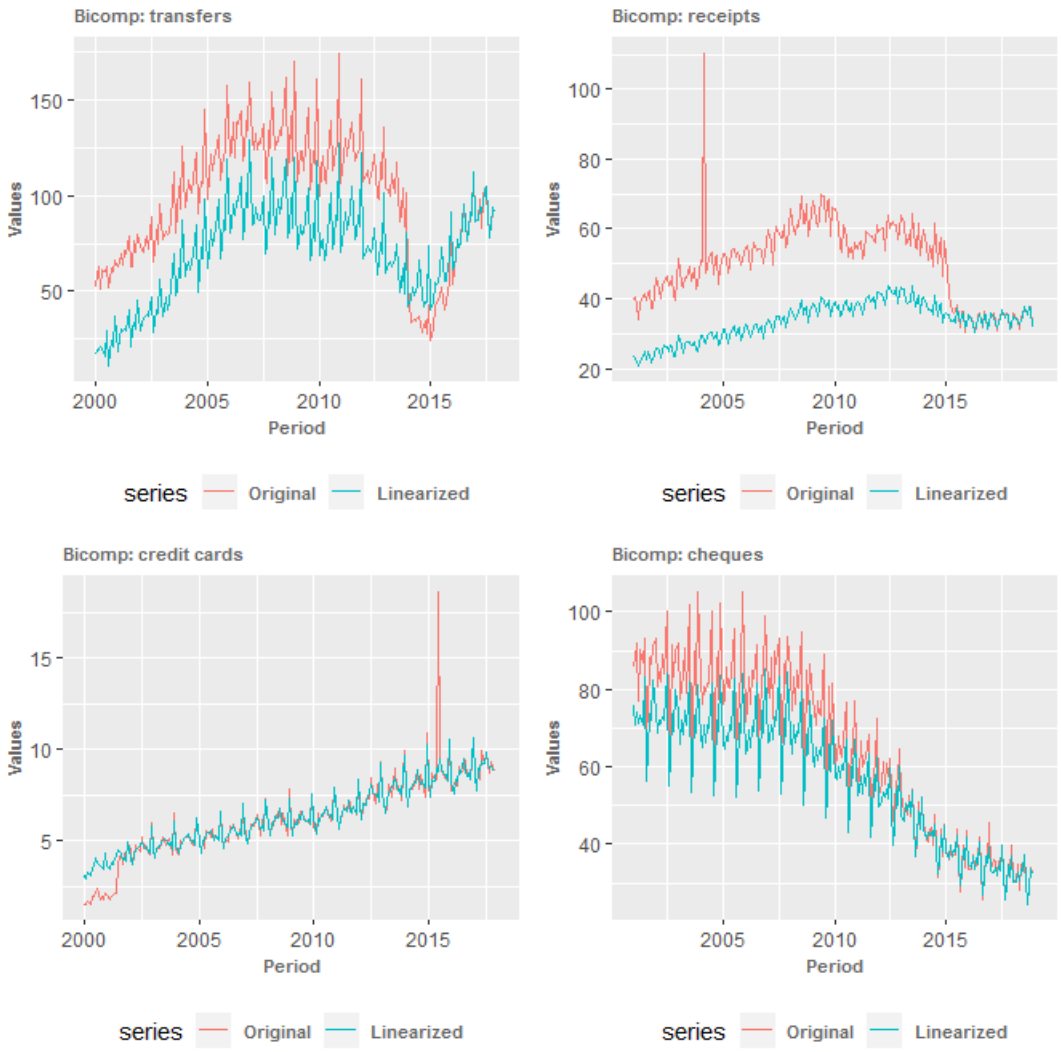

Figure 2.1: BI-COMP component series: original and linearized

Regarding the SARA series the main corrections concern the change in level applied to the wholesale retail and services one, while the Households series is characterize by a large number of outliers, among which the most significant are those of August and October 2002 and May 2003 (Figure 2.3).

Figure 2.2: BI-COMP aggregates and POS series: original and linearized

In the subsequent step a SARIMA model has been estimated using the linearized series and through the SEATS program in Demetra + the series has been decomposed into its trend/cycle seasonal and irregular components. Finally the series have been seasonal adjusted, eliminating the seasonal component. The graphs in Figures 2.4, 2.5 and 2.6 show the result of the seasonal adjustment representing the linearized series in blue and the seasonally adjusted ones in red. All the series of the SARA domain and the BI-COMP series bank transfers, cheques, total and target show a strong seasonality marked by positive peaks in December and with lower values in August. Different seasonality instead is present in credit cards and receipts: the first one has a positive peak both in December and in July, while the second one has a less pronounced seasonality compared to the other components analyzed up to now. As for the series of POS, there is a very pronounced seasonality with a peak in December and a period of decline in February.

Figure 2.3: SARA series: original and linearized

Finally the seasonal adjusted variables has been used in the VAR model to forecast household consumption. Consumption data comes from the quarterly estimates of national accounts in a seasonally adjusted and chain-linked form with reference year 2010. Therefore, the seasonally adjusted data for the series of international payments and anti-money laundering reports were aggregated by calculating their quarterly average. The purpose of the exercise is to evaluate the forecasting ability of these series in terms of improvement compared to the current model used in Istat for short-term forecasts. In particular, forecasts for final consumption are obtained through two phases. In the first, VAR models are estimated on the growth rates compared to the previous period calculated on the seasonally adjusted series of consumption of durable goods and non-durable goods. The models for the two aggregates use the following variables:

- **Consumption of durable goods**: Buying or building a house within the next 12 months, Car Registration, Household purchasing power

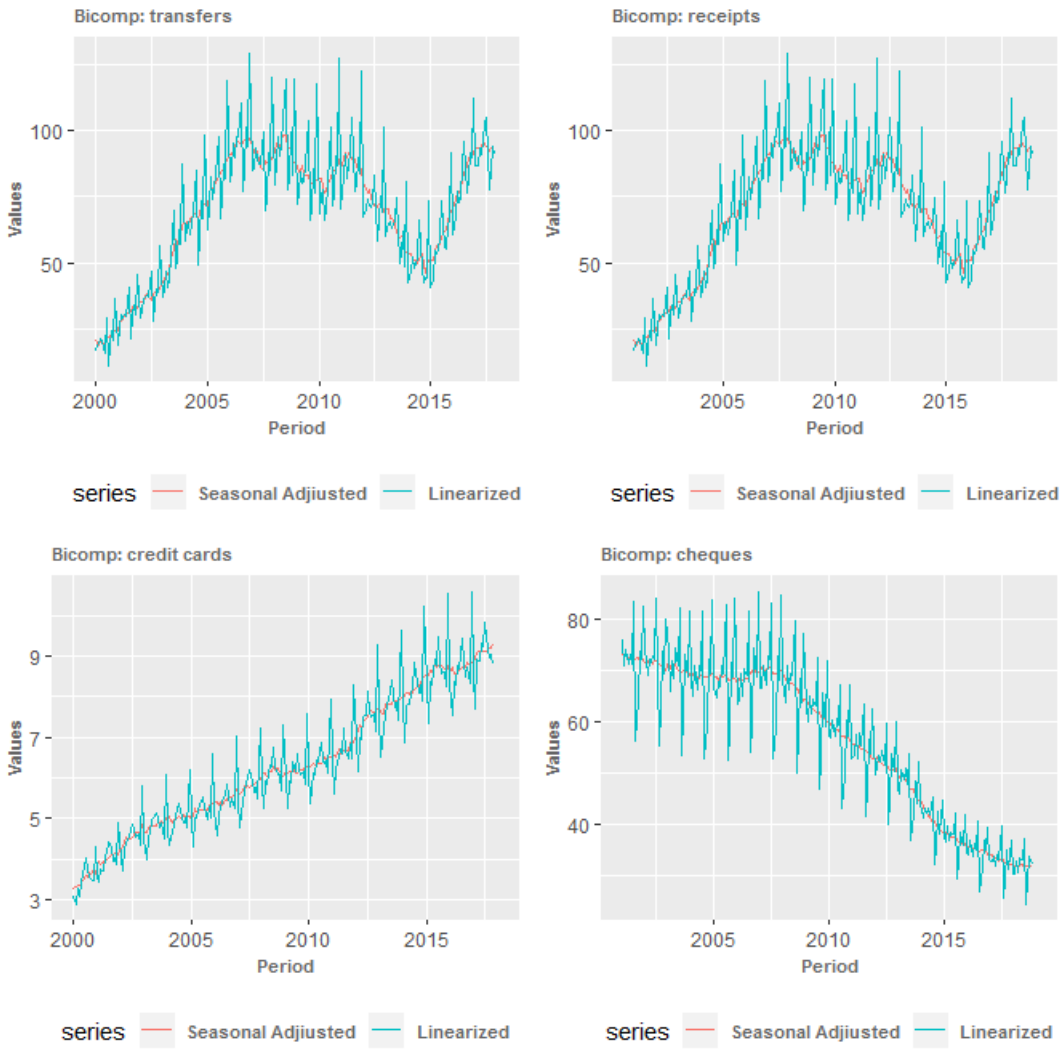

Figure 2.4: BI-COMP components series: seasonal adjusted and linearized

- **Consumption of non-durable goods**: Industrial production of consumer goods, volume of stocks currently held by retail sales companies

In the second phase the forecasts of the two aggregates are combined for the forecast of the growth rate of private consumption. In the previous VAR specification the BI-COMP and SARA series are then added one by one. As first step the contemporary correlation between each BI-COMP, SARA series and the consumption one is evaluated. Figure 2.7 show the correlation matrix with in blue the positive correlations and in red the negative one.

The correlation between the BI series and the private consumption has also been evaluated at different lags (figure 2.8).

Figure 2.5: BI-COMP aggregates and POS series: seasonal adjusted and linearized

This means that for each series the correlation between the private consumption series and the BI series in $t-k$ with $k=0,1,2,3$ has been calculated. From the graphical analysis the most significant correlations are the contemporary and the one between the BI series in t-1 and the private consumption series at time t. In the forecast exercise given the timeliness and correlation characteristic of the BI series, the contemporary and lagged $t-1$ series had been used. The exercise in pseudo real time was conducted as follows:

- the forecasts are one and two steps
- the reference period for the BI-COMP series is $2000:Q1-2014:Q4$, for the SARA and POS series is $2002:Q1-2014:Q4$
- the forecast period instead is $2015:Q1-2017:Q2$

Figure 2.6: SARA series: seasonal adjusted and linearized

The results in table 2.2 show that for the BI-COMP domain, the variables leading to an improvement in the forecasts for private consumption are the cheques variables and the aggregate variable (sum of the total of the BI-COMP variables and the target variable). The improvement, measured as the ratio of the RMSFE and MAFE of the various models estimated on the baseline VAR (used in the Istat procedure), is obtained for both forecasting horizons. The greatest reduction occurs for the aggregate variable (-9.7% and -8.0% in terms of MAFE respectively for $h = 1$ and $h = 2$). For the variables of the SARA domain, on the other hand, the only variable that leads to a reduction in the accuracy measurements of the forecasts is the one which refers to the retail trade for both the forecast horizons in terms of MAFE and only in terms of RMSFE for h = 2 for the presence of higher errors in this latter case.

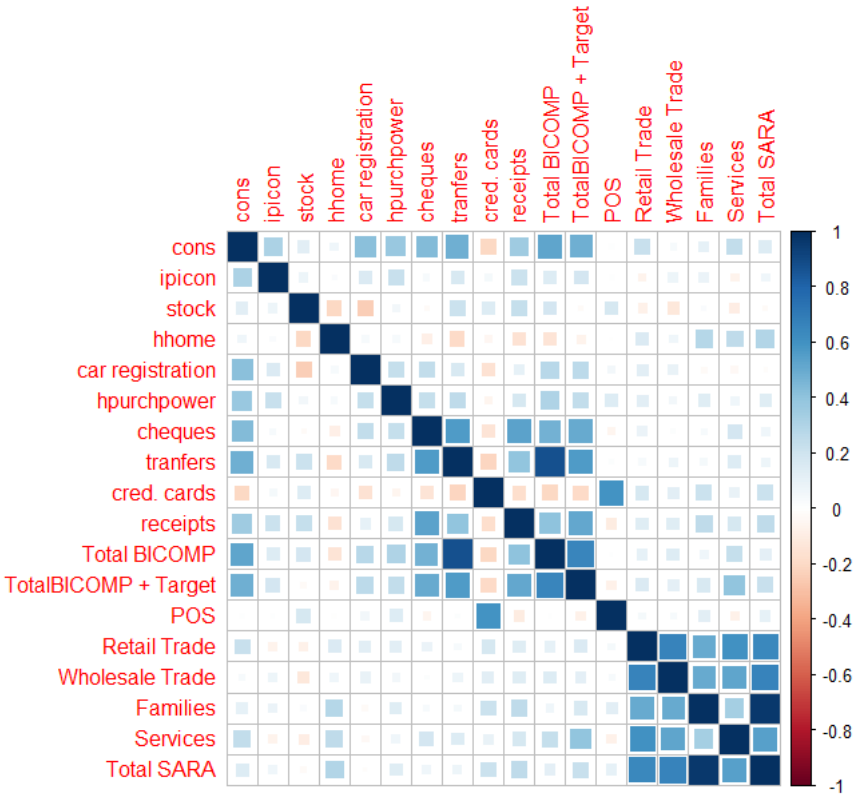

Figure 2.7: Correlation analysis

Table 2.2: RMSFE and MAFE for private consumption forecast

	MAFE			RMSFE	
	h=1	h=2		h=1	h=2
Cheques	-8.5	-3.3		-3.4	-4.1
Transfers	8	0.9		8.7	3.3
Credit Cards	7.5	-0.2		0.6	0
Receipts	10.7	14.4		10.7	13.1
Total BI-COMP	20.1	-6		9.6	-7.8
Target	-9.7	-8		-7.9	-8.1
POS	0	1.9		-1.7	-2.8
Retail Trade	-4.8	-4.6		2	-4.6
Wholesale Trade	5.3	3		3.4	1.2
Families	4.4	-5.7		4.1	-1.8
Services	12.4	-0.9		5	-0.7
SARA Total	10	-4.4		6.5	-2.6

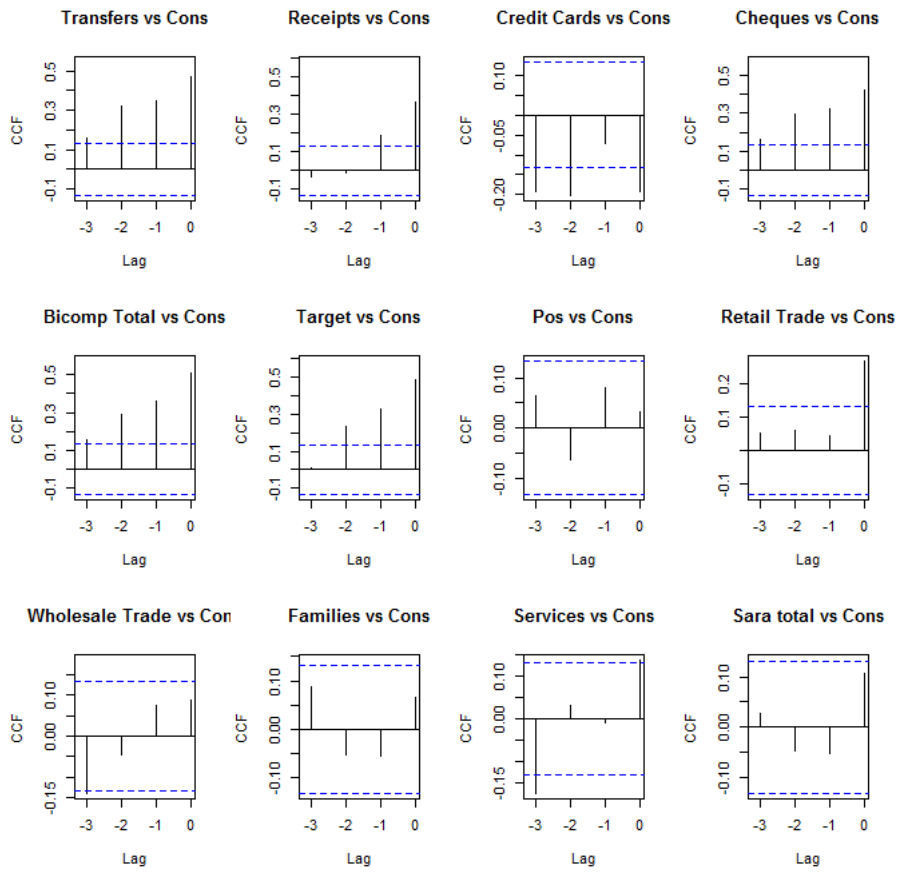

Figure 2.8: Example of a parametric plot

3. Non-traditional data as a primary data source for SDG indicators

3.1. Selection bias of big data sources

This section will focus on the use of big data as primary source to measure the SDG indicators. Chapter 2 of Delivarable 2.1 explores the existing and potential big data sources for the SDG indicators.For example, big data such as from satellite imagery and sensor networks make environment and development indicators increasingly measurable: the *Annual change in forest area and land under cultivation* can be measured with geospatial data; as well as the *Area of forest under sustainable forest management as a percent of forest area*, where also Administrative data can be used. Looking at the SDG Goal 7 on *Affordable and clean energy*, for which, in particular EU has set the indicator *Final energy consumption in households per capita*, we mention the work on big data of the ESSNET BIG DATA with WP3 *Smart Meters* (see for details `https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP3_Smart_meters1`). The WP3 tried to demonstrate by concrete estimates whether buildings equipped with smart meters (electricity meters which can be read from a distance and measure electricity consumption at a high frequency) can be used to produce energy statistics. The deployment of smart meters allows increasing enormously the insights into consumer demand, quality of power and electricity consumption.

Statistical community has immediately recognised the benefits of the usage of big data in many domains, but at the same time has questioned their quality, in this session we will deal in particular with bias that can arise on the estimation based only on big data.

3.2. Preamble

When data used in statistical analyses are not randomly drawn from the target population selection bias may arise. That is, standard estimators and tests may result in misleading inference. The selection bias arises if the probability of a particular observation to be included in the analysis depends upon the phenomenon to be studied; this may occur for a number of reasons:

- Inaccuracy in the frame;
- Non-response;
- Self-selection of respondents/units.

Selection bias is strictly related to the representativeness of the of the data considered for the analysis. Buelens et al. (2014) state: "A subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as selective." Since the selectivity is related to specific aspects (variables), it could be the case that a set of data that is highly selective may nonetheless be useable for some purposes but inadequate for others. The lack of representativeness generates potential source of inaccuracy that are usually referred to as

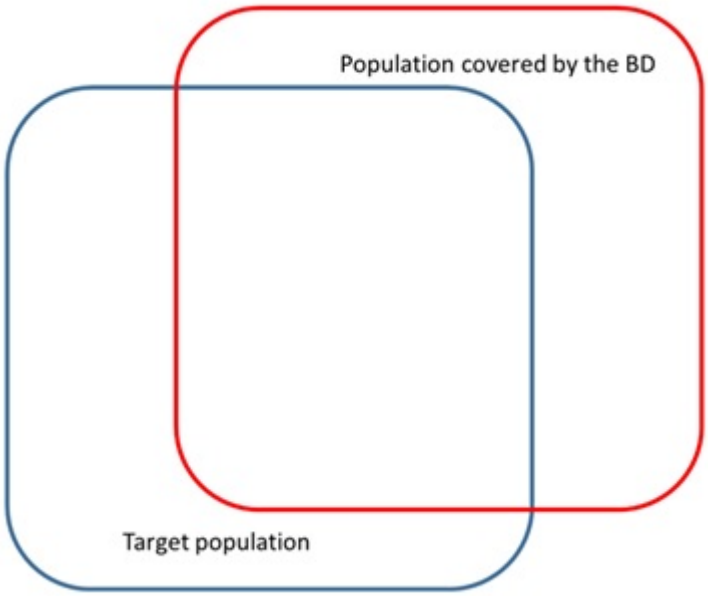

Figure 3.1: Partial overlapping between target population and population covered by the Big Data, under and over-coverage occur.

coverage errors, distinguishing in under-coverage and overcoverage. The former occurs when units belonging to the population of interest are not included in the available data, whereas the latter refers to the situation in which out-of-scope units (including duplicates) are erroneously in the data. Generally speaking, under-coverage and over-coverage may become a key concern if they affect the representativeness of data. Coverage is one of the quality aspects (errors) that affect the accuracy of statistical information. Accuracy is considered as one of the quality dimensions, both for traditional data as survey data, administrative data and Big Data as well (Daas et al. (2009)), Statistics Canada (2002), UNECE Big Data task team (2014).

Generally speaking the accuracy is related to the degree to which the information correctly describes the phenomena of interest and it is usually characterized in terms of error in statistical estimates, traditionally decomposed into bias (systematic error) and variance (random error) components. It is worthwhile noting that the bias introduced by the lack of representativeness (selectivity-bias) compromises the *trust* in statistical results obtained using big data,

The question when a small representative sample may be preferable to a very large non-probabilistic dataset is widely debated in the statistical community (Keiding and Louis (2016), Meng (2018), Pfeffermann (2018), Tam and Kim (2018)). Looking back to our example on the use of smart meters for producing statistics on electricity, grid electricity is supplied by electricity distribution companies who maintain records for billing purposes. However, households may use electricity from non-grid sources thus producing an underestimation when only grids are considered. On the other hand, smart meters may be linked to a business outside the target population. Then, the effect of the selectivity of smart meters might in principle affect the estimates. The figure 3.1 depicts the relationship between the target population and the observed population of smart meters.

3.3. The problem is not new

Selection bias is a concern not only related to big data source and statisticians have faced it from long time. In sampling surveys, a well known cause of selection bias is the undercoverage of the sampling frame, e.g. when samples are drawn from a register of telephone subscribers in countries with low penetration rates for telephone ownership. A second situation producing selectivity is related to the self-selection of respondents in a probabilistic sample affected by non response. Self-selection may also occur with non random surveys, such as opt-in web surveys where respondents may benefit from being included in the survey. See for example Vehovar et al. (2016) . Several proposals have been developed to overcome the bias caused by the selectivity in the above-mentioned situations. The methods can be broadly classified as design based and model based methods:

- Design based
  - (Re)weighting and calibration (including model calibration)
  - The Propensity Score method (Rosenbaum and Rubin (1984)), quasi randomisation techniques, pseudo weights (Puza and ONeill (2006), Tam and Kim (2018), Elliott and Valliant (2017))
  - Benchmarking
  - Sample Matching (Lavallée and Brisbane (2015)) (Rivers and Bailey (2009))
- Superpopulation approach-Model-based estimation
  - Two-steps model parameters prediction (Heckman (1976))
  - Finite population paramenters' prediction (e.g. see Sverchkov and Pfeffermann (2004), Pfeffermann and Sverchkov (2009))
  - Empirical likelihood (e.g. see Feder and Pfeffermann (2015))
  - Semiparametric estimation (Heckman (1990))
  - Bounds on the distribution function (Manski (Manski))

The methods can be also grouped in methods at unit level or domain level according to the type of information is needed for their application.

A short overview of methods in sample surveys can be found in the AAPOR task force final report on non-probability samples AAPOR (2013). The report by (Beresewicz et al. (2018)) gives an overview of the methods and reports some useful applications. See the following section 3.5 for a short summary of the above mentioned methods. Simulation of some methods to correct selectivity are in Buelens et al.

(2015). They used data-generating mechanisms mimicking those often assumed in big data situations. The authors demonstrated the importance of strong auxiliary data that is capable of explaining the data-generating mechanism for a successful adjustment for the selection bias.

3.4. Formalisation of the problem

Let us indicate with $f(Y_i|X_i)$ the probability distribution function of Y given X, S_i be an indicator of the inclusion of the unit in the source or 0 otherwise. Following Rubin (1976) (see also Little and Rubin (1987)) for the mechanism of non-response, we can define the mechanism of inclusion of an unit in a source as ignorable when

$$f(Y_i|X_i, S_i = 1) = f(Y_i|X_i).$$

The ignorability can also be expressed as

$$P(S_i = 1|X_i, Y_i) = P(S_i = 1|X_i)$$

In official statistics often we are interested in the expected values and we can then relax the requirement to

$$E(Y_i|X_i, S_i = 1) = E(Y_i|X_i).$$

When the inclusion in the data is not ignorable then the expected value of the observed data will differ:

$$Bias = E(Y_i|S_i = 1) - E(Y_i) = (1 - p)(E(Y_i|S_i = 1) - E(Y_i|S_i = 0))$$

where p is the probability of inclusion in the data. Meng (2018) has given the following form of the bias raised by selectivity (without explicitly including the auxiliary information X_j in the notation for the sake of simplicity):

$$\frac{E(S_jY_j) - E(S_j)E(Y_j)}{\sigma_S\sigma_Y}\frac{\sigma_S}{E(S)}\sigma_Y$$

Following this reasoning, Meng (2018) identifies three components for the total error:

$$Error = \rho_{SY}\sqrt{\frac{N-n}{n}}\sigma_Y$$

where the first component ρ_{SY} represents the "'data quality"', the second component $\sqrt{\frac{N-n}{n}}$ represents the "data quantity" and the third component σ_Y represents the "'problem difficulty" inherent to the study of the target variable Y . The focus of this section, the data quality, is captured by the data defect correlation ρ_{SY}, that precisely measures both the sign and degree of selection bias caused by the selection mechanism. The factorization of the total error in these three components allows mainly showing the role of selection bias in evaluating the Big Data sources.

3.5. A short summary of methods for selection bias correction

This section reports a very short description of the most common methods for correcting selection bias. The main references are provided for an exhaustive description and further details. As in subsection 3.4, the methods are mainly classified by design-based vs model-based approaches.

(Re)Weighting and calibration This class of methods consists in the application of the standard weighting also to account for the *non-response* or self-selection mechanism (Särndal and Lundström (2005), Haziza and Lesage (2016), Wu and Sitter (2001)).

Quasi randomization In the quasi-randomization approach, pseudo-inclusion probabilities are estimated and used to correct for selection bias. Given estimates of the pseudo-probabilities, design-based formulas are used for point estimates and variances. In this setting, the propensity score is defined as the conditional probability of receiving treatment given the vector of the individual' s covariates and is calculated for each individual. It is often estimated in a logistic regression model. A key issue is how the probability are evaluated. The methods generally require information on nonsample units. One approach is to use a reference survey in parallel to the nonprobability survey. The statistical approach is to combine the reference sample and the sample of volunteers and fit a model to predict the probability of being in the nonprobability sample. For further details on the estimation of the pseudo-weights see Elliott and Valliant (2017). Lee and Valliant (2009) derived a combination of propensity weighting and calibration weighting to account for the bias in volunteer panel web surveys. In the resulting two-step method, the design weights are first adjusted by propensity scores to correct for selection bias due to non-probability sampling, and the adjusted weights are then calibrated to auxiliary variable totals for the target population in order to adjust for coverage bias. Tam and Kim (2018) derived an estimator using pseudo-weights for the special case of binary variable Y (estimates of proportions) that depends on the ratio r of the probability of inclusion in the source conditional on $Y = 1$ over the conditional probability of inclusion given $Y = 0$; the result obtained is the same as in Puza and ONeill (2006). To estimate the ratio r and its standard deviation, a random sample of the target population is required, to jointly observe the target variable Y and the inclusion in the source.

Sampling matching Sample matching is another approach to attempting to reduce selection biases in a nonprobability sample. Matching can be at either aggregate level or individual level. The former consists simply on making the frequency distribution of the nonprobability sample the same as that of the population. Sample matching at individual level consists in matching individuals in the non-proability sample with the individuals in a probability sample on the basis of covariates available in each dataset Rivers (2007); this approach belongs to the class of statistical matching (D'Orazio et al. (2006)). This may be done based on individual covariate values or on propensity scores as described in Rosenbaum and Rubin (1983). Sample matching is similar to nearest neighbour imputation and the estimators consists simply in applying the survey estimator on the probability sample with the target variables observed in the big data source. Lavallée and Brisbane (2015) suggests using GWSW to further expand the scope of sample matching In fact GWSM was initially proposed to deal with the weighting procedure in longitudinal surveys. The problem that the GWSM aims to solve is specifying weights when statistical units change with time, the methods cna be applied in indirect sampling, a setting that might be useful when dealing with secondary sources where the records are not the target statistical units. Lee applies sample matching to correct selectivity in prediction of election results. They noted that due to imperfect matching, it may still be necessary to weight the sample after matching, even if the

ratio of the pool of available observations to the target sample is small, especially if the matching ratio is less than five.

Superpopulation approach, Model-based approach In the superpopulation modelling approach, a statistical model is fitted for a Y analysis variable from the sample and used to project the sample to the full population. The general idea in model-based estimation when estimating a total is to sum the responses for the sample cases and add to them the sum of predictions for nonsample cases. The key to forming unbiased estimates is that the variables to be analysed for the sample and nonsample follow a common model. When the selection is not ignorable, an example of model-based adjustment is proposed in Heckman (1976). The method consists in a two-step estimator based on modelling the selection mechanism and the target variables into two different steps. A crucial assumption is the identifiability of the models. This requires that the two models are dependent at least by one different variable. Heckman (1990) proposed a semiparametric estimation of the sample selection model to relax the normality assumption of the regression errors. Empirical applications are not straightforward. See B orsch-Supan and Winter (2004) for an application for an on-line survey. The problem of identifiability of the estimation of regression in the presence of a *non ignorable* sample selection mechanism is dealt by Manski (Manski) by posing bounds on the conditional support of $f(y|xI = 0)$ or bounding the effect of x on the expected value of Y conditional on X and $I = 1$.

In the presence of informative sampling Sverchkov and Pfeffermann (2004) proposed to adjust the likelihood to base the inference of finite population totals on sample distribution that is a corrected version of the likelihood with sample inclusion probabilities (Pfeffermann et al. (1998)). See Pfeffermann and Sverchkov (2009) for different approaches to modeling and estimating the expected inclusion probabilities conditional on the target variable y and covariates $x's$ In order to stabilise the ML estimation, Feder and Pfeffermann (2015) propose the use of the empirical likelihood paradigm under the population model; they combine it with a parametric model for the response probabilities that contains the outcome variable as one of the covariates and finally they estimate the expectations of the unit weights (inverse of the inclusion probabilities) nonparametrically using kernel smoothing .

Benchmarking- domain level correction approaches The domain-level approach refers to models that assume data aggregated at a certain level. The term d̈omain" does not only mean spatial aggregation (e.g. LAU 1), but also cross-classifications (e.g. cross-classifications by sex and labour status).

We also mention the possibility of combining estimates from a probabilistic sample with estimates from a secondary source, either at aggregate level or at individual level. See Barcaroli et al. (2018) for a comparison of the methods for adjusting estimates based on webscraping combining them with survey data.

Kwang Kim and Wang (2018) propose inverse sampling from big data source to correct its bias. Importance weights are associated to the records of the big data source according to a know auxiliary

variable, a solution is proposed also for the case when only population means of the covariates are known. A sample is then selected proportionally to the importance weights; simple average of the sampled records will provide unbiased estimates of the unknown population mean.

4. Evaluating sustainability through an input-state-output framework in Italy

4.1. Genesis of the I-S-O framework

Every non isolated system (e.g. a cell, a tree, an ecosystem, a person, a city, a production process, a national economy) needs a continuous flow of energy and matter to survive and develop; this flow of resources must be processed by internal elements that form the structure and self-organizing mechanisms of the system; finally, a flow in the form of useful products, but also emissions, wastes, degraded matter and energy, and heat is produced by the system. This is the way in which biological systems and ecosystems behave and, for this reason, a simplified framework has been introduced by Coscieme et al. (2013) to explore the main characteristics of living system/ecosystem dynamics under this multidimensional viewpoint. The system under study can be in fact represented by a so-called Input-State-Output (I-S-O) scheme to summarize the three crucial phases of system dynamics: resource collection (input); resource processing by means of organization (state), and generation of products or services (output).

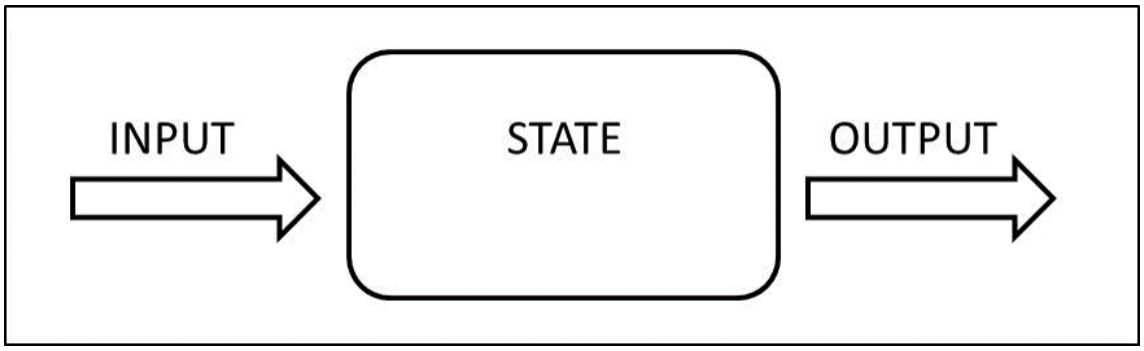

Figure 4.1: the I-S-O framework

Ecosystems can be viewed as thermodynamic systems, open to energy and matter, that self-organize towards higher complexity and organization, create order, and self-maintain far from thermodynamic equilibrium. Furthermore, within a socio-ecological context, that includes the relationship between human activity and the environment, we can see the ecosystems as providers of goods and services that human continuously use independently of market inclusion: these goods and services are commonly defined as 'ecosystems services', which enormously contribute to human welfare (their value at the global level was estimated to be at least 1.8 times the global GDP, see Costanza et al. (1997)).

An Input-State-Output scheme has been used to describe ecosystems (in a socio-ecological context), whose characteristics can be described by the relationship among the three orientors – emergy, eco-exergy, ecosystem services – making it clear that inputs are used up, directly or indirectly, to create and maintain a given system state and/or to produce services in output. In that case we identified orientors as follows:

- *Emergy* (Odum, 1996), is an indicator able to account for the convergence of matter and energy into a system on a common basis (e.g. solar energy), enabling us to quantify and weigh the inputs that feed the system during its evolution. In brief, the emergy flow to a system is representative of the sum of resources feeding it per unit time;

- *Eco-exergy* is a measure of complexity in ecology, as expected to be associated with the presence of more complex organisms, which, in principle, correspond to higher information content (in the form of DNA, RNA, and protein sequences) and greater distance from thermodynamic equilibrium (Jørgensen and Mejer, 1981, Jørgensen, 2008). In brief, eco-exergy reflects the component and the structure of a system;

- *Ecosystem services* are defined as 'the benefits people obtain from ecosystems' (MA, 2005). These benefits can be viewed as 'ecological functions of value to humans' (Fisher et al., 2009). They depend on ecosystem functions and biodiversity but also on users' needs. In brief, the ecosystem service value is a measure of the output of the system.

If we consider a 3D space, determined by three axes, corresponding to the three orientors, a point is identified by the combination of the three measures. In other words, the point (the system under study) is the result of the combination of the values of indicators but the presence of three axes enables not to lose information in a unique aggregate index. The choice of measures/indicators is crucial because it influences results and consequent interpretation. The use of systemic/holistic indicators (eco-exergy to emergy flow ratio) is appropriate because it allow us to understand if the system under study is globally following a path that will take it to a 'better' or to a 'worse' state. On this basis, a categorization of ecosystems has been provided (result illustration of that research goes beyond the scope of this section; for details see Coscieme et al. (2013). In general, different ecosystems have a very different translation capacity of emergy (inputs) into eco-exergy (structure, organization), as well as of eco-exergy into ecosystem services (outputs). That analysis helped identify where, in the input-state-output chain, there can be concerns about an ecosystem's proper functioning and ecosystem service provisions. In other words, it is possible to detect if the system lacks energy/matter inputs, and/or has structural/functional problems, and/or if it can be managed differently concerning the utilization of ecosystem services.

The I-S-O framework is thus a useful tool to consider and evaluate multidimensional aspects of system behaviour using a limited number of indicators. Moreover, as already stated by Fath et al. (2001) and Jørgensen et al. (2007), among other authors, the use of a plurality of 'goal functions', highlighting their complementarity and interdependency, is able to properly detect the dynamic behaviour of complex systems, such as ecosystems, but also human-driven systems like social and economic systems, and a larger system that comprises both in a new emergent network of relations.

4.2. Application of the I-S-O framework to assess sustainability: a 3D representation

Sustainable development 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). Several

representations of sustainability have been produced in recent years, including the traditional one of three intersecting spheres (economic, social and environmental), the intersection of the three circles being where sustainable development is realized (Barbier, 1987). A more advanced representation (Pulselli et al., 2015) is the pyramid in which the mutual relationships among the three dimensions of sustainability is represented (Figure 2a). The base of the pyramid represents the natural assets, which form the crucial inputs to the system; the intermediate level can be viewed as the state of the system, specifically the society and its organization and structure; the top level of the pyramid, is the real economy of the system, that should produce the 'useful' output of the system (Figure 4.2a). If we rotate the pyramid clockwise, we can orient the succession of the stages and confer a logical, physical, relational and thermodynamic order (i.e. environment–society–economy) to the representation (Figure 4.2b): a flow of material and energy inputs, generated by the available stock of Natural Capital, feeds (is captured by) the system. These resources are necessary for the elements of the system (namely, the society and its organizational units) to operate (act, live, survive); the level of organization of the society influences the degree of utility/satisfaction derived from processing/using/consuming resources. An organized society is supposed to be able to achieve better economic results providing outputs from its productive processes. The pyramid can be actually and immediately translated into an I-S-O scheme (Figure 4.2b). Here different combinations of indicators can be used to account for the energy and matter inputs to a system, describe the state organization, and quantify the outputs of the system.

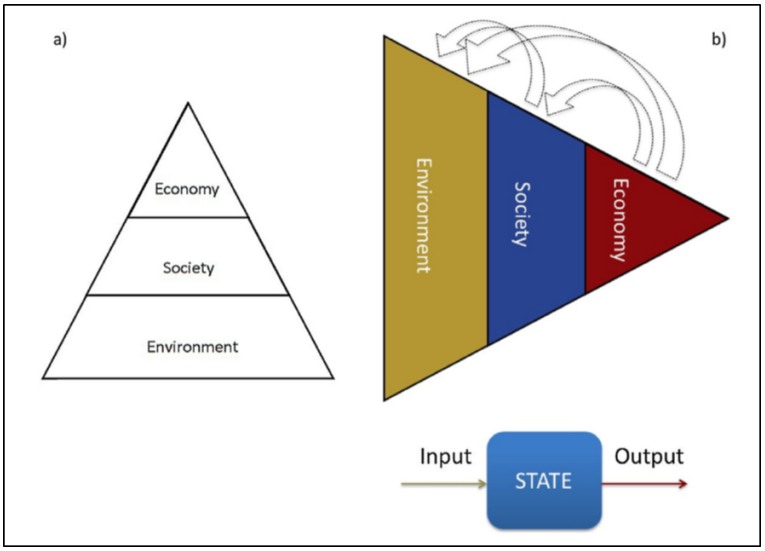

Figure 4.2: A three-storey pyramid representation of sustainable development recognizes a relational and physical order of environment, society and economy. It resembles an input-state-output diagram to investigate economic systems. Feedbacks between compartments are also shown.

The I-S-O framework can be thus successfully applied to investigate economic systems (e.g. national or regional economies) regarding their level of sustainability (see also Bastianoni et al. (2014b)). In general, each part of the input-state-output model should be described with the most appropriate indicator relative to the kind of information needed to assess the various aspects of open system functioning. Regarding national economies, for example, the choice of proper indicators may help identify

emergent properties and the relationships among the three dimensions of sustainability. Improvement and information refinement may derive from data availability (that depends on the type of indicator), statistical computation of indicators and measures (e.g. rankings, quartiles, scales), and data aggregation (e.g. cluster analysis). Several combinations of indicators can be adapted to this structure supporting the integration of different disciplinary approaches. In this case, the three dimensions are not simply juxtaposed, but the logical structure of the pyramid shows how the three compartments work together through relations, interactions, feedbacks, etc.

The input-state-output scheme can be developed by means of a 3D representation deriving from a three-axis diagram, in which the three dimensions are simultaneously, but separately, considered to identify points resulting from the combination of values, without losing information. The indicator values are distributed along three axes, occupying three segments equal in length, respectively, in which the median value is identified to separate low and high domains. The median values are forced in the middle of the segments, in such a way that 8 sub-cubes can be determined to facilitate categorization of systems on the basis of different characteristics (Figure 4.3).

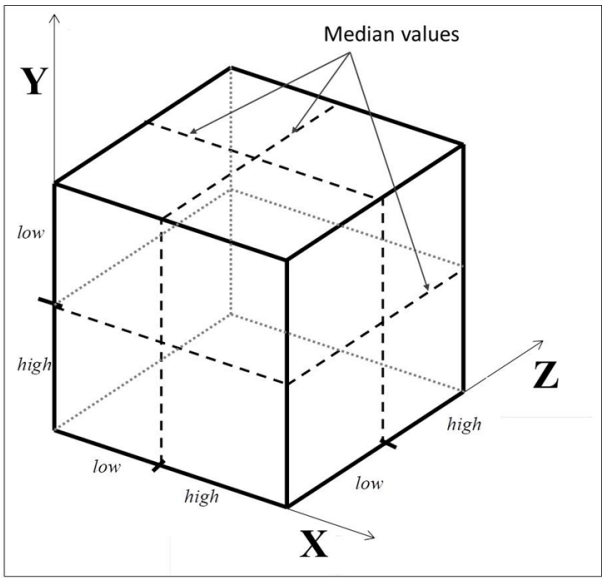

Figure 4.3: A cubic representation derives from a three-axis diagram. Median values of variables X, Y and Z are forced in the middle of the segments. In this way, 8 sub-cubes can be used to categorize different combinations of indicator values.

In sum, considering the current need to identify a system for measuring sustainability, the I-S-O framework enables to represent and monitor sustainability with a trade-off aiming at maximizing information with the minimum number of indicators: the information should be obtained by using indicators representative of the whole system; their number is kept to the minimum to independently depict the three different dimensions of system sustainability, ensuring that each indicator maintains its identity, and complementary informative capacity (Pulselli et al., 2015).

4.3. I-S-O framework application at the national level

In a paper titled 'The world economy in a cube', Pulselli et al. (2015) categorized 99 national economies using the I-S-O framework. To measure the three dimensions, the following indicators have been chosen:

- *Emergy flow per capita* as an input-based indicator: it accounts for all the resources, coming directly and indirectly from the environment, that feed the national system, all expressed in a common unit (solar energy).

- *Gini index of income distribution* as a descriptor of the organization of the state of the economic system: though it is an economics-based measure, it reflects inequality which is crucial to assess the state of a society. In fact, greater income inequality has proven to be related with declining social capital, worsening health status of the population and decreasing chances of moving up the social ladder (Wilkinson and Pickett, 2009, 2018).

- *Gross Domestic Product (GDP) per capita* as a measure of the total economic output: once converted into international dollars using purchasing power parity rates (GDP, PPP), it can be used to make rightful comparisons among national economic performances.

The values of the three indicators are placed along the three dimensions of the cube in Figure 4.3. A number of points, corresponding to countries, can be identified within the 8 sub-cubes, in line with the high or low domain of the indicator values (above or below the median value, respectively). The result are represented in Figure 4.4: the four yellow sub-cubes are almost completely empty, in the others the units (countries) are distributed. The sub-cubes are labelled with nicknames to highlight their main characteristics. For example, the Dissipative space hosts countries that consume resources, are highly organized (equal) and produce GDP. The opposite condition is that of Unevenly Frugal, environmentally and economically poor, with a bad income distribution. More than 55% of the analyzed countries are located in these two groups, representing worldwide inequality under different viewpoints. Evenly Frugal means poor but more equal than other countries. Socially Distracted are those countries in which large material and energy flows are converted into GDP, without paying much attention to society. Empty cubes are ineffective, inconsistent or impossible to realize (e.g. building wealth from nothing), Dematerialization is a kind of modern economic tendency, nevertheless, it is not realized yet because production and distribution of an output is not possible without a physical support.

From the cross-country experiment proposed by Pulselli et al. (2015), the following considerations may be summarized:

- In general, a strong direct relationship between resource use per capita and GDP per capita can be noted; countries presenting concordant values for input and output (high-high or low-low) are 85 out of 99, pointing out how economic growth drives, and depends on, an increasing requirement of energy and matter to be transformed by the economic system.

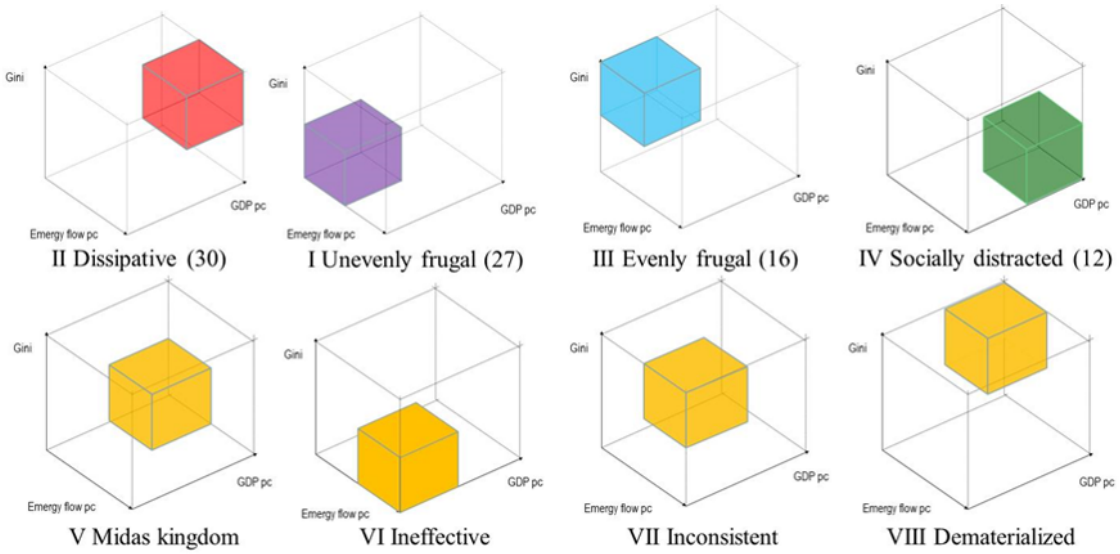

Figure 4.4: Full and empty sub-cubes. Nomenclature has been given on the basis of high and low domains of the three indicators. In brackets the points within each sub-cube. (Source: Pulselli et al. (2015)).

- Independently of per capita emergy and GDP values, we may have very different levels of income distribution within societies. It is worth to highlight that 49 countries out of 99 present high value for the Gini Index, for different combinations of the others indicators. Overall, to successfully manage societal parameters, a punctual political intervention agenda in the social sphere is needed. The abovementioned study by Wilkinson and Pickett (2009) states that social diseases may manifest in both poor and rich countries. Moreover, the 'socially distracted' nations risk to behave like machines designed to transform inputs into outputs without taking care of the members of their communities.

- Inequality does not only emerge as unfair distribution of income, but also as difference in resource availability, opportunities and development possibility. In fact among the 49 countries presenting high Gini index values, 39 show low figures for the input indicator.

- In contrast with many neo-liberal prescriptions that see dematerialization as a path to sustainability, our cross-country results suggest that there may be physical limits for the dematerialization process, which, at least, has not yet happened. Just 2 out of 99 countries belong to the dematerialized sub-cube.

The definition of a partition of the cube in sub cubes, is based on a threshold value, introduced to discriminate between high and low domains for each indicator of the triad. This threshold is artificially forced in the middle of each segment in order to have the three axes/segments (the dimensions of the cube) divided into two equal domains, high and low meaning greater or lower than the median, so that 8 sub-cubes characterized by different combinations of the indicators are visible. The threshold can be considered as a double drawback of the ISO framework, because it is a crisp classification based on an arbitrary threshold (the median in this analysis). This approach forces countries having

values for at least one indicator below and above the median. to belong to different sub cubes. For example, the countries included in the sub cube 'Midas kingdom' and 'dematerialized' such as Greece, Romania, South Africa, Mexico and Brazil show values of emergy per capita and GDP per capita that are not significantly different from the median values, as can be noted in Figure 4.6. In order to overcome the drawback of the arbitrary threshold, a more statistically relevant approach based on the concept of dissimilarity among different countries (instead of simply discriminating between high and low domains), namely, the cluster analysis, has been introduced to categorize national economies according to the I-S-O framework rationale (Pulselli et al., 2015, Neri et al., 2017). Four clusters have been identified. This choice is consistent with the four 'usual' configurations in the cube representation; moreover, no cluster with these unusual input–output configurations has been found. It is worth to highlight on a peculiar cluster, the one including countries with values around the median for the three indicators: the 'environmentally, socially and economically median group of economies' is identified and isolated in a specific cluster. This group of countries cannot be captured by the classification in sub-cubes, because, using the median value as the only classification threshold for each indicator, each country can be characterized only by low (below-median value) or high (above-median value) level of each indicator. Indeed the countries belonging to this cluster were scattered in the 8 cubes.

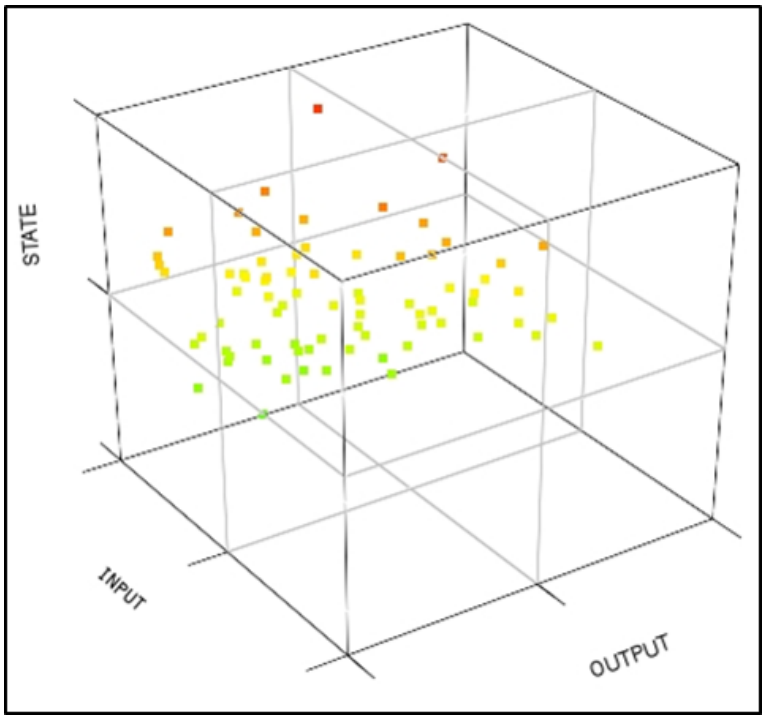

Figure 4.5: Example (A) Visualization of point distribution (countries in the 3D space - the cube)

4.4. I-S-O framework application at the sub-national level: regions and provinces of Italy

The I-S-O framework can be a useful tool for investigation and policy making at the sub-national level. In line with the availability of data, regional and provincial analyses can be performed. A regional application of the I-S-O framework, using the same indicators of the cross-country analysis presented above has been proposed in Italy. The results highlight the diversity among Italian regions and show that diversity among areas within country boundaries emerges in various forms which are

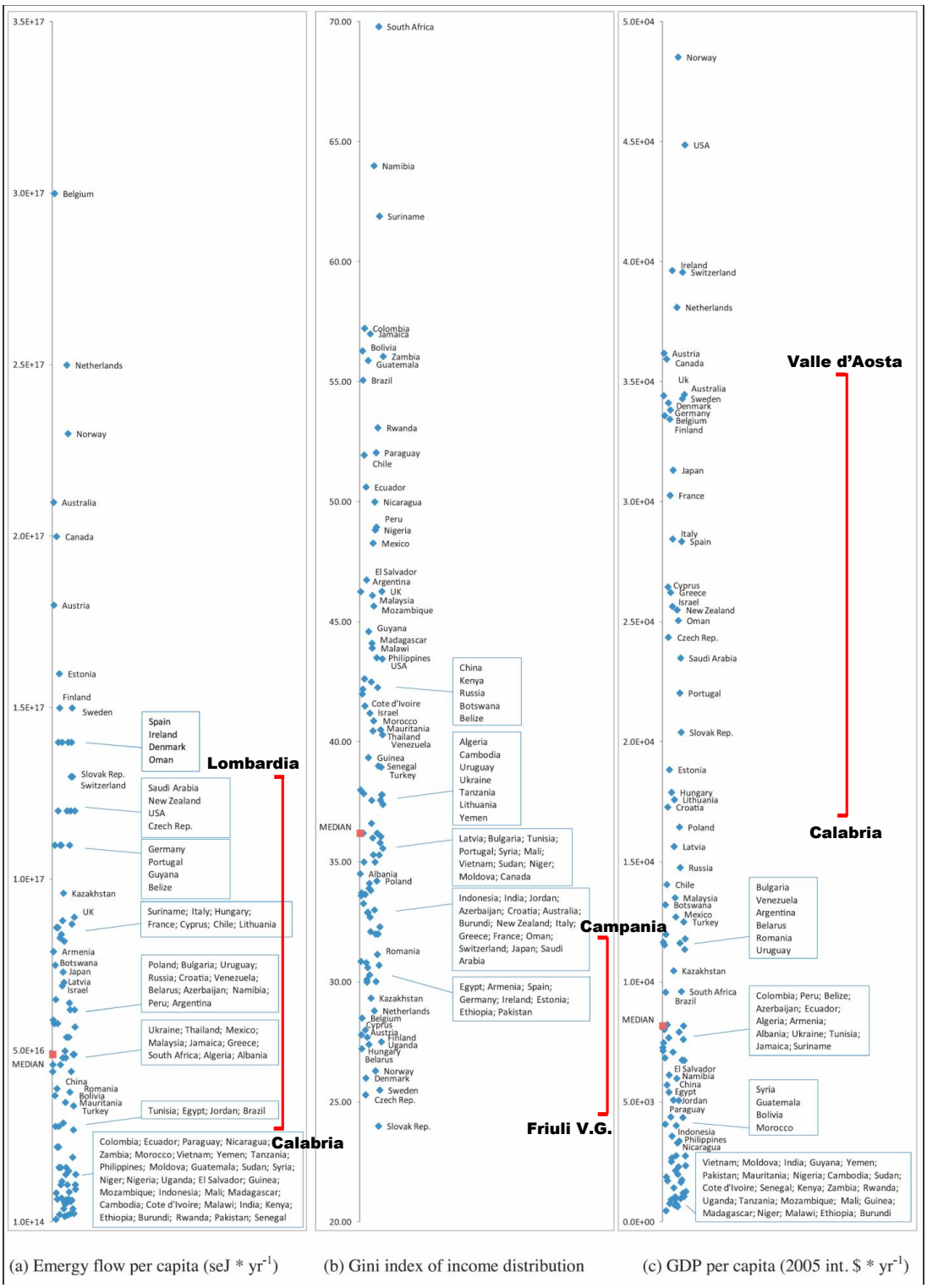

Figure 4.6: Example (B) Representation of the range of Italian regions relative to the range of 99 countries of the world (source: adapted after Pulselli et al. (2015)) .

detected and possibly measured. It is therefore crucial to meet the need of information and tools for a central government to administrate peripheral areas in a sustainable way, especially due to great diversities that often emerge in the three spheres of sustainability. An investigation at the provincial (sub-regional) level in Italy corroborates that point. An I-S-O framework has been built, based on three indicators: energy consumption (electricity and a set of fuel types, converted into CO2 to be aggregated and divided by the area of each province) as input indicator; the employment rate as state indicator; GDP per capita as output indicator. In order to produce a reasonable 'objective' classification of the Italian provinces in terms of the three aspects of sustainability, also overcoming the drawback of a crisp classification, a fuzzy cluster analysis has been conducted: so that each point (province) may belong to two or more clusters with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several clusters are not forced to fully belong to one of the cluster, but rather are assigned membership degrees between 0 and 1: being 0 if the data point is at the farthest possible point from a cluster's center and 1 if the data point is the closest to the center.

In the analysis a problem occurred that frequently happens in real data analysis, the presence outliers. Such a subset, that may be referred to as noise, tends to disrupt clustering algorithms making difficult to detect the cluster structure of the remaining domain points. According to Davé (1991) it can happen that the first k standard clusters are homogeneous, whereas the noise cluster, serving as a 'garbage collector', contains the outliers and is usually not formed by objects with homogeneous. In the present analysis the 'garbage collector' include just one province, Milan, presenting a level of CO2 emission per area more than ten times the average values of all the other provinces and GDP per capita which is nearly double than the average values of all the other provinces. Cluster composition (see Figure 4.7) suggests a high heterogeneity among clusters emphasizing the existence of economic disparities. The analysis conducted confirm the well-known dualism, resulted in a North-South divide in GDP per capita and in labor-market performance, adding a new element: the Southern Italian provinces are homogeneous with respect to the considered characteristics whilst the Northern and Central provinces are not homogeneous even if they belong to the same region. This result suggests that local policies can be better aligned and tailored to specific local opportunities and challenges.

4.5. The added value of the ISO framework

Overall, this approach represents a useful and comprehensive systemic tool for the assessment of country performances. The pure economic representation (based on GDP, profit, cost-benefit logic, convenience, etc.) is the most common to identify the systems, but it is often an over-simplification that is limited to only one aspect of the reality. The I-S-O framework is a 'beside-GDP' monitoring system that considers the informative capacity of various aspects of system behaviour, also providing a synthetic picture of the reality. The ability of this framework to evaluate sustainability can be maximized in the case of dynamic (time series) analyses at both the macro- and micro-system level.

Indeed, the temporal dimension should become crucial in monitoring sustainability. The evolution of a single point (i.e. a country) in the diagram can be monitored when it moves from one to another region of the space within the cube or from a cluster to another, characterized by different input-state-output values. One can wonder whether an optimal trend for a single country exists and must be followed, or exogenous factors, like the economic crisis or the progressive exhaustion of a non-renewable resource,

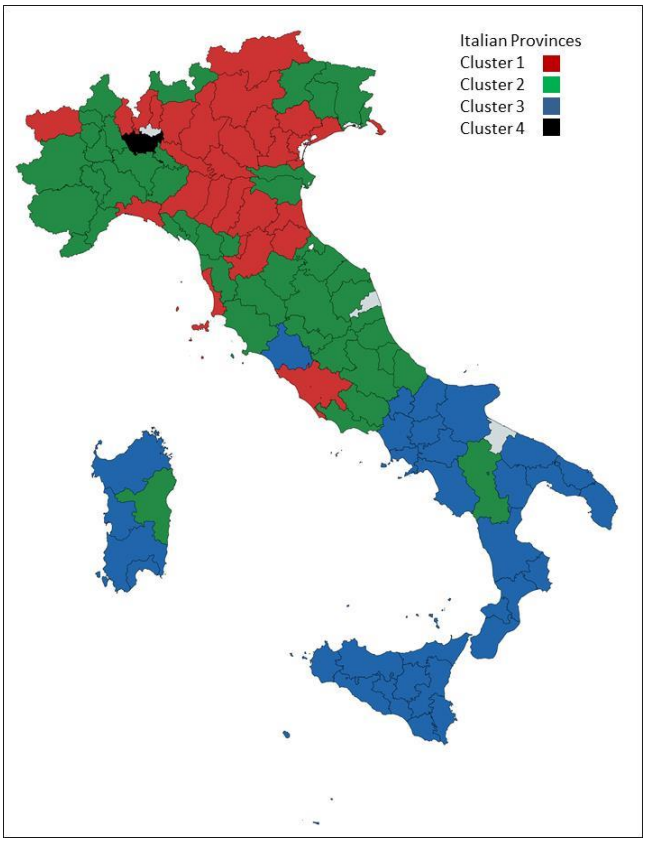

Figure 4.7: Example (C) Clustering and categorization of Italian provinces based on the I-S-O framework and the values of the three indicators adopted. White provinces are Monza-Brianza, Fermo, Barletta-Andria-Trani, not involved in the analysis for unavailability of data.

influence the position of a nation.

The ability of this framework to evaluate sustainability can be maximized in the case of dynamic analyses (see Bastianoni et al., 2014) at both the macro- and micro-system level: policy makers can use this monitoring system as an orientor to pursue sustainability programs. Moreover, this tool can be useful to evaluate the effects of national or sub-national policies in economic but also in social and environmental terms. Further advancements can be also proposed. To give an example, we may refer to the level of interdependency of provinces: in other words, a province with heavy industry will have high environmental impact (high resource consumption or CO2 emission) compared to a province with mainly service oriented activities. Although the latter consumes products produced by other regions, it remains invisible in the input indicator. Such a case would deserve further discussion for a number of reasons including delocalization, inter- and intra-regional (or provincial) distribution of resources and, especially, the degree of responsibility we can assign to every subject involved in the analysis. In particular, the use of indicator values obtained following a consumer-based approach (responsibility assignment based on consumers rather than producers of goods) may greatly change the results of the analysis (Bastianoni et al., 2004, 2014a, Caro et al., 2014).

Finally, the I-S-O framework also enables visualization of results in different ways: the following examples derive from the abovementioned studies. (A) Distribution of points, this visualization (Figure 4.5) can be static (cross-country analysis) or dynamic (time series analysis to study trajectories of points); (B) Regional vs. national diversity (as in Figure 4.6); (C) Map showing categorization of provinces according to cluster analysis (Figure 4.7).

5. Discussion

Official statistics produced by national statistical institutes are traditionally based on sample surveys in combination with design-based inference modes. Over the last decades an increasingly amount of alternative data sources become available, which resulted in the question how these data sources can be used in the production of official statistics. In this report two approaches are considered to use these data in measurement frameworks for well-being and sustainability. The first approach, reported in Chapter 2, is based on model-based inference procedures and combines survey data with related non-traditional data sources in prediction models. In this case the survey data are the dependent data and the non-traditional data are used as covariates. This approach comes from the methodology developed in small area estimation literature. The additional information of non-traditional data sources can be used in different ways. In cross-sectional models they can be used to make more prices regional estimates. If non-traditional data sources can be used to construct time series that are related to repeatedly conducted surveys, then their timeliness and high frequency can also be utilized to improve the timeliness of the sample survey by making more precise now casts with e.g. dynamic factor models at the moment that the observations for the auxiliary series become available during the reference period of the sample survey.

In the last decade there has been a substantial increase in the uses of new forms of data such as remotely sensed and mobile data, and these sources are being used to produce indicators. New forms of data have a number of attractive properties, including wide coverage, public access and dynamic updating. However, they have only begun to be explored for their usefulness in producing official statistics, which typically have a higher quality threshold because of their use in monitoring the effectiveness of governments. The case for using such data is a compelling one when working in countries that lack survey and Census data, but this does not mean that the quality is inherently better – just that the alternatives are worse. The first step is to establish the association between indicators constructed from the new forms of data and the targets of estimation, for example the incidence of poverty and SDG-related indicators. This is a necessary step for the credibility of the outputs if new forms of data are to be used in producing official statistics with model-based or model-assisted methods.

An important aspect when producing official statistics is to assess their quality. The quality measures in the small area estimation literature, and in the literature using new forms of data are generally different, and there is a need to bring these different approaches together to understand the quality of the outputs. There are some outstanding research questions, such as whether new forms of data should be treated as fixed, and if not how to quantify the errors associated with these data sources and account for them in estimation. The coverage of new forms of data is also an aspect that needs exploration - although in many cases it is much wider than survey sources, this does not mean that it is complete, or representative. It is our view that some representative data from surveys are needed too, and that methodological tools should be developed to allow for the combined use of survey and new forms of data.

Defining appropriate spatial scales at which estimates are produced also requires additional research, so as not to push models past the point where they are useful. Over-reliance on a model can be risky in any situation, but when the aim is to produce official statistics which form the basis for policy decisions we need to apply a much stronger precautionary principle to avoid making poor decisions.

Although there have been some important steps in establishing the potential value of new forms of data, it is our view that significant effort is required before we will be able to bring these ideas into the mainstream production of official statistics.

A second approach is to use non-traditional data sources directly to produce offcial statistics about well-being and sustainability. In this case methods must be applied that account for the potential selection bias in these data sources. In Chapter 3 we have reported a list of methods that are proposed in the literature to deal with the bias in the estimates that can arise due to a non random selection of the data, a situation that is common for secondary sources and big data, in particular. These methods have been proposed in the framework of sample data with non response and non ignorable sample selection, for example web-surveys. However, all the methods that were illustrated in 3 need additional information on the records (units) of the big data source and an external auxiliary source that is unaffected by bias to adjust the selectivity. These conditions for the adjustment can be very demanding for secondary data. Sometimes a separate survey has to be implemented. Indeed, this might not be the major issues since the survey is needed only with the aim of adjusting the secondary source and then it has not to be very large. However, the availability of very basic auxiliary information for the records in the big data source (e.g. gender and age) is not always straightforward. Daas et al. (2016) explored the possibility to 'extract' features from twitter accounts that could serve to correct selectivity as described in 3. However, feature extraction will produce a 'measurement errors' on the covariates later used for the selectivity correction, whose effect on the adjustment methods should be further explored. An additional very relevant issue when treating big data sources is the identification of the target statistical units from the records of the secondary source, also in consideration of the fact that records often refer to events and not to the statistical units. The effect of the errors in this process (duplication of units or wrong association to the same statistical units of records) could make more complex the use of the statistical adjustments for bias and introduce other source of bias.

Chapter 3 only referred to bias raised by non-representative sample selection of the target population. In the above discussion we have mentioned other sources of errors that might reduce the benefit of bias adjustment; on the other hand these sources of errors themselves induce bias (and variability), e.g. think to

- The unit error (wrong identification of units)

- Measurement error if the measure observed in the source is different from the target variable and if a biased algorithm is applied for feature extraction.

For example, let assume we measure the 'Annual change in forest area and land under cultivation' with remote sensing data, the Y variable in this case is not directly observed but will be the result

of a processing of the original data that might introduce bias in the classification of the land cover. The given example also introduce another peculiar aspect of big data, i.e. its consistency over time, In fact the data for technical reasons are often subject to changes making comparisons though time also affected by bias.

A standard framework for dealing with these errors in big data is not yet well defined.

The input-state-output framework is characterized by an ordered series of processes that describes the system behaviour highlighting the dependence of an economic system on a level of societal organization and, especially, on environmental resources. The framework is a rational solution for the study of system sustainability, because it incorporates consistency with traditional sectors proposed in sustainability research and is feasible because it is limited to small number of data. The framework encompasses the three sectors that traditionally compose the concept of sustainability and, maintaining the informative capacity of every aspect of system behaviour, it provides a synthetic picture of the reality. All these aspects must be indeed, carefully monitored because the relationships among the three components may result in sustainable or unsustainable behaviours. Regarding the strict link between the economy and the environment, some feedbacks can be identified, for example massive investments in natural capital that enable increased use of renewable environ-mental goods and services without compromising ecosystem functions may have positive effects in the future perspectives and the sustainability of entire national economies. This feature should be taken into account in the further development of the framework. The ability of this framework to evaluate sustainability can be maximized in the case of dynamic analyses at both the macro- and micro-system level: policy makers can use this 'beside-GDP' monitoring system as an orientor to pursue sustainability programs. Indeed, given the triad of indicators, the evolution of a single point (i.e. a country) in the diagram can be monitored when it moves from one to another region of the space within the cube, characterized by different input-state-output values. One can evaluate whether an optimal trend for a single country exists and must be followed, or how exogenous factors, like the economic crisis or the progressive exhaustion of a non-renewable resource, influence the position of a nation in the 3D space. Moreover, this tool can be useful to evaluate the effects of national policies in economic but also in social and environmental terms. In order to make feasible the monitoring of countries over time appropriate information source should be available and this is not at all obvious, especially for environmental indicators.

The results of this report have the following relations with the other work packages of the MAKSWELL project. The small area estimation methods developed in Section 2.1 will be further picked up in WP3 where small area estimation methods will be applied to produce regional estimates for poverty. The time series methods developed in Section 2.2 and 2.3 will be applied in WP4 where dynamic factor models are used for nowcasting time series observed with repeated sample surveys, using non-traditional data sources such as google trends. Correction methods from Chapter 3 will be used in the pilot of WP5, where an indicator will be constructed from a non-traditional data source.

Bibliography

AAPOR, t. F. o. N.-P. S. (2013). *Non probability Sampling: REPORT OF THE AAPOR TASK FORCE on Non-Probability Sampling.*

Aprigliano, V., G. Ardizzi, and L. Monteforte (2017). Using the payment system data to forecast the italian gdp.

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics 122*, 137–183.

Baldacci, E., D. Buono, G. Kapetanios, S. Krische, M. Marcellino, G. L. Mazzi, and F. Papailias (2016). Big data and macroeconomic nowcasting: from data access to modelling. *Luxembourg: Eurostat. Doi: http://dx. doi. org/10.2785/360587*.

Barbier, E. (1987). The concept of sustainable economic development. *Environ. Conserv 14*, 101–110.

Barcaroli, G., N. Golini, and P. Righi (2018). Quality evaluation of experimental statistics produced by making use of big data.

Bastianoni, S., L. Coscieme, and F. Pulselli (2014a). The effect of a consumption-based accounting method in national ghg inventories: a trilateral trade system application. *Frontiers in Energy Systems and Policy 2*, 1–8.

Bastianoni, S., L. Coscieme, and F. Pulselli (2014b). The input-state-output model and related indicators to investigate the relationships among environment, society and economy. *Ecological Modelling 325*, 84–88.

Bastianoni, S., F. Pulselli, and E. Tiezzi (2004). The problem of assigning responsibility for greenhouse gas emissions. *Ecological Economics 49*, 253–257.

Battese, G., R. Harter, and W. Fuller (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 83*, 28–63.

Bell, W. (2005). Some considerations of seasonal adjustment variances. Census bureau. https://www.census.gov/ts/papers/jsm2005wrb.pdf.

Beresewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio, and K. M. (2018). *An overview of methods for treating selectivity in big data sources.*

Bethlehem, J. (2002). *Weighting Nonresponse Adjustments Based on Auxiliary Information. In Survey Nonresponse, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little.* John Wiley & Sons.

Binder, D. and J. Dick (1990). A method for the analysis of seasonal arima models. *Survey Methodology 16*, 239–253.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science* (350).

Boivin, J. and S. Ng (2005). Understanding and comparing factor-based forecasts. *Internatioonal Journal of Central Banking 3*, 117–151.

Bollineni-Balabay, O., J. van den Brakel, F. Palm, and H. J. Boonstra (2017). Multilevel hierarchical bayesian versus state space approach in time series small area estimation: the dutch travel survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(4), 1281–1308.

Boonstra, H. J. and J. van den Brakel (2016). Estimation of level and change for unemployment using multilevel and structural time series models. Technical Report 201610, `https://www.cbs.nl/nl-nl/achtergrond/2016/37/estimation-of-level-and-change-for-unemployment`, Statistics Netherlands.

B orsch-Supan, A. and J. Winter (2004). How to make internet surveys representative: A case study of a two-step weighting procedure. MEA discussion paper series 04067, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy.

Bovi, M., C. Lupi, and C. Pappalardo (2000). *Predicting GDP Components Using ISAE Bridge Equations Econometric Forecasting Model (BEEF)*. ISAE.

Box, G. and G. Jenkins (1989). *Time series analysis: forecasting an control*. Holden-Day, San Francisco.

Buelens, B., J. Burger, and J. Brakel (2015). Predictive inference for non-probability samples: a simulation study. Technical report.

Buelens, B., P. Daas, J. Burger, M. Puts, and J. Brakel (2014). Selectivity of big data. Technical report.

Caro, D., S. Bastianoni, S. Borghesi, and F. Pulselli (2014). On the feasibility of a consumer-based allocation method in national ghg inventories. *Ecological Indicators 36*, 640–643.

Casas-Cordero, C., J. Encina, and P. Lahiri (2016). *Poverty mapping for the Chilean comunas. In Analysis of Poverty Data by Small Area Estimation (ed. M. Pratesi)*. Hoboken: Wiley.

Chan, J. and I. Jeliazkov (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation 1*, 101–120.

Choi, H. and H. Varian (2012). Predicting the present with google trends. *Economic Record 88*, 2–9.

Coscieme, L., F. Pulselli, S. Jørgensen, and S. Bastianoni (2013). Thermodynamics based categorization of ecosystems in a socio-ecological context. *Ecological Modelling 258*, 1–8.

Costanza, R., R. d' Arge, R. de Groot, S. Farber, M. Grasso, and K. Limburg (1997). The value of the worldâ€™s ecosystem services and natural capital. *Nature 387*, 253–260.

Daas, P., J. Burger, Q. Le, O. ten Bosch, and P. M. (2016). Profiling of twitter users: a big data selectivity study. Technical report, CBS Discussion Paper.

Daas, P., S. Ossen, R. Vis-Visschers, and J. Arends-Toth (2009). Checklist for the quality evaluation of administrative data sources. Technical report.

Datta, G., P. Lahiri, T. Maiti, and K. Lu (1999). Hierarchical bayes estimation of unemployment rates for the states of the u.s. *Journal of the American Statistical Association 94*(448), 1074–1082.

Davé, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters 12*, 657–664.

Doornik, J. (2009). *An Object-oriented Matrix Programming Language Ox 6*. Timberlake Consultants Press: London.

D'Orazio, M., M. Di Zio, and M. Scanu (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.

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.

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

Durbin, J. and B. Quenneville (1997). Benchmarking by state space models. *International Statistical Review 65*, 23–48.

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

Fabrizi, E. and C. Trivisano (2016). Small area estimation of the gini concentration coefficient. *Data analysis 99*, 223–234.

Fath, B., B. Patten, and J. Choi (2001). Complementarity of ecological goal functions. *Journal of Theoretical Biology 208*, 493–506.

Fay, R. and R. Herriot (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association 74*(366), 269–277.

Feder, M. and D. Pfeffermann (2015). Statistical inference under non-ignorable sampling and non-response. an empirical likelihood approach. Technical report.

Fisher, B., R. Turner, and P. Morling (2009). Defining and classifying ecosystem services for decision making. *Ecological Economics 68*, 643–653.

Frias-Martinez, V. and J. Virseda (2012). On the relationship between socio-economic factors and cell phone usage. Ictd 2012, march 12-15, 2015, atlanta, ga, usa.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*, 665–676.

Gonzalez-Manteiga, W., M. Lombardia, I. Molina, D. Morales, and L. Santamaria (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation 78*, 443–462.

Grudkowska, S. (2016). Jdemetra+ reference manual version 2.1https. *ec. europa. eu/eurostat/cros/system/files/jdemetra_reference_manual_version_2. 1_0. pdf*.

Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Harvey, A. and C. Chung (2000). Estimating the underlying change in unemployment in the uk. *Journal of the Royal Statisitcal Society, A series 163*, 303–339.

Haziza, D. and É. Lesage (2016). A discussion of weighting procedures for unit nonresponse.

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.

Helske, J. (2017). Kfas: exponential family state space models in r. *Journal of Statistical Software 78*.

Jean, N., M. Burke, M. Xie, M. Davies, D. Lobeil, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science* (350).

Jørgensen (2008). *Exergy. In: JÃ¸rgensen, S.E., Fath, B.D. (Eds.), Encyclopedia of Ecology*. Elsevier, Oxford.

Jørgensen, S., B. Fath, S. Bastianoni, J. Marques, F. Muller, and S. Nielsen (2007). *A New Ecology: Systems Perspective*. Elsevier, Oxford.

Jørgensen, S. and H. Mejer (1981). *Thermodynamics based categorization of ecosystems in a socio-ecological context. In: Dubois, D. (Ed.), Progress in Ecological Modelling.* CEBEDOC, Liege.

Keiding and Louis (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. R. Statist. Soc. A 179, Part 2*, 1–2831.

Knorr-Held, L. and H. Rue (2002). On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*(4), 597–614.

Koopman, S., A. Harvey, N. Shephard, and J. Doornik (2009). *STAMP 8.2; Structural Time Series Analyser, Modeller and Predictor*. Timberlake Consultants Press: London.

Koopman, S., N. Shephard, and J. Doornik (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*. Timberlake Consultants Press: London.

Koopman, S., N. Shephard, and J. Doornik (2009). Statistical algorithms for models in state space form using ssfpack 2.2. *Econometrics Journal 2*, 113–166.

Krieg, S. and J. van den Brakel (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics and Data Analysis 56*, 2918–2933.

Kwang Kim, J. and Z. Wang (2018). Sampling techniques for big data analysis in finite population inference. *International Statistical Review*.

Lavallée, P. and J. Brisbane (2015). Sample matching: Toward a probabilistic approach for web surveys and big data.

Lee, S. and R. Valliant (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research 37*(3), 319–343.

Little, R. and D. Rubin (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.

Lütkepohl, H. (2005). *A new introduction to Multiple Time Series Analysis.* Springer, New-York.

MA (2005). *Millennium Ecosystem Assessment.* Washington DC, Island Press.

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

Marcellino, M., J. Stock, and M. Watson (2003). Macroeconomic forecasting in the euro area; country specific versus area-wide information. *European economic review 47*, 1–18.

Marchetti, S., C. Giusti, and M. Pratesi (2016). The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. *AStA Wirtsch Sozialstat Arch 10*, 79–931.

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 dtata sources. *Journal of Official Statistics 31*, 263–281.

McCausland, W., S. Miller, and D. Pelletier (2011). Simulation smoothing for state-space models: A computational efficiency analysis. *Computational Statistics and Data Analysis 55*, 199–212.

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, No. 2*, 685–726.

Moauro, F. and G. Savio (2005). Temporal disaggregation using multivariate structural time series models. *Econometrics Journal 8*, 214–234.

Neri, L., A. D'Agostino, A. Regoli, F. Pulselli, and L. Coscieme (2017). Evaluating dynamics of national economies through cluster analysis within the input-state-output sustainability framework. *Ecological Indicators 72*, 77–90.

Noor, A., V. Alegana, P. Gething, A. Tatem, and R. Snow (2008). Using remotely sensed night-time light as a proxy for poverty in africa. *Population and Health Metrics* (6:5).

Odum, H. (1996). *Environmental Accounting. Emergy and Environmental Decision Making.* John Wiley and Sons, New York.

Pappalardo, L., S. Rinzivillo, Z. Qu, D. Pedreschi, and F. Giannotti (2013). Understanding the patterns of car travel. *The European Physical Journal 215*, 61–73.

Petris, G. (2010). An r package for dynamic linear models. *Journal of Statistical Software 9*, 163–175.

Pfeffermann, D. (2018). Challenges in the production of official statistics with different ways of data collection.

Pfeffermann, D. and S. Bleuer (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology 19*, 149–163.

Pfeffermann, D. and L. Burck (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology 16*, 217–237.

Pfeffermann, D., A. M. Krieger, and R. Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica 8*, 1087–1114.

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

Pfeffermann, D. and M. Sverchkov (2014). Estimation of mean squared error of x-11-arima and other estimators of time series components. *Journal of Official Statistics 30*, 811–838.

Pfeffermann, D. and R. Tiller (2006). Small area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association 101*, 1387–1397.

Piepho, H. and J. Ogutu (2014). Simple state-space models in a mixed model framework. *The American Statistician 61*(3), 224–232.

Powell, B., G. Nason, D. Elliott, M. Mayhew, J. J. Davies, and J. Winton (2017). Tracking and modelling prices using web-scraped price microdata: Towards automated daily consumer price index forecasting. *Journal of the Royal Statistical Society, Series A 181*.

Prasad, N. and J. Rao (1990). he estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association 85*, 163–171.

Pulselli, F., L. Coscieme, L. Neri, A. Regoli, P. Sutton, A. Lemmi, and S. Bastianoni (2015). The world economy in a cube: A more rational structural representation of sustainability. *Global Environmental Change 35*, 41–51.

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

Rao, J. and I. Molina (2015). *Small Area Estimation*. Wiley-Interscience.

Rao, J. and M. Yu (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics 22*, 511–528.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.

Rivers, D. and Bailey (2009). Inference from matched samples in the 2008 u.s. national elections.

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.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Ruiz-Cárdenas, R., E. Krainski, and H. Rue (2012). Direct fitting of dynamic models using integrated nested laplace approximations - inla. *Computational Statistics and Data Analysis 56*, 1808–1828.

Särndal, C.-E. and S. Lundström (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

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

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

Slud, E. and T. Maiti (2006). Mean-squared error estimation in transformed fay-herriot models. *Journal of the Royal Statistical Society, Series B 68*, 239–257.

Smith-Clarke, C., A. Mashhadi, and L. Capra (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Statistics Canada (2002). Statistics canada's quality 2002 assurance framework. Technical report.

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

Steorts, R., N. Tzavidis, and T. Smith (2018). Bayesian smoothing and benchmarking for small area estimation with application to rental prices in berlin. Research paper.

Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statisitcal Society 97*, 1167–1179.

Stock, J. and M. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*, 147–162.

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

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

Team, R. C. (2017). R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria.

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

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

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. R. Perilla (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statisitcal Society, A series 181*, 927–979.

UNECE Big Data task team (2014). A suggested framework for the quality of big data. Technical report.

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.

van den Brakel, J. and S. Krieg (2016). Small area estimation with state-space common factor models for rotating panels. *Journal of the Royal Statisitcal Society, A series 179*(4), 763–791.

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

Vehovar, V., V. Toepoel, and S. Steinmetz (2016). Non-probability sampling. *The SAGE Handbook of Survey Methodology*, 329–345.

Vosen, M. and T. Schmidt (2011). Forecasting private consumption: survey-based indicators vs. google trends. *Journal of Forecasting 30*, 565–578.

Watmough, G., P. Atkinson, A. Saikia, and C. Hutton (2016). Understanding the evidence base for povert-enviornment relationships using remotely sensed satellite data: An example from assam, india. World development, 78.

Wilkinson, R. and K. Pickett (2009). *The Spirit Level.* Penguin Books Ltd., London.

Wilkinson, R. and K. Pickett (2018). *The Inner Level.* Penguin Books Ltd., London.

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*(453), 185–193.

Ybarra, L. and S. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of canada. *Survey Methodology 34*(1), 19–27.

You, Y., J. Rao, and J. Gambino (2003). Model-based unemployment rate estimation for the canadian labour force survey: A hierarchical bayes approach. *Survey Methodology 29*(1), 25–32.